International Journal of Spatial Data Infrastructures Research, 2011, Vol.6, 168-186

A Review of the Implementation of OGC Web Services across Europe*

Francisco J. Lopez-Pellicer[1], Rubén Béjar[1], Aneta J. Florczyk[1],
Pedro R. Muro-Medrano[1], F. Javier Zarazaga-Soria[1]

[1]Department of Computer Science and Systems Engineering, Universidad
Zaragoza, Spain
{fjlopez, rbejar, florczyk, prmuro, javy}@unizar.es

Abstract

This paper presents the results of an investigation conducted in the spring of 2010 to review the availability across Europe of public Web services conforming to the standard specifications issued by the Open Geospatial Consortium. The descriptive and statistical analysis of 6,544 geospatial network services found might provide insight into the current level of implementation of these services in Europe. These services were discovered with the help of a focused crawler able to discover access points to public geospatial network services. This crawler relies on general-purpose search engines for finding seed URLs related with geospatial network services from which to begin crawling. The work also identifies potential limitations and data biases derived from the methodology. Nevertheless, this kind of strategy might open up new opportunities to complement SDI implementation assessments when exhaustive, periodic and up to date monitoring is required.

Keywords: OGC, crawler, Web services

DOI: 10.2902/1725-0463.2011.06.art8

1. INTRODUCTION

Soon after the introduction in the 1990s of the concept of Spatial Data Infrastructures (SDI) for spatial data coordination, the geospatial information community began to consider SDIs as the new paradigm to manage and access geospatial information. In a context where organizations, administrations and citizens are willing to provide and consume geospatial information (mainly free), SDIs have provided the geospatial information community with the infrastructure that allows the exploitation of the broad possibilities of the Internet for information access. This situation is even more apparent in the European Union due to the existence of the European Directive 2007/2/EC, Infrastructure for Spatial Information in the European Community (INSPIRE) (European Commission, 2007) that implies obligations to the Member States towards the access to geospatial information using the Internet.

Interoperability and open standards are key factors that facilitate any kind of access in the Internet. Therefore, both are key issues for SDI (Yang and Raskin, 2009; Wua et al, 2011). Although the different cookbooks (such as Nebert, 2001, 2004) and legislations (such as the INSPIRE Regulation on Discovery and View Services (European Commission, 2009b)) generally do not force exclusive use of specific open standards for interoperability, the Open Geospatial Consortium (OGC) standards have become the "de facto" standard geospatial Web service stack for SDIs.

On the other hand, the advance in the implementation of SDIs is raising attention to the research on their assessment (see the book of Crompvoets et al, 2008, for a very interesting compendium). Assessment of SDIs is still in its beginning and remains to be seen as a complicated task. Giff and Crompvoets (2008) identify the complex, multifaceted, and dynamic nature of SDIs as the main cause of difficulties to the assessment. The difficulties have led to different assessment approaches, many of them based on the use of qualitative or quantitative indicators. In general, the data that these indicators require for their computation is manually collected by means of surveys, interviews, analysing Web information or reading documentation. This kind of approach has provided reasonable good results to the assessment of SDIs (see Craglia and Johnston, 2004; Crompvoets et al, 2004; Kok and van Loenen, 2005). Nevertheless, these approaches require intensive human intervention for data collection and often depend on the collaboration of third parties for collecting sufficient sampling and accurate information.

The substantial use of the Web and open standards in SDIs open good opportunities for considering the use of machines’ work for assessment, automating the data collection process. That is, it could be now feasible to automate data collection of main technical components of SDIs, including network services, client applications, downloadable datasets, metadata documents, and geoportals. These are the accessible spatial assets of an SDI (Béjar, 2009), which can be defined as any useful or valuable spatial information resource that has been made accessible to the users of an SDI. The automatic collection of accessible spatial assets could reduce the need of human intervention and, in addition, could make it possible to collect more reliable, complete and up to date data.

As it has being pointed out by Giff and Crompvoets (2008), the complexity of an SDI means that it cannot be understood only in terms of the summation of its components. The SDI as a whole produces a value greater than the value of the summation of its individual components (de Man, 2006) and requires a better level of measuring the results of the integration of the components. An important limitation of a machine-based approach could be that it is only focussed on accessible spatial assets. Hence, this approach could not pay special attention to other elements of an SDI, such as the ones related to people, policies, or others (Eelderink et al, 2008). However, the accessible spatial assets are, in some way, the final purpose of the different components of an SDI. It is reasonable to assume that the assessments about many SDI components (individual components and the results of its integration) could at least partially be derived from the assessments about related accessible spatial assets.

With the impulse of INSPIRE, the SDIs in Europe have made accessible a significant amount of spatial assets, enabling their collection with machine-based techniques. The machine-based approach for collecting data for SDI assessment is still today in an early technological and theoretical stage. Despite that, its application could be useful for example as a complement of other assessment approaches (Crompvoets et al, 2008), in the context of the Implementing Rules for Monitoring and Reporting of the INSPIRE directive (European Commission 2009a) to verify monitoring information, or to provide a bit more complete and reliable view of online resources described by the State of Play of SDI in Europe (Vandenbroucke et al, 2010).

Specifically, this paper presents a machine-based approach for the assessment of the number and distribution of OGC Web service (OWS) instances in Europe. The assessment will be focussed on the OWS specifications that have been indicated as possible implementations of INSPIRE compliant network services, like Web Map Service (WMS, view services), Web Feature Service (WFS, download services), Web Coverage Service (WCS, download services), Web Processing Service (WPS, invoke spatial service services and transformation services), and Catalogue Service for the Web (CSW, discovery services).

In addition, this paper presents a method to estimate the number of OWS instances based on capture-recapture estimators (Chao, 2001). This technique has been applied to estimate the size of other kind of Web resources (for example in Lu and Li, 2010), and may enable the derivation of some plausible limits on population size. To the best of our knowledge, we have not found a study that estimates the number of OWS instances in the Web.

The paper is organized as follows. Section 2 provides background on the automated discovery of OWS instances using Web crawlers. Section 3 describes several aspects of the analysis method. Section 4 presents the results of the analysis of the services discovered in the period of April-May of 2010. These results include data and estimations about size, search engines coverage, specifications, patterns of deployment, and geographic location of instances. Section 5 identifies the limitations of the approach. Finally, some conclusions and future work are presented.

2. BACKGROUND

This section introduces a number of important concepts related to the discovery of geographic Web services with Web crawlers.

2.1 OGC Web Services

Since 1994, the OGC issues consensus standards that ease the use and integration of geographic Web services. The OGC has defined standards for the portrayal (WMS) and download (WFS, WMS) of geospatial data, for the discovery of geospatial resources (CSW), and for the remote invocation on geospatial algorithms (WPS) among others. These specifications share a simple HTTP interface with a common *GetCapabilities* operation (Whiteside and Greenwood, 2010). The document returned by a *GetCapabilities* request is the XML document that contains the service description.

The OGC service architecture (Lieberman, 2003) is a service-oriented architecture (see Erl, 2005) where the discovery of OWS instances is based on the publish-find-bind pattern. First, service providers describe the features of each Web service in standardized XML documents named service descriptions or capabilities documents. Then, the service provider publishes these descriptions in registries, or makes them accessible from its geoportal through hyperlinks. Next, service consumers could query for service descriptions in catalogues, or find the service serendipitously by browsing the geoportal. When a user finds a service that fits his requirements, he can use the service description to bind an application to the service and then interact with the service. That is, the service description also includes all the required technical information to interact with the service at once.

2.2 Automated Discovery of OGC Web Services

As it has been pointed out above, the approach taken in OGC for the discovery of network services is to create metadata and use service registries, catalogues or directories for the discovering process. Recently, search engines, such as Google, Bing or Yahoo! have become surprisingly new major sources for the discovery of Web services. For example, Refractions Research (2006) applied basic heuristics derived from the signatures of WMS requests for discovering 695 WMS in Google.

The growing relevance of geographic information in 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 questions the role of catalogue viewers and catalogue services as the sole discovery tools (Turner, 2006; Goodchild, 2007). However, the SDI research world has paid little attention to the automated discovery of Web services. Sample et al (2006) found 761 WMS using and open source crawler. Li et al. (2010) analysed the distribution of WMS 1126 across the globe found using an in-house Web crawler. Similar techniques for building lists of WMS through automated processes are applied in WMS-finder (http://wms-finder.de/), Geopole (http://www.geopole.org/), and Mapmatters (http://www.mapmatters.org/). Variations of these approaches are the OWS Search Engine (http://ows-search-engine.appspot.com/), a meta-search engine of OWS, and the system described in Chen et al (2008), where the detected WMS are registered in a CSW repository.

2.3 Challenges for Automated Discovery

The Web crawling algorithm is quite simple and well documented (see Olston and Najork, 2010). Given a set of seed URLs:

1. The crawler downloads all the Web resources addressed by the URLs.
2. The crawler checks if it can deal with the resource representation returned (e.g. an HTML page, a XML document).
3. The crawler extracts the hyperlinks contained in the representations.
4. A new work cycle begins using the hyperlinks found as seeds for the next iteration.

Effective automated discovery of OWS instances from search engines, geoportals and Web applications goes beyond simple Web crawling. It requires the use of a focused crawler (Chakrabarti et al, 1999). The goal of a focused crawler is to crawl first pages that are relevant for finding fast resources that matches a pre-defined set of topics. In addition, a focused crawler can implement heuristics focused in the discovery of these topics.

The design of an efficient focused crawler for OWS instances faces several challenges:

1. **XML links**. Some services might not have an HTML link to the capabilities document: no HTML document link to them, or the HTML documents exist, but the links are text without link mark-up. However, there are accessible on the Web many valuable XML documents with links to OWS instances: service metadata in metadata repositories. Indexing metadata repositories require the use of deep Web crawlers.
2. **Lack of textual description**. The capabilities document can be indexed by a search engine, but its textual content is scarce, generic, or unrelated with the service. The lack of textual information makes difficult to retrieve it from a search engine. If an OWS capability document does not contain an adequate amount of relevant words and out links, the ranking techniques of search engines can tag the document as no relevant. Hence, an unfiltered keyword based search is prone to have a low precision.
3. **OWS Exception reports**. The link might return an OWS exception report, which is indistinguishable by conventional crawlers from other kind of documents. This barrier requires domain knowledge to construct an appropriate new request from the content of the exception report.
4. **Links from Web applications**. The service might be only accessible through Web applications, such as a Web mapping client. Usually, the out links of a Web application cannot be extracted by current crawling techniques, or, it requires specific domain knowledge that helps to construct an appropriate link from the application code.
5. **Domain knowledge**. Crawling OWS instances requires not only knowledge of OGC standards but also to know how each geospatial community applies these standards.

The first challenge is shared with deep Web resources, that is, web sites whose content is not reachable via hyperlinks and instead can only be retrieved by submitting HTML forms (Chang et al, 2004). The second challenge is related with the low quality of the description of any Web service on the Web. Third and four challenges are related with the specific characteristics of OWS specifications, which differ from mainstream standard Web Services specifications, such as WSDL-based Web services. Finally, the fifth challenge implies that a general-purpose crawler may fail to find OWS instances because it does not implement domain heuristics described in the next section.

3. METHODOLOGY

The data used in this research are collected using an advanced focused crawler that was configured for finding OGC Web services. The seeds of the crawl were the search results of automated queries made to Bing, Google and Yahoo!. Starting from these seeds, the crawler navigates all the accessible pages within a maximum number of links from the seeds. Several heuristics have been used for the discovery of OWS instances not hyperlinked from the explored pages. The heuristics included detection of links in text and the automated generation of *GetCapabilities* requests.

The search of OWSs in general-purpose search engines does not require complex queries or the use of advanced operations. A basic strategy that queries for mandatory terms associated with requests for OWS capabilities (e.g. “request”, “getcapabilities”, “service”) plus additional terms related to the targeted standard service (e.g. “wps”, “wms”, “wcs”) or tasks related with the service (e.g. “coordinate transformation”, “metadata”, “map”) would suffice for finding OWS instances among the first search results. For example, a user looking for a mapping service related with soils can use the query “getcapabilities wms soil italy”. This strategy asks search engines for pages that contain these terms in their text or in their URL. The search results should include pages with links to WMS servers, and WMS service metadata documents that provide soil maps of Italy. The rationale behind this search strategy is to exploit the references in text and links found in geoportals and technical papers to OWS services. A complete analysis about the precision of different alternatives and strategies for searching geospatial Web services is available in Lopez-Pellicer et al (2011).

Each search result and, optionally, the documents that are navigable from the search result are compared against an oracle for checking documents returned by request to OGC Web servers. 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). The oracle checks if the document complies with any OGC service specification. The specifications of reference are the OGC Interface Standards (http://www.opengeospatial.org/standards) and the XML Schemas, DTDs and XML examples for Web services maintained by OGC (http://schemas.opengis.net/). If the oracle identifies an OWS server, several tests are applied to collecting data from the returned documents.

4. RESULTS

This research has crawled in the period of April-May of 2010 three main search engines in Europe (Google, Yahoo! and Bing) with a crawler focused on OWS specifications. The crawler has been configured to crawl resources within a distance of three links from the search results found in search engines to increase the coverage of the research. The findings have been geocoded and filtered to discard non-European services. The crawl has ascertained 6,544 services in Europe.

4.1. Estimation of the Number of Services

Table 1 shows our findings by source. The columns “Google”, “Yahoo” and “Bing” refer to the services present in the responses of the respective search engines. The column “Other” identifies the services found by the focused crawler applying several heuristics. The data for the services found is aggregated as a categorical data format, that is, the frequencies of the same pattern are grouped. Each row aggregates the number of ascertained services present or absent from a source. We use $Z_{s1..sn}$ to denote this number. For example, Z_{0001} is the number of services in Europe discovered by the OWS-focused crawler crawling the Web and not found among the search results returned by search engines, Z_{1010} is the number of services only found by the crawler in the search results returned by the search engines Google and Bing, and Z_{1110} is the number of services only found in search engines.

Table 1: Service Data for Each Search Engine in Europe

Service Lists				*Services*	
Google	*Yahoo!*	*Bing*	*other*		
			X	Z_{0001} =	2,918
		X		Z_{0010} =	144
		X	X	Z_{0011} =	323
	X			Z_{0100} =	33
	X		X	Z_{0101} =	140
	X	X		Z_{0110} =	3
	X	X	X	Z_{0111} =	344
X				Z_{1000} =	6
X			X	Z_{1001} =	56
X		X		Z_{1010} =	10
X		X	X	Z_{1011} =	16
X	X			Z_{1100} =	2
X	X		X	Z_{1101} =	1,135
X	X	X		Z_{1110} =	5
X	X	X	X	Z_{1111} =	1,409
				Total =	6,544

We use this data as input for providing an estimate of the number of public OWS instances in Europe at the time of writing with the help of capture-recapture methods (see Chao, 2001). Capture-recapture is a family of methods used to estimate population size used in different areas, such as ecology, medical research, software engineering and Web research. The simplest capture-recapture model is the Petersen estimate or two-sample model. This model is used to estimate the unknown size of a population with two samples. Using the numbers caught in each samples (the captures |A| and |B|) and the numbers of individuals caught in both samples (the recapture $|A \cap B|$), it is possible to provide an estimate of the total population size assuming that both samples are independent ($|A||B|/|A \cap B|$). Variants of this simple model have been applied in Web research to obtain estimates of the size of the Web using URLs sampled from search engines (Gulli and Signorini, 2005).

Next, we apply capture-recapture methods to estimate the number of public OWS instances in Europe. The input data are the public OWS found by the focused crawler and the search results returned by search engines (Table 1). We compute different capture-recapture estimates, their Standard Error (SE), the confidence interval (CI) and the goodness of fit of some estimators applying methods described in Chao, 2001: pair of samples (Petersen estimate), log-linear models and sample coverage. All of them are computed using the package CARE (http://chao.stat.nthu.edu.tw/ softwareCE.html), and the results are presented in Table 2.

The Petersen estimates use pair of sources, and they range from 3,170 to 4,222, the lower limit of the estimate with worst SE and the upper limit of the estimate with best SE respectively. A Petersen estimate can underestimate or overestimate the results when the sources are dependent. In other words, when a service is found in a source, this event affects the probability for that service of being found in other sources. Two log linear models are also considered: the first estimates the population size under the hypothesis of independence of each source, and the latter estimates the population size under the hypothesis of local dependence between search engines. Dependence between search engines means that they index roughly the same part of the Web. Finally, the sample coverage is a method suitable in scenarios where the coverage among some samples is high.

Table 2: Estimation of the total of public OWS instances in Europe.

Approach	*Model*	$\hat{N}$	*SE*	*CI 95%*
Pair of samples	Petersen (best SE)	3,177	5	3,170 – 3,189
	Petersen (worst SE)	4,131	44	4,049 – 4,222
Log-Linear	Independent search engines	6,589	8	6,584 – 6,617
	Dependent search engines	6,717	19	6,684 – 6,757
Sample coverage	High coverage	6,897	34	6,836 – 6,971

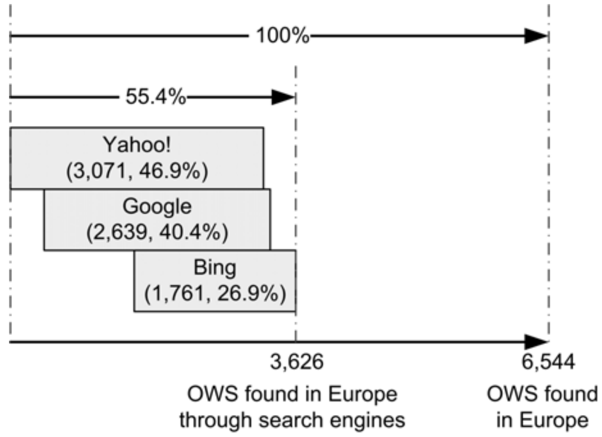
The log-linear model that considers search engines dependent fits the data well and presents an estimate of 6,717 services with a CI 95% of (6,684 – 6,757). The log-linear model that considers independent the sources of services has a lower estimated SE, but it behaves worse when we test the goodness of fit. The goodness of fit of the dependent model is 614.83, meanwhile the goodness of fit of the independent model is 7,412.45. In the sample coverage approach, the coverage among sources is estimated to be 86.6%. This high coverage yields a precise estimate of 6,897 services with a CI 95% of (6,836 – 6,971), and it is close to the log-linear model that considers some dependence among search engines.

The sample coverage and the log-linear model with dependence are good population estimators, but we consider that the log-linear model with dependence is the most adequate because it fits better with the characteristics of the analysed sources.

4.2. Search Engine Coverage in Europe

Yahoo!, Google and Bing, combined index only a 55.4 per cent of the public OWSs found. Their respective coverage is 49.6 per cent, 40.4 per cent and 26.9 per cent (see Figure 1). The percentage of services found simultaneously in the three search engines is 21.6 per cent, which is far from the combined coverage. However, if we consider pairs of search engines, Google and Yahoo overlap an 80.7 per cents of their results, Google and Bing a 53.5 per cent, and Yahoo and Bing a 57.3 per cent. These results suggest that if we cannot find a Web service in a general-purpose search engine, it is likely that we cannot find it in other general-purpose search services. The three main search engines cover only about half of the estimated amount of OWSs in Europe. That is, half of the OGC Web services found are hidden Web resources.

Figure 1: Coverage of OWS Instances by Search Engines in Europe



4.3. Which is the Most Common Specification?

Tables 3 and 4 show our findings. We have detected that 56.7 per cent of OWSs are map services. However there exist a significant 28.7 per cent of inactive services, that is, more than 1,800 on-line services without feature types, coverages, metadata records and processes. Considering this fact, we should highlight that four out of five active OWSs are map services. This has an easy justification because visualization may be the most important feature of geospatial information and the creation and access to images is well supported computers, that is why WMS was the earliest OGC Web service standard.

Table 3: Services Found by Specification

	Total		Active	
Web Map Service	3,712	56.7%	3,705	79.5%
Web Feature Service	1,492	22.8%	735	15.7%
Web Coverage Service	1,156	17.7%	174	3.7%
Catalogue Service	45	0.7%	35	0.7%
Web Processing Service	42	0.7%	17	0.4%
Other (*)	97	1.4%	-	-
	6,544	100.0%	4,666	100.0%

(*) Other OWS specifications, such as Sensor Observation Services.

We have also analysed which are the versions offered by default or advertised by links to the ascertained active services. The most common version implemented in map, feature and coverage services is a deprecated standard. In some circumstances, such as in map and coverage services, the standard specification is far from being an alternative to the deprecated standard.

Table 4: Default Versions in Active Services

	Version	Default	Version	Default	Total coverage
Web Map Service	1.1.1	3,361	1.3.0	321	99.38 %
Web Feature Service	1.0.0	427	1.1.0	308	100.00 %
Web Coverage Service	1.0.0	145	1.1.1	29	100.00 %
Catalogue Service	2.0.2	29	2.0.1	4	94.29 %
Web Processing Service	1.0.0	13	0.4.0	4	100.00 %

We can identify some other findings:

- **Focus on portrayal services**. Today, the main use of public OWS instances is the portrayal of spatial data (WMS). The ratio between active and inactive services in other kind of services might point that, even when it is easy to set up an OWS instance, there is no interest by the service providers to give access to its data through standard services.

- **Low penetration of new standards**. The results suggest that the new versions of the standards have a low rate of implementation in Europe. This is especially significant with map services.
- **Low barriers for deployment**. Geospatial suites allow user to set up quickly OWS servers, for example some commercial suites when the user set up a WMS the system also set up WFS and WCS by default. In many case these WFS and WCS remain inactive.
- **Bad administration practices**. The huge amount of inactive feature and coverage services might indicate that the services are managed by end users, rather than by system administrators, with more sensibility on maintaining and optimizing systems. However, this finding could be considered as uncertain as these situations could be justified by legitimate reasons such as testing purposes.

4.4. Which are the Patterns of Deployment?

The deployment of services can be modelled as hosts publishing services, and services operating with different kind of contents (layers, feature types, coverages, processes and record types, depending on the type of service). We can use this simple model to discover patterns of deployment by analysing the relationships among hosts, services and information types. Table 5 summarizes the distribution of the relation between hosts and services (e.g. a host has 2 web map services), services and contents (e.g. a web map service publishes 2 map layers), and hosts and contents (e.g. a host gives access to 17 map layers). 50 per cent of hosts serve one service, mainly WMS, or two services, two WMS or one WMS with a data access service. A 75 per cent of hosts serve six or less services. A 50 per cent of services offer one or two kinds of information types (mainly map layers). A 75 per cent of services offer less than five information types. The distribution of information types per hosts has a mean greater than the 3rd quartile. That is, 25 per cent of hosts contain most of the served types.

Table 5: Summary of the Distribution of Public Services, Served Types and Hosts found in Europe

	Services per Host	Contents per Service	Contents per Host
Minimum	1.00	0.00	0.00
1st quartile	1.00	0.00	6.00
Median	2.00	2.00	17.00
Mean	11.55	7.30	83.37
3rd quartile	6.00	5.00	64.00
Maximum	1,125.00	948.00	5,749.00

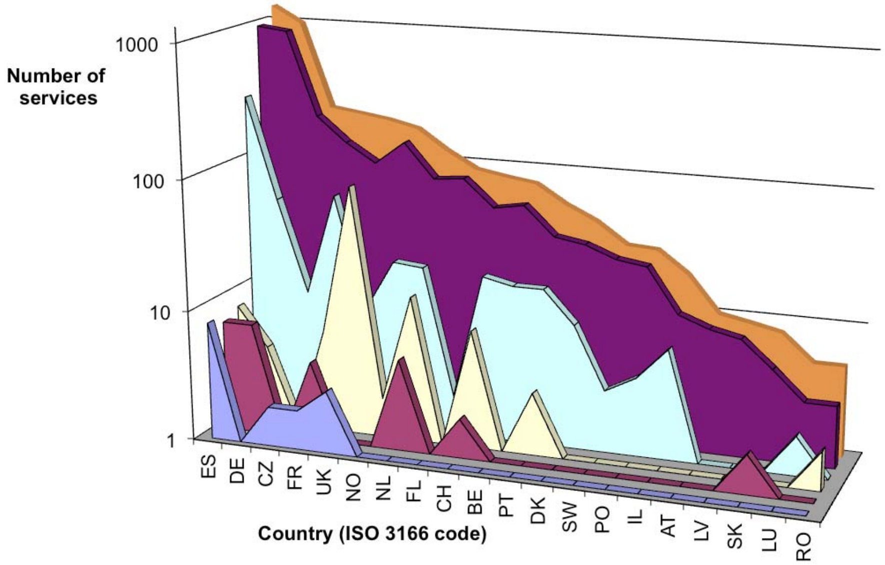
We can identify some patterns:

- **Simple services**. It is quite common to find hosts that serve one or two WMS with two or three layers. Low technical barriers for WMS and servicing data only for publishing purposes might be the causes of this pattern. It is quite common to find that one of the servers does not offer resources. That is, the barrier for creating a service is too low and the procedure as simple that is prone to waste resources into empty services.
- **Service farms**. 25 per cent of hosts services more than 6 different service instances, with a maximum of 1,125 services. This pattern could be caused by hosts hosting third party Web services, server farms dedicated to geospatial Web services, and bad practices such as overloading a host with too many instances of services.
- **Desktop-oriented capabilities documents**. There are WMS servicing a very huge number of layers. A huge number of layers imply a huge and complex XML capabilities document increases the difficult to be managed by Web applications. These services seem to be oriented to consumption by traditional GIS systems rather than Web applications.

4.5. Where Are the Services Found Located?

Table 6 shows active services found in Europe. That table only shows data from countries with at least 5 services found by the crawler. The implementation of services varies from countries, such as Germany (DE) and Spain (ES), with more than 900 public services, to other apparently without public OWS services, such as Bulgaria. The reasons behind this apparent division may lie on political factors (the policies on geographic data publication), geographic factors (the extent of the country), economical factors (the level of economic development), and technological factors (the implementation of geographic information systems in the public administration).

Table 6: Active Services Found in Europe.



	ES	DE	CZ	FR	UK	NO	NL	FL	CH	BE	PT	DK	SW	PO	IL	AT	LV	SK	LU	RO
WPS	8	1	2	2	3	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
CSW	7	7	0	4	0	0	5	1	2	1	0	0	0	0	1	0	0	2	0	0
WCS	8	4	1	6	82	2	13	1	8	0	3	1	0	0	0	1	0	1	0	2
WFS	303	51	10	59	8	20	19	2	18	16	16	9	3	4	7	0	1	0	2	0
WMS	971	910	211	137	99	148	81	85	53	58	35	32	26	24	11	9	8	5	3	3
Total	1297	973	224	208	192	170	119	89	81	75	54	42	29	28	19	10	9	8	5	5

5. LIMITATIONS

This section identifies several considerations that should be taken into account when the results above are analysed. These considerations can be groped in the following topics:

1. **Underestimation of OWS instances**. The assessment only includes information of the indexable Web. Hence, CSW catalogues in geoportals, and geoportals driven by databases fall outside of its business scope. The public OWS services described in these Web resources have a chance of not being indexed by search engines. A better assessment should include at least the OWS instances registered in the CSW catalogues found by the crawler.
2. **Seed location bias**. Search engines rank results taking into account where the query has been made, and answer with indexes available in the nearest datacentre. As in this case, the crawler was located in Spain; it is possible that there are more odds that query answers contain an OWS

service in Spain rather than other countries. The location bias could be avoided by deploying the crawler in hosts distributed in several countries.

3. **Seed language bias**. Many information retrieval techniques have their focus in the Latin alphabet, hindering the discovery of services whose descriptions were available in Greek and Cyrillic alphabets. This might have affected the results from some countries.
4. **Location accuracy**. The geocodification of services is based on the country top-level domain, and if the address is an IP or the top-level domain is generic, the IP is georeferenced. The accuracy of georeferenced IP varies. For example, there are 505 services for which we only know they are in Europe. The accuracy of the localization should increase if the description of the OWS provider found in the capabilities document is taken into account to assign the country.

6. CONCLUSIONS

The capability of important components of the SDI to be accessed and queried automatically, open new opportunities to complement the assessment of SDI like to provide real time reliable information, to facilitate the verification of monitoring information or to complement information difficult to obtain by other sources. Particularly geospatial Web services based on open standards may constitute adequate components to be used for automatic data collection. The reduction of human intervention and advantages related with bigger completeness and better data reliability are interesting properties for assessments that need to be done periodically. Thus, measures like the number of publicly available services in the Web and related properties can be suitable for indicators of the degree of implementation of SDI and be periodically calculated to monitor evolution.

A main topic in this approach is the broad discovery of components. To do that, the use of search engines combined with deep Web crawling techniques has shown to be a good method capable to index an elevate percentage of publicly available OWS in Europe providing a near real-time overview more complete than any other published so far. An estimation of the maximum real size of the set of services has also been provided. Possible biases in the results due to the use of search engines as primary data source have also been identified.

The information contained in this survey also includes technological aspects like versions of standards, use of host services and inactivity of services that could also be considered as indirect indicators, trends or common practices, that could give advice to SDI stakeholders in their related technological decisions. Future work will address these biases and complete the description of OWS with information related to service quality, such as quality of description, uptime and response latency, and technical characteristics of the service, such as service vendor, and SOAP support.

Some characteristics of the use of OWS in Europe have also been identified. Thus the geographic distribution of services, the current unbalanced focus on portrayal services, the low penetration of recent OGC standards, the detection of some problematic practices or the prevalence of simple deployment of services versus more complex solutions, have been brought to light.

This line of work is still in its early stage and much progress has to be done in aspects like heuristics to improve crawling, data mining, characterization of services size and importance, or analysis of properties and relations between components. We think that this approach may play an important role as a complement of other methods of SDI assessment, mainly in a steady state of the SDI development where exhaustive and periodic monitoring may be very informative.

ACKNOWLEDGEMENT

This work has been partially supported by Spanish Government (projects “España Virtual” ref. CENIT 2008-1030, TIN2009-10971 and PET2008_0026), the Aragón Government (project PI075/08), the National Geographic Institute (IGN) of Spain, and GeoSpatiumLab S.L. The work of Aneta J. Florczyk has been partially supported by a grant (ref. AP2007-03275) from the Spanish government. We also have to thank to two anonymous referees for its valuable suggestions that unquestionably allowed improving the paper.

REFERENCES

Baresi, L. and M. Young (2001). Test oracles. Technical Report CIS-TR-01-02, University of Oregon, Dept. of Computer and Information Science, Eugene, Oregon, U.S.A. http://ix.cs.uoregon.edu/~michal/pubs/oracles.html, [accessed 23 June 2011].

Béjar, R. (2009). Contributions to the modelling of Spatial Data Infrastructures and their portrayal services. PhD thesis, Computer Science and Systems Engineering Department, Universidad Zaragoza. Available at https://www.educacion.es/teseo/imprimirFicheroTesis.do?fichero=10608 [accessed 23 June 2011].

Chakrabarti, S., Van den Berg, M., and B. Dom (1999). Focused crawling: a new approach to topic-specific web resource discovery. *Computer Networks*, 31(11-16):1623-1640.

Chang, K. C.-C., He, B., Li, C., Patel, M., and Z. Zhang (2004). Structured databases on the web: observations and implications. *ACM SIGMOD Record*, 33(3):61-70. doi:10.1145/1031570.1031584.

Chao, A. (2001). An overview of closed capture-recapture models. *Journal of Agricultural, Biological, and Environmental Statistics*, 6(2):158-175.

Chen, N., Chen, Z., and J. He, (2008). Interoperable AntSDI — the geospatial information gateway to Antarctica: Architecture and application. In *ISPRS Congress Beijing 2008, WG VIII/8: Polar and Alpine Research*, volume XXXVII-B8, pp. 825-830. ISPRS Archives.

Craglia, M. and A. Johnston (2004). Assessing the impacts of spatial data infrastructures: Methods and gaps. In *7th AGILE Conference on Geographic Information Science, 29 April-1 May 2004,* Heraklion, Greece.

Crompvoets, J., Bregt, A., Rajabifard, A., and I. Williamson (2004). Assessing the worldwide developments of national spatial data clearinghouses. *International Journal of Geographical Information Science*, 18(25):665-689. doi:10.1080/13658810410001702030.

Crompvoets, J., Rajabifard, A., van Loenen, B., and T. D. Fernández, editors (2008). *A Multi-View Framework to Assess SDIs*. Space for Geo-Information (RGI), Wageningen and University and Centre for SDIs and Land Administration, Department of Geomatics, The University of Melbourne.

de Man, W. E. (2006). Understanding SDI: Complexity and institutionalization. *International Journal of Geographical Information Science*, 20(3):329-343. doi:10.1080/13658810500399688.

Eelderink, L., Crompvoets, J., and W. E. de Man (2008). Towards key variables to assess National Spatial Data Infrastructures (NSDIs) in developing countries. In Crompvoets, J., Rajabifard, A., van Loenen, B., and Fernández, T. D., editors, *A Multi-View Framework to Assess SDIs*, pp. 307-324. Space for Geo-Information (RGI), Wageningen and University and Centre for SDIs and Land Administration, Department of Geomatics, The University of Melbourne.

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

European Commission (2007). Directive 2007/2/EC of the European Parliament and of the Council of 14 March 2007 Establishing an Infrastructure for Spatial Information in the European Community (INSPIRE). *Official Journal of the European Union*, 50 (L 108) of 25 April 2007: 1-14.

European Commission (2009a). Commission Decision of 5 June 2009 Implementing Directive 2007/2/EC of the European Parliament and of the Council as regards monitoring and reporting. *Official Journal of the European Union*, 52 (L 148) of 11 June 2009: 18-26.

European Commission (2009b). Commission Regulation (EC) No 976/2009 of 19 October 2009 implementing Directive 2007/2/EC of the European Parliament and of the Council as regards the Network Services. *Official Journal of the European Union*, 52(L 274) of 20 October 2009: 9- 18.

Giff, G. A. and J. Crompvoets (2008). Performance indicators a tool to support spatial data infrastructure assessment. *Computers, Environment and Urban Systems*, 32(5):365-376.

Goodchild, M. (2007). Citizens as sensors: the world of volunteered geography. *GeoJournal*, 69(4): 211-221. doi:10.1007/s10708-007-9111-y.

Gulli, A. and A. Signorini (2005). The indexable web is more than 11.5 billion pages. In *Special interest tracks and posters of the 14th international conference on World Wide Web* (WWW '05). ACM, New York, NY, USA, 902-903. doi: 10.1145/1062745.1062789.

Kok, B. and B. van Loenen (2005). How to assess the success of national spatial data infrastructures? *Computers, Environment and Urban Systems*, 29(6):699-717. Part Special Issue: Urban Data Management Symposium, Chioggia, Italy. doi:10.1016/j.compenvurbsys.2004.02.001.

Li, W., Yang, C., and Yang, C. (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*, 24(8):1127-1147.

Lieberman, J., editor (2003). *OpenGIS Web Services Architecture v0.3*. Number OGC 03-025. Open Geospatial Consortium, Inc. http://portal.opengeospatial.org/files/?artifact_id=1320, [accessed 23 June 2011] .

Lopez-Pellicer, F. J., Florczyk, A. J., Béjar, R., Muro-Medrano, P. R., and F. J. Zarazaga-Soria (2011). Discovering geographic web services in search engines. *Online Information Review*, 35(6). In press.

Lu, J. and D. Li (2010). Estimating deep web data source size by capture—recapture method. *Information Retrieval*, 13(1): 70-95. doi:10.1007/s10791-009-9107-y.

Nebert, D., editor (2001). *Developing Spatial Data Infrastructures: The SDI Cookbook v.1.1*. Global Spatial Data Infrastructure. http://www.gsdi.org/pubs/cookbook/Default.htm, [accessed 23 June 2011]

Nebert, D., editor (2004). *Developing Spatial Data Infrastructures: The SDI Cookbook v 2.0*.Global Spatial Data Infrastructure, http://www.gsdi.org/docs2004/Cookbook/cookbookV2.0.pdf, [accessed 23 June 2011]

Olston, C. and M. Najork (2010). Web crawling. *Information Retrieval*, 4(3):175-246.

Refractions Research (2006). OGC Services Survey. white paper, Refractions Research, Victoria, Canada, http://www.refractions.net/expertise/whitepapers/ogcsurvey/ [accessed 23 June 2011]

Sample, J. T., Ladner, R., Shulman, L., Ioup, E., Petry, F., Warner, E., Shaw, K., and F. P. McCreedy (2006). Enhancing the US Navy's GIDB Portal with Web Services. *Internet Computing, IEEE*, 10(5): 53-60. doi:10.1109/MIC.2006.96.

Turner, A. (2006). *Introduction to neogeography*. O'Reilly Media, Inc.

Vandenbroucke, D., Crompvoets, J., Janssen, K., and D. Biliouris (2010). INSPIRE & NSDI SoP. D4.1 — Summary report regarding the results of the European Assessment of 34 NSDI (first year) — September 2010. Summary Report v4.1.1, Spatial Applications Division, K.U.Leuven Research & Development.

Whiteside, A. and J. Greenwood, editors (2010). *OGC Web Services Common Standard v2.0.0*. Number OGC 06-121r9. Open Geospatial Consortium, Inc. http://portal.opengeospatial.org/files/?artifact_id=38867.

Wua, H., Li, Z., Zhang, H., Yang, C., and S. Shen (2011). Monitoring and evaluating the quality of web map service resources for optimizing map composition over the internet to support decision making. *Computers & Geosciences*. In press.

Yang, C. and R. Raskin (2009). Introduction to distributed geographic information processing research. *International Journal of Geographical Information Science*, 23(5).