Enabling Geolocating via Ontologies[12]

Aneta J. Florczyk [1], Francisco Javier López-Pellicer [1], Rodolfo Rioja [2], Javier Nogueras-Iso [1], Francisco Javier Zarazaga-Soria [1]

[1] Computer Science and Systems Engineering Department
University of Zaragoza
C/ María de Luna, 1, 50.018, Zaragoza, Spain
{florczyk,fjlopez,jnog,javy}@unizar.es
[2] GeoSpatiumLab
C/ Carlos Marx, 4, local der., 50.015 Zaragoza, Spain

Abstract. This paper presents an architecture for compound geocoding Web services built on diverse Web services of geographic information, especially the gazetteer and geocoding services. The proposed architecture uses ontologies as a key element to source selection and data integration. This approach intends to satisfy the user needs for an adaptive geocoding service of general purpose, which might be a core component of geolocating service architecture.

Keywords: Geocoding, Ontology, Geolocating, Service Architecture, Web Service.

1 Introduction

Public authorities charged with the responsibility of dealing with the maintenance of public infrastructure require urban management systems able to manage large amount of resources. Especially in case of emergency situations (e.g. floods, fires), an effective management of resources included in the emergency plan involves a centralized and coordinated management system. These systems often require support from Geographic Information Systems for geocoding street addresses. Thus, there is a need for up-to-date data of high quality (e.g. coverage at the national level, high precision, reliability). However, assurance of data quality is a hard task. The decentralized public administration is responsible for data maintenance, and the spatial information changes continuously, especially in the urban area (e.g. cadastral data).

[12] This work has been partially founded by the Spanish Government through the projects "España Virtual" (ref. CENIT 2008-1030) and TIN200765341, and by the Government of Aragon through the project PI075/08. The work of Aneta J. Florczyk has been partially supported by an FPU Grant from the Spanish Ministry of Science and Innovation (ref. AP2007-03275). The work of Rodolfo Rioja has been partially supported by the Spanish Ministry of Science and Technology (Torres Quevedo program ref. PTQ06-2_0790).

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* [1]. This includes relative description of location as input data [2], and any geographic description as the output, i.e. a point, a polygon or three-dimensional geospatial entity [3].

The most prevalent way of providing geocoding functionality is a geocoding service. Nowadays there are lots of geocoding services with diverse characteristics determined by the quality of geo-service, the provided data and the terms of service (ToS). Today there is no problem in using online geocoding providers but to find and to choose a proper provider.

The main differences among geocoding Web services are caused by the provided data, i.e. the type of content (e.g. point of interest, address). The quality of geo-service is influenced by factors that depend on the typical QoS requirements (e.g. response time, reliability) and the quality of the spatial data.

The geocoding services also differ in their ToS. In general, regardless of service origin, which might be private sector (e.g. Google, ViaMichelin or Yahoo), public sector or volunteer communities (e.g. GeoNames[13], Megalithic Portal[14], Geograph British Isles[15], OpenStreetMap[16]), the services may be divided into three groups due to their ToS: paid access services, free of charge services with restricted use, and services of free use.

The private sector offers the ad-hoc designed paid services, which guarantee the quality of data and service. Free services offered by the public sector or open communities provide less quality than the dedicated ones. Usually, the largest providers offer free access to their address geocoding services with lower quality and some use restrictions. Their ToS restrict the presentation (e.g. the license requires use of the supplier's visualization APIs), prohibit the reuse of data, and have influence on the quality of applications based on that service, i.e. establishing limits, such as rate limit or the maximum number of requests per day.

Finally, we should also consider that new types of geocoding service applications such as support of mobile application, demand supplementary characteristics. Location-based services require the support of geocoding services for tracking of user location and the reverse geocoding at the operation system level (e.g. the Android[17] or GeoClue[18] projects). The availability and capabilities of these services have to be adjusted to the requirements of mobile devices (battery life, cellular network, access to the Web, or GPS availability).

The choice of service is determined by the use case. Free geocoding Web services are appropriate for "geotagging" (i.e. the process of adding the geocoded information to any kind of media) the 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, the systems on which depend public health [4],

---

[13] http://www.geonames.org/
[14] http://www.megalithic.co.uk/
[15] http://www.geograph.org.uk/
[16] http://www.openstreetmap.org/
[17] http://code.google.com/android/
[18] http://www.freedesktop.org/wiki/Software/GeoClue

public security [5] or environmental services [6] require high quality of service and data. For example, quickness and efficacy of fire-fighters depend on the information they possess 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.

The vast heterogeneity of geocoding services and the specific features of geographic data set up the open problem of provider selection. There are many works in the context of the service discovery and selection. Some of proposals need prior service evaluation (e.g. rating agency [7] or user [8] pre-evaluation), but most of the works in this area use typical QoS features [9 – 11] (e.g. end-to-end delay, overall cost, service reliability, availability). Recently, researchers show interest in services of geographic information [12, 13] 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).

Our proposal for compound geocoding architecture attempts to be a first step in the research on the problem of the geocoding service selection via geo-ontologies. This approach uses a framework that allows building hybrid solutions composed of different services which provide geographic information (e.g. geocoding, gazetteer or cadastral service). Each solution applies administrative unit ontology for (i) provider selection and (ii) data integration. 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.

Additionally, integration of multiple geocoding sources and/or services is one of the open issues of research in field of geolocating, which will be presented in the section 2. Our approach aims to develop the fundaments for designing core components of geolocating services.

The paper is organized as follows. Section 2 presents issues that appear in the context of geolocating. The next section reviews the characteristics of geocoding services. Section 4 presents the architecture of the compound geocoder. The following section shows the state in geocoding of urban management applications in Spain, and presents the implementation of the application. Finally, there are presented conclusions and the description of the future work.

2 Geolocating within Urban Systems

Existing geocoding services are generally limited to assign a geographic coordinate to an absolute location such as a street address. However, urban management systems need to geocode location description in a more flexible way. For example, it is common that a citizen who calls an emergency centre does not know the address where he/she is and the descriptive information that provides is ambiguous or even confusing (e.g. "100 meters of a memorial statue in the park, which is situated by Colon street"). In such a situation the user needs to be "geolocated".

Hutchinson and Veenendaal [2] define "geolocating" as the evolution of the geocoding process that permits to assign valid geographic codes to freeform textual descriptions of locations. Conventional geocoding services that work with absolute locations will not be able to determine the coordinates of the place of incident. The best solution is a geolocating service, which might interpret correctly the relative location (e.g. distance, direction), with landmarks (e.g. park, railway station) and features (e.g. coffee shop). Hutchinson and Veenendaal [2] enumerate the following elements that contribute to the concept of geolocating: query syntax flexibility, user interaction, user context consciousness and complex site representation. It is important to stress that the authors argue for a new methodology, however, they do not include details of how one would implement it.

The principal design goals of the architecture proposed in this paper are flexibility and extension facility. Due to the fact that the geolocating process demands spatial information of wide range of types, the compound approach seems to be suitable for the architecture model of a geolocating system.

3 Characteristics of Geocoding Service

The proposed architecture allows the service selection according to the use case requirements. Therefore, the proper description of each source is vital for the behaviour of the whole system. Fig.1 presents the properties that are used in the strategy for provider evaluation, which is essential for the service selection and the decision making process.

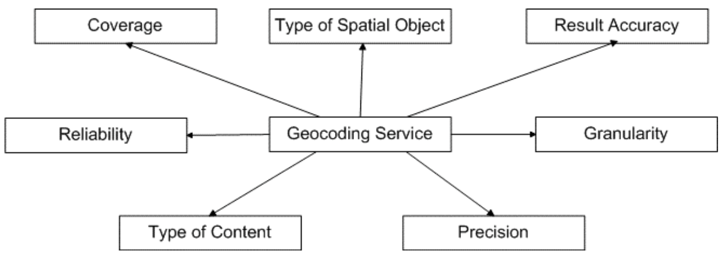

Fig. 6. Characteristics of geocoding service.

The main features of each geocoding service are the spatial *coverage*, the *type of content* and the *type of spatial object*. The first two of them are always given by the provider or are indicated by the name of the service. The first one defines the area in which offered data are situated and, usually, this area corresponds with the jurisdictional geographic object, the concept defined in administrative unit ontology. The second feature, the type of content, strictly depends on the georeferenced types of features. The last one indicates the list of provided types of spatial object, such as point, polygon or 3D entity. For example, the cadastral service of Spain (Servicio de Catastro de España), 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 address geocoded via point, which is provided by the service description.

Another feature is the *result accuracy* and should not be misinterpreted as “data accuracy”, the term commonly used in literature to describe spatial data accuracy. Result accuracy is defined for each source and is derived 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. Let us define the search data model for the address search in Spain. According to the general approach [14], it will contain at least province, municipality, zip code, street name and number. The data model of the cadastral service of Spain maintains this minimum (the accuracy is 'number' or simply 'address'), but the data model of the geographic information service does not have streets and the search accuracy will be 'municipality'.

From the analysis of spatial data, it is possible to obtain two additional indicators (of range 0 - 1): **reliability** and **precision**. The reliability indicates the capacity of representation of elements of physical world in the content. The service that offers all elements of real world has a reliability value equal to '1'. The indicator of precision informs about the average *positional error* [15] of the whole dataset. It is important to note that this indicator may be influenced by difference between the provided spatial object and the search one. For example, using cadastral data for the address geocoding, there will be a decrease in spatial data precision.

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 an additional indicator, called **granularity**, which respects the level of detail.

4 Architecture Overview

The compound geocoding architecture uses different geographic information services such as gazetteer services, geocoding services and cadastral services. Each source has to be properly described as the descriptors values are the clues for source selection, which determines the adequate functionality of the system. The result accuracy is estimated via comparison of the search model and the data model of each source. This requires a well defined model of searched information, possibly with an administrative unit ontology [16]. Also, the data model of each provider has to be described with the help of the administrative unit ontology that permits the mapping of data models.

The main elements of the compound geocoding architecture (see Fig. 2) are an input data processor component, a core component and a mediator component. The input data processor performs the pre-processing of input data. The steps in this phase of geocoding are common techniques among geocoders [2]: cleaning, parsing and standardizing.

The core element is responsible for the whole process of source selection and data evaluation. The rules used by the source selector component and the data evaluator component are implemented in the decision maker component. The rules apply the search criteria that are defined in the same terms as the source characteristics. The search criteria are provided by the application context and/or the user requirements. The first evaluation of the results is done in the mediator and then the data evaluator performs posterior evaluation according to the predefined criteria of search and the characteristics of the content (e.g. data precision, content type).

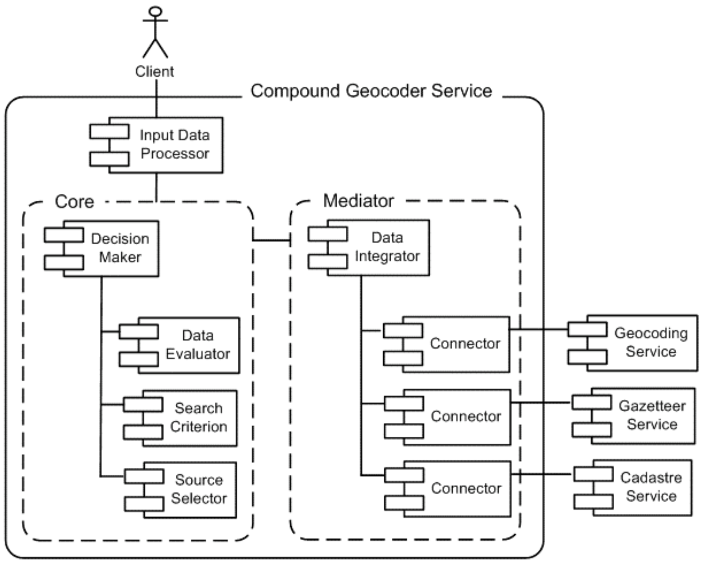

Fig. 7. The main components of the compound geocoding architecture.

The mediator component 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 harmonisation which consists of data mapping and coordinate transformation if necessary.

This architecture allows geocoding different types of named places and due to complementing data from one source with data from the others; it improves the 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.

5 Address Geocoding in Urban Management of Spain

In Spain, in public sector, there are several alternative services in the field of geocoding of street addresses, such as cadastral services, CartoCiudad[19], Geopista[20],

[19] http://www.cartociudad.es/ignnomenclator/
[20] http://www.geopista.com/

or services provided by local authorities (e.g. IDEZar Street Gazetteer Service[21] of the Zaragoza municipality, IDEZarSG).

The national cadastral service is offered by the *Dirección General del Catastro*[22], DGC (an administrative registry which belongs to the Ministry of the Treasury). The service provides updated data and spatial representation of parcels of high quality. However, the spatial data precision is lower for address search (i.e. parcel centroid), and the data access is uncomfortable and might result confusing (definition of the province and then the municipality).

The alternative providers are Geopista or CartoCiudad. The first one uses data that comes from the DGC. Although the second one offers services with data coming from the harmonisation and integration of digital data produced by several official suppliers, the main source of CartoCiudad is again DGC. As both of them use stand-alone data bases, in a short time there will be problems of data reliability. Moreover, they do not provide any update procedures.

The other alternative might be the services of street data provided by local administration, such as the IDEZarSG. These services are characterized mainly by high precision. However, the level of granularity differs depending on the area (e.g. urban centre, town) and there are areas without coverage, such as new highways or suburban areas.

In Spain, the geocoding services in the public sector are not adequate for urban management systems, for there is lack of update procedures, reliability, and precision or even simplicity of use (step strategy). One of the solutions to this problem might be the data integration from different sources via compound service architecture.

Table 1. 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.

	Coverage	Content Type	Result Accuracy	Reliability	Precision
IDEZarSG Service	Municipality of Zaragoza	Street Data	Number	0.98	1
GoogleMaps Service	World	Address	Number	0.96	0.99
Cadastral Service of Spain	Spain	Parcel	Number	1	0.95
IGN Concise Gazetteer Service	Spain	Geographic Features	Locality	1	not applied

The implementation of a compound service for address street geocoding uses the following geocoding services: the IDEZarSG Service, the GoogleMaps Service, the Cadastral Service of Spain, and the Concise Gazetteer Service of the National Geographic Institute (*Instituto Geográfico Nacional*, IGN). The characteristics of each one are described in the table 1.

[21] http://idezar.unizar.es/SRW/client.html

[22] http://www.catastro.meh.es/

The first step is definition of a search model. The street address model is described by the administrative unit ontology that permits assignation of specific legal roles to administrative unit entities, in this case, the address role. In Spain, the address is defined via province, municipality, zip code and address line, i.e. the compound of street name, street type, number with additional information if necessary, e.g. floor and door. As the zip code is not supported by any of the plug-in sources, the data disambiguation requires extending the official address model with locality and suburb fields.

The domain data model and the source data models are defined to satisfy the address search model. The administrative unit ontology is used to define each model that helps in the mapping between them as Fig.3 and Fig.4 show.

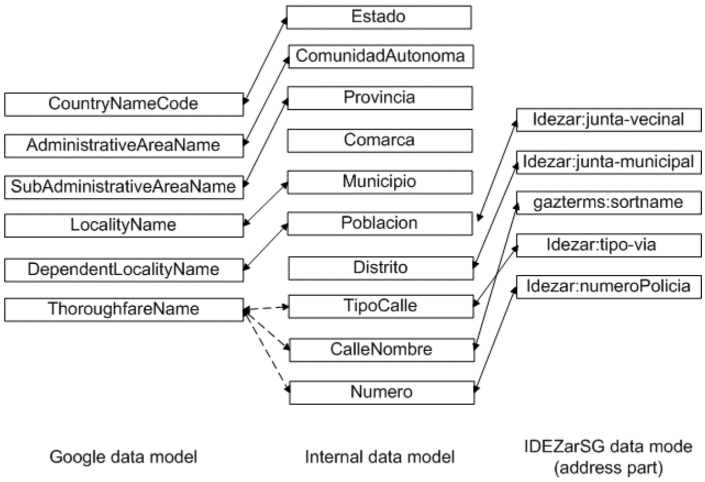

Fig. 8. Data model mapping between the domain model, the Google data model and the IDEZarSRW

The proposed architecture uses several geographic information suppliers to obtain the highest service quality. The features of the compound service allow eliminating the problems indicated in the previous section, such as the need for data update, manual step strategy of search, or omissions in coverage or granularity.

Currently, an instance of this service is used as the geocoding service of the official web of the Zaragoza municipality council[23]. The main data source is the IDEZarSG service but the result data is complemented with the data from other sources, which improves the reliability of the service.

[23] http://idezar.zaragoza.es/callejero/index2.jsp

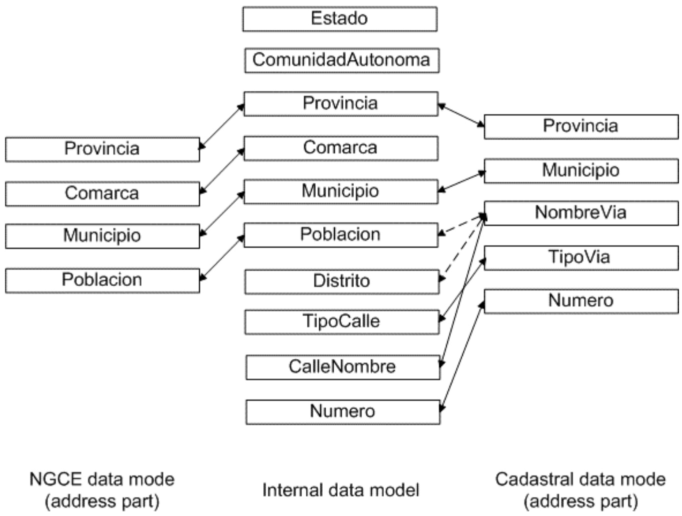

Fig. 9. Data model mapping between the domain model, the address part of the MNE data model and the cadastral service data model

6 Conclusions and Future Work

The compound geocoding architecture ensures the improvement of geocoding results (e.g. reliability, spatial data quality) thanks to the use of different geographic information suppliers. The use of multiple sources involves the development of ontology for describing the geocoding sources. Moreover, the use of ontologies yields an advanced architecture in terms of extensibility, flexibility and adaptability.

The framework for geocoding service selection permits to develop a methodology to geocode diverse categories of data types (e.g. spatial features, points of interest), which is an essential functionality of a geolocating service. In this context, there is a strong demand for a generic search model that involves the formalization of search model. Additionally, knowledge integration system adds new issues to the gazetteer concept. Usually, the spatial data of gazetteers is obtained from geocoding processes and comes from diverse data sources. Nevertheless, data models of gazetteers do not offer information about spatial data accuracy and its origins.

The principal disadvantage of this approach is the need of implementation of a pluggable connector for each source. Although the implementation permits adapting to changes, the effort of connector creation is significant.

As a future work, we will study the current techniques in Web service interoperability that permits automatic discovery and use of the geographic data providers. An effort will be made to create the formal definition of the ontology for geo-service description that includes the features enumerated in this paper. As the next step, we will apply the ontology reasoning for improvement of the service selection.

References

1. Goldberg, D.W., Wilson, J.P., Knoblock, C.A: From Text to Geographic Coordinates: The Current State of Geocoding. URISA Journal. 19(1), pp. 33--46 (2007)
2. Hutchinson, M., Veenendall, B.: Towards Using Intelligence to Move from Geocoding to Geolocating. In: Annual URISA GIS in Addressing Conf., Austin, TX. (2005)
3. Beal, J.R.: Contextual Geolocation, A Specialized Application for Improving Indoor Location Awareness in Wireless Local Area Networks. In: 36th Annual Midwest Instruction and Computing Symposium, Duluth, Minnesota. (2003)
4. Bonner, M.R., Han, D., Nie, J., Rogerson, P., Vena, J.E., Freudenheim, J.L.: Positional Accuracy of Geocoded Addresses in Epidemiologic Research. Annals of Epidemiology. 17, pp. 464--470 (2003)
5. Ratcliffe, J.H.: Geocoding crime and a first estimate of a minimum acceptable hit rate. Int. Journal of Geographical Information Science. 18, pp. 61--72 (2004)
6. Ratcliffe, J.H.: On the accuracy of TIGER type geocoded address data in relation to cadastral and census areal units. IJGIS. 15, pp. 473--485 (2001)
7. Sriharee, N.: Semantic Web Services Discovery Using Ontology-Based Rating Model. In: WI '06. IEEE Computer Society, pp. 608--616 (2006)
8. Manikrao, U.S., Prabhakar, T.: Dynamic Selection of Web Services with Recommendation System. In: NWESP '05. IEEE Computer Society, 117 (2005)
9. Yu, T., Jay Lin, K.: Service Selection Algorithms for Composing Complex Services with Multiple QoS Constraints. In: 3rd Int. Conf. on Service Oriented Computing, pp. 130--143 (2005)
10. Wang, X., Vitvar, T., Kerrigan, M., Toma, I.: A QoS-aware Selection Model for Semantic Web Services. In: ICSOC, Springer, pp. 390--401 (2006)
11. Tsesmetzis, D.T., Roussaki, I.G., Papaioannou, I.V., Anagnostou, M.E.: QoS Awareness Support in Web-Service Semantics. In: Int. Conf. on Internet and Web Applications and Services. (2006)
12. Fallahi, G.R., Mesgari, M.S., Rajabifard, A., Frank, A.U.: A Methodology Based on Ontology for Geo-Service Discovery. World Applied Sciences Journal 3(2) pp. 300--311 (2008)
13. Lan, G., Huang, Q.: Ontology-based Method for Geospatial Web Services Discovery. In: Int. Conf. on Intelligent Systems and Knowledge Engineering. (2007)
14. Walker, R.: A General Approach to Addressing. In: ISO Workshop on Address Standards: Considering the issues related to an international address standard. (2008)
15. Christen, P., Churches, T., Willmore, A.: A Probabilistic Geocoding System Based on a National Address File. In: Australasian Data Mining Conference, Cairns. (2004)
16. López-Pellicer, F.J., Florczyk, A.J., Lacasta, J., Zarazaga-Soria, F.J., Muro-Medrano, P.R.: Administrative Units, an Ontological Perspective. In: ER Workshops. (2008) To appear in: LNG&C