Compound Geocoder: get the right position

Aneta Jadwiga Florczyk[1], Francisco Javier López-Pellicer[1], David Gayan[1],
Pedro Rodrigo-Cardiel[2], Miguel Ángel Latre[1], Javier Nogueras-Iso[1]
[1]Department of Computer Science and Systems Engineering,
University of Zaragoza
{florczyk,fjlopez,dgayan,latre,jnog}@unizar.es
[2]GeoSpatiumLab s.l. {prodrig@geoslab.com}

Abstract

Nowadays, there is no problem in accessing to geocoding providers but in choosing the proper one. The application requirements determine the selection of the service in a context where the user needs an answer and is not often interested in knowing where to find the right information.

This paper presents an architectural approach for compound geocoding Web services built above diverse Web Services with spatial content, especially gazetteer and geocoding services. The diversity in scenarios of geocoding usage requires an adaptive geocoding service. The proposed architecture satisfies this user requirement.

Keywords: Geocoding, Georeferencing, Web Service, Geo-QoS, Compound Architecture

1. INTRODUCTION

There is no single definition for the geocoding concept in the literature over time. In this paper, *geocoding* means the "act of turning descriptive locational data such as a postal address or a named place into an absolute geographic reference" [Goldberg et al, 2007]. Nowadays, the geocoding systems do not only deal with simple addresses but also with descriptions of relative locations [Hutchinson et al, 2005]. For example, there is no restriction in the spatial description, and it might be a point, a polygon or a three-dimensional geospatial entity [Beal, 2003].

The geocoding functionality can be provided by a batch application, a library or a Web Service, being the latter the most popular way. Currently, there are lots of geocoding services, each with its own characteristics. The main differences among the geocoding Web Services are determined by the type of content (addresses, points of interest, historical names, etc.) and the coverage (country, municipality or the world). Each service has a different QoS which is influenced by factors that depend not only on the typical Web Service QoS requirements (response time, reliability) but also on the quality of the spatial content.

In general, the services could be divided into three groups due to the Terms of Service (ToS): the paid access services, free services with restricted use, and free use services. This classification is independent of the service origin, which might be the private sector, public sector or open communities.

The private sector is the main provider of dedicated paid services. Users pay for these services because there is a guarantee of data and service quality. Free access service, provided by the public sector or open communities is free of fees but provides less quality than the dedicated one. Usually, the largest suppliers (e.g. Google, ViaMichelin or Yahoo) offer also free access to their address geocoding service with some restrictions (e.g. lower data quality). These services are accessible via company Web pages and/or APIs. Their ToS restrict the presentation (e.g. the license requires use of the supplier's visualization APIs), forbid the reuse of data, and might even detriment the overall quality of the applications built on base on that service by establishing access limits, such as rate limit or the maximum number of requests per day (see table 1).

Table 1: ToS comparison among some principal vendors of geocoding services.

Service	Access restrictions	Request limits (per day)	Other restrictions
Google Maps	Access code	15.000	Display, reuse, Request rate
Yahoo Maps	IP	5.000	Display, reuse
Via Michelin	Access code	1.000	Display, reuse
GeoNames	IP	50.000	Reuse

Geocoding requirements are in continuous evolution. For example, the support of mobile application demands supplementary characteristics for geocoding services. Location-based services (LBS) are the key for mobile applications that require the support of geocoding services for tracking of user location and the reverse geocoding. For example, the project Android[1] or GeoClue[2] are based on the geocoding and reverse geocoding services. Such geocoding service has to be adjusted to the requirements of mobile devices (e.g. need for energy, cellular network, access to the Web, or GPS availability).

The constant change of the requirements and the vast heterogeneity in geocoding services set up the problem of the supplier selection. For example, free geocoding Web services are appropriate for *geotagging*, i.e. the process of adding the geocoded information to any kind of media, local news or incidents (e.g. water supply shortage, planned roadwork), because such information does not require high quality geocoding services or spatial data. On the other hand, systems on which depend public health [Bonner et al, 2003], public security [Ratcliffe, 2004] or environmental services [Ratcliffe, 2001] require high quality

[1] http://code.google.com/intl/en/android/

[2] http://www.freedesktop.org/wiki/Software/GeoClue

services and data. For example, quickness and efficacy of fire-fighters depend on information such as the characteristics of the building in fire (e.g. number of floors, shape, location of entrances and the accessibility, nearby buildings) or the localization of fire hydrants.

There are many works in the context of the service discovery and selection. Some proposals in this area need prior service evaluation (e.g. rating agency [Sriharee, 2006] or user [Manikrao et al, 2005] pre-evaluation), but most of the works in this area use typical QoS features [Yu et al, 2005, Wang et al, 2006, Tsesmetzis et al, 2006] (e.g. end-to-end delay, overall cost, service reliability, availability). Recently, researchers are showing interest in services of geographic information [Lan et al, 2007, Fallahi et al, 2008] but they include only basic concepts (e.g. coverage) and do not exploit specific characteristics of geographic data in discovery and selection processes (e.g. reasoning based on coverage, quality of geographic objects).

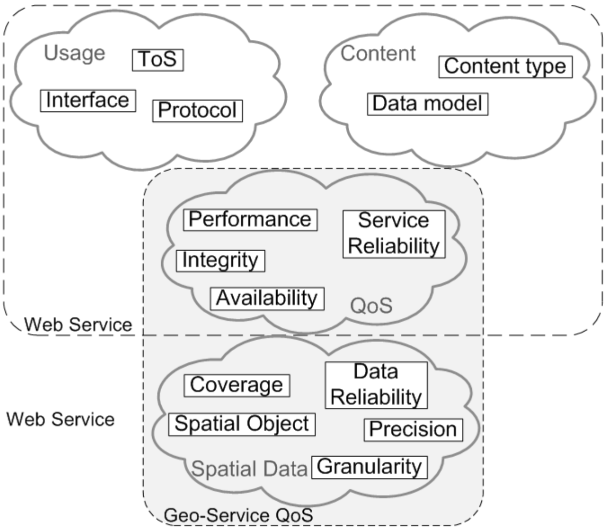
One of the principal issues raised in this paper is service selection which considers the particular characteristics of services that provide geographic information. Our proposal for compound geocoding architecture is a framework based on service selection. With the help of geo-ontologies, it allows building hybrid solutions composed of services specialized in different kinds of geographic information (e.g. geocoding, gazetteer or cadastral service). This approach may increase the flexibility and adaptability of applications. In case of the public services, it provides access to different services, national and local, in a transparent manner and ensures the use of updated data.

The rest of this paper is organized as follows. The second section describes the features of services which provide geographic data, which are the base for service selection task. Then, the proposed compound geocoding architecture is presented. The forth section presents the current geocoding solutions offered by public authorities in Spain, and the application of the proposed architecture for improving the geocoding service of Zaragoza city council. Finally, some conclusions are drawn and future work is outlined.

2. CHARACTERISTICS OF GEOCODING SERVICES

Each use case has its own expectation about geocoding functionality. To be able to offer a proper geocoding Web Service, it is necessary to identify its principal characteristics, which would allow users or machines to compare and evaluate them. For this reason, we attempt to distinguish common features of the Web Services with spatial content (in short, *georeferencing Web Services*). Figure.1 presents the properties that have been recognized. In this paper, we will only focus on the features determined by the spatial content, therefore, the general term, *georeferencing service*, is used.

Figure 1: The characteristics of the Georeferencing Web Service.



The main features of each georeferencing service are the spatial **coverage**, the **content type** and the **type of spatial object**. The first two of them are always given by the provider as metadata or are indicated implicitly by the name of the service. The coverage defines the area in which offered data are located. This information can be provided by means of a geographic description based on coordinates (e.g. minimum bounding rectangle), by a concept defined in administrative unit ontology or in a thesaurus, by a place name from a gazetteer or by free text. The type of content strictly depends on the georeferenced types of features. The type of spatial object indicates the list of provided types of spatial object, such as point, polygon or 3D entity. For example, the National Cadastre Service of Spain[3], as the name of the service indicates, has the coverage of Spain and offers coordinates of parcels. Google Maps has world coverage and its type of content is street addresses geocoded via point, which is provided by the service description.

It is possible to obtain two additional indicators (of range 0 - 1) from the analysis of spatial data: the **data reliability** and the **precision**. The data reliability indicates the capacity of representation of elements of physical world in the content. The service that offers all elements of the real world has a data reliability value equal to '1'. The indicator of precision informs about the average *positional error* [Christen, 2004] of the whole dataset. It is important to note that this indicator may be influenced by the difference between the provided spatial object

[3] http://ovc.catastro.minhac.es/

and the searched one. For example, when using cadastral data for the address geocoding, there will be a decrease in spatial data precision.

In addition, the indicator of the level of the detail might be determined on base of spatial data analysis. Usually, the reliability of a street address service varies in function of the area relevance, for example a new suburb might be even omitted. Such a feature might be indicated by the **granularity**.

Another feature is the **result accuracy**. It should not be misinterpreted as *data accuracy* which is the term commonly used in literature to describe spatial data accuracy. Result accuracy is defined for each source and is extracted from the analysis of the source data model and the search data model. It indicates the level of overlapping of the search and data models. For example, *address search model* (i.e. the search data model for the address search) in Spain should contain, at least, ‘province’, ‘municipality’, ‘zip code’, ‘street name’ and ‘portal’ [Walker, 2008]. There are different address providers in Spain but each adapts this general model to its business requirements. The data model of the National Cadastre Service of Spain keeps all elements of address search model, so the *result accuracy* is 'portal'. On contrary, the Concise Gazetteer Service of National Institute of Geography[4] (IGN Concise Gazetteer) does not provide street data and the *result accuracy* will be 'municipality'.

3. ARCHITECTURE

A compound geocoding architecture can use different geographic information services (e.g. gazetteer, geocoding and cadastral services). The functionality of the system built upon this architecture strictly depends on values of the service features introduced in the previous section. Therefore, the proper evaluation of each source is crucial for behaviour of whole system. It could be done on base of several tests which procedures have to be conformed with the service feature description. For definition of the model of searched information (*domain data model*) and the data model of each provider only one domain ontology should be applied. This will facilitate the data integration through the mapping of data models, and the estimation of the *result accuracy* for each source. Application of the ontologies about administrative units [López-Pellicer et al, 2008] will permit proper use of the *coverage* feature.

The main elements of the compound geocoding architecture (see figure. 2) are:

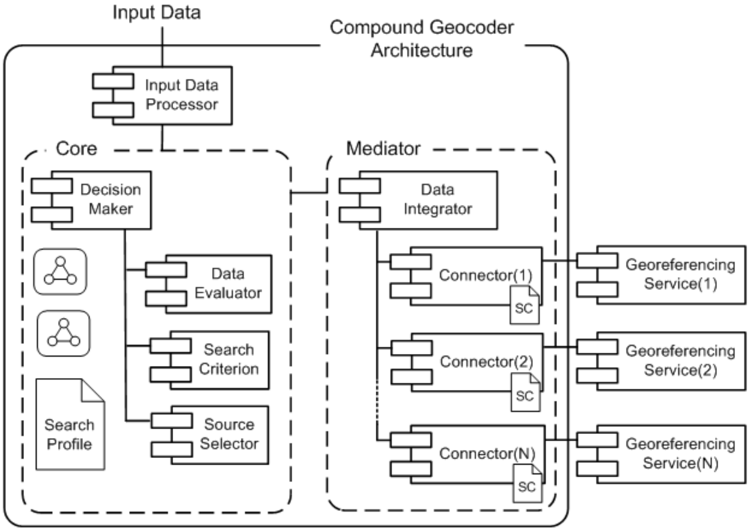
- **Input data processor component.** This component is responsible for performing the pre-processing of text from the *input data*. The steps in this

[4] http://www.idee.es/show.do?to=pideep_gazetteer_search.ES

phase of geocoding are common techniques among geocoders [Hutchinson et al, 2005]: cleaning, parsing and standardizing.

- **Core component.** The *core* component is responsible for the whole process of source selection (*source selector* component) and result data evaluation (*data evaluator* component). This process is based on rules implemented in the *decision maker* component. These rules apply search criterions (*search criterion*) and service characteristics (*SCs*) associated with each connector and obtained from previous source tests. Also the search criterions are defined in the terms of the service features described in the section 2 of this paper, and they are provided by the application context (*search profile*) and/or the user requirements (part of the *input data*).
- **Mediator component.** It consists of pluggable service *connectors* and a *data integration* component. The main advantage of the *connectors* is the abstraction from communication protocol, invocation styles or interfaces used. The data integration component is responsible for data harmonization which consists of data mapping and coordinate transformation if necessary.

Figure 2: Compound Geocoding Service Architecture (*SC – Service Characteristics*).



This architecture allows geocoding different types of named places and, due to complementing data from one source with data from others, it improves the data reliability and the data precision. It also gives the user more freedom in deciding the search strategy. Due to the access to several services, the search strategy may select the best response of the entire system, the best answer for each source or the best answers from a chosen source. In addition, as the details of implementation are hidden in the mediator, this allows incorporating any type of georeferenced data.

4. USE CASE: GEOCODING FOR PUBLIC AUTHORITIES IN SPAIN

Public authorities need geocoding tools for diverse applications. On the one hand, the official Web pages publish geocoded information and become an important source of centralized knowledge for citizens. In this case, result data obtained via the usage of the geocoding services of low quality are satisfactory. The same services are not appropriate for the urban management systems that support the public services of civil protection (heath centers, firefighters). Especially, the systems that manage emergency situation (e.g. floods) require from Geographic Information Systems up-to-date data of high quality (state coverage, reliability, high precision).

In the public sector in Spain, the task of provide spatial data appropriated for urban management systems is rather tough. The responsibility for maintenance of the urban content is decentralized, however the decision about publication of spatial data of new urban areas is taken earlier at state level, which is reflected in central government datasets. As a result, local administration has data of the highest precision but of lower reliability, when, compared to the services that use the content of central government, e.g. the Service of Cadastre Data of Spain.

There are some proposals of geocoding services supported by public authorities at state level, e.g. the National Cadastre Service of Spain or CartoCiudad Service[5]. The first one is characterized by the best reliability among the other existing geocoding services at state level in Spain, but, due to the fact, that its content type is *parcel*, the precision for address geocoding is decreased. The CartoCiudad combines the spatial contents provided by diverse public institutions (i.e. General Direction of Cadastre, Postal Office, National Institute of Statistic, and General Direction of National Institute of Geography) and from local authorities. The principal disadvantage of the CartoCiudad services is lack of the update procedure and gaps in coverage. Additionally, both of this proposals share the problem of the uncomfortable search as it is necessary to indicate the search area (province and municipality).

Most common are the Web Services offered by local authorities at their portals, e.g. the Street Data Web Service of Zaragoza city council (IDEZar SG). These services are characterized by a high data precision although granularity may vary depending on the area (i.e. urban centre, village) and, usually, there are lacks in coverage, e.g. motorways or new urban zones.

4.1. Implementation and results

[5] http://www.cartociudad.es

In Spain there are several alternatives of geocoding services in public sector with disparate QoS as described above. The compound geocoder architecture can improve the results of geocoding by mixing different kinds of geocoder services according to application requirements. This section presents an application of the proposed architecture to implement the geocoding service for the municipality of Zaragoza to provide better functionality that the geocoding system based on the IDEZar SG.

The main characteristic of the IDEZar SG is that it covers only the urban area of the city. One of the best providers for rural areas in Spain is the geocoder from the National Cadastre Web Service. However, both systems fail when the address identifies a flat on a condominium or is an address on a motorway. A commercial provider, such as Google Maps, can resolve (or interpolate) the location for these kind of addresses. Finally, none of the previously cited providers is able to provide vernacular names as they rely on official names. However, these names are maintained by the National Institute of Geography and are accessible via IGN Concise Gazetteer. Theoretically, summing up the advantages of each of these services it should be possible to obtain better results. According to this reasoning, the implementation uses all these services (see table 2).

Table 2: Differences among the Spanish selected georeferencing services

Service	Protocol	Interface	Standard Data Model	CRS
IDEZar SG	SOAP (HTTP)	Open (SRW)	No (exclusive of the Zaragoza City Council)	EPSG:23030
GoogleMaps	GET (HTTP)	Closed	Some (e.g. xAl)	EPSG:4326
National Cadastre Web Services	SOAP (HTTP)	Open	Yes (National)	Several
IGN Concise Gazetteer	GET/POST (HTTP)	Open (WFS)	Yes (National)	EPSG:4230

The compound geocoder logic depends on the QoS characteristics of each third party services. Therefore, the first step in the compound service development is to evaluate each service provider to infer its QoS parameters about coverage, content, precision, reliability and accuracy. The details are shown in table 3.

Table 3: Characteristics of the selected services in the context of address search in municipality of Zaragoza. The values are obtained via series of provider tests according to the description presented in section 3.

Service	Coverage	Content Type	Result Accuracy	Reliability	Precision
IDEZar SG	Municipality of Zaragoza	Street Data	Portal	0.98	1.00
GoogleMaps	World	Address	Portal	0.96	0.99

Service	Coverage	Content Type	Result Accuracy	Reliability	Precision
National Cadastre Web Services	Spain	Parcel	Portal	1.00	0.95
IGN Concise Gazetteer	Spain	Geographic Features	Locality	1.00	N/A

The compound geocoder service admits queries in the free form text (e.g. “C/ mayor, 20", “coso", “plz España"). Then, it identifies an address pattern and transforms the string query in a query request to a general address model derived from the configuration of the compound geocoder. Depending on the query and the configuration, one or several services are queried. Each service connector knows how to translate the query to the particular query model the service, to query and to translate its answer to the general address model. The query results are filtered first in each connector and then in the compound geocoder logic module.

Today, there is an instance of this service applied to the management of addresses in the Zaragoza city council. This implementation uses IDEZar SG as its main provider, however, the use of other data sources has increased the perceived quality of the geocoded addresses.

5. CONCLUSION AND FUTURE WORK

This paper presents the problem of geocoding service selection and an approach to this problem based on a compound geocoding architecture. We have detected that the main issue in this approach is the selection and measuring of selection indicators. The principal design goals of the proposed architecture are flexibility and extension facility. The advanced application of geocoding, i.e. geolocating [Hutchinson et al, 2005] which deal with freeform textual description of location, demands spatial information of wide range of types. Therefore, the compound approach seems to be suitable for the architecture model of a geolocating system.

Future work will be focused on the development of an appropriate methodology for the evaluation of georeferencing services, which would require a metadata model for describing these services. Next steps will require the use of statistical methods to evaluate the responses and to improve the measures of granularity and precision. With this information, it would be feasible to develop techniques for comparing different services of georeferencing. As the principal disadvantage of the proposed approach is the high cost of the implementation of the connectors, future work will focus on this issue as well. There will be effort dedicated to employ the recent advances in the research on service interoperability, ontology alignment and reasoning

ACKNOLEDGMENTS

This work has been partially supported by the Spanish government through the projects “España Virtual” (ref. CENIT 2008-1030) and TIN2007-65341; the Aragon government thought the project PI075/08; the National Geographic Institute of Spain (IGN); and GeoSpatiumLab s.l.. The work of Aneta Jadwiga Florczyk has been partially supported by a grant (ref. AP2007-03275) from the Spanish government.

REFERENCES

Beal, J.R. (2003): “Contextual Geolocation: A Specialized Application for Improving Indoor Location Awareness in Wireless Local Area Networks”, *Proceedings of the 36th Annual Midwest Instruction and Computing Symposium, 2003, Duluth, Minnesota, USA*.

Bonner, M.R., Han, D., Nie, J., Rogerson, P., Vena, J.E. and Freudenheim, J.L. (2003). Positional Accuracy of Geocoded Addresses in Epidemiologic Research, *Annals of Epidemiology*, 17: 464-470.

Christen, P., Churches, T. and Willmore, A. (2004). “A Probabilistic Geocoding System Based on a National Address File”, *Proceedings of the 3rd Australasian Data Mining Conference, Decembre, 2004, Cairns, Queensland, Australia*.

Fallahi, G.R., Mesgari, M.S., Rajabifard, A. and Frank, A.U. (2008). A Methodology Based on Ontology for Geo-Service Discovery, *World Applied Sciences Journal*, 3(2): 300-311.

Goldberg, D.W., Wilson, J.P. and Knoblock C.A. (2007). From Text to Geographic Coordinates: The Current State of Geocoding. *Journal of the Urban and Regional Information Systems Association*, 19(1): 33-46.

Hutchinson, M. and Veenendall, B. (2005). “Towards using intelligence to move from geocoding to geolocating”, *Proceedings of the 7th Annual URISA GIS in Addressing Conference, August 2005, Austin, TX, USA*.

Lan, G. and Huang, Q. (2007). “Ontology-based Method for Geospatial Web Services Discovery”, *Proceedings of the International Conference on Intelligent Systems and Knowledge Engineering (ISKE 2007), October 15-16, 2007, Chengdu, China*.

López-Pellicer, F.J., Florczyk, A.J., Lacasta, J., Zarazaga-Soria, F.J. and Muro-Medrano, P.R. (2008). “Administrative Units, an Ontological Perspective”, *Proceedings of the ER 2008 Workshops (CMLSA, ECDM, FP-UML, M2AS, RIGiM, SeCoGIS, WISM) on Advances in Conceptual Modeling: Challenges and Opportunities, 2008*, pp. 354-363. Springer.

Manikrao, U.S. and Prabhakar, T. (2005). “Dynamic Selection of Web Services with Recommendation System”, *Proceedings of the International*

Conference on Next Generation Web Services Practices, 2005, pp. 117. IEEE Computer Society.

Ratcliffe, J.H. (2001). On the accuracy of TIGER type geocoded address data in relation to cadastral and census areal units, *International Journal of Geographical Information Science*, 15(5): 473-485.

Ratcliffe, J.H. (2004). Geocoding crime and a first estimate of a minimum acceptable hit rate, *International Journal of Geographical Information Science*, 18(1): 61-72.

Sriharee, N. (2006). “Semantic Web Services Discovery Using Ontology-Based Rating Model”, *Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence, 2006,* pp. 608-616. IEEE Computer Society.

Tsesmetzis, D.T., Roussaki, I.G., Papaioannou, I.V. and Anagnostou, M.E. (2006). “QoS Awareness Support in Web-Service Semantics”, *Proceedings of the International Conference on Internet and Web Applications and Services (AICT-ICIW'06), 2006*, pp. 128. IEEE Computer Society.

Walker, R. (2008). “A General Approach to Addressing”, *Proceedings of the ISO Workshop on Address Standards: Considering the issues related to an international address standard, May, 2008, Copenhagen NV, Denmark*.

Wang, X., Vitvar, T., Kerrigan, M. and Toma, I. (2006). ”A QoS-aware Selection Model for Semantic Web Services”, *Proceedings of the 4th International Conference on Service Oriented Computing, Chicago, IL, USA, December 4-7, 2006*, pp. 390-401. Springer.

Yu, T. and Jay Lin, K. (2005). “Service Selection Algorithms for Composing Complex Services with Multiple QoS Constraints”, *Proceedings of the 3rd International Conference on Service Oriented Computing, Amsterdam, The Netherlands, December 12-15, 2005*, pp. 130-143. Springer.