The current issue and full text archive of this journal is available at
www.emeraldinsight.com/1468-4527.htm

Discovering geographic web services in search engines

Francisco J. Lopez-Pellicer, Aneta J. Florczyk, Rubén Béjar, Pedro R. Muro-Medrano and F. Javier Zarazaga-Soria
Advanced Information Systems Laboratory (IAAA), Department of Computer Science and Systems Engineering, Universidad Zaragoza, Zaragoza, Spain

Received 15 September 2010
Accepted 6 February 2011

Abstract

Purpose – There is an open discussion in the geographic information community about the use of digital libraries or search engines for the discovery of resources. Some researchers suggest that search engines are a feasible alternative for searching geographic web services based on anecdotal evidence. The purpose of this study is to measure the performance of Bing (formerly Microsoft Live Search), Google and Yahoo! in searching standardised XML documents that describe, identify and locate geographic web services.

Design/methodology/approach – The study performed an automated evaluation of three search engines using their application programming interfaces. The queries asked for XML documents describing geographic web services, and documents containing links to those documents. Relevant XML documents linked from the documents found in the search results were also included in the evaluation.

Findings – The study reveals that the discovery of geographic web services in search engines does not require the use of advanced search operators. Data collected suggest that a resource-oriented search should combine simple queries to search engines with the exploration of the pages linked from the search results. Finally the study identifies Yahoo! as the best performer.

Originality/value – This is the first study that measures and compares the performance of major search engines in the discovery of geographic web services. Previous studies were focused on demonstrating the technical feasibility of the approach. The paper also reveals that some technical advances in search engines could harm resource-oriented queries.

Keywords Search engines, Geographic information, Web services, Discovery, Open geospatial consortium, Information retrieval, Geographic information systems

Paper type Research paper

Introduction

Users can search for a specific web site, for information, or to obtain web resources. Broder (2002, p. 6) classified these searches as transactional, whose purpose is "to reach a site where further interaction will happen". Rose and Levinson (2004) included this kind

This work has been partially funded through the EuroGeoSource project (project number 250532) from the European Union's ICT Policy Support Programme, as part of the Competitiveness and Innovation Framework Programme, the Spanish Government (projects "España Virtual" ref. CENIT 2008-1030 and TIN2009-10971), and GeoSpatiumLab SL. The work of Aneta J. Florczyk has been partially supported by a grant (ref. AP2007-03275) from the Spanish government. This work reflects only the authors' views and the European Union is not liable for any use that might be made of information contained therein.

Online Information Review
Vol. 35 No. 6, 2011
pp. 909-927

1468-4527
DOI 10.1108/14684521111193193

of search within a broader category and coined the term resource queries. Resource searches are those searches where the goal is to get access to an interactive resource, or to collect a list of interactive resources for later use. The most common queries with a resource goal are those related to the downloading of music and movies. Resource queries can target very specific domain resources such as malware, source code, web forms and web services (see Long, 2007). Nevertheless resource-oriented searches seem to be viewed as less worthy of attention than other types in the research about the performance of search engines. Rose and Levinson (2004, p. 14) expressed their belief that "resource searches are a relatively neglected category in the search engine world".

The advances in search technology have changed how people search for resources. Several studies (Brophy and Bawden, 2005; Norris *et al.*, 2008; Lewandowski, 2010) confirm and measure the ability of some search engines to provide access and replace full-text search in some digital library systems. The geographic information community, where the discovery of geographic resources is based on the digital library metaphor (see Béjar *et al.*, 2009), has started to evaluate the potential of search engines to conduct some distributed discovery tasks. The cause of this interest is that today many of the available discovery systems are pilot projects (see Vandenbroucke *et al.*, 2008; Khalsa *et al.*, 2009). There are studies about the ability of search engines in the discovery of geographic web services (Bartley, 2005; Refractions Research, 2006; Sample *et al.*, 2006; Lopez-Pellicer *et al.*, 2010; Wenwen *et al.*, 2010). However their findings are questionable, as they seem to be based on anecdotal evidence. The available literature has a strong bias towards Google, and it often does not disclose measures of performance except for the amount of services discovered.

The purpose of this study is to measure the performance of the public search application programming interfaces (APIs) of the three main commercial search engines – Bing (formerly Microsoft Live Search), Google and Yahoo! – in discovering geographic web services. With the required caution when generalising the results obtained from the search APIs, the collected data could help to address the following questions not answered by the available literature:

- How do the search engines perform?
- Which is the best-suited search engine?
- Which is the best discovery strategy?

This paper is structured in five parts. First, concepts related to geographic web services and previous studies on searching geographic web services are presented. Next the methodology of the study is described. Then the results are presented and examined. After that several aspects of the methodology and the findings are discussed. In conclusion the findings are summarised and their implications are identified.

Background

In this section we introduce a number of important concepts related to the discovery of geographic web services.

Geographic web services

The effective use of geographic resources is of critical importance in a knowledge-based economy (see Longley *et al.*, 2005). It is essential for many activities to have web access to geographic resources. From the response to a disaster

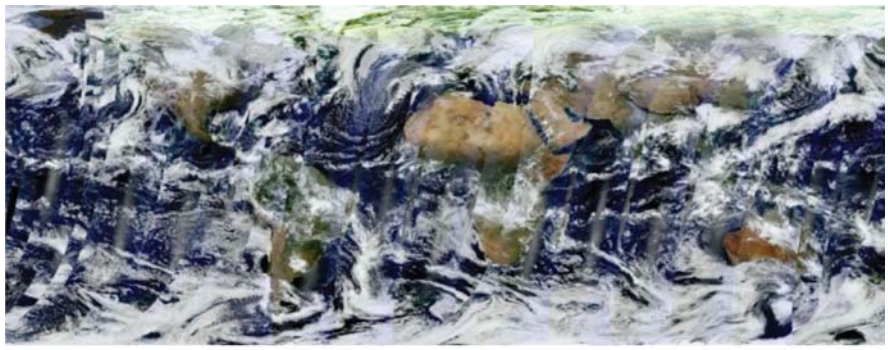
to the decision to develop a new business, access to up-to-date online geographic information might be the difference between success and failure. Geographic web services are the interfaces that geographic data providers make available for discovering, visualising, evaluating, processing and accessing geographic data.

Since 1994 the Open Geospatial Consortium (OGC) has provided standards that ease the use and integration of geographic web services. The OGC service architecture (Lieberman, 2003; Whiteside, 2007) is service-oriented (see Erl, 2005). The discovery of OGC services is based on the publish-find-bind pattern. First service providers describe the features of each web service in standardised XML documents named service descriptions or capabilities documents. Then these descriptions are published in registries, or made accessible through hyperlinks. Next service consumers search service descriptions in catalogues and on the web. When a user finds a service that fits their requirements, they can use the service description to bind an application to the service. The service description also includes all the technical information required to interact with the service.

The most relevant OGC standard is the Web Map Service (WMS) interface standard. The WMS standard was first published in 2000 and its last review was in 2006 (Beaujardiere, 2006). The WMS standard defines a simple HTTP interface for requesting pictures of geographic data, i.e. maps, from one or more distributed geospatial data sources. The response is returned in a standard image format that can be displayed in web browsers (Figure 1).

Searching for geographic resources in metadata catalogues

The discovery of data and services in the geographic information community is founded on the digital library metaphor (Nebert, 2004). A group of providers of geographic resources agrees to set up a coordinated set of online catalogue services and interactive catalogue viewers for searching and managing metadata about their geographic data and services. These services allow distributed searches. Catalogue services can send user requests to other catalogue services and aggregate all the results with the results obtained from a local metadata repository. Unlike metasearch engines distributed catalogues do not use special algorithms to standardise results. Distributed catalogues rely on shared metadata schemas. Catalogue services and catalogue



Notes: This global, continuously updating, image of the Earth can be accessed at http://onearth.jpl.nasa.gov/wms.cgi through a WMS interface. Daily Planet is used as alternative to Google Maps in some websites (e.g. http://exploreourpla.net/explorer/). It can be found in any search engine by searching for "wms daily planet"

Figure 1.
Daily Planet

viewers are offered to the public as core parts of institutional portals named geoportals (Maguire and Longley, 2005). The most important geoportals that provide discovery services are Geospatial One-Stop (http://gos2.geodata.gov/) in the USA and INSPIRE geoportal (www.inspire-geoportal.eu/) in the European Union. There are many other geoportals with different thematic and spatial coverage, and they vary in the numbers and kinds of providers.

Figure 2 depicts a scenario where a user interacts with a client application and a catalogue viewer for searching geographic services and data. The user employs the client application for solving an unspecified task that requires access to geographic resources. On behalf of the user the client application can access a distributed catalogue service, discover servers that offer data and services, and then access them. As an alternate task flow the user might use the catalogue viewer to discover resources, and then configure the client application to use those resources.

The use of catalogues for advertising geographic web services is a costly extra effort that providers sometimes neglect (Nogueras-Iso *et al.*, 2005). Nevertheless the literature on finding geographic resources often proposes high-cost, high-profile, heavy-requirement solutions based on service catalogues that range from CQL based services (Doyle *et al.*, 2001) to ontology-driven registry brokers (Liping *et al.*, 2005).

State of the art in searching geographic web services using search engines

The growing relevance of geographic information to search engines – especially since the successful release in 2005 of Google Maps – and the development and use of geographic resources outside the geographic information community, challenges the role of catalogue viewers and catalogue services as the sole discovery tools (Turner,

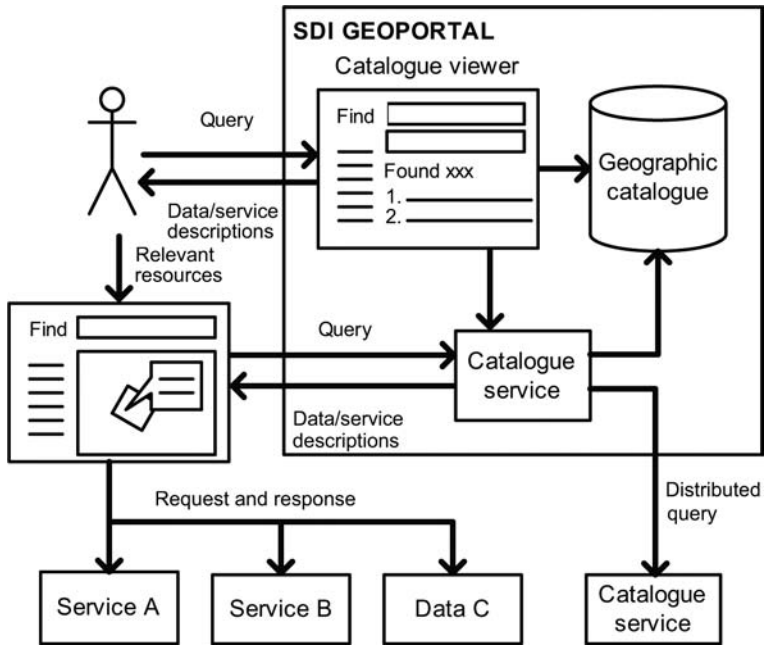


Figure 2.
Possible interactions of a user with a catalogue viewer and a catalogue service for the discovery of geographic resources

2006; Goodchild, 2007). Similar shifts have also happened outside the geographic information community. For example the studies of Bachlechner *et al.* (2006), Yan *et al.* (2007), and Al-Masri and Mahmoud (2008) question the role of registries for finding web services based on SOAP and consider search engines as a viable alternative.

The available literature reports that geographic web services can be discovered using search engines. However the literature has a strong bias towards the WMS standard and Google. Bartley (2005) searched for WMS services in Google, looking for documents with WMS specific parameters in their URL, and documented only 100 services. Refractions Research (2006) described a process for discovering WMS services using the Google APIs and ascertained 615 services. Sample *et al.* (2006) detailed a WMS focused crawler that starts its crawl from relevant sites found by querying the Google API. They did not give the number of WMS services found in the results provided by Google. Wenwen *et al.* (2010) analysed the relation between some terms and hyperlinks to a WMS service in search results from Google. They found that 16 percent of the top 400 search results containing the phrase "Web Map Service" returned at least one hyperlink to a WMS service. Lopez-Pellicer *et al.* (2010) described a crawler that queries Bing, Google and Yahoo! for several types of OGC web services. This crawler discovered 3,626 services in Europe.

Methodology

Description of the scenario

Geographic web services include standard services, such as OGC web services, and proprietary services, typically found in commercial products. This study restricts the analysis to OGC web services because OGC standards are the preferred choice for the development of interoperable geographic services (see Nebert, 2004 and its live version at www.gsdidocs.org/GSDIWiki). Figure 3 presents a scenario where a user uses a search engine for searching geographic services. This scenario is conceptually similar to the one described for catalogue services: the user uses the client application to solve

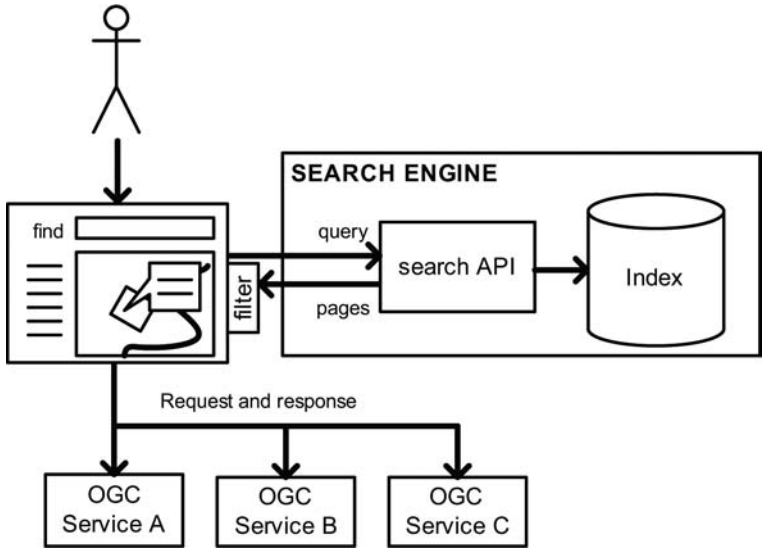


Figure 3.
Possible interactions of a user with a search engine API for the discovery of OGC web resources

an unspecified task that requires access to geographic resources. To fulfil this task the client application can query the search engine through its APIs on behalf of or as instructed by the user. Next the application should filter search results, retrieve candidate documents, detect XML descriptions of services, extract binding information for the most relevant, and then interact with the geographic service.

We will evaluate two discovery strategies in this scenario:

(1) A basic strategy that searches for mandatory terms associated with requests for OGC service metadata, e.g. "service", "request", "getcapabilities" (see Whiteside, 2007), plus additional terms related to the task, e.g. "atlas", "water", "towns". For example a user looking for an atlas can use the query "atlas getcapabilities". The basic strategy looks for OGC service metadata requests (e.g. www.inspire-geoportal.eu/discovery/csw? service = CSW&request = GetCapabilities describes a geoportal catalogue service) and web pages with links that encode requests for OGC service metadata (e.g. http://geoserver.org/display/GEOS/Available + WMS + and + WFS + servers contains a list of map and data services). The rationale behind this search strategy is to exploit geoportals with links to several OGC services, usually through catalogue viewers (Maguire and Longley, 2005).

(2) An expert strategy well documented in the literature that just attempts to find OGC service metadata (Bartley, 2005; Refractions Research, 2006; Sample *et al.*, 2006). The expert strategy refines the basic strategy by restricting the query to documents whose URLs match the pattern of a request for the capabilities document of an OGC service. For example a query in Google with the "inurl" operator followed by the word "getcapabilities" (e.g. "atlas inurl:getcapabilities") will restrict the results to documents containing that term in the URL.

There are three search goals in the above scenario that can be evaluated:

(1) *List candidate services*. The user expects a list of documents about OGC services. The documents can be service descriptions or documents with links to service descriptions. For example an environmentalist can do an exploratory search of pollution maps offered by OGC services and pages with links to these maps for later use.

(2) *Interact with services*. The user needs to interact with relevant OGC services found in the search response in order to achieving an objective. For example an emergency manager might search for URLs of OGC services offering real-time images and data of a flood to use them immediately in an application.

(3) *Discover services*. The user collects a ranked set of candidate web services for an unspecified purpose. These web services can be found in the search response, or by navigating the links in the pages found in the search response. For example an analyst might collect OGC services from a search engine for building a thematic collection of services about urban planning. The analyst measures the success of the query, taking into account the number of different OGC services found, including those found by navigating the links.

The performance of each search engine will be analysed by considering all the possible combinations between the discovery strategies and the search goals.

Data collection

Data were collected from automated queries made to Bing, Google and Yahoo! between 29 and 30 July 2010. The queries were made through their free search APIs. The search engines were queried with two sets of 1,000 queries. The first query set was named "basic", and it represented the basic strategy described above. The second query set was named "expert", and it represented the expert strategy described above. For each query only the first 50 results were used for the study.

Each query in the basic set contained two terms: getcapabilities, which represents the intent of a user to obtain OGC services, and a term from a domain vocabulary that represents a topic constraint. The term getcapabilities is a mandatory value that appears in the URL of a HTTP GET GetCapabilities request, which is the OGC service operation that returns the XML document that describes the service.

The queries in the expert set contained the same 1000 queries but the term getcapabilites is now prefixed by the inurl operator (Google and Yahoo!) or the inanchor operator (Bing). When the inurl operator is included in a query Google and Yahoo! restrict the results to documents containing the term getcapabilities in their URLs. Bing does not currently offer the inurl operator. The inanchor operator in Bing restricts the results to documents where the anchor text of an incoming link contains the term getcapabilities. Geoportals often use as anchor text in an OGC service link the URL of a GetCapabilities request (e.g. the Spatial Data Infrastructure of Spain, www.idee.es/) or the term getcapabilities (e.g. the Swiss geoportal, www.geo.admin.ch/). In this context the inanchor operator and the inurl operator should yield the same results.

The domain vocabulary is derived from a collection of 9,370 OGC service descriptions found worldwide (see Lopez-Pellicer *et al.*, 2010). Each XML document in this collection was parsed in order to extract words used for describing the service. The title and the subjects of the service description were converted into a bag of words. The tokenisation rules were similar to those that search engines use. Words commonly ignored by search engines are also excluded except for those with semantic relevance in the geographic domain. The resulting vocabulary of 6,553 words contains topics (e.g. population, elevation, petroleum), technical terms (e.g. satellite, orthophoto, buoy), product names (e.g. Corine, Landsat, MODIS), place names (e.g. London, California, Australia) and providers (e.g. USGS, IGN, ESA). A simple random sample without replacement of 1,000 terms was selected for the queries. A query set of 1,000 queries produces a considerable amount of data but it is small enough to avoid the constraints imposed by some search engines on long series of automated queries.

Identification and relevance of results for each search goal

Each search result and, optionally, the documents that are navigable from the search result are compared against an "oracle" for checking OGC services. An oracle is a mechanism by which someone might test properties that a product should have, providing a pass/fail judgment (see Baresi and Young, 2001). Our oracle checks whether the document complies with any OGC service specification. The specifications of reference are the OGC interface standards (www.opengeospatial.org/standards) and the XML schemas, DTDs and XML examples for web services maintained by OGC (http://schemas.opengis.net/). Next a relevance score is assigned to the search result. The relevance score is computed differently for each search goal.

- *List candidate services*. A search result is relevant or true positive (TP) with a relevance score of 1 when it is an OGC service description or a web site that include links to OGC service descriptions. Otherwise it is considered a false positive (FP). Its precision is computed as usual (see Sebastiani, 2002).
- *Interact with services*. A search result is a TP with a relevance score of 1 if it is an OGC service description. Otherwise it is considered a FP. Its precision is computed as usual.
- *Discover services*. A search result is a TP if it is an OGC service description or a web site with links to OGC service descriptions not found in results with higher ranks. Otherwise it is considered a false positive. The relevance score is proportional to the number of new OGC service descriptions. We define the metrics pseudo-relevance, and pseudo-precision, for the discovery task as follows:

$$\tilde{r}(s_k) = \begin{cases} n, & s_k \text{ contains links to } n \text{ new OGC service descriptions} \\ 1, & s_k \text{ is a new OGC service description} \\ 0, & \text{otherwise} \end{cases}$$

$$a_i = \sum_{k=1}^{i} \tilde{r}(s_k) \cdot TP(s_k)$$

$$b_i = \sum_{k=1}^{i} FP(s_k)$$

$$\tilde{P}_i = \frac{a_i}{a_i + b_i}$$

In this study the computable characteristic is being either an OGC service description or a document containing links to OGC service descriptions. We can assume that a manual human judgement would consider these two characteristics as necessary but not sufficient for a relevant result. Hence the estimates of precision obtained in this study could be interpreted as an estimate of the maximum precision obtained with a manual human judgement of the results.

Results

The results include an analysis of the overall results, the unique results across all queries, the query size distribution, the cross-coverage of search engines, and the precision. Absolute precision and pseudo-precision values cannot be computed because the number of relevant documents in relation to each query is unknown, and the search engines limit the size of the search response. We obtain estimates of precision using the microaveraging and macroaveraging methods (see Sebastiani, 2002). Microaveraging precision is estimated by summing all results. Macroaveraging precision estimates the precision of each query, and only counts queries with non-empty results.

Overall results

Table I contains a summary of the overall results of the two strategies. As expected the overall results show that it is feasible to find OGC web services and pages linking to

Basic query	Bing		Google		Yahoo!	
Pages with links to geographic services	13,952	70.7%	2,383	51.5%	11,986	51.5%
Geographic service descriptions	631	3.2%	418	9%	2,551	11%
Noise	5,154	26.1%	1,827	39.5%	8,722	37.5%
Total results	19,737	100%	4,628	100%	23,259	100%
Discovered geographic services	23,688		5,679		27,408	
Expert query	Bing (inanchor)		Google (inurl)		Yahoo! (inurl)	
Pages with links to geographic services	211	14.2%	97	3.5%	614	10.3%
Geographic service descriptions	1,173	78.7%	2,695	95.9%	5,162	86.9%
Noise	107	7.2%	19	0.7%	162	2.7%
Total results	1,491	100%	2,811	100%	5,938	100%
Discovered geographic services	1,173		2,699		5,359	

Note: This table only considers the top 50 results. The operator employed in the expert query is between parentheses in the corresponding column

Table I.
Number of results and discovered resources in 1,000 queries

them in the evaluated search engines. There is sufficient statistical evidence to suggest that the overall proportions of relevant pages, resources and noise are not equal in the analysed search engines for both query sets (basic: $\chi^2_{\text{obs}} = 2066.90$, $\wp$-value ≈ 0, and expert $\chi^2_{\text{obs}} = 333.86$, $\wp$-value ≈ 0). The percentage of OGC service documents found using an expert strategy ranges from 95.9 percent in Google (inurl operator) to 78.7 percent in Bing (inanchor operator). If the goal of our search is to interact with services, an expert query provides more direct relevant hits than a basic query. Nevertheless if we take into account the services that can be found following the hyperlinks of the returned pages, the best approach is a basic query. For example the mean average of discovered services per query in Yahoo! is 27.4 using simple queries compared with 5.3 obtained with expert queries. Yahoo! outperforms Google in the number of results and relevant results returned in both strategies. If we consider only basic queries Bing gets a number of relevant results similar to that of Yahoo! and outperforms Google. These results imply that Yahoo! and Bing cover this part of the web better than Google, or that the criteria used by Yahoo! and Bing for limiting the length of search responses are less restrictive than the criteria used by Google. Yahoo! performs even worse than Google with expert queries. Bing's inanchor operator returns geographic service descriptions but it cannot replace the missing inurl operator.

Unique results

Table II shows the overview of the unique results across all queries. The statistical evidence suggests that the proportion of unique relevant pages, resources and noise continue to be different for each search engine (basic: $\chi^2_{\text{obs}} = 676.44$, $\wp$-value ≈ 0, and expert $\chi^2_{\text{obs}} = 26.46$, $\wp$-value ≈ 0). The table includes the ratio between overall and unique results for the same category in each search engine (o/u). From the o/u ratio we can conclude that the same result might be found from 1.8 to 5.2 times across the responses of the same query set. The value of the o/u ratio for the expert strategy is always greater than that of the basic strategy, notably for Google and Bing. As queries were made in sequence it is possible that Google and Bing cache answers to queries with operators for their reuse. This search engine optimisation is known as query locality (Baeza-Yates *et al.*, 2007). A search engine can use a cache of recent answers to

Table II.
Number of unique results and discovered resources in 1,000 queries

Basic query	Bing		o/u	Google		o/u	Yahoo!		o/u
Pages with links to geographic services	4,214	70.1%	3.3	1,372	53%	1.7	5,226	55.3%	2.3
Geographic service descriptions	221	3.7%	2.9	214	8.3%	1.9	1,347	14.2%	1.9
Noise	1,576	26.2%	3.3	1,002	38.7%	1.8	2,882	30.5%	3.0
Total results	6,011	100%	3.3	2,588	100%	1.8	9,455	100%	2.5
Discovered geographic services	3,272			2,055			5,036		
Expert query	Bing (inanchor)		o/u	Google (inurl)		o/u	Yahoo! (inurl)		o/u
Pages with links to geographic services	26	9.1%	8.1	15	2.7%	6.5	191	8.7%	3.2
Geographic service descriptions	257	89.6%	4.6	537	96.9%	5.0	1,987	90.4%	2.6
Noise	4	1.3%	26.7	2	0.4%	9.5	21	0.9%	7.7
Total results	287	100%	5.2	554	100%	5.1	2,199	100%	2.7
Discovered geographic services	257			538			2,015		

Note: This table only considers the top 50 results. The operator employed in the expert query is between parentheses in the corresponding column

speed up query computation. In other words the search engine can return search results from a local cache rather than from its main index. The presence of inurl and inanchor doubles the o/u ratio in Google and Bing respectively but not in Yahoo!. This fact suggests that the caching policies of Bing and Google are more aggressive when a query includes operators. Yahoo! presents the best absolute results. For the basic strategy it returns an amount similar to the sum of Bing's and Google's results, and outperforms them with the expert strategy.

Query result size

Table III presents the histogram of query result size (i.e. how many results are returned). The table also includes the average size of query results. Yahoo! has the largest average size for both kinds of queries (23.26 in basic and 5.94 in expert mode). Google fails to answer 81.1 percent of the basic queries. This high result is even more remarkable compared to the results of Bing (10.4 percent) and Yahoo! (4.9 percent). The use of an operator reduces the average size of the results. Empty responses soar in Bing when the inanchor operator is used. Surprisingly the inurl operator behaves differently in Yahoo! and Google: it increases empty query results in Yahoo! and reduces them in Google. A second noticeable result is a notable peak in the distribution of basic query results around the 40-49 level in Bing and Yahoo!.

Cross-coverage of relevant results

The cross-coverage of unique relevant results gives an approximate idea of the opportunity costs related to the exclusive use of a single search engine. Table IV shows the cross-coverage of unique relevant results per search engine. Using a basic strategy 81.6 percent of the relevant results are found only in one search engine. If we consider the interactive goal, i.e. only OGC service descriptions are relevant, the value rises to 90.6 percent. However if we observe the discovery goal, which takes resources accessible from indexed pages into account, the value drops to 52.3 percent. We can assume that search engine bots can index these resources. Therefore 52.3 percent should be a value near to the upper limit of OGC service descriptions indexed by only one search engine. For expert queries the percentage of relevant results found only in one search engine is around 88 percent no matter which evaluation factor is applied. The results suggest that using only one search engine has an elevated opportunity cost because half of the relevant results are returned exclusively by one search engine.

	Basic query			Expert query		
	Bing	Google	Yahoo!	Bing	Google	Yahoo!
0	104	811	49	615	672	189
1-9	390	72	356	353	240	629
10-19	119	25	132	24	51	101
20-29	60	13	76	3	16	34
30-39	46	12	41	2	5	19
40-49	112	8	199	3	3	21
50 +	169	59	147	0	13	7
Avg.	19.74	4.63	23.26	1.49	2.81	5.94

Table III.
Query result size

Found in			Basic query			Expert query		
Bing	Google	Yahoo!	List	Interact	Discovery	List	Interact	Discovery
	X	4,825	1,202	2,240	1,890	1,709	1,736	
X		878	143	354	357	350	351	
X	X	213	52	239	150	143	143	
X			2,739	121	597	130	107	106
X		X	1,201	81	1,213	108	106	107
X	X		161	7	118	15	15	15
X	X	X	334	12	1,344	30	29	29
Total			10,351	1,618	6,105	2,680	2,459	2,487
Indexed by one			8,442	1,466	3,191	2,377	2,166	2,193
Indexed by two			1,575	140	1,570	273	264	265
Indexed by all			334	12	1.344	30	29	29

Table IV.
Cross-coverage of relevant unique results for different search goals

Note: This table only considers the top 50 results

Precision

Figures 4 and 5 present micro and macro average precisions for each combination of query set and search goal for each search engine. Google is the most precise service for searches using operators. For example the microaverage precision of the top 10 results in the "interact" goal is 94.6 percent compared with 88.4 percent for Yahoo! and 78.6 percent for Bing. Bing's inanchor operator cannot replace the missing inurl operator but it is precise enough for this kind of search. The precision values computed for the "discovery" goal are quite similar to the interact goal for expert queries. As the expert queries focus on XML documents there is little chance of finding a page with hyperlinks to geographic web services.

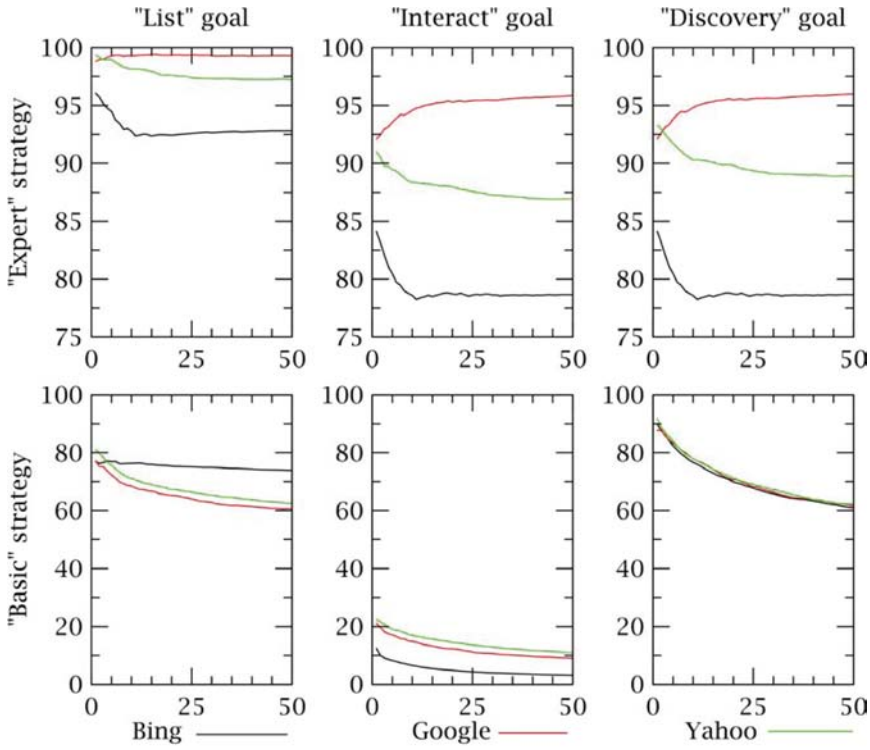
The response precision drops for basic searches. Yahoo! performs to a certain extent better than Google. Bing is better than Google and Yahoo! in "list" queries but worse in interact queries. The list goal is more precise than the interact goal as expected. The shape of the microaverage discovery curve is identical for the three search engines. It seems that the discovery goal is more precise than the list strategy for Google and Yahoo!. We interpret that some search engines tend to prefer documents with many links, even if the links point to XML documents.

Discussion

Methodology

The ever-changing world of search engines led us to perform an automated evaluation of the API results instead of a manual evaluation. As the work of Shang and Li (2002) shows, manual precision evaluation is accurate but also subjective and time-consuming. Its subjectivity makes manual precision a measure that "cannot adapt well to the ever changing search engines and WWW" (Shang and Li, 2002, p. 160). Manual evaluation might provide promising findings, but they become out-dated when the technology of commercial search engines changes.

The literature offers several examples of automatic evaluation of search engine results, e.g. Chowdhury and Soboroff (2002), Shang and Li (2002) and Can *et al.* (2004). We believe that analysing the performance of commercial search engines in well defined search tasks related to a domain is a low-cost way to provide up-to-date

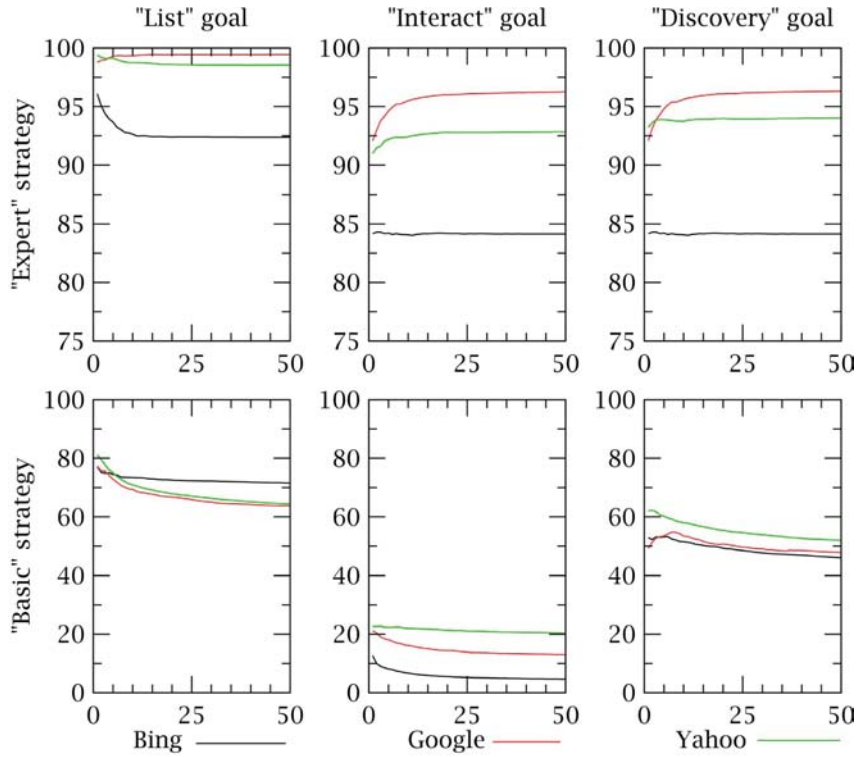


Note: First and second columns are standard precisions, third column is the pseudo-precision

Figure 4.
Microaverage precision for each combination of query set and goal

estimates of the upper limits of their effectiveness as search engines for that domain. It also provides the searcher with a tool for collecting comparable data to track the evolution of search engines. For example our study was performed weeks before Yahoo! officially transitioned its search backend to the Microsoft search platform in the USA and Canada (25 August 2010). As the evaluation is automated we can repeat the study and compare the effect of the search agreement in the performance of Yahoo!.

The decision to perform an automated evaluation is also related to the decision to use the search APIs instead of the web user interfaces (WUIs) provided by search engines. WUIs are not designed to be queried through mediators. This is not only a technical issue. The terms of use of the search WUIs of commercial search engines establish legal limitations on automatic queries and the use of "screen scraping" techniques for extracting the results. It is fair to question whether WUIs and APIs provide the same results. McCown and Nelson (2007, p. 317) examined this issue and concluded that "researchers may need to use caution when generalizing their results obtained from the API to those results that the common user sees using the WUI". In fact WUIs and APIs have two different market targets. WUIs target users who use a web browser as the user agent for querying the search engine; APIs target user groups that use the same user agent, often a web application developed for supporting the user group, that queries the search engine on behalf of each user, even autonomously.



Note: First and second columns are standard precisions, third column is the pseudo-precision

Figure 5.
Macroaverage precision for each combination of query set and goal

Automated tools help the researcher to further analyse the search results. For example in this study we consider documents linked from the results. These data cannot be evaluated with typical IR measures. We have introduced a pseudo-relevance function for the evaluation of the discovery task. The rationale of this measure is the same as the relevance function used in the three-level scoring method developed by Shang and Li (2002) for evaluating web search engines. The three-level scoring method relevance function considers the positive and negative effects in the relevance of the content of related links and duplicate links respectively found in result pages. Our approach is different, as our score is proportional to the number of links that point to new discovered resources.

Results

We have studied the performance of the search engines for searching service descriptions using two query sets: basic and expert. Most (60-75 percent) of the responses to the basic queries and even more (92-100 percent) of the responses to the expert queries were about geographic web services. It is feasible to discover at least one geographic service that matches a query using the basic strategy when the search engine returns some results. Table V presents a summary of the results comparing

strategies with search goals. Each strategy is evaluated in terms of outcome, or number of services discovered, and precision. Each cell has an assigned value between 1 (best result) and 5 (worst result) derived from the results detailed above. Our study compares the strategy recommended in the literature with a quite basic strategy. The expert strategy performs worse than a relatively simple strategy if we consider the number of services found. From the collected data the expert strategy should be used only if the user does not want to perform a further exploration of the results. Basic queries are more appropriate for listing and discovering geographic web services.

We have also analysed which search engine is the best suited for the task. Table VI presents a summary of the findings. Search engines are sorted by their performance for each discovery strategy and search goal. If we consider the outcome Yahoo! is the best performer for each combination, followed by Google and Bing. The analysis of the results suggests that the inferior outcomes of Google and Bing could be caused by query computation optimisation, such as query locality. If we consider precision Google is the best performer in expert queries. The high precision of Google in expert queries could explain why the geographic service discovery literature focuses on Google. The results show that Bing's inanchor operator does not work as well as the inurl operator in this scenario.

Conclusion

This study provides some measures that evaluate the performance of the APIs of several search engines for the discovery of geographic web services. These results should not be generalised to the results that a user could see using the web user interface of a search engine. This study shows that geographic web services can be found in any search engine without the use of advanced operators. A quite simple query strategy based on mandatory terms found in geographic web service standards found 60 percent of relevant results, i.e. descriptions of geographic web services or pages with links to descriptions of geographic web services. The absolute results state that Yahoo! might be the best search engine for this task. Google and Bing have fewer

Strategy	Factor	List	Interact	Discovery
Basic	Outcome	1	5	2
	Precision	2	5	2
Expert	Outcome	4	4	4
	Precision	1	1	1

Table V. Comparison of query strategies versus search goals

Strategy	Factor	List	Interact	Discover
Basic	Outcome	Y > B > G	Y >> B = G	Y >> B > G
	Precision	B > Y = G	Y > G > B	Y > G = B
Expert	Outcome	Y >> G > B	Y >> G > B	Y >> G > B
	Precision	G = Y > B	G > Y > B	G > Y > B

Notes: B, G, and Y refer to Bing, Google and Yahoo! respectively

Table VI. Performance of search engines for each combination of search strategy and search goal

unique results. However the use of only one search engine is discouraged because half of the services found were located by only one search engine. Optimisations for speeding up query computation, such as query locality or the criteria used for limiting response size, could explain the smaller numbers of results from Google and Bing for this kind of query. Therefore we suspect that resource searches continue to be a neglected category in the search engine world. Google, Bing and Yahoo!, the latter to a lesser extent, could have optimised their systems for information queries. However search engines would be more serviceable for resource searches if users could disable some of those optimisations in their profiles.

This work shows that it is feasible to use automated queries to evaluate the performance of search engines for some resource searches. Future research will use this approach for comparing the performance of major search engines versus distributed geographic catalogues for the discovery of different types of geographic resources. This research will provide insights about how to use major search engines to improve geographic metadata catalogues and the strategies for the discovery of geographic resources.

References

Al-Masri, E. and Mahmoud, Q.H. (2008), "Investigating web services on the world wide web", *Proceedings of the 17th International Conference on World Wide Web*, ACM, New York, NY, pp. 795-804.

Bachlechner, D., Siorpaes, K., Fensel, D. and Toma, I. (2006), *Web Service Discovery – A Reality Check, Technical Report*, DERI Galway, Galway, 17 January.

Baeza-Yates, R., Gionis, A., Junqueira, F., Murdock, V., Plachouras, V. and Silvestri, F. (2007), "The impact of caching on search engines", *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM, New York, NY, pp. 183-90.

Baresi, L. and Young, M. (2001), *Test Oracles, Technical Report CIS-TR-01-02*, Department of Computer and Information Science, University of Oregon, Eugene, OR.

Bartley, J. (2005), Mapdex: an online index of Web Mapping Services, available at: www.mapdex.org/data/Mapdex_MAGIC_2006_public.ppt (accessed 14 September 2010).

Beaujardiere, J.D. (2006), *OpenGIS® Web Map Server Implementation Specification, OGC 06-042*, Open Geospatial Consortium, Wayland, MA.

Béjar, R., Nogueras-Iso, J., Latre, M.A., Muro-Medrano, P.R. and Zarazaga-Soria, F.J. (2009), "Digital libraries as a foundation of spatial data infrastructures", *Handbook of Research on Digital Libraries: Design, Development, and Impact*, IGI Global, Singapore, pp. 382-9.

Broder, A. (2002), "A taxonomy of web search", *ACM SIGIR Forum*, Vol. 36 No. 2, pp. 3-10.

Brophy, J. and Bawden, D. (2005), "Is Google enough? Comparison of an internet search engine with academic library resources", *Aslib Proceedings*, Vol. 57 No. 6, pp. 498-512.

Can, F., Nuray, R. and Sevdik, A.B. (2004), "Automatic performance evaluation of web search engines", *Information Processing and Management*, Vol. 40 No. 3, pp. 495-514.

Chowdhury, A. and Soboroff, I. (2002), "Automatic evaluation of world wide web search services", *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM, New York, NY, pp. 421-2.

Doyle, A., Reed, C., Harrison, J. and Reichardt, M. (2001), *Introduction to OGC Web Services, White Paper*, Open Geospatial Consortium, Wayland, MA, 30 May.

Erl, T. (2005), *Service-Oriented Architecture: Concepts, Technology, and Design*, Prentice Hall, Upper Saddle River, NJ.

Goodchild, M.F. (2007), "Citizens as sensors: the world of volunteered geography", *GeoJournal*, Vol. 69 No. 4, pp. 211-21.

Khalsa, S., Nativi, S. and Geller, G. (2009), "The GEOSS interoperability process pilot project (IP3)", *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 47 No. 1, pp. 80-91.

Lewandowski, D. (2010), "Google Scholar as a tool for discovering journal articles in library and information science", *Online Information Review*, Vol. 34 No. 2, pp. 250-62.

Lieberman, J. (2003), *OpenGIS® Web Services Architecture, OGC 03-025*, Open Geospatial Consortium, Wayland, MA, 18 January.

Liping, D., Peisheng, Z., Wenli, Y., Genong, Y. and Peng, Y. (2005), "Intelligent geospatial web services", *Proceedings of IGARSS 2005 IEEE International Geoscience and Remote Sensing Symposium*, IEEE, Piscataway, NJ, pp. 1229-32.

Long, J. (2007), *Google Hacking for Penetration Testers*, Syngress, Burlington, MA.

Longley, P.A., Goodchild, M.F., Maguire, D.J. and Rhind, D.W. (2005), *Geographic Information Systems and Science*, John Wiley & Sons, Chichester.

Lopez-Pellicer, F.J., Béjar, R., Florczyk, A.J., Muro-Medrano, P.R. and Zarazaga-Soria, F.J. (2010), State of play of OGC web services across the web, paper presented at the INSPIRE 2010 conference, Kraków, 22-25 June, available at: http://inspire.jrc.ec.europa.eu/events/conferences/inspire_2010/presentations/80_pdf_presentation.pdf (accessed 14 September 2010).

McCown, F. and Nelson, M.L. (2007), "Agreeing to disagree: search engines and their public interfaces", *Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries*, ACM, New York, NY, pp. 309-18.

Maguire, D.J. and Longley, P.A. (2005), "The emergence of geoportals and their role in spatial data infrastructures", *Computers, Environment and Urban Systems*, Vol. 29 No. 1, pp. 3-14.

Nebert, D. (2004), Developing spatial data infrastructures: the SDI cookbook, available at: www.gsdi.org/docs2004/Cookbook/cook-bookV2.0.pdf (accessed 14 September 2010).

Nogueras-Iso, J., Zarazaga-Soria, F.J. and Muro-Medrano, P.R. (2005), *Geographic Information Metadata for Spatial Data Infrastructures: Resources, Interoperability, and Information Retrieval*, Springer-Verlag, Berlin.

Norris, M., Oppenheim, C. and Rowland, F. (2008), "Finding open access articles using Google, Google Scholar, OAIster and OpenDOAR", *Online Information Review*, Vol. 32 No. 6, pp. 709-15.

Refractions Research (2006), *OGC Services Survey, White Paper*, Victoria, Canada, 9 October.

Rose, D.E. and Levinson, D. (2004), "Understanding user goals in web search", *Proceedings of the 13th International Conference on World Wide Web, WWW '04*, Vol. 04, ACM Press, New York, NY, pp. 13-19.

Sample, J.T., Ladner, R., Shulman, L., Ioup, E., Petry, F., Warner, E., Shaw, K. and McCreedy, F. (2006), "Enhancing the US Navy's GIDB portal with web services", *IEEE Internet Computing*, Vol. 10 No. 5, pp. 53-60.

Sebastiani, F. (2002), "Machine learning in automated text categorization", *ACM Computing Surveys (CSUR)*, Vol. 34 No. 1, pp. 1-47.

Shang, Y. and Li, L. (2002), "Precision evaluation of search engines", *World Wide Web*, Vol. 5 No. 2, pp. 159-73.

Turner, A. (2006), *Introduction to Neogeography*, O'Reilly Media, Sebastopol, CA.

Vandenbroucke, D., Janssen, K. and Van Orshoven, J. (2008), INSPIRE state of play: development of the NSDI in 32 European countries between 2002 and 2007, paper presented at Tenth International Conference for Spatial Data Infrastructure, St. Augustine, 25-29 February, available at: www.gsdi.org/gsdiconf/gsdi10/papers/TS1.5paper.pdf (accessed 14 September 2010).

Wenwen, L., Chaowei, Y. and Chongjun, Y. (2010), "An active crawler for discovering geospatial web services and their distribution pattern – a case study of OGC Web Map Service", *International Journal of Geographical Information Science*, Vol. 24 No. 8, pp. 1127-47.

Whiteside, A. (2007), *OGC Web Services Common Specification, OGC 06-121r3*, Open Geospatial Consortium, Wayland, MA, 9 February.

Yan, L., Yao, L., Liangjie, Z., Ge, L. and Bing, X. (2007), "An exploratory study of web services on the internet", *Proceedings of IEEE International Conference on Web Services (ICWS2007), IEEE*, IEEE, Piscataway, NJ, pp. 380-7.

About the authors

Francisco J. Lopez-Pellicer holds MS and PhD degrees in Computer Science and a MS degree in Economics and Business, all from the University of Zaragoza, where he has been a teaching assistant since 2007. He has focused his research efforts on improving the use of geospatial semantics within the multidisciplinary area of spatial data infrastructures. His current research interests are the development of geospatial ontologies, vocabularies and gazetteers, the discovery and indexing of geographic web resources, and the publication of geographic information on the semantic web. He is author and co-author of several articles published in journals, books and national and international conference proceedings. He has also contributed to various R&D open tenders and technology transfer agreements, both national and European. Francisco J. Lopez-Pellicer is the corresponding author and can be contacted at: fjlopez@unizar.es

Aneta J. Florczyk holds a MS degree in Computer Science from the Czestochowa University of Technology (Poland). Her research interests cover the multidisciplinary area of spatial data infrastructures, especially knowledge discovery and geospatial services on the semantic web. Within this context she has co-authored publications in conference proceedings and has collaborated on several R&D projects.

Rubén Béjar holds a MS degree in Computer Science and a PhD degree, both from the University of Zaragoza, where he is a Tenured Assistant Professor in the Computer Science and Systems Engineering Department. He has co-authored dozens of papers on geographic information systems and spatial data infrastructures, and has been the main researcher in more than 20 public and privately funded R&D projects. He has given lectures in several universities, has taken part in the development of ISO standards and has contributed as an expert to the development of land cover data models for the National Geographic Institute of Spain and the United Nations Food and Agriculture Organisation.

Pedro R. Muro-Medrano holds MS and PhD degrees in Industrial Engineering from the University of Zaragoza. He has worked in private industry for two years and has held different visiting research positions at the Carnegie Mellon University's Robotics Institute (Pittsburgh, PA), the University of Maryland (College Park) and the US National Institutes of Health (Bethesda, MD). He has 27 years of experience with R&D activities in software development and engineering and he has been a Professor of Computer Science at the University of Zaragoza since 1988. He is co-author of more than 150 national and international papers published in books, journals and conference proceedings. He has participated in more than 130 national and international R&D projects, he has been principal investigator in more than 60 projects, and has registered the intellectual co-property of 21 software programs. He is the head of the Advanced Information Systems Laboratory in the Computer Science and Systems Engineering Department and the Engineering Research Institute of Aragón at the University of Zaragoza. The Laboratory

is a technology based multidisciplinary R&D group with 34 full time staff with MS or PhD degrees and three with BS degrees.

F. Javier Zarazaga-Soria holds a MS degree in Computer Science from the University of Valencia and a PhD degree from the University of Zaragoza. He did his Master's thesis at the Road Safety Engineering Laboratory (University of London). In 1994 he started collaborating with the Advanced Information Systems Laboratory and then he joined as a scholar. In 1996 he became an official scholar with the Education Department of Spain and began work as an assistant teacher at the University of Zaragoza. He became an Associate Professor in 2003. He is co-author of more than 50 national and international papers published in books, journals and conference proceedings and he has participated in many national and international R&D projects, funded by public and private companies.