A Toponym Resolution Service following the OGC WPS Standard *

Susana Ladra, Miguel R. Luaces, Oscar Pedreira, and Diego Seco

Database Laboratory, University of A Coruña
Campus de Elviña, 15071 A Coruña, Spain
{sladra, luaces, opedreira, dseco}@udc.es
Contact person: Miguel R. Luaces

Abstract. In the research field of Geographic Information Systems (GIS), a cooperative effort has been undertaken by several international organizations to define standards and specifications for interoperable systems. The Web Processing Service (WPS) is one of the most recent specifications of the Open Geospatial Consortium (OGC). It is designed to standardize the way that GIS calculations are made available to the Internet.
We present in this paper a WPS to perform *Toponym Resolution*. This service defines two geospatial operations. The first operation, *getAll*, returns all possible geographic descriptions with the requested name ordered by a relevance ranking. The second operation, *getMostProbable*, filters the result and returns only the most probable geographic description. Furthermore, both operations can be parameterized according to the level of detail needed in the result.

Keywords: Web Services, Open Geospatial Consortium, Web Processing Service, Toponym Resolution

1 Introduction

The research field of Geographic Information Systems [1] has received much attention during the last years. Recent improvements in hardware have made the implementation of this type of systems affordable for many organizations. Furthermore, a cooperative effort has been undertaken by two international organizations (ISO [2] and the Open Geospatial Consortium [3]) to define standards and specifications for interoperable systems. This effort is making possible that many public organizations are working on the construction of spatial data infrastructures [4] that will enable them to share their geographic information.

* This work has been partially supported by "Ministerio de Educación y Ciencia" (PGE y FEDER) ref. TIN2006-16071-C03-03, by "Xunta de Galicia" ref. PGIDIT05SIN10502PR and ref. 2006/4, by "Ministerio de Educación y Ciencia" ref. AP-2006-03214 (FPU Program) for Oscar Pedreira, and by "Dirección Xeral de Ordenación e Calidade do Sistema Universitario de Galicia, da Consellería de Educación e Ordenación Universitaria-Xunta de Galicia" for Diego Seco.

The OGC is a consensus standards organization that is leading the creation of standards to allow the development of interoperable geospatial systems. One of the most recent specifications standardized by the OGC is the Web Processing Service (WPS) [5] (version 1.0.0 of this standard was published on June, 2007). The WPS specification defines a mechanism by which a client can submit a spatial processing task to a server to be completed. In other words, this specification standardizes the way that GIS calculations are made available in Internet. In this paper, we briefly summarize the most important characteristics of the specification and we present a *Toponym Resolution* service developed according with its interface.

Toponym Resolution is a task related to mapping a place name to a representation of the extensional semantics of the location referred, such as a geographic latitude/longitude footprint [6]. This task has been widely used in *Geographic Information Retrieval* (GIR), *question answering*, or *map generation*. The research field in GIR [7] has appeared a few years ago as the confluence of *Geographic Information Systems* [1] and *Information Retrieval* [8]. The main goal of this field is to define index structures and techniques to efficiently store and retrieve documents using both the text and the geographic references contained within the text. Therefore, the documents have to be annotated with the toponyms mentioned in the text. This task has recently been automated, achieving near-human performance using machine learning [9]. However, annotating the documents with the toponyms mentioned in the text is not enough when the documents have to be spatially indexed. In this case, place names must additionally be related to a correlate in a model of the world (for example, using its coordinates in latitude/longitude). A *gazetteer* could be used to obtain these *geo-references*.

A gazetteer is a geographic dictionary that contains, in addition to location names, alternative names, populations, location of places, and other information related to the location. However, Gazetteers are not enough to fully automate the geo-referencing task because they provide the toponyms and the coordinates associated with a place name without any measure of relevance. This problem is related with the *referential ambiguity*. For example, *London* is the capital of the United Kingdom but it is a city in Ontario, Canadaa too. Given the question *where is London*, a Gazetteer would return both locations without giving any hint of which of them is more appropriate.

Furthermore, gazetteers do not usually provide geometries for the location names other than a single representative point (its coordinates). But, sometimes, the real geometry of the toponym is needed. In [10], the authors describe a spatial index structure where the nodes of the structure are connected by means of inclusion relationships. Therefore, each non-leaf node stores, as well as the toponym, the bounding box of the geometry. For such an application, the authors need a service that returns not only the most probable location, but also its complete geometry to build the spatial index. In this paper, we present a service to perform *Toponym Resolution*. This service provides an operation to obtain all the possible geographic descriptions for a toponym ordered by a rank-

ing of relevance. Moreover, the service provides an operation to obtain only the most probable geographic description. Both operations can be parameterized according to the level of detail needed in the result (i.e., whether a single representative point is enough or a complete geometry is needed). In accordance with the current trend in GIS, these operations, or spatial processes, are offered as a service according with the WPS specification. The rest of the paper is organized as follows. We first describe some related work in Section 2. In addition to that, Section 3 resumes the main characteristics of the WPS specification. Then, in Section 4, we present the general architecture of the system and describe its components. After that, in Section 5, some implementation details are described. Finally, Section 6 presents some conclusions and future lines of work.

2 Related Work

Gazetteers are considered one of the most important components in Spatial Data Infrastructures [4]. A gazetteer service returns information about places in response to queries using their identifiers (e.g., location names). This information typically contains geographic data, such as the coordinates, social statistics, etc. The international OGC specification *Gazetteer Profile of WFS* (WFS-G) [11] standardizes the functionalities that may implement a gazetteer. Service metadata, operations, and types of geographic entities are defined in this specification. The main differences between the WFS-G and WFS specifications are:

- The gazetteer structure is described in an additional section of the document describing service metadata.
- All the geographic entities defined in a WFS-G are subclass of the predefined *SI_LocationInstance*. Therefore, geographic entities share a set of basic attributes and can define other attributes specifically designed for the concrete application.

Many free resources have been published in Internet that provide gazetteer functionalities, geographic ontologies, etc. *Alexandria Digital Library* [12], *Getty Thesaurus of Geographic Names* [13], or *GeoNames* [14] are some examples of these resources. However, none of them define a service following the WFS-G specification.

An important drawback of the gazetteers is that they do not usually provide a complete geographic description of the location returned by a query. There are several cartographic resources that can be used to complete the information provided by the gazetteers. Global Administrative Unit Layers (GAUL) [15] and *Vector Map* (VMAP) cartography [16] are two interesting resources because they provide a complete and updated cartography of the world. However, this cartography is not usually offered by gazetteers. Instead, only a single representative point is returned for each location queried.

Gazetteers are a key component in the task of *Toponym Resolution*. The goal of this task is to obtain the *referent* of the place names. The work of Leidner

[6] in this task is focused on the research field in Geographic Information Retrieval (GIR). Several papers describe the architecture of GIR systems and the NERC+R (Named Entity Recognition and Classification with Resolution) task is shared in most of the proposals. Recognizing the toponyms in the texts of the documents and relate these toponyms to correlate in a model of the world is the main goal of this task. Some papers that deal with different aspects of this problem in the context of GIR have been published in the last years [17] [18] [19]. Web-a-where [17] uses *spatial containers* in order to identify locations in documents, MetaCarta (the commercial system described in [18]) uses a natural language processing method, and STEWARD [19] uses an hybrid approach. A common drawback of gazetteers when applied to this task is that, given a location name, gazetteers provide a list of toponyms that is not ordered by relevance. Therefore, the user of the gazetteer must find a method to order the list of results.

3 OGC Web Processing Service

The Web Processing Service (WPS) [5] is one of the most recent specifications of the OGC. This standard defines a mechanism by which a client can submit a spatial processing task to a server to be completed. Recently, some papers have appeared that review the specification and propose several examples of its usage [20] [21]. In this section we briefly resume the most important characteristics of the specification.

As said above, the WPS specification is centred in the communication between the server and the client. An XML-based protocol using the POST method of HTTP has been defined to perform this communication. Furthermore, requests can also be expressed in Key-Value-Pairs (KVP) encoding using the GET method of HTTP. In addition to the communication protocol, the specification defines three operations:

- *GetCapabilities*. This operation is common in many OGC specifications. The response is a XML document with two main parts: *ServiceIdentification* and *ProcessOfferings*. The first one is shared with other OGC specifications and it describes information of interest regarding the service provider. The second one lists all the processes offered by the service.
- *DescribeProcess*. After a client parses the GetCapabilities response, it has a list of the processes offered by the service. The operation *DescribeProcess* can then be used to request more information about each of them. This operation receives the process identifier as a parameter and returns a XML document that describes all the characteristics of the process such as the title, abstract, etc. Moreover, a full description of the process input parameters is provided in order to allow the client to understand the way in which the process is invoked. Finally, the response document also describes the output format of the process.
- *Execute*. Finally, clients have enough information to request for the execution of a process. The response of the operation *Execute* is a XML document with

information about the *status* of the process, inputs that were used, and the output. The output can be a simple literal (for example, a numerical result or the url where a complex document is accessible) or a complex output (for example, a feature collection description in GML [22]).
Geospatial processes can be very complex and they can take a long time to complete (in terms of hours, days, or even weeks). Therefore, these processes must be performed in an asynchronous way. The specification defines the *status* description in the XML document in response to a *Execute* request for this purpose. The value *ProcessAccepted* indicates that the process was correctly received. *ProcessStarted* indicates that the server is performing the process. *ProcessSucceeded* indicates that the process is finished, and therefore, the result is ready. Finally, *ProcessFiled* indicates that a problem appeared in the execution of the process.

One of the most attractive characteristics of this specification is that it can be applied to an unlimited number of cases. All geospatial process can be offered in Internet following this specification. However, there are some issues that must be considered to decide whether to define a process as a WPS or not. First, complex processes that take a long time to complete are the best candidates to be implemented as a WPS. However, if the complexity of the process is low and the main part of the time is taken up by managing a lot of data stored locally and not on the server, the process can be completed more effectively locally. Second, WPS have the advantages of all the general-purpose web services. One of the most important is that the service is centralized. Therefore, a WPS is appropriate for the deployment of new processes that are under active development. WPS developers can release new versions simply updating the version of the process implementation in the server. Finally, WPS makes possible to create advanced services by means of the *orchestration* of several services.

Recently, several frameworks and implementations of the OGC WPS have appeared to make it usage easier. However, most of them are implementing the 0.4.0 version of the standard. We use the *52 North WPS* [23] framework in the implementation of our system. The 52 North Web Processing Service enables the deployment of geo-processes on the web. It features a pluggable architecture for processes and data encodings. The implementation is based on the version 1.0.0 of the specification.

4 System Architecture

Fig. 1 shows our proposal for the system architecture of a *Web Processing Service* (WPS) to perform *Toponym Resolotuion*. The architecture has two independent layers: the *WPS layer* and the *Toponym Resolution layer*.

The WPS layer is at the top of the architecture. As we noted before, the 52 North WPS implementation [23] has been used in this work. In [24], the authors present the 52 North WPS architecture and an example of use for a generalization process. This architecture is quite simple. There is a *Request Processor*

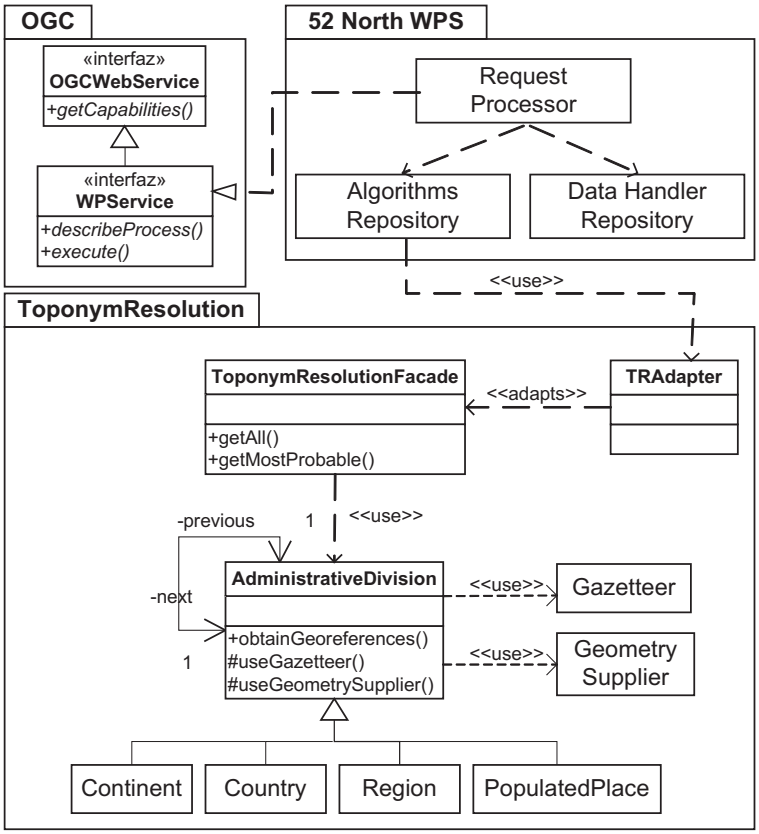

Fig. 1. System Architecture

that manages the communication protocol with the clients. This *Request Processor* implements the OGC WPS specification and it encapsulates all the details related to the communication protocol. In order to achieve a high extensibility, the implementation of a 52 North WPS is organized in *repositories* that provide dynamic access to the embedded functionality of the WPS. For each geospatial process offered by the service, an algorithm must be defined in the *Algorithms Repository*. For example, the algorithms repository in [24] is composed of several generalization algorithms. In our implementation, we adopt the notion of *repository* and we design an intermediary component to adapt the geospatial algorithm to the interface of the specific repository (see the design pattern *Adapter* in [25]). Therefore, our particular implementation of the algorithms does not depend excesively on the details of the *Request Processor*.

The bottom part of the figure shows the architecture of the Toponym Resolution component. The *TRAdapter* class represents the *adapter* between this component and the algorithms repository of 52 North. The adapter uses the

ToponymResolutionFacade that is a *Facade* [25] that provides a simplified interface of the component. This facade defines two public operations: *getAll* and *getMostProbable*. The first one returns all the possible geographic descriptions in response to a place name. There are two main differences between this operation and the functionality offered by a gazetteer. First, our implementation can be configured to obtain the real geometry, the *bounding box*, or a single representative point. Second, we provide these descriptions ordered by a relevance ranking. The second operation returns only the most probable geographic description.

The implementation of both operations uses a hierarchy of *Administrative Divisions* to perform the process that defines four levels of administrative divisions (*Continent*, *Country*, *Region*, and *PopulatedPlace*). The implementation is easily extensible because several design patterns were used to obtain a robust and extensible architecture. First, the hierarchy follows the pattern *Chain of Responsibility*. Therefore, the class that represents each administrative level is in charge of a part of the whole process and it delegates the rest on the next level. Two *chains of responsibility* are configured in the system. The first one is used to go down the hierarchy finding place names. Once a toponym is found, the second chain is used to go up the hierarchy in order to return the *complete path* that fully describes the toponym in the hierarchy. For example, if the requested place name is *A Coruña*, the complete path is composed by the geographic descriptions of *Europe*, *Spain*, *Galicia*, and *A Coruña*. Furthermore, algorithms to obtain the georeferences were designed following the pattern *Template Method*. Therefore, the superclass (*AdministrativeDivision*) defines the general algorithm and the concrete steps may be changed by the subclasess. These steps define how the *Gazetteer* and the *Geometry Supplier* are queried in each level. In the section 5, we present more details about the concrete algorithm implemented in the system to retrieve and rank toponyms.

5 Implementation

As we said in the previous section, the *Toponym Resolution* layer uses a *Gazetteer* and a *Geometry Supplier* in order to obtain the geographic descriptions. In our test implementation we use *Geonames* [14] that provides a geographical database available under a creative commons attribution license. This database contains more than two million populated places over the world with their latitude/longitude coordinates in WGS84 (*World Geodetic System 1984*). All the populated places are categorized so that it is possible to classify them into the different administrative division levels that are defined by the architecture (continents, countries, regions, and populated places).

However, Geonames (and Gazetteers in general) does not provide geometries for the location names other than a single representative point. But for our system we need the real geometry of the location name. We define a *Geometry Supplier* service to obtain the geometries of those location names. As a base for this service we used the *Vector Map* (VMap) cartography [16]. VMap is an updated and improved version of the National Imagery and Mapping Agency's

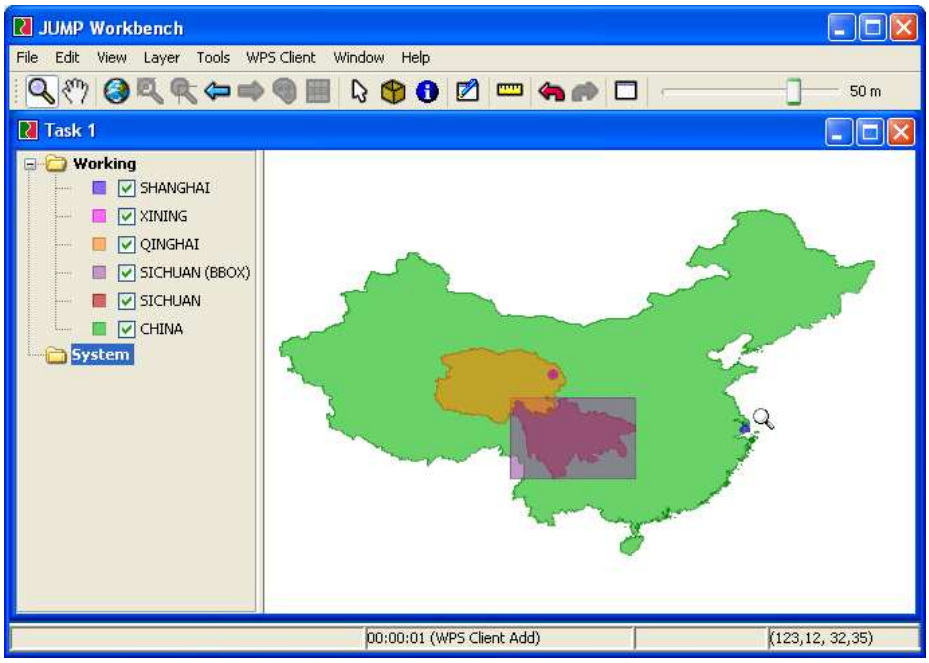

Fig. 2. Examples of results using the WPS plugin for JUMP

Digital Chart of the World. It supplies first and second level administrative division geometries in a proprietary format. However, there are free tools that can create *shapefiles* from that format, such as FWTools [26]. We have created a PostGIS [27] spatial database with these *shapefiles* and we have done several corrections and improvements over this database.

Even though our test implementation uses Geonames and VMap, it has been designed so that these components are easily exchangeable. All accesses to these components are performed through generic interfaces that can be easily implemented for other components.

The spatial operations defined by the hierarchy in the architecture combine both services to geo-reference location names. Each level contains a connection to the gazetteer and to the geometry supplier in order to retrieve the data needed by the process. In other words, subclasses in this hierarchy change the abstract methods of the superclass to implement real queries to both services.

Furthermore, the algorithm to obtain geo-references is implemented in two steps each of them using one of the *chain of responsabilities* defined in the hierarchy. In the first step, each level obtains from the gazetteer all the locations with the requested name. After that, in the second step, the system builds the complete path of geographic descriptions from bottom to top. For instance, if the requested location name was London, in the first step the system obtains

at least two locations with this name. After that, it returns the paths *United Kingdom*, *England*, *London* and *Canada*, *Ontario*, *London*. Finally, to elaborate a relevance ranking of the results (or to return the most relevant result) the algorithm computes a measure of relevance for each result. This measure combines the length of the path, the population of the place, a weighting factor depending of wheter the place is a capital, main city, etc. Most of these data come from the gazetteer.

Fig. 2 presents some examples using the generic WPS plugin for JUMP [28] developed by 52 North. JUMP provides a graphic user interface for viewing and manipulating spatial data-sets. The architecture of this tool is very extensible and it defines a mechanism of extension based on plugins. 52 North developed a plugin for JUMP that implements a generic WPS client. We use this client to test the Toponym Resolution WPS.

One can see in the figure the result of several requests to the *getMostProbable* operation. All of these requests were executed with the parameter *full_path* set to *false*. If this parameter is set to *true*, the WPS returns the geographic description of all the nodes in the path (continent, country, etc.). Furthermore, all the requests, except the layer namely *SICHUAN(BBox)*, were executed with the parameter *bounding_box* set to *false*. The bounding box, instead the real geometry, is returned by the WPS if this parameter is set to *true*. The requested place names are, from bottom to top, *China* (a country), *Sichuan* (a province of China), the bounding box of this province, *Qinghai* (a province of China), *Xining* (the capital of Qinghai), and *Shanghai* (the host city of the conference W2GIS 2008).

6 Conclusions and Future Work

We have presented in this paper a system to perform *Toponym Resolution*. The interface of this system defines two spatial operations *getAll* and *getMostProbable*. The first one returns all the geographic descriptions with the requested place name ordered by a relevance ranking. The second one filters the result and it returns only the most relevant geographic description with the requested name. Furthermore, both operations can be customized with two parameters. The *bbox* parameter is used to obtain the bounding box of the geometries instead of the real geometries. The *full_path* parameter is used to obtain the full path that represents the requested place name instead of the leaf node of this path. Moreover, following the current trend in GIS, we developed a Web Processing Service (WPS) to offer both *getAll* and *getMostProbable* operations as processes that can be performed through the Internet.

Future improvements of this WPS are possible. Many times, intrinsic features of the toponyms (such as population or administrative level) are not enough to decide the most relevant result in a certain *context*. For example, if the place name *Santiago* appears in a document with other place names such as *Atacama*, or *Magallanes*, the document describes regions in Chile. However, if *Santiago* appears with *Madrid* or *Barcelona*, the document describes places in Spain. We are

currently working on a new operation that can be invoked with more than one place name. The result of this operation must be the most probable geographic descriptions to each place name. This operations can be very useful in the research field of *Geographic Information Retrieval* (GIR). Therefore, another line of future work involves integrating this WPS in the architecture of GIR systems. Furthermore, changes in the algorithms are needed to improve the performance of the system. Finally, we plan on exploring other gazetteers and cartographies to determine the way that they affect the performance of the system.

References

1. Worboys, M.F.: GIS: A Computing Perspective. CRC (2004) ISBN: 0415283752.
2. ISO/IEC: Geographic Information – Reference Model. International Standard 19101, ISO/IEC (2002)
3. Open GIS Consortium, Inc.: OpenGIS Reference Model. OpenGIS Project Document 03-040, Open GIS Consortium, Inc. (2003)
4. Global Spatial Data Infrastructure Association: Online documentation. Retrieved May 2007 from http://www.gsdi.org/.
5. Open GIS Consortium, Inc.: OpenGIS Web Processing Service Implementation Specification. OpenGIS Standard 05-007r7, Open GIS Consortium, Inc. (2007)
6. Leidner, J.L.: Toponym Resolution in text: "Which Sheffield is it?". In: Proceedings of the the 27th Annual International ACM SIGIR Conference (SIGIR 2004), Sheffield, UK (2004) Abstract, Doctoral Consortium.
7. Jones, C.B., Purves, R., Ruas, A., Sanderson, M., Sester, M., van Kreveld, M., Weibel, R.: Spatial information retrieval and geographical ontologies an overview of the SPIRIT project. In: Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. (2002) 387 – 388
8. Baeza-Yates, R., Ribeiro-Neto, B.: Modern Information Retrieval. Addison Wesley (1999)
9. Zheng, G., Su, J.: Named entity tagging using an hmm-based chunk tagger. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. (2002) 209 – 219
10. Luaces, M.R., Paramá, J.R., Pedreira, O., Seco, D.: An ontology-based index to retrieve documents with geographic information. In: Proceedings of the 20th International Conference on Scientific and Statistical Database Management (SSDBM). Lecture Notes in Computer Science, Hong Kong (2008)
11. Open GIS Consortium, Inc.: Gazetteer Profile of WFS (WFS-G) Specification. Opengis project document, Open GIS Consortium, Inc. (2006)
12. Library, A.D.: Gazetteer. Retrieved September 2007 from http://www.alexandria.ucsb.edu/gazetteer/.
13. Getty, T.: Getty Thesaurus of Geographic Names. Retrieved September 2007 from http://www.getty.edu/research/conducting_research/vocabularies/tgn/.
14. Geonames: Gazetteer. Retrieved September 2007 from http://www.geonames.org.
15. Food and Agriculture Organization of the United Nations (FAO): Global Administrative Unit Layers (GAUL). Retrieved September 2007 from http://www.fao.org/geonetwork/srv/en/metadata.show?id=12691.
16. National Imagery and Mapping Agency (NIMA): Vector Map Level 0. Retrieved September 2007 from http://www.mapability.com.

17. Amitay, E., Har'El, N., Sivan, R., Soffer, A.: Web-a-where: geotagging web content. In: Proceedings of 27th annual international ACM SIGIR. (2004) 273 – 280
18. Rauch, E., Bukatin, M., Baker, K.: A confidence-based framework for disambiguating geographic terms. In: Proceedings of the HLT-NAACL 2003 workshop on Analysis of geographic references. (2003) 50 – 54
19. Lieberman, M.D., Samet, H., Sankaranarayanan, J., Sperling, J.: STEEWARD: Architecture of a Spatio-Textual Search Engine. In: Proceedings of the 15th ACM Int. Symp. on Advances in Geographic Infomation Systems (ACMGIS07). (2007) 186 – 193
20. Michaelis, C.D., Ames, D.P.: Evaluation and implementation of the ogc web processing service for use in client-side gis. Geoinformatica (2008)
21. Cepický, J.: Ogc web processing service and it's usage. In: Proceedings of the 15th International Symposium GIS Ostrava. (2008)
22. Open GIS Consortium, Inc.: OpenGIS Geographic Markup Language (GML) Encoding Standard. Opengis standard, Open GIS Consortium, Inc. (2007)
23. 52 North: Geoprocessing. Retrieved December 2007 from http://52north.org/.
24. Foerster, T., Stoter, J.: Establishing an ogc web processing service for generalization process. In: Proceedings of the Workshop of the ICA Commission on Map Generalisation and Multiple Representation. (2006)
25. Gamma, E., Helm, R., Johnson, R., Vlissides, J.: Design Patterns: Elements of Reusable Object-oriented Software. Addison-Wesley (1996)
26. FWTools: Open Source GIS Binary Kit for Windows and Linux. Retrieved September 2007 from http://fwtools.maptools.org.
27. Refractions Research: PostGIS. Retrieved June 2007 from http://postgis.refractions.net.
28. The JUMP Project: JUMP Unified Mapping Platform. Retrieved January 2008 from http://www.jump-project.org/.