TESIS DE LA UNIVERSIDAD DE ZARAGOZA

2012

22

Aneta Jadwiga Florczyk

Search improvement within the geospatial web in the context of spatial data infrastructures

Departamento
Informática e Ingeniería de Sistemas

Director/es
Zarazaga Soria, Francisco Javier
López Pellicer, Francisco Javier

ISSN 2254-7606

Prensas de la Universidad
Universidad Zaragoza

Universidad Zaragoza
1542

Tesis Doctoral

SEARCH IMPROVEMENT WITHIN THE GEOSPATIAL WEB IN THE CONTEXT OF SPATIAL DATA INFRASTRUCTURES

Autor

Aneta Jadwiga Florczyk

Director/es

Zarazaga Soria, Francisco Javier
López Pellicer, Francisco Javier

UNIVERSIDAD DE ZARAGOZA

Informática e Ingeniería de Sistemas

2012

SEARCH IMPROVEMENT WITHIN THE GEOSPATIAL WEB IN THE CONTEXT OF SPATIAL DATA INFRASTRUCTURES

Aneta Jadwiga Florczyk

PhD DISSERTATION

RESEARCH ADVISORS

Dr. Francisco Javier Zarazaga–Soria
Dr. Francisco Javier López–Pellicer

May 2012

Computer Science and Systems Engineering Department
Universidad de Zaragoza

Acknowledgments

I would like to thank all members of the IAAA research group of the University of Zaragoza, to which I had pleasue belong during the thesis development, for their collaboration and support in any aspect, and especially, my research advisors, F.Javier Zarazaga–Soria and F.Javier López–Pellicer for their endless forbearance and patience.

Also, I would like to express my gratitude to members of the IFGI of the University of Muenster, Germany, and every other person I have met there, who made effort to make my stay a wonderful experience, which has contributed to my profesional and personal development. I feel specially grateful to proffesor Werner Kuhn, Tomi Kauppinen and Patrick Maué for their help in this development.

I would like to thank the people that have reviewed this thesis. Despite all of their help, I take full responsibility for any errors or omission herein.

Finally, I would like to thank particularly my parents for their unconditional support during the work on this thesis. They always managed to find the right words to encourage me to follow the path I have chosen, even if this meant personal sacrifices for them.

Contents

List of Tables

List of Figures

Listings

List of Algorithms

Nomenclature

CI	Cyberinfrastructure
CSDGM	Content Standard for Digital Geospatial Metadata
DC	Dublin Core
DL	Digital Library
ESDIN	European Spatial Data Infrastructure with a Best Practice Network
FGDC	Federal Geographic Data Committee
GCI	Geospatial Cyberinfrastructure
GEOSS	Global Earth Observation System of Systems
GIBO	Geographic information–bearing objects
GIR	Geographic Information Retrieval
GIS	Geographic Information System
GML	Geography Markup Language
GPS	Global Positioning System
GUI	Graphical User Interface
HTML	HyperText Markup Language
HTTP	Hypertext Transfer Protocol
INSPIRE	Infrastructure for Spatial Information in the European Community
IR	Information Retrieval
ISO	International Organization for the Standardization

ISO/TC 211	ISO Technical Committee 211 Geographic information/Geomatics
LBS	Location based service
MBBOX	Minimum Bounding Box
NAP	North American Profile of ISO 19115: Geographic Information - Metadata
NER	Named Entity Recognition
NSDI	National Spatial Data Infrastructure
OASIS	The Organization for the Advancement of Structured Information Standards
OGC	Open Geospatial Consortium
OWL	Web Ontology Language
OWS	OGC Web Service
QoS	Quality of Service
RDF	Resource Description Framework
RDFS	RDF Schema
SDI	Spatial Data Infrastructure
SDTS	Spatial Data Transfer Standard
SE	Search engine
SEIS	European Shared Environmental Information Space
SKOS	Simple Knowledge Organization System
SLA	Service–level agreement
SOA	Service Oriented Architecture
ToS	Terms of Service
UDDI	Universal Description, Discovery and Integration
UNGIWG	Unated Nations Geographic Information Working Group
UNSDI	United Nations Spatial Data Infrastructure
URI	Universal Resource Identifier

WSDL	Web Service Definition Language
WSQF	OASIS Web Service Quality Factor
WSQM	OASIS Web Services Quality Model

What magical trick makes us intelligent? The trick is that there is no trick. The power of intelligence stems from our vast diversity, not from any single, perfect principle.

Minsky, 1986

Chapter 1

Context and research issues

Geographic information technologies facilitate the integration of scientific, social and economic data through space and time in spatially enabled societies (see Figure 1.1). Williamson et al. (2010) defines a society as Spatially Enabled Society when "*location and spatial information are regarded as common goods made available to citizens and businesses to encourage creativity and product development*". The Spatial Data Infrastructure (SDI) concept is a milestone in these societies. An SDI is a System of Systems (Béjar, 2009) that promotes the economic development, improves stewardship of natural resources, and protects the environment. Nebert (2004) provides the following definition for SDI:

> "*The relevant base collection of technologies, policies and institutional arrangements that facilitate the availability of and access to spatial data. The SDI provides a basis for spatial data discovery, evaluation, and application for users and providers within all levels of government, the commercial sector, the non–profit sector, academia and by citizens in general*".

There are many ongoing initiatives on establishing SDIs at local, national, regional or global levels. Table 1.1 presents some examples. However, management, sharing and use of geographic information within those SDIs require standardisation efforts. Therefore, the SDI community usually adopts existing standards about geographic information to achieve these goals.

The main international standardisation bodies that deal with geographic information are the International Organization for the Standardization (ISO) Technical Committee 211 'Geographic information/Geomatics' (ISO/TC 211), and the Open Geospatial Consortium[3] (OGC). ISO/TC 211 is responsible for the ISO geographic information series of standards, which aim to establish a structured set of standards for information concerning objects or phenomena that are directly or indirectly associated with a location relative to the Earth.[4] The OGC is an international industry

[3] www.opengeospatial.org/
[4] http://www.isotc211.org/

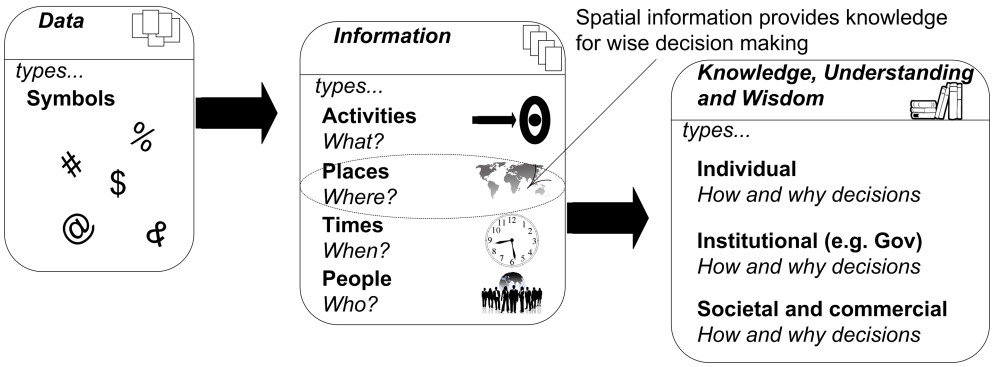

Figure 1.1: Spatial dimension of information in information management (source: (Wallace, 2007)).

Level	SDI	Area	Note
Global	United Nations Spatial Data Infrastructure (UNSDI)	Unated Nations	"*A voluntary network of UN specialized agencies, programmes and funds*", established in March 2000 (UNGIWG, 2007). Unated Nations Geographic Information Working Group (UNGIWG) is the coordination body.
Regional	Infrastructure for Spatial Information in the European Community (INSPIRE)	European Union	Established by Directive 2007/2/EC (EC, 2007a) of the European Parliament and of the Council of 14 March 2007 "*to support Community environmental policies, and policies or activities which may have an impact on the environment.*"[1]
National	National Spatial Data Infrastructure (NSDI)	USA	Established by The White House (1994) to "*reduce duplication of effort among agencies, improve quality and reduce costs related to geographic information, to make geographic data and to establish key partnerships with states, counties, cities, tribal nations, academia and the private sector to increase data availability.*"[2] The Federal Geographic Data Committee (FGDC) is the coordination body.
National	Infraestructura de Datos Espaciales de España (IDEE)	Spain	Jefatura de estado (2010), on infrastructure and geographic information services in Spain, transpotes the INSPIRE directive.

Table 1.1: Examples of SDIs.

consortium of 439 companies, government agencies and universities participating in a consensus process to develop publicly available interface standards.[5] The OGC standards specify well–defined interfaces for spatial data services to ensure interoperability across information communities. As ISO and OGC are liaised, it often results in virtually identical standards (Kresse and Fadaie, 2004).

The adoption of open OGC standards for the implementation of Geospatial Web services in SDIs (Nebert, 2004) has favoured the development of a public, open and interoperable Geospatial Web. Table 1.2 presents the main OGC Web Service (OWS) interface specifications. López-Pellicer

[5] http://www.opengeospatial.org/ogc

Name	Current version	Objective
Web Map Service (WMS)	1.3.0 (de la Beaujardiere, 2006)	Portrayal
Web Feature Service (WFS)	2.0 (Vretanos, 2010b)	Download features
Web Coverage Service (WCS)	2.0.0 (Baumann, 2010)	Download coverages
Catalogue Service for the Web (CSW)	2.0.2 (Nebert et al., 2007)	Discovery
Web Processing Service (WPS)	1.0.0 (Schut, 2007)	Remote invocation

Table 1.2: OGC Web Service interface specifications relevant in this work.

(2011) describes the Geospatial Web as "*the collection of Web services, geospatial data and metadata that supports the use of geospatial data in a range of domain applications*". In the present thesis, Web services, geospatial data and metadata that belong to the Geospatial Web are called Geospatial Web resources. Such definition embraces a variety of resources that bear geographic information (Goodchild and Zhou, 2003). Geospatial Web resources may be encoded using open standards, closed standards and proprietary formats. These resources include online systems which support capturing, storing, analysing, managing, and presenting data with a geospatial dimension (e.g. ESRI ArcGIS Server). This kind of system is also known as online Geographic Information System (GIS). Finally, spatial browsing systems (for example Google Maps) also belong to the Geospatial Web.

Proper support for finding information should be considered a prerequisite for an effective information–based community. Therefore, every community, including the SDI community, is encouraged by their users and stakeholders to develop their own approach to the task of searching for information with regard to the characteristics of the domain. Any other activity in the community tightly depends on its effectiveness.

1.1 Context

The most prominent examples of formal SDI programs are driven by national or federal officially recognised governments. However, there are other ongoing initiatives, such as the Global Earth Observation System of Systems[6] (GEOSS) or the European Shared Environmental Information Space[7] (SEIS), which aim to make an abundance of geographic information about the environment available through OGC services. SDIs and these initiatives support each other with their own unique emphases. An SDI focuses on data collection, data sharing and data reuse. GEOSS, for example, is an initiative which considers, analyses, and integrates isolated Earth observation systems that have been maintained by involved nations (Bai et al., 2009), and its aim is to build a system of systems for global Earth observations to provide systematic monitoring and assessment of nine social

[6] http://www.earthobservations.org/geoss.shtml
[7] http://ec.europa.eu/environment/seis/index.htm

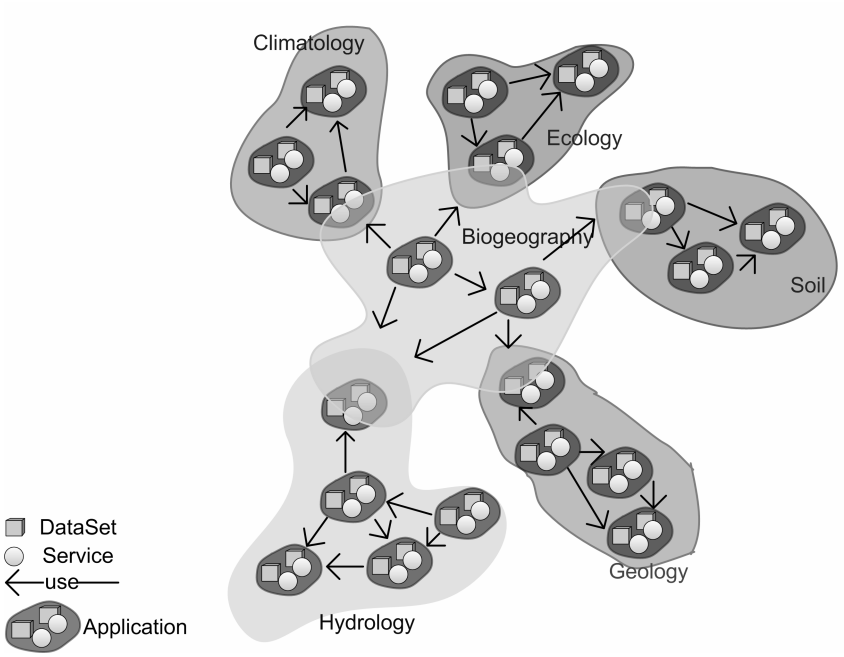

Figure 1.2: Augmentation property of Geospatial Cyberinfrastructure that supports domain and cross–domain research. (Other dependencies among the domains are not shown to avoid overloading the presentation).

benefit areas (disasters, health, energy, climate, water, weather, ecosystems, agriculture, biodiversity) (Mitsos et al., 2005). All these initiatives contribute to the development of the Geospatial Cyberinfrastructure (GCI). GCI has its origins in Cyberinfrastructure (CI), a generic information infrastructure, to collect, archive, share, analyse, visualise, and simulate data, information, and knowledge (NSF, 2003, 2007), which is especially required to support data and computation intensive scientific fields (Ellisman, 2005). Therefore, a GCI can be defined as a combination of geospatial data resources, network protocols, computing platforms, and computational services together to perform data–intensive applications focused on geospatial information within the geographic information community and across science and business domains (Yang et al., 2010). Since current practices in geospatial science provide cross–cutting geospatial analysis and modelling within and across many scientific domains (Yang et al., 2010; Wang and Zhu, 2008), it can be advanced that GCI promotes cross–domain e–science. Diverse scientific fields produce new outcomes that frequently state new research questions. New data and novel inquiry approaches raise new demands on geospatial science. Therefore, new dedicated GCI–based solutions are needed to process and integrate geospatial information to support new requirements, for example to enable managing user–generated information (Díaz et al., 2011). As a result, a GCI based solution utilises an integrated architecture that builds upon past investments to share spatial data, information, and knowledge (Yang et al., 2010).

Figure 1.2 outlines the idea of GCI, where the resources generated by a dedicated solution support the e–science performed within a domain and across domains. For example, research in hydrology might require outcomes (data, process and tools) from the previous studies in the same field, and a research in the biogeography field might need to use results from various domains.

The enablement of intensive computing infrastructures offers new opportunities for e–science communities (e.g. the environmental sciences communities (Giuliani et al., 2011)), which contributes to SDI goals. However, one of the important issues in promoting new e–science is information and knowledge sharing across boundaries of the community or organisation (Longueville, 2010). In this aspect, a geoportal is employed as one of the building blocks of an SDI (Bernard et al., 2005; Maguire and Longley, 2005). It can be understood as "*a web site considered to be an entry point to geographic content on the web or, more simply, a web site where geographic content can be discovered*" (Tait, 2005). Using the service taxonomy from ISO/DIS 19119 (Percivall, 2002), geoportals can be classified as (1) *human interaction services* (i.e. services for management of user interfaces, graphics, multimedia, and for presentation of compound documents) and are closely related to (2) *model/information management services* (i.e. services for management of the development, manipulation, and storage of metadata, conceptual schemas, and datasets), and also (3) the *system management services* to support services for authorisation, authentication or e–commerce. As for technical aspects of geoportals built on open standards, the OGC provides a discussion paper on Geospatial Portal Reference Architecture (OGC GP–RA): *"The Geospatial Portal Reference Architecture documents a 'core' set of interoperability agreements that provide instructions for bridging the gaps between different organizations and communities that have heretofore shared geospatial information only with great difficulty. The portal addresses technical interoperability between diverse systems and it also helps address 'information interoperability' between groups whose content has been created with different data models and metadata schemas."* (Rose, 2004) Figure 1.3 presents the idea of the OGC GP-RA and Figure 1.4 shows the OGC PG-RA services distribution. The OGC GP–RA defines an abstract architecture for interoperable geoportals, which has been widely adopted in Web-based geospatial system, and specifies four classes of Web Services that are required to implement an efficient SDI geoportal using related interoperability specifications:

- *Portal Services.* These services offer an entry point to discover and access data as well as management and administration capabilities;
- *Catalog Services.* These services provide information about data and services;
- *Portrayal Services.* These services are used to process geospatial information and prepare it for presentation to the user by offering mapping and styling capabilities;
- *Data Services.* These services concentrate on data access and processing.

The above architecture is based on Service Oriented Architecture (SOA) principles (Erl, 2005), i.e.

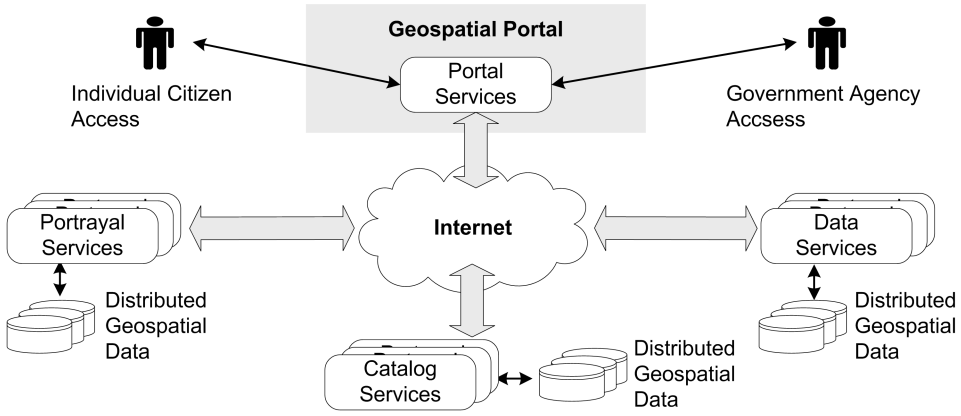

Figure 1.3: Geospatial Portal Reference Architecture (source: Percivall (2002)).

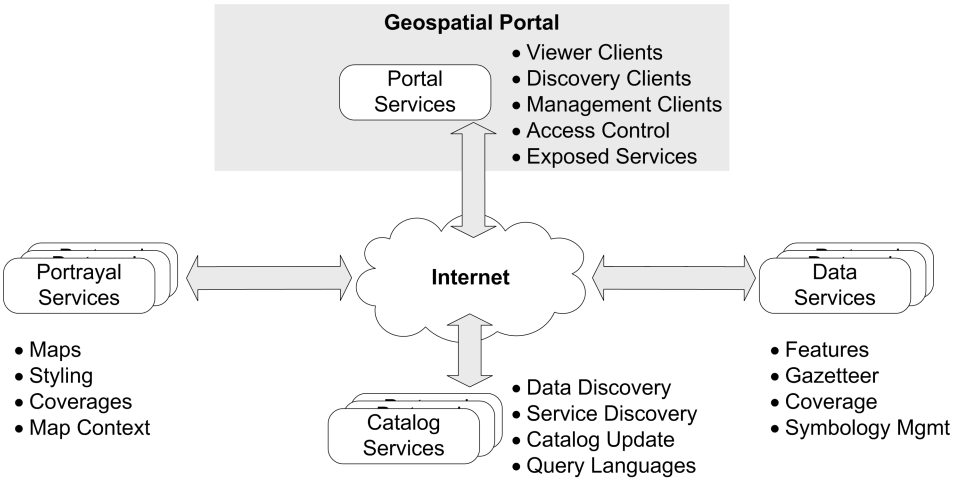

Figure 1.4: Geospatial Portal Reference Architecture Services Distribution (source: Percivall (2002)).

the popular *publish–find–bind* and *self–describing service* patterns. All OGC Web Services are self–describing by providing a `GetCapabilities` operation that returns a *Capabilities* document that describes the service endpoints, the operations exposed, as well as information about the geospatial data that they supply.

1.2 Motivation

This thesis aims to determine some basic search patterns that are relevant to the SDI community considering the characteristics of geographic information and the geospatial community. These patterns are used to identify approaches for the improvement of search in the Geospatial Web from the SDI perspective.

A search for specific information is an important issue in any information system. In the SDI context, *catalogue services*[8] are discovery and access systems for geographic information that use indexed and searchable metadata against which intelligent geospatial searches can be performed within or across SDI communities (Nebert, 2004). The metadata profiles applied in SDIs use different standards as their base. For example INSPIRE Implementing Rules on metadata (INSPIRE DTM and EC/JRC, 2010) identifies ISO 19115:2003 (ISO, 2003), and FGDC recommends Content Standard for Digital Geospatial Metadata (CSDGM) (FGDC, 1998a) and North American Profile (NAP) of ISO 19115:2003 (ANSI, 2009). Although metadata standards vary across SDIs, *crosswalks* enable the translation of this information in order to make it conform to a selected metadata standard or profile (Nogueras-Iso et al., 2004; Batcheller, 2008; Khoo and Hall, 2010). For example, Nogueras-Iso et al. (2004) presents a transformation between ISO 19115 Core and Dublin Core (DC) metadata standards. The DC metadata standard is the outcome of an open metadata initiative for the description of cross–domain information resources (Powell et al., 2007) that has become an ISO Standard (ISO Standard 15836:2009 (ISO, 2009)). Crosswalks enable users to perform different search strategies among different SDIs, such as federated or centralised search. For example, Leite et al. (2006) propose a Web–based GIS architecture where the central point is a catalogue that offers federated search over distributed catalogues. Another proposal from the geospatial community is a multi–catalogue search engine that searches across various catalogues that implement the OGC Catalogue Service standard (Nebert et al., 2007) and offers an integrated result (Li et al., 2011).

Traditionally, Web resources, from Web pages to SDI geoportals, have not received much attention from the geospatial community. Metadata about these resources, often potentially interesting for users and stakeholders, is seldom found in SDI catalog services. However, indices point out that the status quo may change. The number of geoportals available online is growing, a fact that paradoxically increases the difficulties of searching for geospatial information. Geoportals are the visible part of a GCI that can be exploited by a potential user, which will try to discover such *areas of interest* (i.e. geoportals) for their future exploration. In addition, there are resources generated by experts in the field of geographic information that are published on the Web but are not published for their discovery in SDI catalogues or geoportals. Moreover, some works indicate importance of geographic information generated by communities of Web users (i.e. *neogeography* (Turner, 2006), *naïve geography* (Egenhofer and Mark, 1995) or *Volunteered Geographic Information* (VGI) (Goodchild, 2007)) as important source of information for SDI, or at least a complementary source (Craglia et al., 2008; Keßler and Bishr, 2009).

These new requirements and the augmentative character of the content produced by the geospatial community require the identification of new approaches for the improvement of search in the Geospatial Web. Web crawling and the Semantic Web have been considered as relevant for this

[8] Different names are used in the geospatial community when referring to such a system, for example, *catalogue services* (OpenGIS Consortium), *Spatial Data Directory* (Australian Spatial data Infrastructure), and *Clearinghouse* and the *Geospatial One–Stop Portal* (FGDC).

purpose.

Today, search engines (SEs) are popularly employed to perform search activity in the Web. In general, a search engine, in response to a user query, returns an ordered list of informative items on potentially relevant Web resources (e.g. Web pages, images) retrieved from a repository. Three main types might be distinguished: directory–based, crawler–based or hybrid (Inthiran et al., 2010). A directory–based SE is based on curated lists (e. g. Yahoo[9]). The directories existing in the Web usually ensure better precision but Web cover is lower, and frequently domain–oriented. In this sense, geospatial catalogues can be classified as a directory–based SEs.

In contrast to a directory–based SE, the repository of a crawler–based SE is created and maintained automatically. Automatic Web crawlers (called also robots or spiders) are responsible for finding, analysing and indexing the Web resource. Considering web resource accessibility through a crawler–based SE, the Web might be divided into three layers:

- *Surface Web.* It is the part of the Web which resources are reachable by web crawlers via hyper-links (Bergman, 2001);
- *Deep Web.* The deep Web is composed of the online databases whose data are exposed via forms, applications or web services (Bergman, 2001);
- *Invisible Web.* The rest of the Web is not easily indexable or does not have valuable content (Price and Sherman, 2001).

Domain portals found through a general SE require additional exploration by users to find required resources (Green, 2000). For example, a geoportal is an example of domain portal which requires further interaction to be performed by user in order to access the deep Web resources (Figure 1.5). This is a characteristic of domain–niches, the domains whose resources are of low interest for a general search engine. In other words, the required effort of discovering, indexing and supporting proper search functionality is not compensated by the interest of a typical Web user. Therefore, in terms of specialisation, search engines might be divided into (1) general SEs (for example Google Search Engine[10]) that offer horizontal search, and (2) domain–oriented SEs (e.g. Google Scholar[11]) that offer vertical search. The domain–oriented SEs are optimised to support specific characteristics of the domain in terms of thematic search (e.g. in context of medicine (Chakrabarti et al., 1999)), kind of web resource (e.g. web service (Al-Masri and Mahmoud, 2008), picture, video), web cover (i.e. controlled list of web addresses), or combinations of these characteristics.

There are works that present a dedicated crawler capable of retrieval and indexing Geospatial Web services (López-Pellicer et al., 2011e; Li et al., 2010). Such crawler can expose deep Web resources of the Geospatial Web (López-Pellicer et al., 2010c). In addition, the research in Web search engines

[9] http://dir.yahoo.com/
[10] http://www.google.com/
[11] http://scholar.google.es/

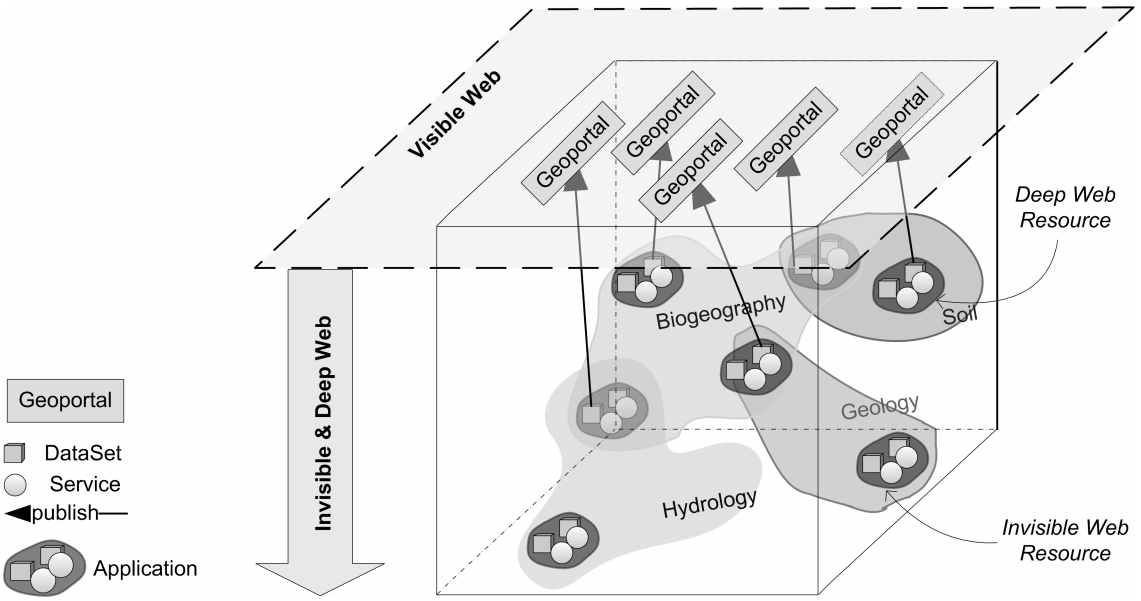

Figure 1.5: Geoportals as access points into multidisciplinary GCI cube for exploring research results.

has a long tradition in handling the dynamic aspects of Web content, and the aspects of user search by means of a search engine has been investigated intensively by the Web community (Broder, 2002; Rose and Levinson, 2004).

Another research community which technological advances are relevant to the search approaches used in the geospatial community is the Semantic Web community. These technologies are of particular importance for the geospatial community, because the Semantic Web aims to improve the reflection of human reasoning in sharing and processing information contained in the Web resources using automated tools. Such approach requires creation of ontologies – *"a formal, explicit specification of a shared conceptualisation"* (Gruber, 1993). The semantics that capture the cognitive content of Web resources might be presented in different ways. The easiest way is to add simple metadata, e.g., specially designed tags in XML–based format. The semantics might also be represented as data models via other Web resources that provide conceptual structures, for example as the Resource Description Framework (RDF) (Klyne and Carroll, 2004). The most complex but also the most rich in meaning are ontology–based semantics expressed in the form of RDF+RDFS (RDF Schema) (Brickley and Guha, 2004) or the Web Ontology Language (OWL) (W3C OWL Working Group, 2009).

The combination of RDF documents and the Hypertext Transfer Protocol (HTTP) has gained considerable interest in the Semantic Web community, as it allows publishing structured data on the Web as Linked Data (Bizer et al., 2009), and offers logic references to any related resources. The potential of the created Web of Data consists of the identification of a concept via a dereferenceable Universal Resource Identifier (URI) that permits retrieving the description of a concept from Web

as an RDF document. This document may contain references to other documents about the same concept (i.e. identifies its instances) or states the logic relation with other concepts referenced via their URIs. In this simple manner it is possible to create a web of interlaced concepts.

Also geographic information has to be accompanied with a knowledge backbone to be appropriately handled due to its peculiarities (Egenhofer, 2002). Therefore, the proper semantic description of geographic information seems to be the first step in the improvement of its usage. A methodology for referencing plain–text annotations to a backbone ontology is being considered by the geospatial information industry (Maué et al., 2009). These additional annotations might be added on three levels, (1) resource metadata (e.g., an OWS *Capabilities* document), (2) data model (e.g. a Geography Markup Language (GML) Application Schema (Portele, 2012)), and (3) data entities (e.g., a GML file). The formal specifications of concepts from the reference ontologies can be used then for tasks such as semantics–based information retrieval during workflow definition process. The semantic–based solutions have been applied successfully to improve searching within a catalog service, for example by search query expansion to add synonyms or to support multilingual queries (Latre et al., 2009), or event to enable on–demand delivery of geospatial information and knowledge (Yue et al., 2011).

1.3 Problem statement

Marvin Minsky (Minsky, 1986) acknowledges that there is no single, perfect principle, and vast diversity enables intelligence. Although he refers to human mind, this statement can be easily brought in other contexts. For example, Spatially Enabled Society requires the applications which use a variety of resources, including spatial data and services offered by SDIs. Also e–science consumes those resources and produces new outcomes that might be issue of incorporation to the existing infrastructures. This aspect of SDI advances some relevant issues. First, an SDI becomes a part of to a broader digital geographic information community, and search task might involve non–SDI resources. Next, the profile of a user of SDI resources evolves as well. There is a general assumption on expert dedication of SDI resources (Boes and Pavlova, 2008). However, the advances in semantic–based solutions dedicated to SDIs (Maué, 2008; Maué and Schade, 2009; Janowicz et al., 2009, 2010) and initiatives that promote an open and massive usage of geospatial data (e.g. Digital Earth (Craglia et al., 2008)), encourage non–expert users to join the SDI user community. Additionally, the Web is used as the technological solution for SDIs. This fact and open standards used by SDIs give possibilities to provide SDIs with an enhanced search for geospatial resources based on the practices of the Web community.

Thus, the vast diversity of SDI resources and the need of support for intelligent queries is the basis for the problem statement of this thesis. There is a hypothesis on search improvement within the Geospatial Web in the context of SDI that this thesis addresses.

"In order to improve searching for geospatial information and resources in the context of SDI, it is necessary to develop and provide systems which are able to use semantics and content–based heuristics for exploiting the published resources in an automatic manner."

1.4 Research questions

Searching is the process of trying to locate something specific (Beale, 2006). Therefore, considering the perspective of the geographic information community, there might be identified two main categories of searching:

- a search for information about a geographic feature, and
- a search for geographic information–bearing objects (GIBO) (i.e. data sets that relate to well–defined areas or footprints on Earth's surface (Goodchild and Zhou, 2003)), which includes search for GIBOs related to a geographic feature.

In general, the world can be represented using field–based and object–based views (Goodchild et al., 2007), where a geo–field (i.e. a surface or coverage) represents a set of continuous characteristics of a natural phenomenon, and a geo–object (i.e. entity or feature) represents a set of discrete characteristics of a natural phenomenon. The field/object distinction is closely related to human perception of the world, which is populated with discrete objects (e.g. a named place) and the environment properties are perceived as continuously varying fields (e.g. noise) (Couclelis, 1992). These distinctions in human perception influences the way of searching for information related to a location in the world. Application of the object–based view during search task can be related to the search by place names or textual descriptions of locations (e.g. address).

Human perception and SDI characteristics determine searching tasks. In general, there are three basic search patters that might be performed:

- *Searching for information about a feature.* It requires searching through discovered datasets and involves recognition of the searched feature in the collection.
- *Searching within an SDI for resources that provide information about a feature.* It requires searching through an SDI catalogue.
- *Searching among SDIs for resources that provide information about a feature.* It requires searching through a variety of SDI catalogues.

In the first scenario, a user has to access a variety of geographic information sources in order to find the required information (see the left site of Figure 1.6). A system that offers enhanced search for information about a geographic feature should be able to access different resources. Considering other search scenarios, the resources should belong to distinct SDIs, and Non–SDI resources should

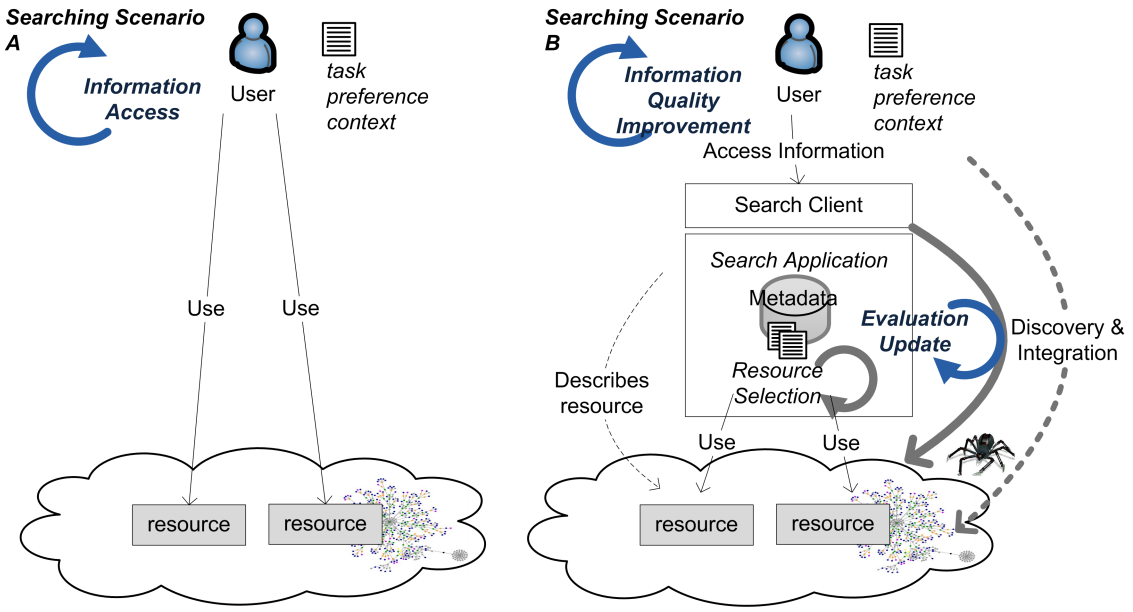

Figure 1.6: Scenario for searching for information about a feature within various resources, according to a general approach (on the left) and its enhanced version (on the right).

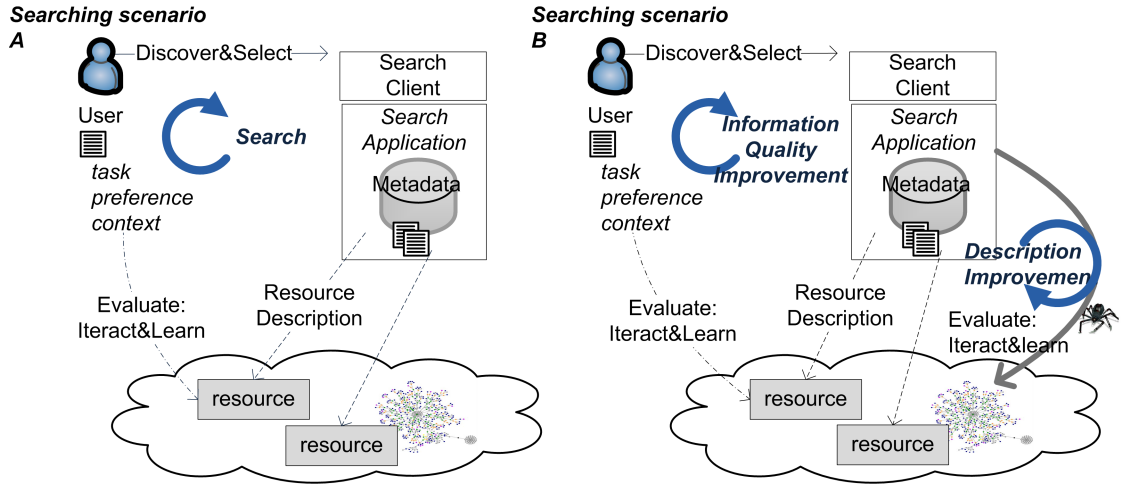

Figure 1.7: Scenario for searching within an SDI for resources that provide information about a feature, according to a general approach (on the left) and its enhanced version (on the right).

be used as well. Additionally, user requirements should be considered when performing search (see the right site of Figure 1.6).

The second scenario corresponds with a typical search task within an SDI (see the left site of Figure 1.7). Here, a user accesses a catalogue to search for resources. Then, the discovered resources are evaluated and selected eventually. The searching results can be improved by applying semantic

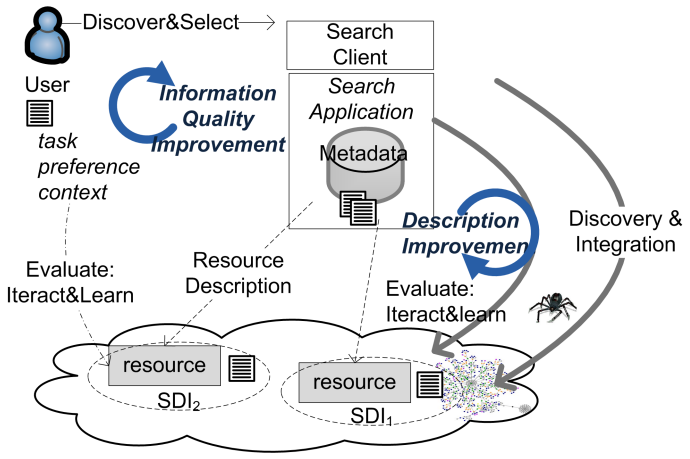

Figure 1.8: Scenario for searching for resources among different SDIs.

and content–based methods to extract additional metadata that permit developing approaches to improve the user experience (see the right site of Figure 1.7).

In the last scenario, a user searches for resources across the boundaries of different SDIs. The SDI approach is based on a catalogue which offers an integrated search over several catalogues from different SDIs. In this work, the Web context is considered. New catalogues can be discovered using typical approaches from the Web community, i. e. exploiting search engines to discover at least geoportals for their further exploration (as in the second scenario). In such scenario, the searching can be improved by enabling a system which is able to discover, analyse and index relevant geographic information resources found on the Web. Geoportals are entrance points to SDIs which can be exploited by automatic crawlers (see Figure 1.8). Additionally, non–SDI resources can be considered or these which are not registered within an SDI catalogue traditionally (e.g. a Web site of provider of geographic information resources).

A search system based on crawling the Web for the discovery of geographic information resources has some important advantages. First of all, the experience of the Web Service community shows that the publishing-finding-binding pattern applied by SDIs might not be successful, and crawler–based system can handle this issue in some degree. However, this problem can be mitigated by the policy–based character of an SDI, whose members are required to provide (i.e. publish and register) a minimal set of services (e.g. INSPIRE). Additionally, such system can discover and index some relevant geographic information resources which do not belong to any SDI (i.e. to be published in any catalogue), and they might be produced by experts of the geospatial communities or other communities of relevance.

Searching within various SDIs in the style of a web search engine seems to be a requirement which is caused by natural evolution in SDIs. Taking into account the incremental character of spatio-temporal data and e-science results (which might be available as raw data), the number of

resources produced by the live SDI community have an augmenting character as well. The user community, even if restricted to specialists (Craglia, 2007; Boes and Pavlova, 2008), is quite broad and not all of them have to be geographic information professionals who "*normally know what data are available and where*" (Gould, 2007). Advances in offering technological solutions to non–experts from a concrete domain (for example, grids applications, semantic frameworks that are dedicated to non-domain experts), increase number of potential users.

1.5 Methodology

The work is divided into separate although complementary applications which investigate possibilities for search improvement in the identified search scenarios in the context of SDI. The applied systematic methodology is related to software engineering. First, the problem is presented and analysed. The solution proposed to the problem is the result of a cyclic incremental development process, which is composed of:

1. **Analysis**. During the analysis process, the relevant research literature is reviewed to identify advances related to the research issue.
2. **Problem specification**. It provides a rationale of the motivations or the challenges for the research question.
3. **Conceptualisation**. A solution is proposed during the conceptualisation stage.
4. **Implementation**. The conceptualisation guides the implementation developments.
5. **Evaluation**. The evaluation applies the implementation to a concrete problem and evaluates its usefulness.

1.6 Scope

This research work has the following scope:

- **Feature–based Search**. Although Goodchild et al. (2007) identifies two possible views of the world (i.e. the geo–field and geo–object based), the search task is limited only to feature–based approach (e.g. searching for information about '*Madrit*').
- **Evaluation method**. Research literature on evaluation of Information Retrieval (IR) systems identifies four main theories about IR systems (Järvelin, 2011): ranking theory, search theory, information access theory, and information interaction theory. Figure 1.9 presents them as part of the nested evaluation frameworks. In this work, only the two first theories (ranking and search) are considered. The algorithms and methods developed are evaluated using the

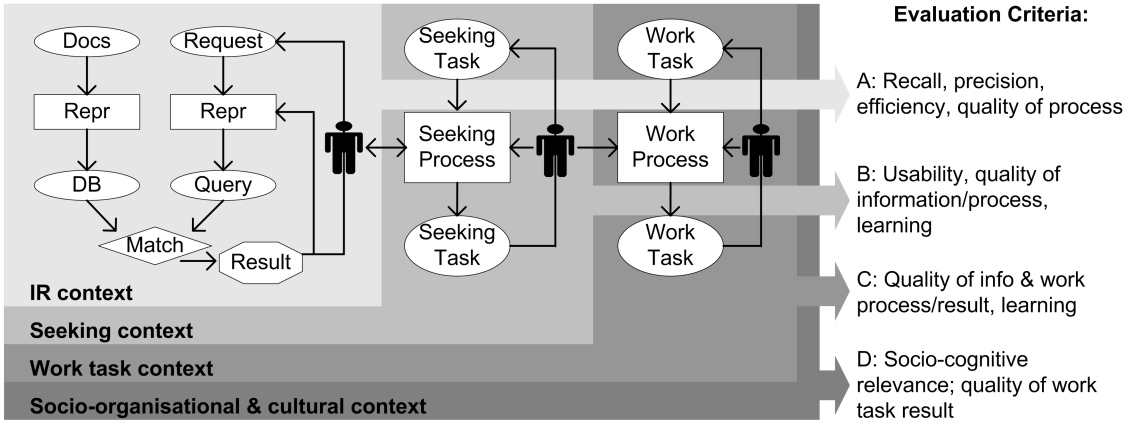

Figure 1.9: Nested evaluation frameworks of an IR system (source: Järvelin (2011)).

ranking theory. One of the systems developed in this work is evaluated using ranking and search theory. Two other systems are used within real applications. In such case, an empirical study of a prototyped version has not been considered.

- **Semantic support for search in SDI**. One of the relevant research questions is the semantic support in searching within and among catalogues of SDI. For example, the support for thematic search involves vocabulary mappings and/or query expansion. This is not targeted in this work. Here, semantic technologies are used as supporting tools to provide demanded functionality and the conceptualisation process is a secondary activity. These technologies are limited to D2R server and Apache Jena toolkit.

- **Knowledge representation**. In this work, several knowledge representation systems have been used. These systems are mainly simple SKOS (Simple Knowledge Organization System (Isaac and Summers, 2009), a W3C standard for porting knowledge organisation systems to the Semantic Web) vocabularies and `RDF/XML` graphs.

- **SDI resources**. In general, SDI resources that follow OGC specification have been in the focus of this work. However, there are some exceptions. In the case of the compound geocoding application, services from different SDIs have been integrated, whose interfaces do not necessarily follow the corresponding OGC specifications. A non–SDI resource has been used as well. The system for automatic generation of geographic metadata for Web resources focuses on Web sites of OWS providers. Web sites are also considered by the Geospatial Web Search Engine.

- **Web metadata**. A prototype of the system for automatic generation of geographic metadata for Web resources is limited to HyperText Markup Language (`HTML`) documents compliant with the HTML 4 W3C Recommendation (Raggett et al., 1999). It uses the header section

to extract the *meta* elements. Although there exists another manner to embed semantic information within a Web page (e.g. microformats[12]), the presented content–based methods only process the HTML elements.

1.7 Contributions

This work aims to define basic search patterns within the Geospatial Web from the SDI perspective. These patterns are used to show the need for systems that apply semantic and content–based heuristics to enable enhanced search.

- First, this thesis identifies three approaches for searching information in the context of SDI.
- Second, this thesis develops a compound architecture for searching information about a geographic feature within several datasets.
- Third, this thesis describes two approaches that enhance the task of searching for resources within an SDI catalogue.
- Fourth, this thesis presents a system for the automatic generation of geographic metadata for Web resources.
- Fifth, this thesis presents a Geospatial Web Search Engine tailored to the needs of a non–expert user.

1.8 Thesis structure

Due to the interdisciplinary nature of this thesis, each Chapter may include a background section. They are organised as follows.

Chapter 2 describes an approach for improving the search for information about a geographic feature. It introduces the concept of *geocoding* and related terms, such as *georeferencing*, *geotagging* and *geolocating*. Aspects of the quality of a service–based geocoding system are identified (i.e. quality of Web service, quality of spatial data and quality of geocoding process). The application developed for address geocoding is deployed in a real environment.

Chapter 3 discusses the improvement of an SDI catalogue by applying semantics and content–based methods. Two different approaches are presented and applied in a real application. The first one shows how semantic Web technologies can be used to improve the precision of spatial search. Linking approaches from the semantic Web community and the Geospatial Web community are discussed. The method proposed is based on the idea of abstraction of a geographic feature from its spatial definition. Modelling the spatial representation of a geo–concept and the idea of geoidentifier

[12] http://microformats.org/

are showed, and used to create the *administrative geography of Spain*. The second part is devoted to the automatic identification of orthoimages offered by a WMS service. A general classification of content published by WMS layer is presented. The method developed uses heuristics based on analysis of the published content in order to provide additional semantic annotations. In addition, the work outlines characteristics of content published by WMS services at the end of 2010.

Chapter 4 presents an architecture for the automatic generation of geographic metadata for Web resources. First, the methods of generation of metadata from the Web community and the geospatial community are contrasted, and geographic metadata in `HTML` Web pages are examined. A method for geographic coverage estimation of Web pages is proposed as well. An empirical study shows that straightforward heuristics can provide this information automatically when a publisher does not offer it. Moreover, empirical study provides a brief overview of the characteristics of the OWS publishers, and reveals the current practices in the geospatial community in the provision of metadata for Web pages.

Chapter 5 presents a Geospatial Web Search Engine which is intended to support non–expert users (i.e. users who lack expertise in the geospatial domain) in achieving their search goals. Searching for Geospatial Web resources is discussed in more detail from the Web community perspective, and the issue of search for Web Services is treated as a starting point.

Finally, Chapter 6 presents the central result of this thesis, summarises the contributions, and ends with suggestions for future work.

Chapter 2

Enhanced search for a geospatial entity

2.1 Introduction

This Chapter presents an enhanced search for information about a geographic feature within different information sources, i.e. Geospatial Web resources. Geocoding is the principal functionality of a system which supports text–based searches within the geospatial domain (Section 2.2). Therefore, this functionality is the target of this work. Existing geocoding systems are generally limited to assign a geographic coordinate to an absolute location, such as a street address. However, many applications (e.g. urban management systems) need to georeference a location description in a more flexible way. Different types of information about a geographic feature may be required when the search task is performed (e.g. a polygon or a point as the spatial object), and the search system should be able to adapt to such changes. A solution might be the usage of various sources of information. In such systems, the quality of Web Service (i.e. QoS for the Web), the quality of spatial data, and the quality of geocoding process influence system behaviour (Section 2.4). A compound architecture to support georeferencing is developed that integrates existing geospatial data services by using patterns (Section 2.5). Then, a framework for address geocoding is developed (Section 2.6), and the implemented application is applied in the real environment (Section 2.6.4). Finally, some conclusions are presented in Section 2.7.

2.2 Terminology

The term *georeferencing* means the act (and processes) of "*relating information to geographic location*" (Hill, 2006). As "*whatever occurs, occurs in space and time*" (Wegener, 2000) almost any

information may contain reference to space. Although the object of georeferencing can be almost anything, the Digital Library (DL) community identifies two main means to refer to locations: the informal and the formal (Hill, 2006). The first one, which is used in ordinary discourse, means referring to locations by using place names (i.e. *toponyms*). The formal representations are geospatial footprints, based on longitude and latitude coordinates or other spatial referencing systems (SRS). In this thesis, the informal reference can be a textual reference to a geographic feature (e.g. an address, a place name), or a concept that can be related with a specific location in the established permanent manner (e.g. telephone number prefix), and any textual descriptions of a location including relative locations (Hutchinson and Veenendall, 2005). As for formal reference, it may be a direct or indirect reference. The direct formal reference is a footprint. A geographic reference system is a source of indirect formal references, i.e. *geo–codes*, which can be transformed into corresponding footprints. For example, an ISO country code (ISO, 2007a) (e.g. `ESP`) can be used as a *geo–code* if its geographic reference system is available (e.g. a geospatial dataset that contains footprints associated with ISO country codes). A reference system is crucial for appropriate interpretation of coordinates as well. For example, longitude and latitude coordinates can have WGS84 or ETRS89 as their spatial reference system. Therefore, a footprint should have reference system associated as well.

Geocoding is a kind of georeferencing, and in this work, it means *the act of turning textual description of location into a formal geospatial reference, where formal geospatial reference is a footprint (direct geocoding) or a geo–code that allows the user to identify the footprint unambiguously via the associated geographic reference system (indirect geocoding)*. There are a lot of different definitions of the term *geocoding*. Nevertheless, the definition used in this thesis is close to those used in Margoulies (2001), where geocoding process "*transforms a description of a feature location, such as a place name, street address or postal code, into a normalised description of the location, which includes a coordinate geometry.*" Figure 2.1 shows the general idea and an example of direct and indirect geocoding.

Although the type of geocoded information can be diverse (Margoulies, 2001), availability of georeferenced datasets and geocoding techniques determine the geocoding target. Address has been the target of study in the field of geocoding for a long time (Ratcliffe, 2001; Bonner et al., 2003; Davis et al., 2003; McElroy et al., 2003; Cayo and Talbot, 2003; Yang et al., 2004; Bakshi et al., 2004; Whitsel et al., 2004). There are systems that permit geocoding to other classes of geospatially related objects such as toponyms (Kimler, 2004), host IP or telephone numbers. In general, any geospatial dataset that gathers textual representations of geographic features or geospatially related objects can be used for geocoding.

Existing geocoding systems are generally limited to assign a geographic coordinate to an absolute location such as a street address. However, many applications (e.g. urban management systems) need to geocode location descriptions in a more flexible way. For example, it is common that a citizen

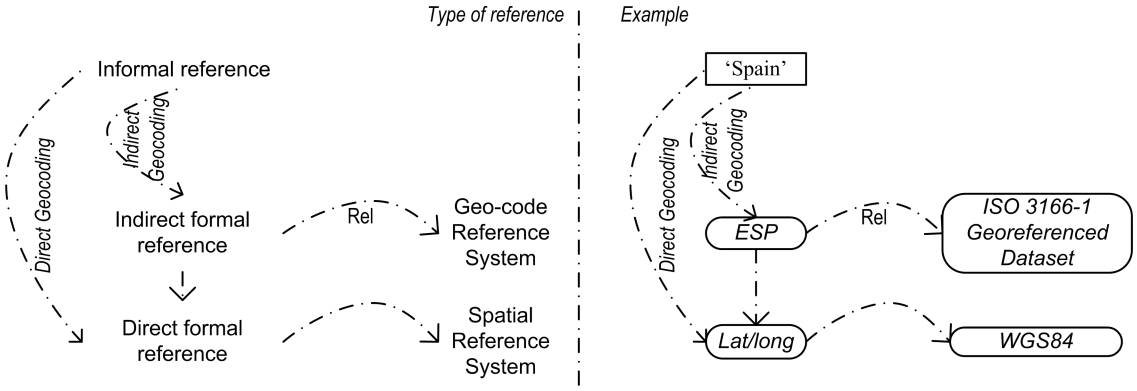

Figure 2.1: Direct and indirect geospatial representation.

who calls an emergency centre does not know the address where he/she is and the descriptive information that the citizen provides is ambiguous or even confusing (e.g. "100 meters from a memorial statue in the park, which is situated by 'Colon' Street"). In general, according to Kimler (2004), the process of georeferencing an unstructured text with geographical references involves *geoparsing* (i.e. extracting geographical references from texts) and geocoding (i.e. resolving geographical references). However, in such a situation the user needs to be *geolocated* in order to estimate unknown position by using known position of other objects. Long before the existence of the Global Positioning System (GPS) geolocation has been used by sailors by fixing their relative position according to the sun, stars, and landmarks (Beal, 2003).Hutchinson and Veenendall (2005) define *geolocating* as a process that permits users to assign valid geographic codes to free–form textual descriptions of locations. Conventional geocoding services that work with absolute locations will not be able to determine the coordinates of the place of incident. The best solution is a geolocation service, which might correctly interpret the relative location (e.g. distance, direction), with landmarks (e.g. park, railway station) and features (e.g. coffee shop). Hutchinson and Veenendall (2005) enumerate the following elements that contribute to the concept of geolocating: query syntax flexibility, user interaction, user context consciousness and complex site representation. It is important to stress that the authors argue for a new methodology, however, they do not include details of how one would implement it.

Although some authors understand geolocating as evolution of geocoding (Goldberg et al., 2007), here this process is seen as a kind of georeferencing. Such process involves not only geocoding parsed location descriptions but it has to handle spatial relationships among identified objects, i.e. topological ("inside of"), metric ("500 meters from"), fuzzy ("near"), ordered ("in front of") (Lo and Yeung, 2006).

2.3 Geocoding Web services

The most prevalent way for providing direct geocoding functionality is a geocoding service. One of the early proposals for standardisation of the geocoding service is the OGC Geocoder Service (Margoulies, 2001). Currently, geocoding functionality has become part of the OpenLS core services (Mabrouk et al., 2008), where it is defined as "*A network–accessible service that transforms a description of a location, such as a place name, street address or postal code, into a normalized description of the location with a Point geometry (...).*" This specification permits the implementation of the traditional geocoding functionality. Since WPS services allow some geospatial processing functionality to be enclosed, a geocoding service can be also provided in conformance with the OGC WPS specification.

Today, most of the geocoding service providers use their own specifications, despite the existence of the open standards. Nevertheless, geocoding Web services are not difficult to use. The problem is rather finding and selecting the right supplier. The type of geocoded object and the terms of service (ToS) are the principal characteristics considered by a Web user. In general, regardless of service origin, which might be private sector (e.g. Google[1], ViaMichelin[2] or GeoNames (Wick, 2012)), public sector (e.g. SwissSearch[3]) or volunteer communities (e.g. opengeocoding.org (Behr and Rimayanti, 2008)), the services may be divided into three groups due to their ToS: a community access service (e.g. SwissSearch, the Swiss cantons only allow the Web sites of the federal government to use the addresses search service), paid access services (e.g. GeoNames for professional users), gratis services but with some restrictions (e.g. Google), and gratis services of free use (Behr and Rimayanti, 2008).

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

Also, it should be considered that new types of geocoding service applications, such as support for mobile application, demand supplementary characteristics. Location based services (LBSs) require the support of geocoding services in tracking of user location and the reverse geocoding at the level of operating system (e.g. the Android[4] or GeoClue[5] projects). The availability and capabilities of these services have to be adjusted to the requirements of mobile devices (e.g. battery life, a cellular network, access to the Web).

[1] http://maps.google.com/
[2] http://www.viamichelin.co.uk/
[3] http://api.geo.admin.ch/main/wsgi/doc/build/services/sdiservices.html
[4] http://code.google.com/android/
[5] http://www.freedesktop.org/wiki/Software/GeoClue

The choice of service is determined by the use case. Free geocoding Web services are appropriate for *geotagging* (i.e. the act of adding the geographic metadata to any kind of on–line resource) of the local news or incidents (e.g. water supply shortage, planned roadwork) because such information does not require geocoding services of high quality. On the other hand, the systems on which depend public health (Bonner et al., 2003), public security (Ratcliffe, 2004) or environmental services (Ratcliffe, 2001) require high quality of service and data. For example, the speed and efficacy of fire–fighters depend on the information they possess such as the location, accessibility and characteristics of the building in fire (e.g. number of floors, shape, location of entrances and the accessibility to the building, nearby buildings) and fire hydrants.

The vast heterogeneity of geocoding services and the specific features of geographic data set up the open problem of provider selection. There are many works in the context of the service discovery and selection. Some proposals need prior service evaluation (e.g. a rating agency (Sriharee, 2006) or a user (Manikrao and Prabhakar, 2005) pre–evaluation), but most of the works in this area use typical QoS features (Yu and jay Lin, 2005; Wang et al., 2006; Tsesmetzis et al., 2008). The research community has also shown interested in services of geographic information (Fallahi et al., 2008; Lan and Huang, 2007) but they include only basic concepts (e.g. coverage) and do not exploit specific characteristics of geographic data in discovery and selection processes (e.g. reasoning based on coverage, quality of geographic objects).

2.4 Quality of geocoding service

Quality is defined as "*totality of characteristics of a product that bear on its ability to satisfy stated and implied needs*" (ISO/TC 211, 2002). In this work, the quality factors of a geocoding Web service are considered. Therefore, the recommendations from the Web Service community and the geospatial community should be considered. Additionally, some specific issues related to the geocoding process will be discussed.

2.4.1 Web service

The principal standardising organisations that work on standards related with Web services are the W3C[6] and the Organization for the Advancement of Structured Information Standards (OASIS, 2012) (OASIS). W3C's primary activity is "*to developing protocols and guidelines that ensure long-term growth for the Web. W3C's standards define key parts of what makes the World Wide Web works*".[7] OASIS is "*a not–for–profit consortium that drives the development, convergence and adoption of open standards for the global information society*" (OASIS, 2012). W3C provide a draft document that describes QoS requirements for Web Services (Lee et al., 2003). However, a recent

[6] http://www.w3.org/
[7] http://www.w3.org/Help/

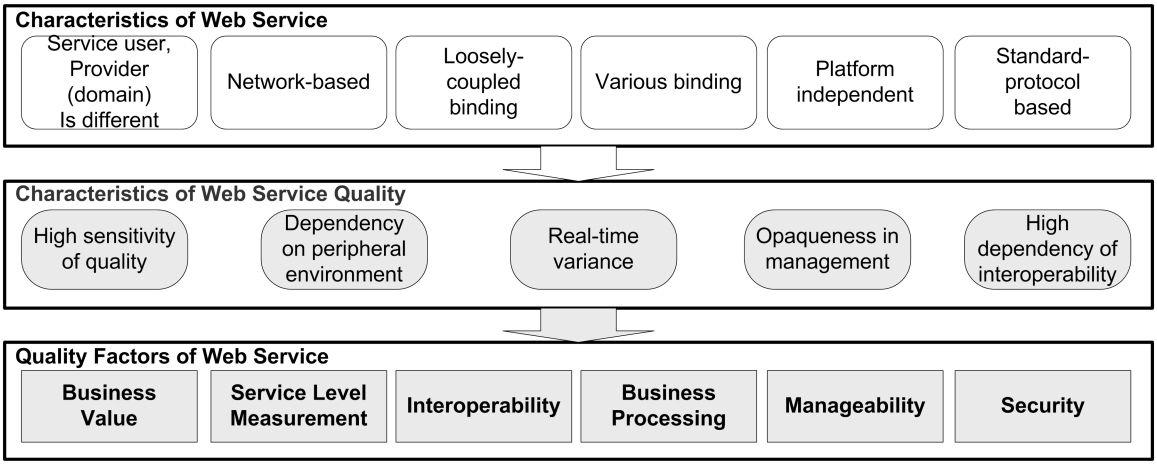

Figure 2.2: Extracting Quality Factors of Web Service (source: (Kim et al., 2011)).

and more complete model of QoS for Web services is the Web service quality factor (Kim et al., 2011) (WSQF), a standard approved by the OASIS Web Services Quality Model (WSQM) Technical Committee. This standard defines common criteria to evaluate quality levels for interoperability, security, and manageability of services. A WSQF refers to *"a group of items which represent web service's functional and non–functional properties (or values) to share the concept of web services quality among web service stakeholders.* Functional quality reflects how well it complies/conforms to a given design, based on functional requirements or specifications. Non–functional quality refers to how that a service meets non–functional requirements that support the delivery of the functional requirements, such as robustness or interoperability, and the degree to which the service was produced correctly. The WSQFs have been induced from the basic characteristics of Web services (Figure 2.2), and are composed of business value quality, service level measurement quality, interoperability quality, business processing quality, manageability quality and security quality. Depending on business perspective or system perspective, they can be categorised into two groups: (1) the business quality group and (2) the system quality group (Figure 2.3).

There is a proposal (Mabrouk et al., 2009) for an extension of the WSQM dedicated to the service environments such as user mobility and context awareness of application services. It provides an overview of several ontologies:

- QoS Core ontology (based on an obsolete version of Web Services Quality Description Language (WSQDL) (Lee and Kim, 2006) published by the WSQM Technical Committee in 2007).
- Infrastructure QoS ontology.
- Service QoS ontology (based on an obsolete version of OASIS WSQM published in 2007).
- User QoS ontology.

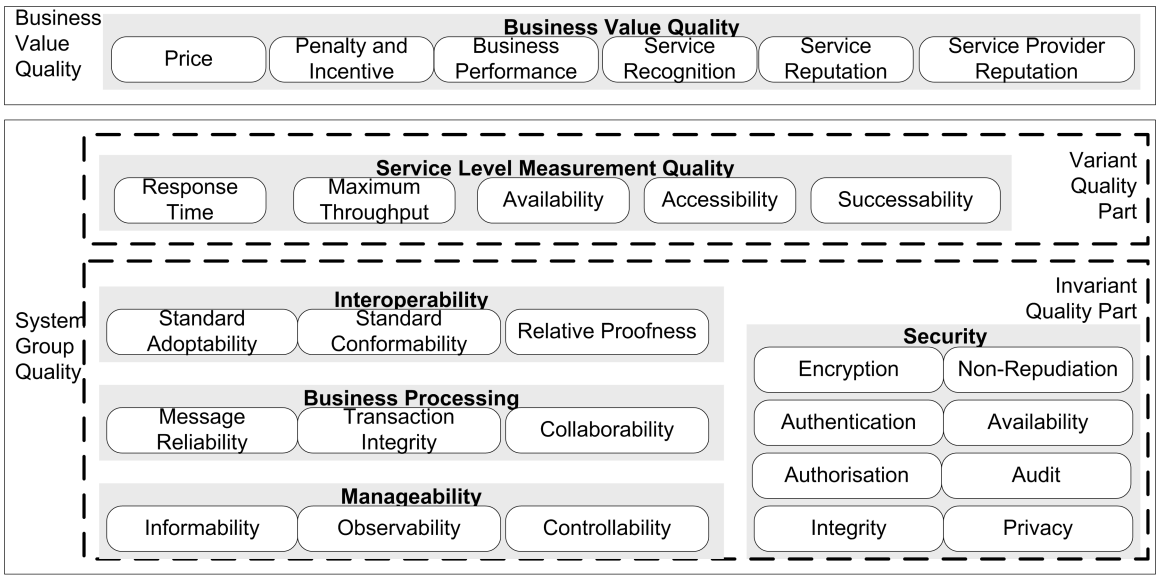

Figure 2.3: Structure of Web Services quality factor (source: (Kim et al., 2011)).

The service QoS ontology considers dynamic capabilities and domain–specific qualities as well. The proposed model can be a general framework to provide the appropriate ground for several engineering capabilities including QoS requirements engineering and QoS–based service engineering (e.g., service discovery, Service–level agreement (SLA) management, monitoring). Although the model is built on obsolete works of OASIS, the discussed issues remain relevant.

Quality of service is also referred in SDI. INSPIRE provide an Implementation Rules document that refers to service quality: the Network Services Performance Guidelines (EC, 2007b). It tries to define minimum performance criteria for the INSPIRE Network Services. Although not all technical specifications of the Network Services are currently available, these guidelines are intended to be applicable to all and to be compatible with the Network Services Architecture Document (INSPIRE NS DT, 2008). These general guidelines have been defined by revising existing standards and recommendations (including the described above), and identify the following attributes of QoS for the INSPIRE Network Services: Performance, Reliability, Capacity, Availability, Security, Regulatory, and Interoperability. The concepts definitions correspond with those provided in Lee et al. (2003).

2.4.2 Spatial data quality

van Oort (2005) identifies several reasons for concerns about spatial data quality issues, as follows:

- *There is an increasing availability, exchange and use of spatial data.*
- *There is a growing group of users less aware of spatial data quality.*

- *GIS enable the use of spatial data in all sorts of applications, regardless of the appropriateness with regard to data quality.*
- *Current GIS offer hardly any tools for handling spatial quality.*
- *There is an increasing distance between those who use the spatial data (the end users) and those who are best informed about the quality of the spatial data (the producers).*

There are intensive standardisation works on handling quality within the geospatial domain. The ISO standards, for example, provide quality principles and define specific concepts (ISO 19113 (ISO/TC 211, 2002)), define principles for quality evaluation (ISO 19114 (ISO/TC 211, 2003)), and provide description of quality assessment methodologies (ISO 19138 (ISO/TC 211, 2006)). As for data quality, ISO 19157 standard revises ISO 19113, ISO 19114 and ISO 19138, and defines a set of measures for the spatial data quality elements identified in ISO 19113. At the time of writing this thesis the standard is still under development (i.e. at enquiry stage[8]).

Sharing and reusing spatial data require paying special attention to quality of spatial data, therefore this issue is relevant in any SDI. Quality has to be considered from different perspectives in an SDI. There might be different viewpoints used to describe quality (Garvin, 1988; Jakobsson, 2006). In his thesis (Jakobsson, 2006), Jakobsson argues that quality management viewpoints are important from the SDI's point of view. He discusses geographic quality concepts using these four viewpoints (Figure 2.4):

- *Production–centred viewpoint.* This perspective focuses on the variations in the production process where the most common measure is the number of defective or non–conforming products.
- *Planning–centred viewpoint.* This perspective focuses on the characteristics of products.
- *Customer–centred viewpoint.* This perspective focuses on the value of products and services to the customer.
- *System–centred viewpoint.* This perspective takes into account all stakeholders who are influenced by the organisation or its products oriented quality.

In the case of INSPIRE, Article 17 of the INSPIRE Directive (EC, 2007a) says: "*Each Member State shall adopt measures for the sharing of spatial data sets between its public authorities ... for the purposes of public tasks that may have an impact on the environment.*" Therefore, an effort within INSPIRE is dedicated to the development of some methods for assessing, measuring, reporting and controlling spatial data quality. These aspects have been considered by the European Spatial Data Infrastructure with a Best Practice Network (ESDIN) project[9] supported by eContent+ programme.

[8] http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=32575
[9] http://www.esdin.eu/

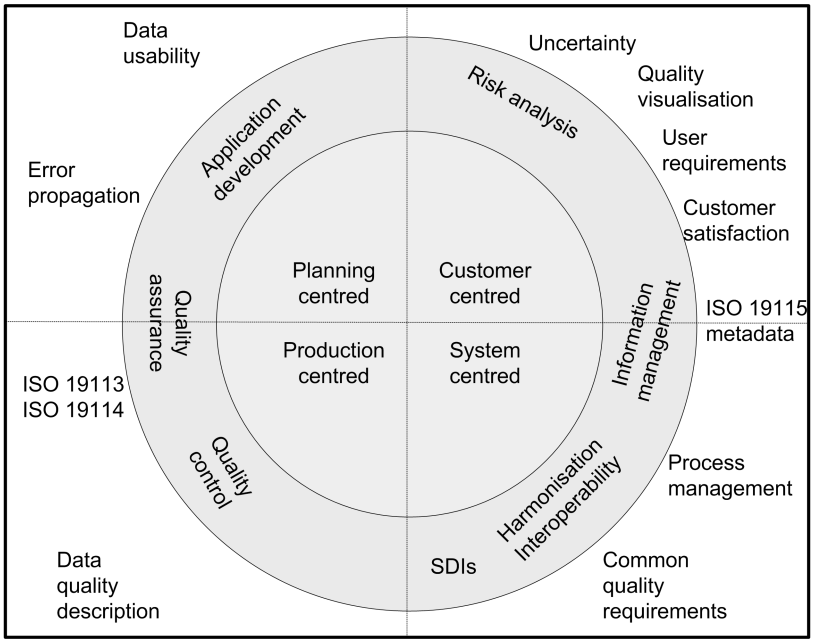

Figure 2.4: Different approaches to geographic information quality from the quality management viewpoint (source: (Jakobsson, 2006)).

Data quality is also considered by the FGDC CSDGM standard (which incorporates the Spatial Data Transfer Standard (SDTS) (FGDC, 1998b) that contains a section on spatial data quality elements) and the NAP standard.

There is a remarkable agreement among the documents on the elements of spatial data quality. Each of the standards that approach that question describes the same core elements:

- *Attribute (Thematic) Accuracy.* CSDGM and SDTS use term "*Attribute Accuracy*", and ISO 19115 refer to the same content as "*Thematic Accuracy*". It can be defined as "*an assessment of the accuracy of the identification of entities and assignment of attribute values in the data set.*" (FGDC, 1998a).

- *Completeness.* It refers to "*information about omissions, selection criteria, generalization, definitions used, and other rules used to derive the data set*" (FGDC, 1998a).

- *Lineage.* It refers to the "*information about the events, parameters, and source data which constructed the data set, and information about the responsible parties*" (FGDC, 1998a).

- *Logical Consistency.* It refers to "*an explanation of the fidelity of relationships in the data set and tests used.*" (FGDC, 1998a).

- *Positional Accuracy.* It is "*an assessment of the accuracy of the positions of spatial objects*" (FGDC, 1998a).

- *Temporal Accuracy.* It is usually defined as "*accuracy of the temporal attributes and temporal relationships of features*" (ISO/TC 211, 2002)

2.4.3 Geocoding quality

Geocoding quality has been investigated intensively in the fields related with spatial analysis, and there is a lot of work on the evaluation of geocoding system accuracy. In general, the typical metrics used to determine fitness-for-use are as follows (Goldberg et al., 2010):

- *Match–rate.* It can be compared to the precision and recall of IR systems, and it refers to the number of input addresses that a geocoding system was able to match.

- *Match type.* It is the level of geographic feature matched to e.g., parcel centroid, postal code, street address.

- *Match certainty.* This value describes the level of similarity and/or likelihood of a match between the input address and the address associated with the matched feature input derived either probabilistically or deterministically.

- *Spatial accuracy.* It is defined usually as an average values of distance and direction "from truth" by comparing computed output locations and known locations for a subset of the data.

However, these metrics do not properly represent the uncertainty value of geocoding results in order to be adequately used in spatial analysis (Zandbergen, 2009). A typical geocoding process is composed of (1) data cleaning, (2) feature matching and (3) feature interpolation. Figure 2.5 outlines a generalised workflow of this process. As geocoding process has influence on geocoding quality, each component of the geocoding workflow has influence on geocoding quality:

- The quality of the input data which is the text describing a location.

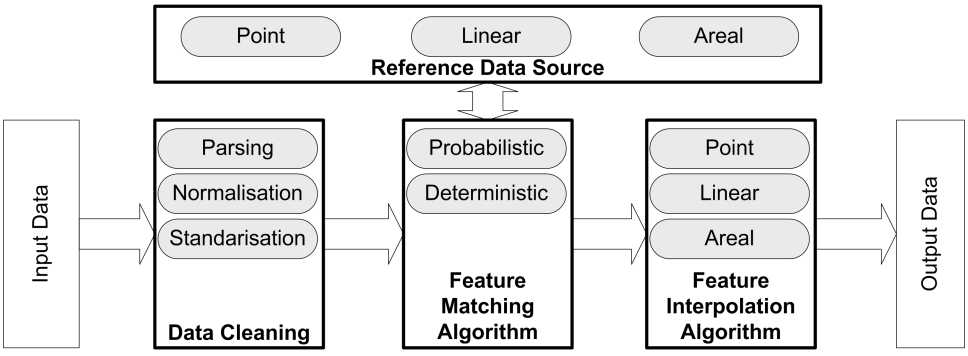

Figure 2.5: The Generalised abstraction of the geocoding process (source: (Goldberg, 2008)).

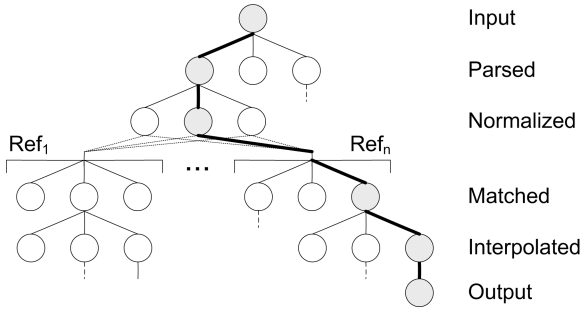

Figure 2.6: The alternative paths in geocode production (source: (Goldberg et al., 2010)).

- The input data parsing and normalisation algorithms which identify the pieces of the input text and transform them to standard values (Davis and Fonseca, 2007).
- The feature matching algorithms which identify candidate matching geographic features in the reference data sources.
- The feature interpolation algorithms.

Some works attempt to quantify an overall propagated uncertainty value of geocoding results (Davis and Fonseca, 2007; Zandbergen, 2009; Goldberg et al., 2010). An interesting approach to geocoding accuracy is presented in Goldberg et al. (2010). A geocoding process is seen as "*a decision tree with multiple potential outcomes at each level (transformation) that guide the set of choices available to all subsequent levels*" (see Figure 2.6). The authors provide discussion on quantitative accuracy metrics to describe the quality of geocoded data from the perspective of the spatial certainty, and propose to report the results of the geocoding process as spatial probability distributions. This method assumes availability of several quantitative factors to describe the spatial–temporal aspects of accuracy and uncertainty for each component. However, the authors admit that the quantities are not well defined, which is stated as future work.

2.5 Compound geocoder

Geolocating process demands spatial information of a wide range of types. Therefore, a compound approach seems to be a suitable architecture model to provide an enhanced geocoding functionality. The proposed compound geocoding architecture allows the building of hybrid solutions composed of different services (e.g. geocoding, gazetteer or cadastral service) enhanced with a geocoding layer if necessary. The task of geocoding service selection exploits geographic ontologies. The Administrative Unit Ontology (AUO) (López-Pellicer et al., 2008) is used for (i) provider selection and (ii) data integration. This approach may increase the flexibility and adaptability of applications. In the case

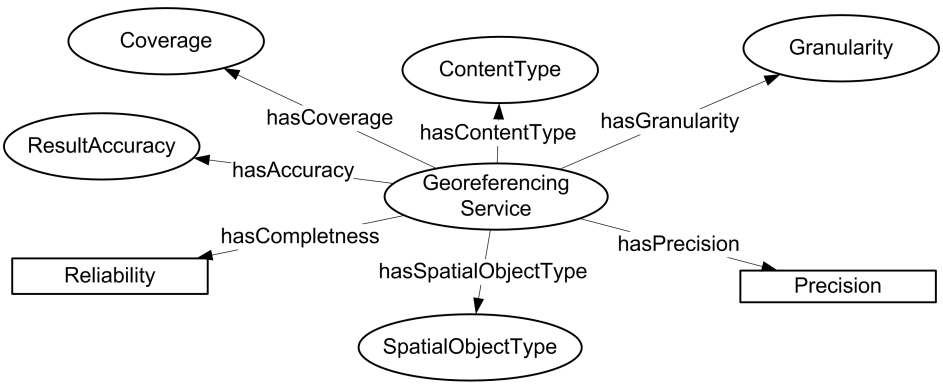

Figure 2.7: The selected characteristics of geocoding service.

of the public services, it provides access to different services, national and local, in a transparent manner and ensures the use of updated data.

First, the characteristics of geospatial services considered in this work for service selection are presented. Then, a patter–based method for service integration is briefly outlined. Finally, the architecture proposed is outlined.

Geocoding service quality

The proposed architecture allows the service selection according to the use case requirements. Therefore, the proper description of each source is vital for the behaviour of the whole system. Figure 2.7 presents the features to consider during the evaluation of providers. The obtained values are stored in RDF files with corresponding semantic annotations. These values are essential for the service selection and the decision making process.

The main features of each geocoding service are (1) the *coverage*, (2) the *type of content* and (3) the *type of spatial object*. The first two features are always given by the provider or might be indicated by the name of the service. The first one defines the area in which offered data are situated. Usually, this area corresponds with the unit or set of units from political division of territory, because the majority of georeferencing Web services are provided by public administrations from local and central level. The solution proposed is based on the use of AUO as the ontology of political organisation of the territory. Using instances of classes which are specialisations of *jurisdictional geographic object* concept from AUO as value of the coverage allows reasoning on the semantic relations among them, rather then relying on their spatial objects. For example, the province of Zaragoza has a set of municipality members and is part of Spain. All municipalities which belong to Zaragoza province are also members of the Aragon Autonomous Community. If searching for services which provide data from the province of Zaragoza, system will also consider those services whose coverage equals the coverage of to Aragon, Spain or the world, or even might return a set of services whose

coverage combination corresponds with the required coverage, e.g. local services of all municipalities of province of Zaragoza.

The type of content strictly depends on the types of geographic feature provided by service. The last one, the type of spatial object, indicates the list of provided types of spatial object, such as point, polygon or 3D entity. For example, the Cadastre Service of Spain[10] (*Servicio de Catastro de España*), as the name of the Web service indicates, has the coverage of Spain and offers centroids of parcels. Google Maps, according to the online documentation, has world coverage and its type of content is street address geocoded via point.

From the analysis of spatial data, it is possible to obtain two additional indicators (of range 0 – 1): (5) the *reliability* and (6) the *precision*. The reliability indicates the capacity of representation of elements of physical world by the content. The service that offers all elements of the real world has a reliability value equal to '1' (100%). The indicator of precision informs about the average positional error of the whole dataset. It is important to note that this indicator may be influenced by the difference between the provided spatial object and the search one. For example, using cadastral data for the address geocoding, there will be a decrease in spatial data precision.

The above presented factors can be used to express the quality of a spatial dataset as a result of comparison with a baseline dataset. Such relative estimation requires semantic accuracy of compared objects. This aspect is related with (4) the *result accuracy*, a typical geocoding quality factor (i.e. the *match type*). It should not be misinterpreted as "*data accuracy*", the term commonly used in literature to describe positional accuracy. Result accuracy is estimated for each source via analysis of the source data model and the application domain model. It indicates the capacity of source to fulfil the domain model, and is represented via the last domain model field which the source might score.

Frequently, values of the reliability and / or the precision of spatial content can vary in function of the area of relevance (e.g. new suburbs might be even omitted). Such feature, called the *granularity*, might be obtained from the exhaustive evaluation of spatial content along with its semantic analysis (e.g. by distinguishing types of geographic features as: cities of population more than 500.000, cities of population more than 100.000, towns of population more than 10.000, villages, and hamlets).

Mediator layer

Services selected to be used as the external resources for information search may vary in terms of technological solutions used by providers. Adding a new, common functionality to a service or over a set of existing services, requires creation of a *mediator service*. Depending on user needs and characteristics of the service which should be adapted, one of the following design patterns might be used to develop a mediator service: adapter or façade. The adapter pattern allows the user to access to functionality of an object via known interfaces and/or adopting message channel (Gamma et al.,

[10] `http://ovc.catastro.meh.es/`

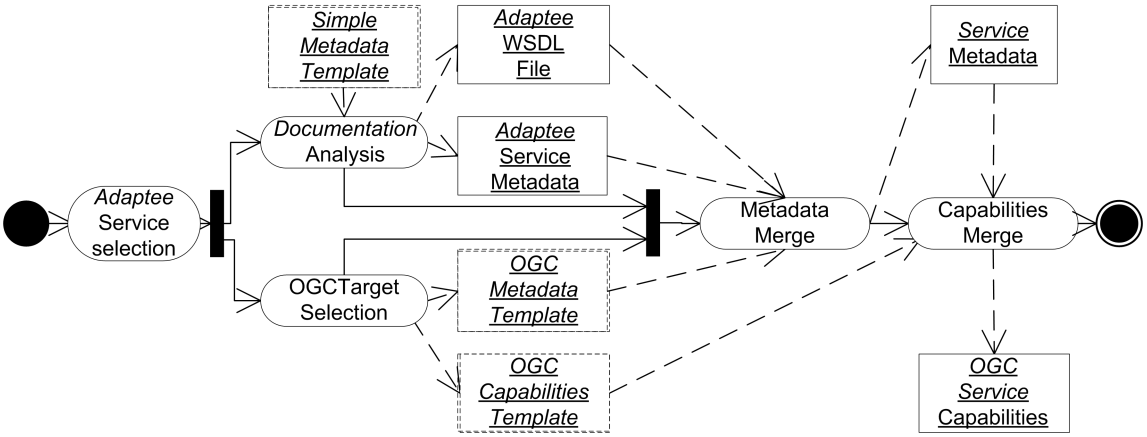

Figure 2.8: Workflow for the metadata generation of a mediator service.

1994; Hohpe and Woolf, 2003). In the majority of cases, the mediator service will implement this pattern to encapsulate the invocation of the original service according to specific requirements. In addition, it may hide the logic of multiple requests or error handling. The façade pattern provides "*a unified interface to a set of interfaces in a subsystem. Façade defines a higher–level interface that makes the subsystem easier to use.*" (Gamma et al., 1994). In terms of services, this approach addresses subsystems which combine several services and allows their logic to be concealed.

For generalisation purposes, an abstract method for service integration into an OGC mediator layer is proposed. The method applies the adapter or façade design patterns to create a mediator service. The implementation of a mediator service requires the identification of the source service (or a set of services) (i.e. the *Adaptee Service*) and the functionality which should be provided by the target service. The target service is an OGC service which may offer the desired functionality. The method produces several metadata documents which will be used as guidance during implementation of the mediator, i.e. the *Capabilities* document that conforms to the selected specification and some additional service metadata (e.g. the ToS of the *Adaptee Service*). The first step of the proposed method is the selection of a base service (or services) to be adapted (see Figure 2.8). A simple metadata template should be filled with information found in the available documentation of the chosen service. Although this might sound rather straightforward, in practice it is not so simple. Popular Web services (e.g. the Google service family) lack documentation in a standardised format, for example a Web Service Definition Language (`WSDL`) document. Commonly, they only provide human–readable documents, an API, or sample code that the service provider publishes for developers to use. Less frequently a `WSDL` description is provided. If there is not a `WSDL` file, this method requires the creation of one during the documentation analysis. The created *Adaptee WSDL File* and the fulfilled service metadata template that produces the *Adaptee Service Metadata* are the result of the documentation analysis.

The task of the target service selection might be performed independently from the documentation analysis. In the most general case, the characteristics of the *Adaptee Service* and the desired functionality will provide enough information. The selected specification allows development of a set of template files, i.e. the *OGC Capabilities Template* file and the *OGC Metadata Template* file. In this work the OGC WPS has been chosen as the base to offer a common geocoding interface.

The next step is the creation of the mediator service metadata by merging the simple metadata file and the WSDL file of the *Adaptee Service*. The resulting dataset metadata (i.e. the *ServiceMetadata*) will be used with an *OGC Capabilities Template* in order to create the *Capabilities* document of the target service. The merging process is semi–automatic, and the resulting files should be revised to check if they fulfil the OGC and target functionality requirements. The *Capabilities* document generated in this workflow is used as a guide for the implementation of the mediator service.

Architecture

The compound geocoding architecture can use different sources of geographic information, such as gazetteer, street data, geocoding or cadastral services. Each source has to be enhanced with a geocoding capacity (here, a generic connector to the geocoding mediator is provided) and described in terms of the introduced characteristics of geocoding service whose values are provided as semantic annotation of service. The proper description of sources is vital for the behaviour of the whole system due to the fact that the obtained values are used as the clues for the source selection, which determines the adequate functionality of the system. The result accuracy feature is defined via comparison of the application domain model of searched information with the response data model of the service. This requires a well defined domain model or a set of them. To simplify the data integration tasks all data models, domain and source data models, should be described with help of one domain ontology. The main elements of the compound geocoding architecture (see Figure 2.9) are:

- *Inputdata Processor*. This component is responsible for performing the preprocessing of text from the input data. The steps in this phase of geocoding are common techniques among geocoders: cleaning, parsing and standardising.
- *Core*. This component implements the geocoding logic of the system. It is responsible for the whole process of the source selection and the result data generation.
- *Resource Manager*. This component is responsible for the management of mediators.
- *Geocoding Layer*. It is a set of mediators which provide geocoding functionality over the selected geospatial Web services.

The *Resource manager* contains information on available mediators (the *SC Repository*) which have been registred previously (the *Reg/Load Mng*). The source selection task (the *Service Selector*) is

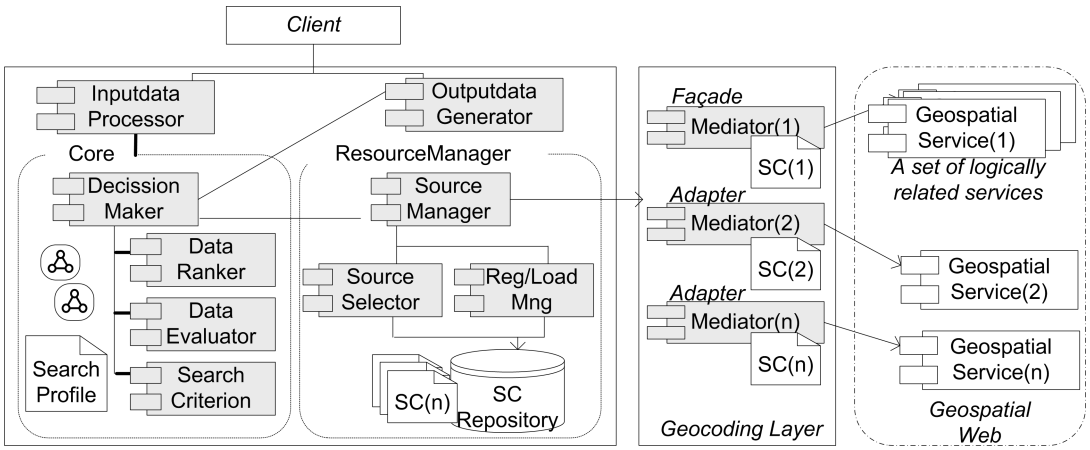

Figure 2.9: Overview of the Compound Geocoding Service Architecture.

based on rules defined in the *Decision Maker*. These rules apply several search criteria (the *Search Criterion*) to reason on the source annotations (the *service characteristics, SCs*) registered in the system. Search criteria are also defined in the terms of the service characteristics and they might be provided by (1) the application context (the *search profile*) and / or (2) the user requirements as part of the input data. An example of search criterion is:

```
hasCoverage(Zaragoza-Province) and hasPrecision > 0.9 and hasReliability > 0.5
```

The strategy of services selection can relax rules of the search constraints, e.g. by decreasing the value of precision and / or reliability, if there is not any service available which conforms to the requirements or the response data are not satisfactory. A generic connector is responsible for the communication with the selected geocoding mediator. It is loaded and configured (the *Reg/Load Mng*) as the result of the selection process.

Mediators ensure the abstraction from the communication protocols, invocation styles or interfaces used by the selected geospatial services. They are responsible for data harmonisation which consists of data models mapping and, if it is required, coordinate transformation. In practice, each mediator is a geocding service, and may require the implementation of some additional techniques related to the matching and ranking used in typical geocoding process in order to retrieve relevant results.

This architecture allows different types of named places to be geocoded, and, permits complementing data from one source with data from others in order to improve the reliability and the precision of the system. It also gives the user more freedom in deciding the search strategy. For example, users can decide in runtime which service should be accessible, or which search strategy should be applied (the best response of the entire system, the best answer for each source or the best answers from a chosen source, etc). In addition, as the mediator layer exposes one, common interface,

adding new resources (i.e. mediators) does not demand any changes in the system implementation.

2.6 Address geocoding for urban management

In Spain, there are some proposals of geocoding services supported by public authorities at state level, e.g. the Cadastre Service of Spain[11] (i.e. a set of Web services), or CartoCiudad services[12]. The first one is characterised by the best reliability among the other existing geocoding services at state level in Spain, but, due to the fact that its content type is parcel, the precision for address geocoding is decreased. The CartoCiudad combines the spatial contents provided by diverse public institutions (i.e. General Direction of Cadastre, Postal Office, National Institute of Statistic, and General Direction of National Institute of Geography) and from local authorities. The main disadvantage of the CartoCiudad services is the lack of the update procedure and gaps in coverage. Additionally, both these proposals share the problem of the uncomfortable search as it is necessary to indicate the search area (province and municipality) and requires the definition of workflow for address geocoding. It is more common to find Web Services offered by local authorities at their portals, e.g. the Street Data Web Service of Zaragoza city council (IDEZar SG). These services used to be characterised by a high data precision. However, their granularity may vary depending on the area (i.e. urban centre, village) and, usually, there are lacks in coverage in areas such as motorways or new urban zones.

Address geocoding is the most important functionality offered by any urban GIS, which is able "*to locate addresses, in any form employed by the population, in a quick and efficient way*" (Davis et al., 2003). An example may be Zaragoza, the fifth largest metropolitan area of Spain. The urban management systems of the local administration of Zaragoza municipality require address geocoding functionality in numerous applications and tasks from diverse areas, such as management of local incidents (e.g. traffic cuts, water or electric supply shortage), event management (e.g. demonstration, match or concert), street map of local administration portal (council of Zaragoza) or Web service to support the point of interest guide for mobile devices. All those applications have to deal with the problems mentioned above. The proposed architecture aims to overcome them. The implementation of a generic compound geocoding service provides instances which might be used in diverse applications of different requirements. These instances adapt to each environment according to the constraints provided by the search profile. This approach reduces significantly the development costs and improves reliability and precision of response.

The first step of the development of compound service for address geocoding in Zaragoza municipal area is the identification of the address domain model. This task requires the extension of the Administration Unit Ontology of Spain, AUSpain (López-Pellicer et al., 2008) with entities from the Aragon Autonomous Community (e.g. *municipality of Zaragoza*, *province of Teruel*, *comarca of*

[11] http://www.sedecatastro.gob.es
[12] `http://www.cartociudad.es/visor/`

Aranda). Then, several Geospatial Web services are selected to populate the mediator layer, and a set of geocoding mediators is created. Each mediator is evaluated in order to provide the geocoding service characteristics used for service selection. Finally, a geocoding logic is defined.

2.6.1 Data model

The domain ontology of the AUSpain ontology is the application of the AUO, therefore the introduced individuals, which represent the units of territorial division, might be used as the coverage description of service. Additionally, as addresses are bound to political organisation of the territory, the concepts of the AUSpain ontology were applied for data integration via mapping process. According to the general approach (Walker, 2008), the data model of the street address in Spain contains at least *province*, *locality*, *zip code*, *street address* (street name, portal number and other elements as floor or letter of door). Most of the Web services that might be used as street address source do not provide zip code. Thus, the main address data model has been extended (*autonomous community* and *comarca*) and modified to disambiguate the search results (*municipality* and *district*).

As this implementation is dedicated to the municipality of Zaragoza, it is possible to initialise some fields of the domain model, and the default values are used to restrict the results from sources during the data integration process. Figure 2.10 presents an overview of the evolution of the domain data model.

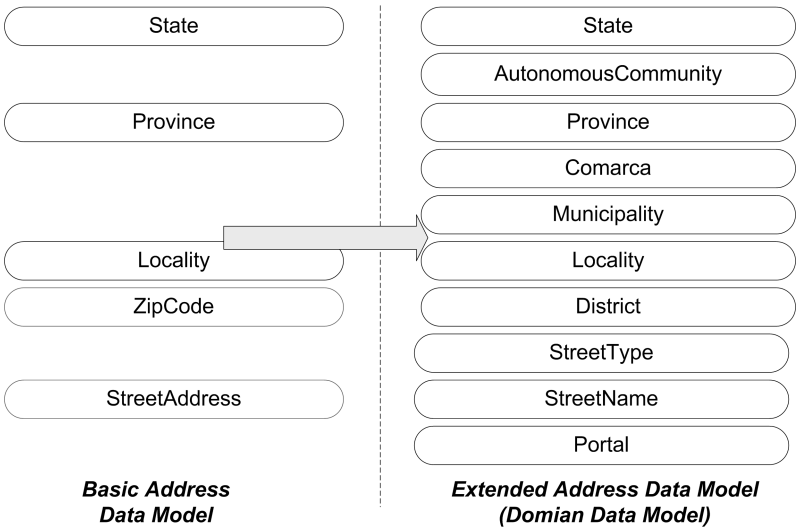

Figure 2.10: Domain data model evolution.

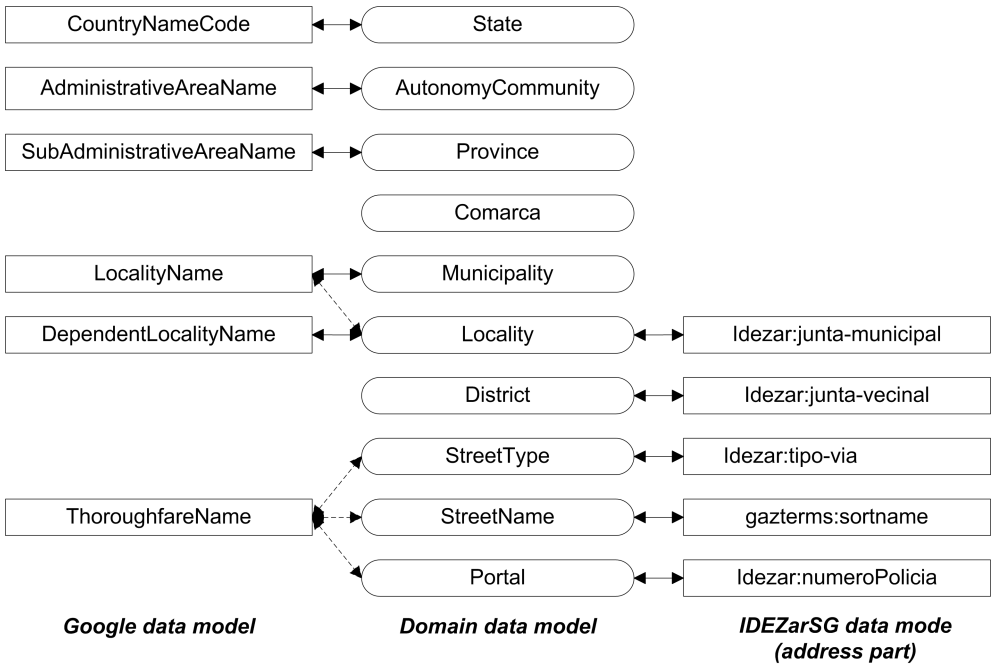

Figure 2.11: Mapping domain data model to data models of Google and IDEZarSG services.

2.6.2 Geospatial resources

The compound service implemented uses the following services: a set of the Cadastre services of Spain[13], a set of CartoCiudad services, Google geocoding service[14], and IDEZar SG service. A set of geocoding services (i.e. mediators) has been created by applying the method to populate the façade layer. For example, in the case of the Google geocoding service (a typical developer–oriented Web Service) the online documentation[15] is descriptive and provides API and examples of use. Also the ToS[16] has to be analysed. The `WSDL` document and service metadata are the results of the documentation analysis. A simple tool has been developed that automatises the creation of the service metadata file and the *Capabilities* document of the future mediator. This tool performs a simple mapping from the corresponding fields of the `WSDL` and the other metadata files used in the method. However, it should be pointed out that the resulting files need human revision to add missing elements.

Each mediator service exposes a common geocoding interface and uses the domain data model to return geocoded features. Figure 2.11 and Figure 2.12 present the mappings between the domain

[13]The services' endpoint were `http://ovc.catastro.meh.es` and `http://ovc2.catastro.meh.es` during the period of the system development. Currently, new versions of these services are available at `http://www.sedecatastro.gob.es` and `http://www1.sedecatastro.gob.es`, respectivelly.

[14]The Geocoding API V2 (`http://maps.googleapis.com/maps/geo?`), used in the developed system, is obsolete and replaced by version 3 (`http://maps.googleapis.com/maps/api/geocode/output?`).

[15]http://code.google.com/intl/es-ES/apis/maps/documentation/services.html

[16]https://developers.google.com/maps/terms

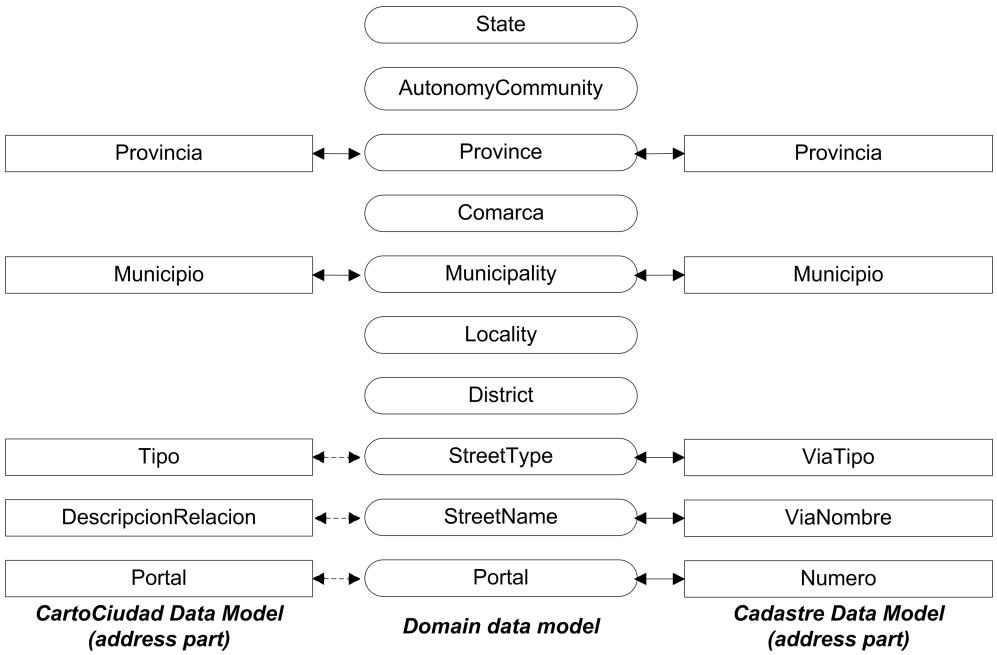

Figure 2.12: Mapping domain data model to data models of CartoCiudad and Cadastre services.

model and the source models. The data model of the CartoCiudad and the Cadastre services are simplified and show only the address part obtained from requesting various services as defined in workflows.

The service features considered by this implementation are coverage, precision, reliability and result accuracy. The rest of the features described in this paper are omitted in the selection strategy due to the fact that all used sources provide points as spatial objects and the searched information is of one type (street address). Determination of values of the basic service features, such as coverage and the spatial object, requires the analysis of the service online documentation. The reliability and the precision metrics are obtained from a set of evaluation tests performed against reference datasets. The choice of the reference datasets is of great importance. The Cadastre Service of Spain, being the official land and property registry, has been taken as the reference dataset of the reliability tests. The IDEZar SG service is characterised by high quality and its coverage corresponds with the target application coverage. Therefore this service has been chosen for the precision reference dataset. The values of reliability and precision are obtained using statistical methods: the reliability is calculated on the basis of the average hit error and precision on the basis of the mean square error. The result accuracy is obtained from the data models mapping. Table 2.1 shows values used in this implementation. Listing 2.1 shows an example of a service characteristics file which describes a created mediator service that uses a set of the Cadastre Services of Spain. The process of evaluation of this service is designated for a system dedicated to address geocoding.

Listing 2.1: Example of RDF description of a mediator service which uses a set of the Cadastre Services of Spain.

```
@prefix st: <http://idee.unizar.es/SW/services/geoservicesotopology#>.
@prefix sc: <http://idee.unizar.es/SW/services/geoservice/characteristics#>.
@prefix gsa: <http://idee.unizar.es/SW/services/geoservice/addressextent#>.
@prefix scat: <http://idee.unizar.es/SW/services/geoservice/servicecat#>.
@prefix gsc: <http://idee.unizar.es/SW/services/geoservice/category#>.
@prefix ag: <http://idee.unizar.es/SW/Onts/AU/AUSpain.owl#>.
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>.

scat:catastrospain2008 rdf:type gsc:AddressGeocodingService;
        rdfs:label "Servicio de Catastro"@es;
        sc:reliability "1";
        sc:precision "0.8";
        sc:resultAccuracy gsa:portal;
        sc:spatialObject st:point;
        sc:coverage ag:ESP_STATE.
```

Parameter/Web service	*Coverage*	*Reliability*	*Precision*	*ResultAccuracy*
IDEZarSG Service	municipality of Zaragoza	0.98	1 (baseline dataset)	Portal
GoogleMaps Service Service	World	0.96	0.99	Portal
Cadastral Service of Spain	Spain (baseline source)	1	0.85	Portal
CartoCiudad services	Spain	0.90	0.98	Portal

Table 2.1: Geocoding parameters of used Web services.

2.6.3 Geocoding framework

The developed service has been applied as a generic component in the urban management systems of Zaragoza city council. Figure 2.13 presents an overview of the applied domain data model in this framework. The instances of this service are dedicated to diverse applications whose requirements are stated by the profile descriptors.

Figure 2.14 shows an overall view of the compound architecture of geocoding service and some examples of types of client applications that use the service. The unified access to georeferencing Web services (e.g. the Cadastre Service) is offered by connectors that implement the *AdvancedGeocoder* interface and hide the request workflow in the case of the Cadastre and the CartoCiudad services. This interface is implemented also by the *Geocode Wrapper* which accesses INE local database, the information about census area units in Spain. The data are stored locally due to the fact that it is not provided through any Web service and the data must be downloaded as CSV files from National

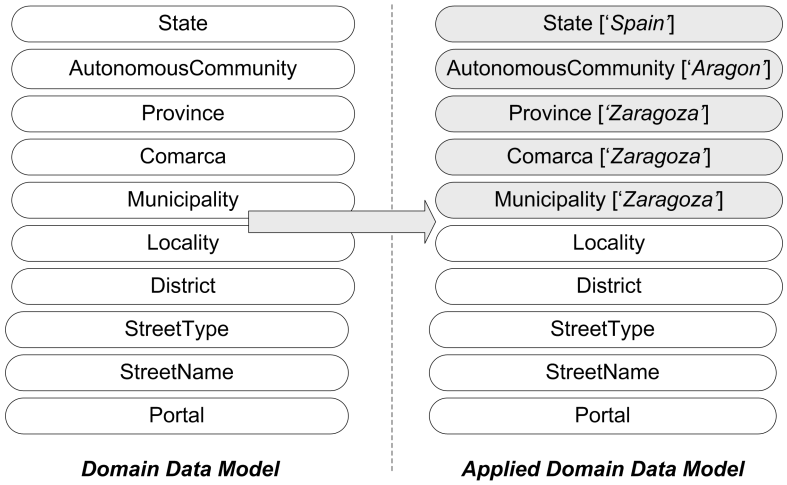

Figure 2.13: Application of the domain data model.

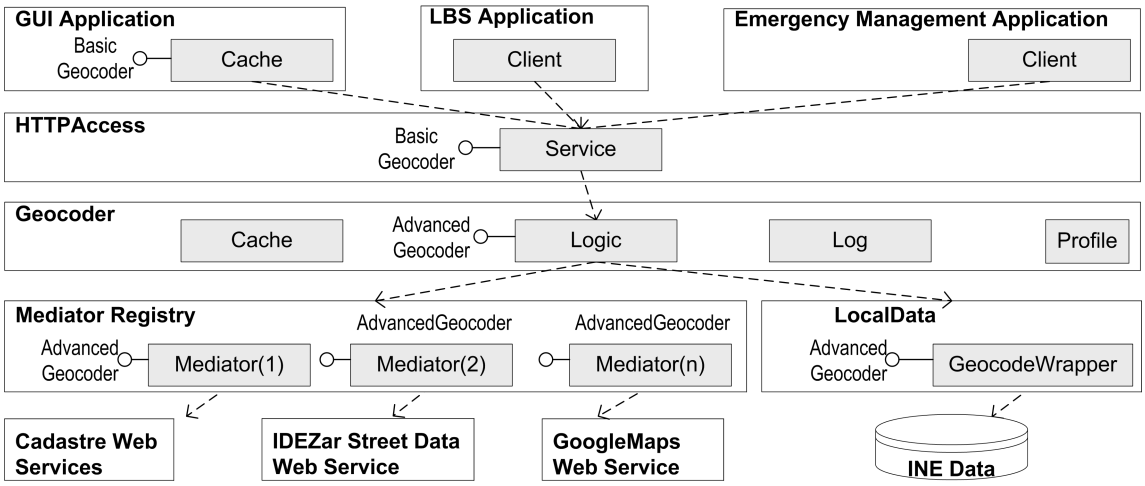

Figure 2.14: Layer view of the compound geocoding service and client components.

Statistics Institute Web page[17] (*Instituto Nacional de Estadística*, INE). As the INE information is official and complete it is therefore used as a reference point in case of result data ambiguity. The cache component improves the performance of the system and the logger (*Log*) permits tracking of its behaviour and offers feedback information on user request profile which is a valuable hint in the cache strategy. The simplified *BasicGeocoder* interface is offered via `HTTP` protocol and it is the request point for the majority of client applications.

[17] http://www.ine.es/

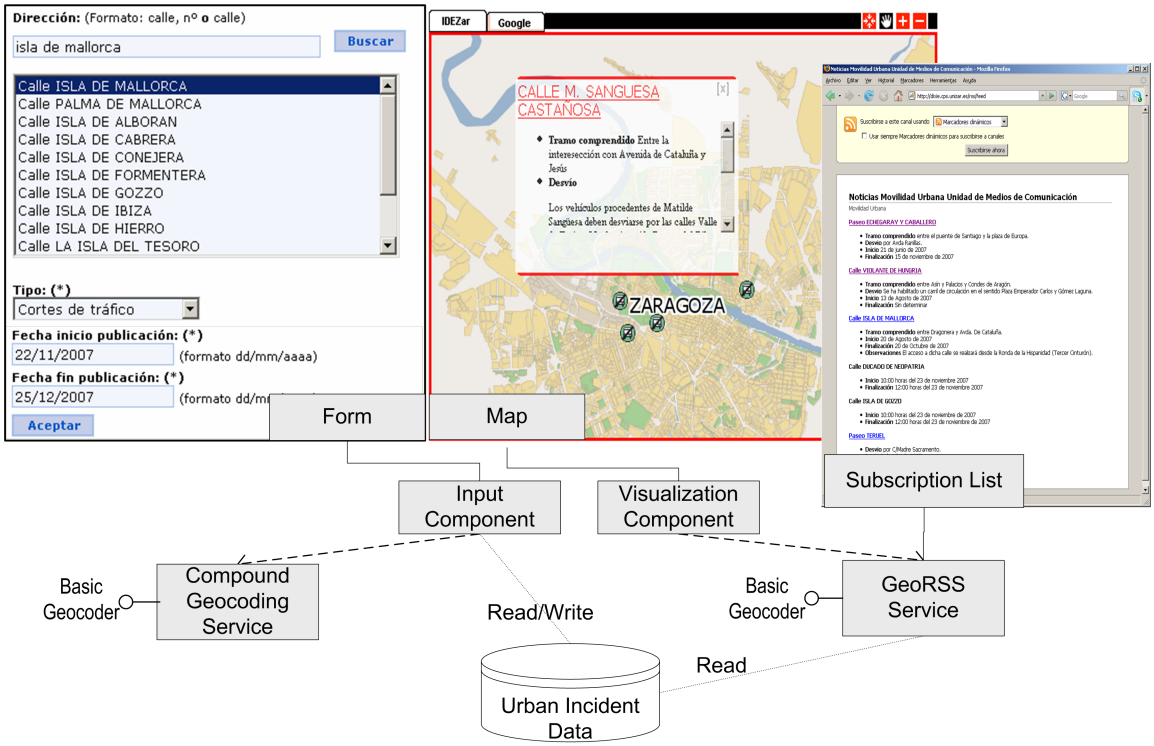

Figure 2.15: Introduction and publication of urban incidents.

2.6.4 Application

An example of these applications is the system to support management of local incidents in Zaragoza locality used by the city council. This system is used by the council workers to introduce the local incident information, which is provided to citizens via an official online portal. Figure 2.15 shows the general overview of the application architecture along with the screenshots of its Graphical User Interface (GUI) components. The form of the *Input Component* allows the user to introduce a new incident and the results of address geocoding are presented as an ordered list ordered via recommendation based on street name and type matching. The selected, normalised street address, its spatial coordinates (a point) and information about the incident (time interval, incident type, etc.) are saved in the local incident database, that is then used by the GeoRSS service (Reed, 2006), which publish new information on inscription lists. This information might be also visualised on the local incident map of the official Web page of Zaragoza city council.

2.7 Summary

This Chapter has proposed an approach to enhanced search for a geospatial entity from the perspective of traditional geocoding. The compound geocoding proposed ensures the improvement of geocoding results (e.g. reliability, spatial data quality) thanks to the use of different geographic information suppliers. The use of multiple sources involves the development of ontology for describing the geocoding sources. Moreover, the use of ontologies yields an advanced architecture in terms of extensibility, flexibility and adaptability.

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

The principal disadvantage of this approach is the need of implementation of a mediator for each source. Although the implementation permits adaptation to changes (e.g via pluggable connectors), the effort of mediator creation is significant.

Chapter 3

Content–based generation of the semantic characterisation of geospatial resources

3.1 Introduction

This Chapter presents methods for semantic characterisation of geospatial resources that go beyond the specifications from geospatial community. It shows applications where the search result can be improved by more precise resource descriptions. The descriptions are generated automatically with consideration of the content as a geospatial resource. As a variety of possible approaches might be regarded within the geospatial community, two examples of such applications are investigated. The first (Section 3.2) focuses on providing more precise spatial objects when interacting with end users. Here, semantic technologies are applied to improve the precision of the spatial search within a services catalog. The second application (Section 3.3) is dedicated to imagery layer identification in WMS services. The method developed has been implemented as a WPS service. The prototypes developed are applied in applications where usability is affected (Section 3.4). Finally, some conclusions are presented in Section 3.5.

3.2 Geoidentifier

One of the main advantages of applying linking–based approaches for referencing geographic features is the maintenance of the abstraction from their spatial representations. Usually, geographic features are characterised by the blurriness of their footprints. Topological elements of the physical world usually lack well defined conceptual boundaries and, consequently, this influences their spatial

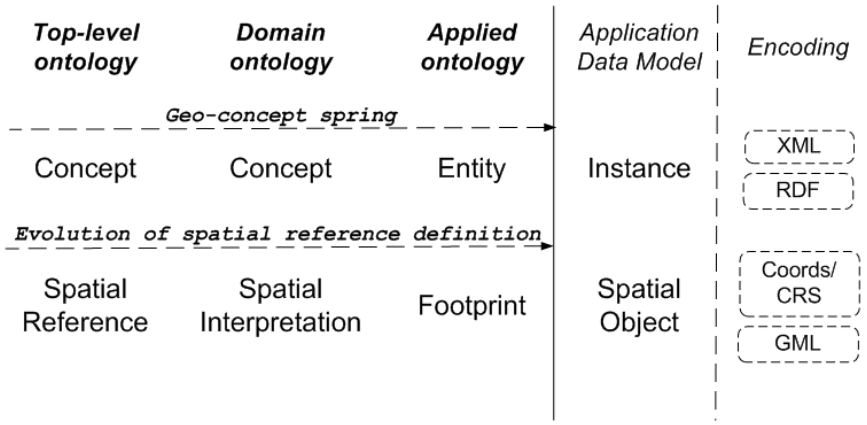

Figure 3.1: Modelling the spatial representation of a geo–concept.

Spatial Reference	A geo–concept from top–level ontology represents a general understanding of a concept. The definition of its spatial realisation is quite difficult and comes as a fuzzy definition (e.g. a spatial interpretation of a river)
Spatial Interpretation	Definition of spatial realisation of a geo–concept linked to a concrete domain.
Footprint	The definition of a perfect footprint of a geo–concept entity (linked to a concrete domain). Its computational representation might not exist.
Spatial Object	A computational representation of the Footprint. It is not a unique identifier of a geo–concept entity. It permits the visualisation of it and reasoning on spatial relations.

Table 3.1: Modelling the spatial object.

definition (e.g. rivers, chains of mountains). What is more, the computational representation of a geographic feature footprint is limited due to the resolution limits. Any geographic feature may be represented via different spatial objects which depend on the system application, the data model, and the applied technological solution. For example, a point is the best choice as a location reference while a polygon is for landscape visualisation. Therefore the spatial object should be interpreted as one of the possible representations of an entity. Figure 3.1 presents the global view of spatial reference definition, and Table 3.1 provides an overview of the process of conceptualisation.

The objective of this work is to present how an SDI might take advantage of best practices from the Semantic Web. An *administrative geography* (AG), i.e. an ontology of political organisation of the territory, published as Linked Data will be introduced as a relevant element of an SDI, since it permits (1) reasoning on logic relations among administrative units, and (2) accessing different representations (including footprints) of a geographic feature. Such ontology can improve functionality of a services catalogue deployed in an SDI by applying geographic reasoning. One of the principal characteristics of an OWS service is its geographic extent provided by the publisher as part of its

descriptive metadata (i.e. `getCapabilities` response). According to the INSPIRE Metadata Implementing Rules (INSPIRE DTM and EC/JRC, 2010), the geographic location is defined as *minimum bounding box* (MBBOX). The descriptive metadata are used by the discovery service to answer the user requests, and a spatial restriction is one of the requestable parameters. Since services from SDI are provided by public administrations, frequently, their geographic extent corresponds with the administrative area of the provider. The usage of MBBOX introduces false positive in the catalog response because the administrative areas are not rectangular. A catalog provided with knowledge about the hierarchy of administrative units and accurate spatial objects of each administrative area might increase considerably the precision and recall of the requests, which is especially important for applications based on on–the–fly data integration.

The rest of the Chapter is organised as follows. First, the state of art in linking geographic features in the Semantic Web community (starting from Linked Data approaches to deal with geographic information, and ending with examples of semantic geographic platforms existing on the current Web) and the geospatial community. Next, the infrastructure for the improvement of services catalog is presented.

3.2.1 Linking geographic features

Geographic features are managed differently in the Semantic Web and in the Geospatial Web. The first community treats geographic features as the additional contextual information, which might link to some other concepts. For the Geospatial Web, the concept of geographic feature is the core element of a geographic platform. However, the Geospatial Web has focused on the interoperability issues maintaining the boundaries among the concepts from different geographic sources.

The Semantic Web

As the contents of SDIs are concealed for common Web users, the Web community has created proper solutions applying successful linking approaches, such as geographic Web platforms (e.g. semantic geocoding service of GeoNames (Wick, 2012)) and for publishing geo–data (e.g. LinkedGeoData (Auer et al., 2009b)). Also, the Linked Data has been applied successfully in spatial solutions (e.g. DBpedia Mobile (Becker and Bizer, 2008, 2009)). Both approaches, standardised OGC services and Linked Data, might complement each other. For example, a created ontology of geographic features might link instances of the same geographic feature across different sources. This framework would provide an integrated view of geographic features rich in logic and spatial information. The richness of geographic feature description, direct or provided by linked sources, might be helpful in dealing with conflation for database integration, the well known problem from the database field (Dolbear and Hart, 2007). Additionally, such unified ontology might be used for defining and publishing complementary logic relations among geographic features (e. g., a representation of different territorial organisation) instead of creating new instances of OGC services. As

for the Semantic Web community, the publication of geographic ontologies by official providers (i.e. a public administration organ) according to Linked Data principles might be a valuable source of references, for example the Administrative Geography of Great Britain (Goodwin et al., 2008).

Nowadays, it is possible to generate Linked Data from a variety of data sources, and support a wide variety of data representation and serialisation formats. There are two main approaches for generation of RDF linked data from existing Web content and deploy them on the Web. The first one maps the data gathered in relational data bases to one or more ontologies / schemas producing RDF Linked Data, and the other approach uses RDFizers, i.e. tools capable of extracting data from one or more sources, which then are mapped. D2R Server (Bizer and Seaborne, 2004) is one of the first tools for publishing the content of database. OpenLink Virtuoso (Erling and Mikhailov, 2009) is more complete technological solution and it supports both approaches. The most popular way of accessing to Linked Data is via a SPARQL endpoint or a REST Web service (Alarcón and Wilde, 2010). Linked Data clients vary from simple RDF Browser to a graph visualisation. Also, there are techniques for requesting RDF data from different endpoints providing a transparent on–the–fly view to the end user (Langegger et al., 2008).

The most important initiative related to creating and publishing interlinked contents on the Web is the W3C Linking Open Data (LOD) project (W3C SWEO, 2012; Heath and Bizer, 2011) founded in January 2007. In May 2007, the number of LD datasets in the LOD cloud diagram was 12 and in September 2011 it rose to 295 (Cyganiak and Jentzsch, 2011). Since on–line contents involve geographic information, geographic features have become part of Linked Data datasets, such as DBPedia, where geographic information has been extracted from Wikipedia (Auer et al., 2008; Becker and Bizer, 2009). In May 2007 there were only six geographic datasets in the LOD cloud diagram, in July 2009 eight, and in September 2011 the number of geographic datasets rose to 31 providing more than 6 billion RDF triples (Cyganiak and Jentzsch, 2011). For example, the LinkedGeoData (Auer et al., 2009b; Stadler et al., 2011) and GeoNames (Wick, 2012) datasets aim at adding geo–semantic meaning to the Web. LinkedGeoData offers a Linked Geo Data Knowledge Base with RDF descriptions of more than 350 million spatial features from the OpenStreetMap database. GeoNames provides a set of REST Web services to access geographic features (about 8 million unique features). From the point of view of the LOD community, a core of interlinked data about geographical locations is in DBPedia since both geographic datasets are linked to it (Heath and Bizer, 2011). However, it is because of the historical reason (i. e. DBPedia is one of the first LD datasets) and the applied linking paradigm (i. e. unidirectional view).

An example of applying Linked Data principles in location based solution is the DBpedia Mobile (Becker and Bizer, 2009), a location–aware client for the Semantic Web for mobile devices. The current user location is used to extract corresponding datasets from the underneath DBpedia database which are interlinked with various other location–related datasets.

An interesting proposal is Triplify (Auer et al., 2009a) which supports circle–based spatial requests. This system directly uses the DB Views model as a base for creating the RDF documents and URLs of published datasets, which facilitate the development. The underlying data base is responsible for processing spatial and semantic queries which are encoded explicitly in the requested URL. The spatial query permits users to retrieve the geographic features located in a circular region defined via a point and radius added to the requested URL. This system enables limited spatial query notwithstanding the fact that the Semantic Web techniques can not take advantage of this facility.

The idea of linking geographic features to create the Web of Data influenced the development of geographic Web platforms, such as Yahoo! GeoPlanet (YDN, 2012) or GeoNames (Wick, 2012). Both of them belong to a new branch of geocoders, the *semantic geocoders*, which return URIs to identify uniquely the named places instead of standardised textual description and location reference (e.g. a point). Although, they use the idea of linking, they are not following the pure Linked Data approach. GeoPlanet uses URIs to identify the named place which permits users to retrieve its semantic description, however, the platform uses a simple XML file instead of RDF. Additionally, the important spatial relations (e.g. *child*, *neighbour*, *siblings*) are encapsulated into the URI definitions. GeoNames is *almost* Linked Data based. It also uses unique identifiers of concepts to identify the named places. The RDF description of features contains spatial relations defined in the published OWL reference ontology[1]. However, GeoNames distinguishes the *concept* from the *descriptive document*. The feature (i.e. concept) is identified via a URI but the geonames server uses 303 redirection to display its location on a map. The RDF description is available by adding the */about.rdf* at the end of the feature URI[2].

The spatial requests supported by the presented solutions of the current Semantic Geo Web are based on a branch of predefined spatio–logic relations (e.g. *near–by*, *belongs–to*, *child*, *siblings*). Since it is impossible to express all spatial relations among geographic features via definition of logic relations, the Semantic Web needs to use a spatial representation of features. Currently, the spatial objects usually used in the Semantic Web are limited to points or MBBOX eventually. The complex spatial requests that mix spatial objects and spatial relations defined in a rich ontology still remain the open issue.

The Geospatial Web

The requirement of unique identifiers for geographic features, *geoidentifiers*, in the geospatial community has been presented from its beginning. Any Geospatial Web framework which publishes information about geographic features uses unique identifiers (within this framework at least), and it might be seen as a source of geoidentifiers. Therefore, any gazetteer (Hill, 2006) or OGC Web Feature Service from an SDI might be such a source of geoidentifiers. Currently, there are several

[1] http://www.geonames.org/ontology/ontology_v3.01.rdf

[2] For example, the town '*Embrun*' in France has associated these two URIs: http://sws.geonames.org/3020251/ and http://sws.geonames.org/3020251/about.rdf

instances of WFS services in the SDI of Spain, which frequently contain instances of the same geographic feature. The services model the geographic feature in different manners and use different identifiers that usually are derived from keys of relational databases; therefore, there are problems of entity identification among different contents, and as a consequence, common problems of data integration.

One of the earlier proposals from the OGC community to apply common geographic identifiers for linking purposes is a framework based on Geolinked Data Access Service (Schut, 2004a) and Geolinking Service (GLS) (Schut, 2004b). This approach is dedicated to publishing geographically linked information (e.g. statistical data) separately from its spatial representation (spatial objects). Since this proposal is based on dataset merging from different sources by using a linkage field defined in the sources, it fixes geographic data to only one source of spatial representations and the geoidentifiers are used only as syntactic links. In practice, it can be seen as a technological facilitation for data publishers. The proposal evolved into the OGC Geographic Linkage Service (GSL) (Schut, 2009) (the result of the Geolinking Interoperability Experiment[3]), and finally into the OGC Georeferenced Table Joining Service (TJS) standard (Schut, 2010).

A proposal for providing an integrated view across distributed services is the EuroGeoNames project implemented as an INSPIRE compliant service (Jakobsson and Zaccheddu, 2009). Apart from defining the data model to be followed by all community members, it provides some rules for identifier definition. The named place identifier has to be composed of (1) its name, (2) the two–letter ISO 3166 code (ISO, 2007a) and (3) a code generated according to the Base36 encoding system (Oscarsson, 2001) (e.g. "2YC67000B"). However, these identifiers still remain unique only in this distributed gazetteer.

Interlinking the corresponding instances of the same geo–concept entity across different providers might be interesting for the Geospatial Web. It might be the way of adding logic relations among geo–concept entities and avoid the necessity of providing a new separate platform. For additionally, it can be useful for the improvement of search in an SDI.

3.2.2 Spatial search

Using linking principles for search in the geospatial domain has some advantages. One of the advantages is the accessibility to multiple representations or models of a geo–concept entity. It includes access to other spatial objects.

Traditionally, when searching with explicit spatial restrictions (i.e. using spatial objects), MBBOXES are used. Some applications support circle–based spatial queries (Auer et al., 2009b; YDN, 2012). Nevertheless, the circle defined might be seen as an approximation of a MBBOX. During the retrieval process, spatial indexes created on georeferenced elements are exploited. The most popular indexes used in spatial data bases apply boxes due to performance reasons (Manolopoulos et al.,

[3]http://www.opengeospatial.org/projects/initiatives/geolinkie

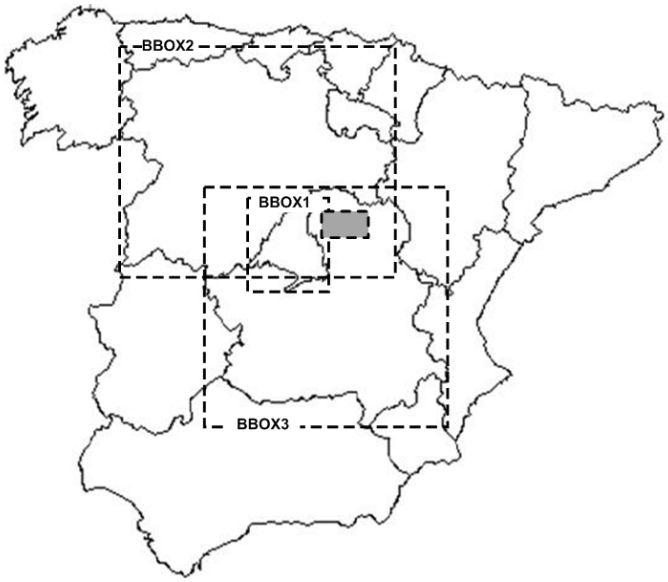

Figure 3.2: Overlapping MBBOXES of administrative areas (a case from Spain).

2006).

Indexing and searching geospatial resources by means of MBBOXES usually introduce false positives. It happens because the footprints of geographic features are not rectangular habitually. For example, the administrative boundaries of Spain are far away from being rectangular and their MBBOXES overlaps significantly. Figure 3.2 shows the issue of overlapping MBBOXES of administrative areas. The shadowed rectangle represents the MBBOX from a spatial search request. In this scenario the search application would return resources whose geographic extent corresponds with `$BBOX1`, `$BBOX2` or `$BBOX3` when, in reality, it should provide only those resources whose MBBOX corresponds with `$BBOX3`.

The precision of spatial search might be improved by operating on more precise spatial objects. A more promising approach could be application of the identifiers from an *administrative geography* which not only provides footprints but also permits reasoning on logic relations among concepts.

3.2.3 Administrative geography of Spain

The existing domain ontologies for modelling political organisations of a territory are usually based on the *part–of* relation (parental relation). Such model is not flexible enough to scale the complexity of the territorial organisation of countries, which apart from main division units (e.g. municipality, province and autonomous community in Spain) has to involve the units of different status (e.g. autonomous cities of Ceuta and Melilla, or associations of administrative units in Spain). Therefore, it is required usually a dedicated administrative ontology is usually required (e.g. Ordnance Survey)

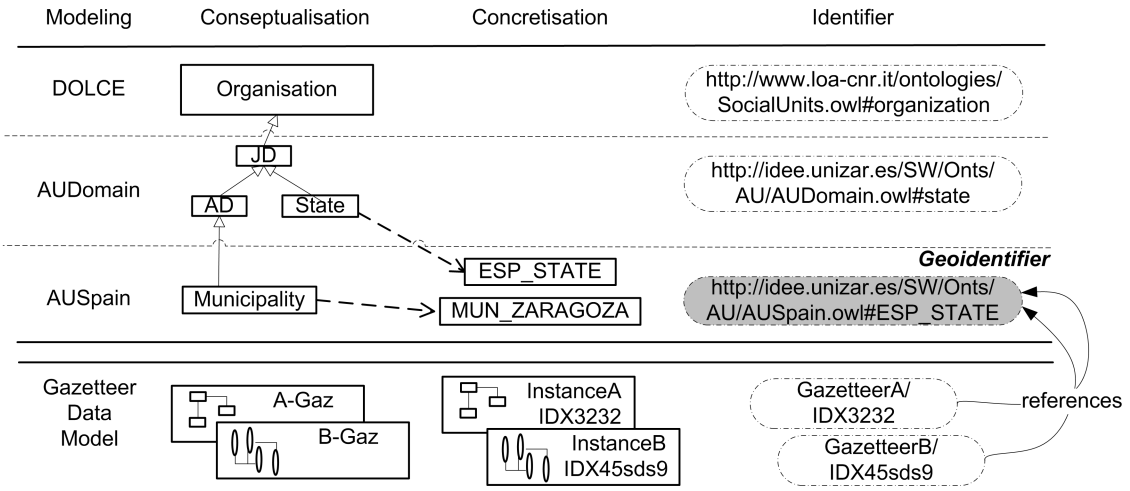

Figure 3.3: Geoidentifiers and modelling the spatial representation of a geo–concept.

to model the political organisation of territory of a country.

There is a variety of OGC WFS services which publish information about administrative unit entities within the Spanish SDI. These services are provided by central and local authorities. Some of them offer administrative boundaries with different resolution (e.g. the *Infraestructura de datos Espaciales de España–WFS* service, IDEE–WFS[4]) separately for each level of administrative division, and others, such as the gazetteers focused on gathering named places (e.g. *IDEE–WFS–Nomenclator–NGC*[5]), contain administrative units among published features. However, neither of these models permit the expression of the full administrative model that exists in Spain.

One of the contributions of this work is an early design of the *administrative geography* for Spain using Linked Data. The Administrative Unit Ontology has been proposed as the domain ontology (López-Pellicer et al., 2008). Apart from *part–of* and *has–part* relations, this domain ontology defines the *is–member–of* and *has–member* relations to distinguish the association of the administrative units whose spatial representation might overlay the boundaries of direct parental units. An example of such association might be the *comarca* of Aragon Autonomous Community which groups municipalities. Each municipality might lie in boundaries of only one province; however, one comarca might aggregate municipalities from different provinces.

The D2R Server has been used to publish the administrative geography of Spain, and to generate a data dump. The national Gazetteer, *IDEE–WFS–Nomenclator–NGC*, has been used as the reference source to extract the entities of the *administrative geography*. Although this model does consider logical relations among features (e.g. *parent–child*), the location of features in the administrative structure is defined indirectly by their names offered via *LocationEntity* element, whose

[4]http://www.idee.es/IDEE-WFS/ogcwebservice?
[5]http://www.idee.es/IDEE-WFS-Nomenclator-NGC/services?

Listing 3.1: Example of RDF description which represents the Zaragoza municipality.

```
@prefix au: <http://idee.unizar.es/sw/gsw/ont/2008/au-spain.owl#>.
@prefix agont: <http://idee.unizar.es/sw/gsw/ont/2009/agont.owl#>.
@prefix ag: <http://idee.unizar.es/sw/gsw/ont/2009/ag/>.
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>.
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>.
@prefix skos: <http://www.w3.org/2009/08/skos-reference/skos.rdf#>.

ag:ZaragozaMun rdf:type au:Municipality;
        rdfs:label "Zaragoza"@es,"Saragossa"@en;
        au:part-of ag:Spain, ag:AragonComunidad;
        au:member-of ag:ZragozaComarca;
        skos:relatedMatch <http://www.idee.es/IDEE-WFS/ogcwebservice?SERVICE=WFS
            &VERSION=1.1.0&REQUEST=GetFeature&MAXFEATURES=1&NAMESPACE=xmlns(
            ideewfs=http://www.idee.es/wfs)&TYPENAME=ideewfs:BDLL1000Municipio&
            FILTER=%3CFilter%20xmlns:ideewfs=%22http://www.idee.es/wfs%22%3E%3
            CPropertyIsEqualTo%3E%3CPropertyName%3Eideewfs:nombre%3C/
            PropertyName%3E%3CLiteral%3EZARAGOZA%3C/Literal%3E%3C/
            PropertyIsEqualTo%3E%3C/Filter%3E>;
(. . .)
        agont:bond-100 <http://www.idee.es/IDEE-WFS/ogcwebservice?SERVICE=WFS&
            VERSION=1.1.0&REQUEST=GetFeature&MAXFEATURES=1&NAMESPACE=xmlns(
            ideewfs=http://www.idee.es/wfs)&TYPENAME=ideewfs:BDLL1000Municipio&
            FILTER=%3CFilter%20xmlns:ideewfs=%22http://www.idee.es/wfs%22%3E%3
            CPropertyIsEqualTo%3E%3CPropertyName%3Eideewfs:nombre%3C/
            PropertyName%3E%3CLiteral%3EZARAGOZA%3C/Literal%3E%3C/
            PropertyIsEqualTo%3E%3C/Filter%3E>.
```

structure contains concepts from the territorial organisations of Spain (e.g. *autonomous community*, *province*, *municipality* or *island*). Since the published data are not complete (e.g. there is no assignation of comarca names to municipalities), the INE online catalog (*Instituto Nacional de Estadística*, the National Statistics Institute of Spain) has been used to complement the data. Then, the result data has been linked to their corresponding instances from different WFS services via the *skos:relatedMatch* relation. This approach has produced the *administrative geography of Spain* published as Linked Data, and one of its advantages might be the maintenance of the references to different instances across the SDI of Spain. Figure 3.3 presents the idea of the *administrative geography of Spain* as source of geoidetifiers, and Listing 3.1 shows the model used on an example of Zaragoza municipality.

An example of application of an *administrative geography of Spain* published as Linked Data for the improvement of search in an SDI can be an enhanced services catalog.

3.2.4 OGC services catalog application

The services catalog is one of the elements of an SDI. Since it is responsible for service discovery, its functionality determines the reusability of the offered services in the SDI and might be improved by applying best practices from the Semantic Web. This work proposes a framework, where the *administrative geography* published as Linked Data is one of the core elements. A services catalog dedicated to supporting the visualisation applications based on an on–the–fly data integration is presented as a use case. The principal advantage of this approach is reflected by the improvement of the functionality of the end application.

Coverage issue

In INSPIRE, the discovery service allows users to search for geospatial resources, i.e. datasets and services. The search requests can contain a restriction on the geographic extent of searched resources. The spatial restriction is provided in the form of a MBBOX. Frequently, the geographic extent of published data and service corresponds with the coverage of an individual from a geographic ontology (e.g. *Europe* from a geographic region ontology, *European Union* from a political organisation ontology). Therefore, using MBBOX to describe available resources usually introduces false positives in the collection of results.

In the SDI of Spain the coverage of published service frequently corresponds with an administrative unit area of provider (e. g., council of Zaragoza). This characteristic can be exploited to extend the service description which is maintained by a catalog and to add a geoidentifier of the corresponding administrative area if identified. In this way, a precise spatial object can be used instead of the MBBOX.

Architecture and implementation

The usage of geographic feature identifiers requires an annotation of the registered resources in a catalog with corresponding geoidentifiers. Creation of metadata of registered services is one of the characteristics of the services catalog deployed in the SDI of Spain (Nogueras-Iso et al., 2009). Its architecture has been extended with the *Knowledge Content (KC)* that is responsible for the service search process (see Nogueras-Iso et al. (2009) for the description of the services catalog architecture). Figure 3.4 presents the main elements of the *KC* component. The *KC* uses two `RDF` dataset sources: the *Administrative Geography (AG)*, i.e. the *administrative geography of Spain*, and the *Service Description Register (SDR)* which contains `RDF` serialisation of the registered service description. The reference ontologies, i.e. *Administrative Geography Ontology (agont)* and the *Service Description Ontology (svont)*, are applied during the reasoning process. The concepts from the *administrative geography* (1) are linked to the entities from the reference WFS service, the source of boundary spatial definitions, and (2) the `URIs` of the administrative units are used in the service description

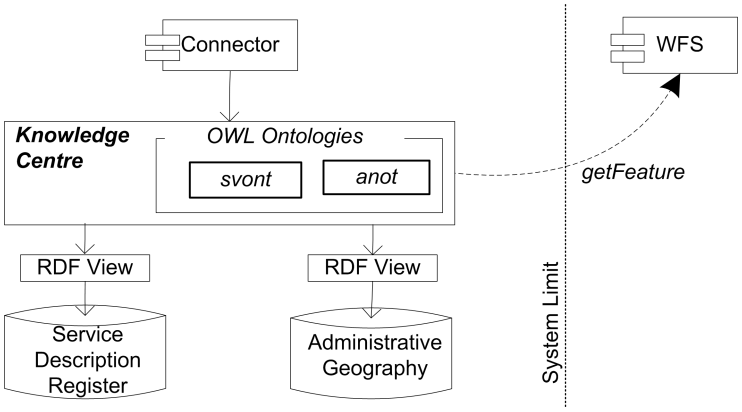

Figure 3.4: Administrative geography as support for services catalog.

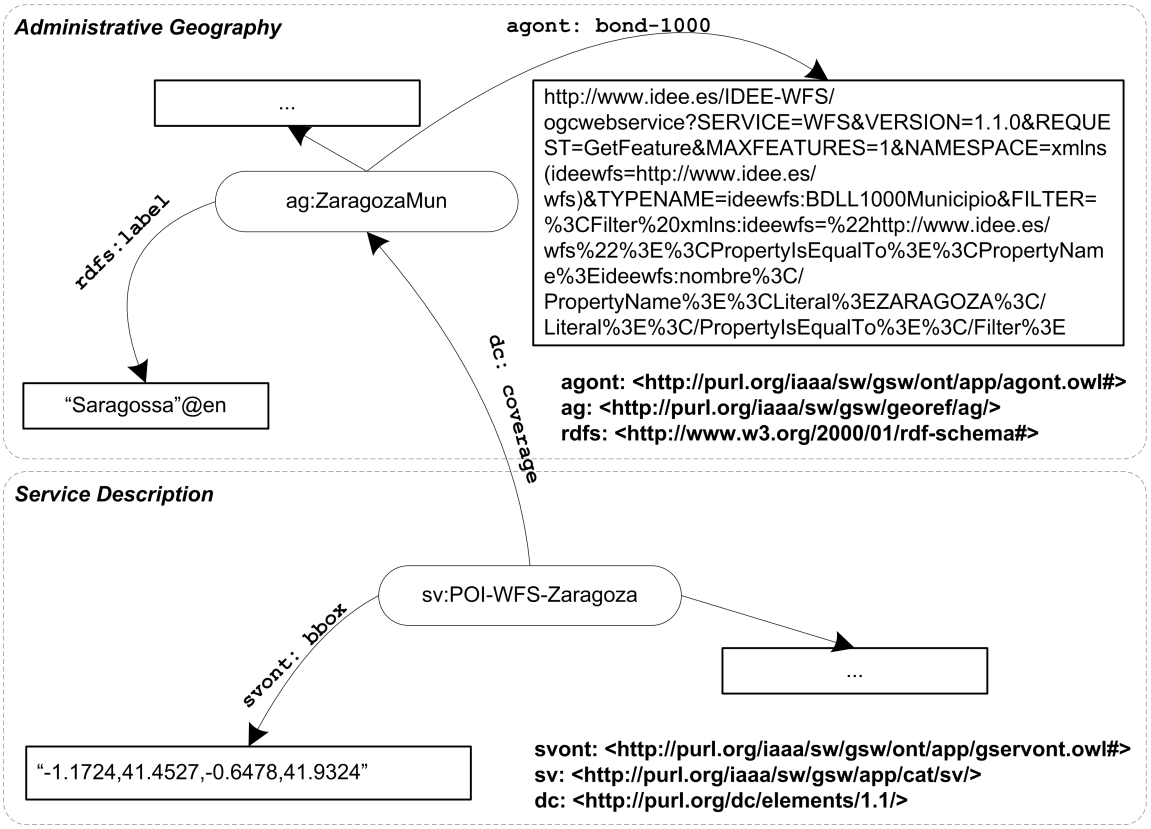

Figure 3.5: Relations between the service description and the *administrative geography*.

as an indication of the geographic extent (*dc:coverage* property). Figure 3.5 represents an example of the link between the *administrative geography* and the service description.

During the registration of a new service, the catalog uses the `getCapabilities` response to

create a proper description of the service. One of the elements is the service geographic extent which is expressed via a MBBOX. The metadata model of the description has been extended with the *geoidentifier* metadata in order to contain a URI from the *administrative geography* of Spain. This element is not an ISO19119 element; therefore, it is neither visible to users nor published by the OGC CSW. The geoidentifier value is obtained from the analysis of the service MBBOX offered by the provider. This MBBOX is used to request the *KC* to identify the *prime administrative unit* (i.e. the most extensive administrative unit that fits within the MBBOX). Validation of the result consists in checking if the service provides any data from the disjoint area of the MBBOX and the prime administrative unit coverage (i.e. a set of retrieval tests). If it is impossible to identify the prime administrative unit (e.g. in the case of the hydrography service of the Ebro river basin, some data lies in France as well) or the validation fails, the URI of the *Non* concept (i.e. the disjoint concept with *administrativeUnit*) is returned. The service registration ends with deployment of the registered service description in the RDF container.

The searching process for services with an MBBOX restriction exploits the semantic links between the semantic description of services and the features from the AG. A simplified example of a request (i.e. only the spatial part) which is consumed by *KC* is shown in Listing 3.2. The request pattern uses as the input the searched MBBOX ($BBOX) expressed as literal (e.g. "1.16311, 41.0937, 1.7132, 41.6686"). The first part of the request looks for those services whose coverages (*dc:coverage*) point to the *administrativeUnits* whose boundaries are in an interaction with the searched MBBOX (i.e. within it or intersects it). The boundary of an *administrativeUnit* is defined via the *bond–1000* property containing a request which retrieves the corresponding feature from a WFS service. The feature is retrieved and its spatial object is extracted automatically by applying the profile instructions. The second part of the request looks for those services whose geoidentifier is defined as *Non*. They are filtered in a similar way as in the previous version of services catalog, i.e. by comparing the service MBBOX and the requested one.

The *KC* component has required the implementation of the spatial functions, such as *intersect*, or *within*, known form spatial databases. Therefore, the Apache Jena framework (Apache Software Foundation, 2012) has been chosen to deploy RDF datasets because the Jena ARQ (i.e. a proprietary extension to SPARQL RDF Query language (Harris and Seaborne, 2012)) allows the implementation of such additional functionality.

3.3 Semantics of the WMS layer

Orthoimages are essential in many Web applications in order to facilitate the background context that helps to understand other georeferenced information. For instance, in the field of disaster management, satellite data play an increasingly important role in supporting decision making (Meisner et al., 2009; Kwan and Ransberger, 2010). Rapid data integration and visualisation are essential

to make data accessible and convey them in an easier–to–perceive way (Iosifescu-Enescu et al., 2010). Especially when presenting information to a non–expert audience, visualisation of the data improves the understanding of the situation at hand. Other applications of remote sensing products cover change analysis for monitoring and tracking the type and rate of landscape changes (Julea et al., 2010), urban environment modelling (Krauss et al., 2007), or assessing geospatial information quality (Skirvin et al., 2004).

With the constant improvement of technologies in high–resolution satellite remote sensors, GPS systems, databases and geoprocessing sources, there are nowadays increasing amounts of imagery and gridded data. Additionally, thanks to the development and increasing importance of SDI, the availability and accessibility of these data through standardised and interoperable Web services have increased exponentially in the last years. The OGC WMS and WCS service specifications, provide the means for the implementation of services which offer visualisation and download of imagery data in well–known formats such as HDF–EOS (Larry Klein, 2007), GeoTIFF (Ritter and Ruth, 2000), DTED (NGA, 1996), NITF (DoD, 2006) or GML. However, it can be observed that there are two main problems that become obstacles for the access to imagery data on the Web. On the one hand, SDI catalogues (the services provided by SDIs to locate data and services) deployed at national, regional or local levels do not necessarily register all the services providing access to imagery data on the Web. On the other hand, it is not easy to automatically identify whether the data offered by a Web service is directly imagery data or not. With respect to the first problem, the discovery of Web services not directly subscribed in SDI catalogues, some researchers have proposed different strategies based on crawling the Web (Li et al., 2010; López-Pellicer et al., 2011e). However, the second problem, the automatic categorisation of the content offered by services has received little

Listing 3.2: SPARQL request pattern for service selection via an MBBOX.

```
1   PREFIX   svont: <http://purl.org/iaaa/sw/gsw/ont/app/gservont.owl#>
2 PREFIX agont: <http://purl.org/iaaa/sw/gsw/ont/app/agont.owl#>
3 PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
4 PREFIX dc: <http://purl.org/dc/elements/1.1/>
5 PREFIX sf: <http://purl.org/iaaa/sw/gsw/app/gfun>
6 SELECT ? s WHERE  {     ?s rdf:type svont:service;
7         dc:coverage ?x.
8 ?x rdf:type agont:au;
9         agont:bond-1000 ?g.
10 FILTER(sf:INTERACT(?g, $BBOX))    }UNION
11 SELECT ?s
12 WHERE {
13 ?s rdf:type svont:service;
14         dc:coverage agont:Non;
15         svont:bbox ?g.
16 FILTER(sf:INTERACT(?g, $BBOX))
17 }
```

attention until now.

The purpose of this work is to investigate this second problem and propose a method for the automatic analysis of WMS services in order to detect if they contain imagery data, i. e. aerial photos or orthorectified satellite images of high resolution. WMS services are considered because their availability is higher on the Web. Currently, the number of WCS service instances is very low in comparison with WMS services (López-Pellicer et al., 2011a). Additionally, as WCS services provide full access to data, complex issues about access rights usually restrict their public deployment.

Various heuristic methods for the automatic analysis of WMS services have been investigated in this work. First, a description–based method has been developed. It exploits information offered by the service provider. Then, a content–based method has been investigated. The characteristics of the information that should be detected (i. e. orthoimages) derive from the requirements on radiometric and spatial resolution of sensors that produce it. Therefore, the effort focuses on deterministic methods that exploit image features generated by these kinds of sensors. Additionally, an effective algorithm for the gathering of spatial information from OGC WMS services is an important issue of this content–based method. In the context of this work, an effective retrieval algorithm for content–based analysis means an algorithm for collecting a representative set of fragments of an image (i. e. not empty image fragments), which enables layer analysis.

The procedure developed has been published as a geoprocessing service, which accepts the URL of a WMS *Capabilities* document and returns a list of those of its layers which offer orthoimages and the scale at which the images have been discovered. This geoprocessing service has been applied within the *Virtual Spain* project[6] to develop a catalogue of orthoimages and associated applications.

The rest of this section is structured as follows. First, the existing related work is outlined. Next the method proposed for the analysis of WMS services is presented. Then some experiments are performed over a crawled collection of WMS services and the efficiency of this method is discussed. Finally, the publication of the proposed method as a Web service compliant with the OGC WPS specification is presented.

3.3.1 Related work

The accessibility of geospatial resources influences any geospatial–based tasks related to geospatial processing. For example, on–demand data production (Mansourian et al., 2008) assumes discoverability of proper geospatial resources as an input. The discovery of resources within an SDI is based on the DL paradigm (Béjar et al., 2009). For instance, Li et al. (2011) present a framework for searching over multiple standardised catalogues, which can provide an integrated search across local and regional SDIs. However, not all Web resources valuable for these kinds of geospatial processes are part of an SDI, e. g. OGC Web services offered by a public administration but not published within any SDI, or resources produced by Web users. In these new scenarios, the automatic discovery

[6]España Virtual project (Virtual Spain project): http://www.españavirtual.org/

of geospatial resources on the Web has recently gained interest within the geospatial community (Li et al., 2010; López-Pellicer et al., 2011e). This new approach might provide users with a wider range of geospatial resources. Although standardised geospatial services are thought to be self–descriptive, some researchers indicate that there is a lack of good practices. For example, it is common to find inconsistencies between metadata in registries and the metadata offered directly by the service (Wu et al., 2010). As automatic discovery on the Web relies mainly on resource analysis, it might motivate providers to ensure more accurate descriptions. For example, in the case of OGC services, it might be supposed that if the discoverability of a resource depends on the quality of its *Capabilities* document, the providers would put more effort in the future to generate more valuable information in comparison with the current practices. In the context of this work, WMS instances crawled on the Web are used as input in the experiments to test the performance of the method proposed for filtering orthoimages. As these services have been found on the Web and are deployed by different providers, they represent a realistic scenario that combines heterogeneous metadata of different quality levels.

Although resources are usually described in conformance with a specific description model, this does not necessarily imply that all potentially interesting information to the user is provided. For example, the WMS *Capabilities* document does not contain a field to state the data presentation scale, i. e. the scale range for which a service renders a map. The capabilities specification just includes an optional field with the range of scales for which it is appropriate to generate a map of a layer. Although this information has different semantics (i. e. it refers rather to the data resolution), it might be assumed that the service renders a map at least at these scales. Additionally, as this field is optional, it may happen that some service providers have included this information in free text fields. In this case, a support tool (e. g. for a keyword–based or semantic–enabled search) is necessary, or an additional effort is required from a potential client (e. g. manual visualisation) to estimate the resource utility. Content–based analysis can be useful to extract additional information from a resource, which is not directly provided within its description. The automatic discovery of geospatial resources based on content analysis instead of relying on keywords in the metadata is gaining interest. For instance, Zhang et al. (2010) present a semantic–based application that is able to perform an intelligent content–based search of Web Feature Services (WFS). Although this approach is not directly applicable for non–textual resources such as imagery, image analysis techniques can be used. Image classification and annotation has been extensively investigated in the field of IR, and machine learning approaches have been commonly applied in this area (Baharudin et al., 2007; Gonzalez-Garcia et al., 2007; Sinha and Jain, 2008). These techniques have been also applied to remote sensing imagery for the classification of multispectral imagery over a semi–urban area (Alonso and Malpica, 2008), or to support the update of existing land use databases (Kressler et al., 2005). The heuristics of the method proposed include both techniques for processing the WMS *Capabilities* document, and techniques for analysing the content of the layers offered by a

WMS service instance.

Finally, it must be noted that any technique for the content–based analysis of a WMS service requires an effective algorithm for collecting a representative set of fragments of an image, instead of analysing the whole extent of an image. The geographic extent defined via an MBBOX is commonly accepted by the geospatial community (including WMS providers) for indexing and coarse resource selection. However, the type of the geographic feature (e. g. a forest, a named place), the spatial object used to represent it (e. g. multipolygons, icons) and the dispersion of spatial data within the dataset have influence on the efficiency of an application that searches for geographic information using a MBBOX. As presented in the previous Section, the semantic annotation of the geographic extent of OWS instances can improve the accuracy in the discovery of WMS services and their layers. Although the WMS services provided as input for the method proposed are only annotated by bounding boxes and it is not always possible to infer a more accurate description of their geographic extent, the algorithm for the collection of image fragments takes into account the variability in the dispersion of data by including a space division strategy.

3.3.2 The method

Overview

The development of appropriate heuristics requires a prior analysis of the characteristics of the studied phenomenon or process studied. For this purpose, the differences between orthoimage layers and the other layer types have been studied. The lessons learned have been used to develop the deterministic heuristics.

A WMS produces spatially referenced maps which are dynamically rendered in a graphical format. The geographic information used for rendering imagery is organised into logical units, i. e. layers, which might be organised as a hierarchical structure. The `GetMap` operation allows a user to request a single layer, a parent–layer, or any combination of those via explicit indication of layers in the request.

According to the content offered by a layer, five general categories can be identified:

- *Orthoimage Layers.* These layers offer orthoimages of high quality such as low altitude aerial photography or satellite orthoimages (normal colour images of high spatial and radiometric resolution) (Schmitt and Stilla, 2010).
- *Non–orthoimage Satellite Layers.* These layers offer non–orthoimage satellite imagery, e. g. radar or laser products.
- *Coverage Data Layers.* These layers offer images produced from coverage data representing continuous phenomena, for example a temperature map produced as an interpolation of sensor measurements (Iosifescu-Enescu et al., 2010), or a coastal ocean model (Schoenhardt et al., 2010).

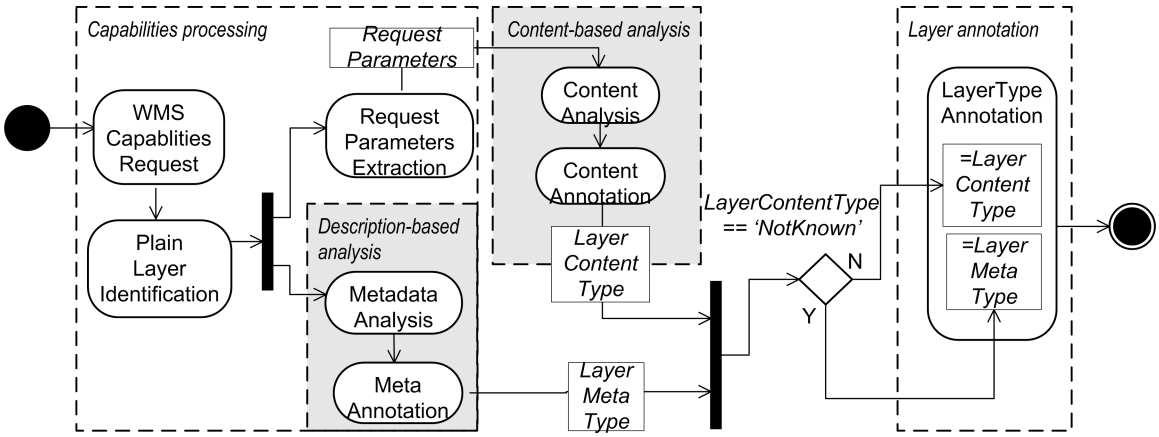

Figure 3.6: Overview of the method for the orthoimage layer detection.

- *Toposheet Layers.* These layers offer digitised but not vectorised toposheets, for example scanned maps.
- *Vector Layers.* These layers offer vector datasets, for example country boundaries or icon–based feature representations.

Various features of existing WMS services have been taken into account in order to detect the discriminating characteristics of an orthoimage layer. The analysis has considered descriptive features (i. e. the keywords used and declared scale in the *Capabilities* document), behavioural features (i. e. the presentation scale at which data are visualised by a viewer) and physical features (i. e. the characteristics of the layer image fragments, for example, number of colours). The proposed methodology for layer analysis is a combination of the two different approaches as depicted in Figure 3.6. First, the methodology proposed analyses the metadata contained in the WMS *Capabilities* document (the *MetadataAnalysis*). The second part of the method gathers a sample collection of image fragments to explore the content served by the WMS service (the *ContentAnalysis*). If it is possible to apply the image analysis to the collected samples, the result of the metadata analysis is not considered during the final annotation of the evaluated layer (the *LayerAnnotation*).

Description–based analysis

The description–based approach considers the OWS *Capabilities* document as an information source to identify the keywords that appear with high frequency in the description of orthoimage layers. Although the *Capabilities* document might be extended by a service provider, only the metadata elements that follow the OGC WMS specification have been considered during this analysis.

In the case of the WMS *Capabilities* document, the offered metadata are split into four groups:

- *General Service Metadata.* This group gathers general service metadata fields (e.g. *title*,

abstract, *keywords*, *bbox*).

- *Parental Layer Metadata.* This group gathers descriptive elements from the parental path in the hierarchical layer structure (e.g. *name*, *title*, *abstract*, *keywords*, *bbox*).
- *Layer Metadata.* This group gathers descriptive elements of the layer which might be inherited from the parental path (e.g. *name*, *title*, *abstract*, *keywords*, *bbox*).
- *Layer Metadata Document.* This group contains more precise layer description defined via a link to a standardised metadata document that describes the presented data.

It is necessary to identify the deterministic set of reference keywords per each metadata group g that distinguishes an orthoimage layer. Having a keywords set which is common for a group of orthoimage layers ($K_{oi}(g)$) and a keywords set which is common for a group of non–orthoimage layers ($K_{noi}(g)$), the deterministic set of reference keywords is the difference of $K_{oi}(g)$ and $K_{noi}(g)$, $K_{oi} \setminus K_{noi}(g)$ (i.e. the set of all elements of $K_{oi}(g)$ that do not belong to $K_{noi}(g)$).

The deterministic sets of reference keywords are then used to analyse the metadata of the evaluated layer. Figure 3.7 shows the workflow of the description–based analysis. First, the keywords are extracted per each metadata group from the *Capabilities* document of the analysed layer (the *GeneralServiceMetaKeywExtraction*, the *ParentalLayerMetaKeywExtraction*, the *LayerMetaKeywExtraction*, the *LayerMetaDocKeywExtraction*). If any keyword of the extracted set matches with the corresponding reference keywords, the layer is assumed to be an orthoimage layer. Otherwise, it is a `NonOrthoimage` layer.

As for the keywords extraction procedure, it processes separately the metadata groups. First, the `XML` document is parsed and the text from those fields is extracted and converted into a set of words. The applied tokenisation rules are similar to those that search engines use: (1) the transformation to lowercase, (2) the use of spaces and punctuation as word separators, and (3) the exclusion of words with encoding errors. Digits, one and two letter words, and words commonly ignored by search engines are also excluded, except those with semantic relevance in the geographic domain. Then, only the words of the highest frequency are considered.

Content–based analysis

Any content–based analysis requires the development of heuristics for effective data collection. In the case of a WMS service, an effective retrieval procedure refers to a process for making requests to the service and collecting a set of representative fragments of the image layer, instead of referring to the full extent of the layer. This set of image fragments allows the development of techniques for content evaluation, in this case, identifying an orthoimage layer via colour–based tests.

Figure 3.8 shows the workflow of the content–based analysis. First, the template of the image fragment request (`GetMap`) is defined using the information obtained from the OWS *Capabilities*

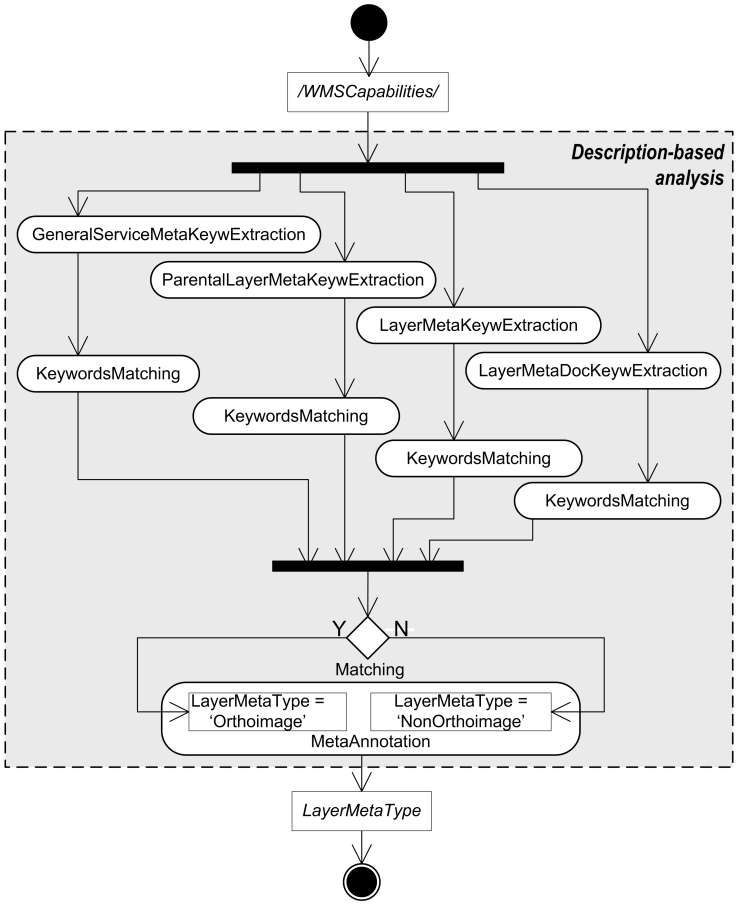

Figure 3.7: Workflow of the description–based analysis.

document. Then, the *collection procedure* is invoked. If a valuable image fragment collection has been gathered, the image analysis is performed over the collection. It combines the *colour test* and the *pixel test*. The result of the content–based analysis can be `Orthoimage` or `NonOrthoimage` according to the result of the *image analysis procedure*. If the image fragment collection is not valuable, the layer is annotated as `NotKnown`. The following subsections present the procedure for gathering a valuable image fragment collection (Section 3.3.2), and the details of the proposed heuristics for image analysis (Section 3.3.2).

Collection procedure The goal of the *collection procedure* developed is to gather a set of valuable image fragments from a WMS layer to perform the image analysis. A valuable image fragment is a valid response from a WMS service whose number of colours (N_{ImgC}) is bigger than an established threshold (T_{NImgC}), i. e. $N_{ImgC} > T_{NImgC}$. In the case of orthoimage layers, a one–colour image fragment is assumed to be non–valuable ($T_{NImgC} = 1$). It is probable that this condition causes

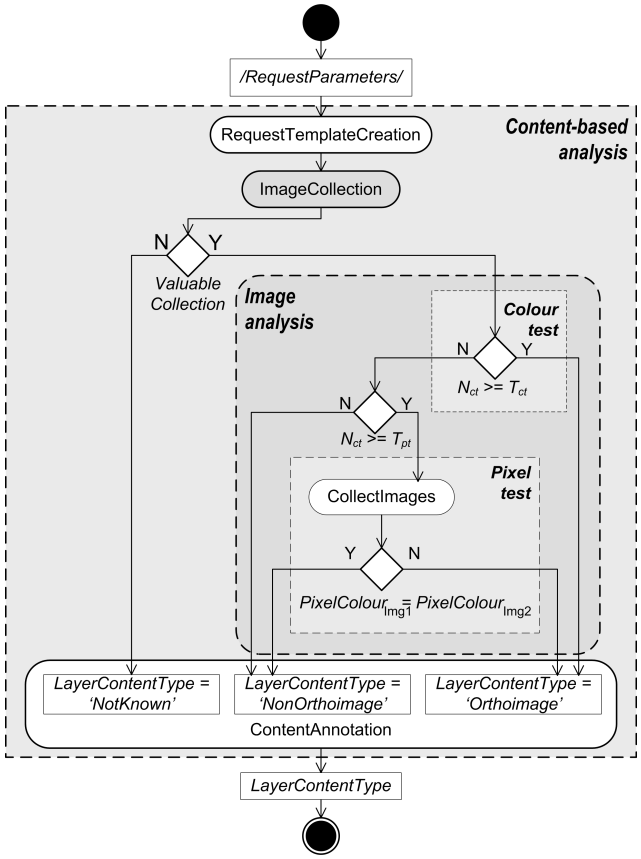

Figure 3.8: Workflow of the content–based analysis.

erroneous interpretation of an image with vector data, e. g. when an image fragment represents the inside part of a polygon represented by one colour. However, in this work orthoimages are targeted, and the risk of estimating a vector layer as an empty layer (i.e. no positive responses) is acceptable.

The *collection procedure* generates and invokes automatically series of `GetMap` requests for a layer which is the object of evaluation (the *layer name*, $IDLayer$). The definition of the request parameters should ensure homogeneous characteristics of images in the collection, i. e. image type, image size and scale. The image format (F_{Img}), for example JPEG or PNG, cannot differ in order to reduce the variability caused by image generation processes, which could distort the image analysis results. The requested image size, i. e. the number of pixels per height and width (H_{Img} and W_{Img} respectively), should not change either for uniformity of the evaluated collection. As the image size is constant, the scale of a WMS image is determined by the dimensions of the bounding box specified in the request ($BBox_r$). For every `GetMap` request a random $BBox_r$ must be randomly generated within the geographic extent of interest (the *work BBox*, $BBox_w$) while preserving the same scale (the *request scale*) and preventing deformation of representation.

The scale plays also a significant role in gathering a valuable image fragment collection, as the layer may offer data only for a specific scale range. The *request scale* should be a *presentation scale* supported by a layer, i. e. a scale for which the layer offers spatial information (i. e., it produces non–empty images). Therefore, a scale change should be considered if a valuable image fragment is not found at the selected scale. The *Capabilities* document does not specify a field to declare the *presentation scale* in the sense it is used here. It allows defining the range of scales for which it is appropriate to generate a map of a layer. Additionally, a provider may not specify this information and the scale recommendation can appear within a free text description of the layer or even in the layer name. In this case, pattern–based searching is used to extract this information.

Other issue that should be considered when designing the collection algorithm is the layer geographic extent. The *Capabilities* document informs about the geographic extent for which the layer offers spatial data, i. e. the bounding box of the layer ($BBox_{Layer}$). However, this information might not be accurate enough to get a valuable collection. The $BBox_{Layer}$ indicates the extent covered by data, but it does not mean that any request within this search area will provide a valuable response. The layer may offer data only for a part of the defined $BBox_{Layer}$. For example, especially in the case of vector data, data might be widely dispersed and their relative size might be considerably tiny. This complicates the development of an automatic procedure for collecting valuable images. If the area covered by data is very small compared to the search area, it can be difficult to find any data, and the task collecting images of the whole area at a given scale might not be efficient. Therefore, it is necessary to restrict the number of requests, and if the collecting task has not gathered a valuable collection, the search area is divided into several parts (N_D). Then, another series of requests are performed per each new search area to ensure better distribution of spatial requests. This division–requesting task can be repeated if necessary. Figure 3.9 shows a sample layer which requires several division iterations to find a valuable image collection.

Although the iteration can be repeated any times, the number of repetitions should depend on the $BBox_{Layer}$, because the method is developed to work on a sample collection of image fragments and not on the whole layer geographic extent. The performance aspect must be considered as well, especially for those bounding boxes of layers producing an image collection of considerable dimension (e.g. Europe) at a selected scale. For this reason, in this work, the number of iterations is limited.

Algorithm 1 presents the algorithm of the developed *collection procedure*. It applies the division–requesting task and considers the change of the *request scale*. Additionally, the algorithm requires the following auxiliary functions to define several thresholds:

- $N_D(i)$ – a function that defines the number of areas into which the $BBox_{Layer}$ is split (i. e. the number of $BBox_w$) in each iteration (i):

$$N_D(i) = 4^{i-1};$$

- $N_{Rw}(i)$ – a function that defines the number of requests per each $BBox_w$ (at iteration i):

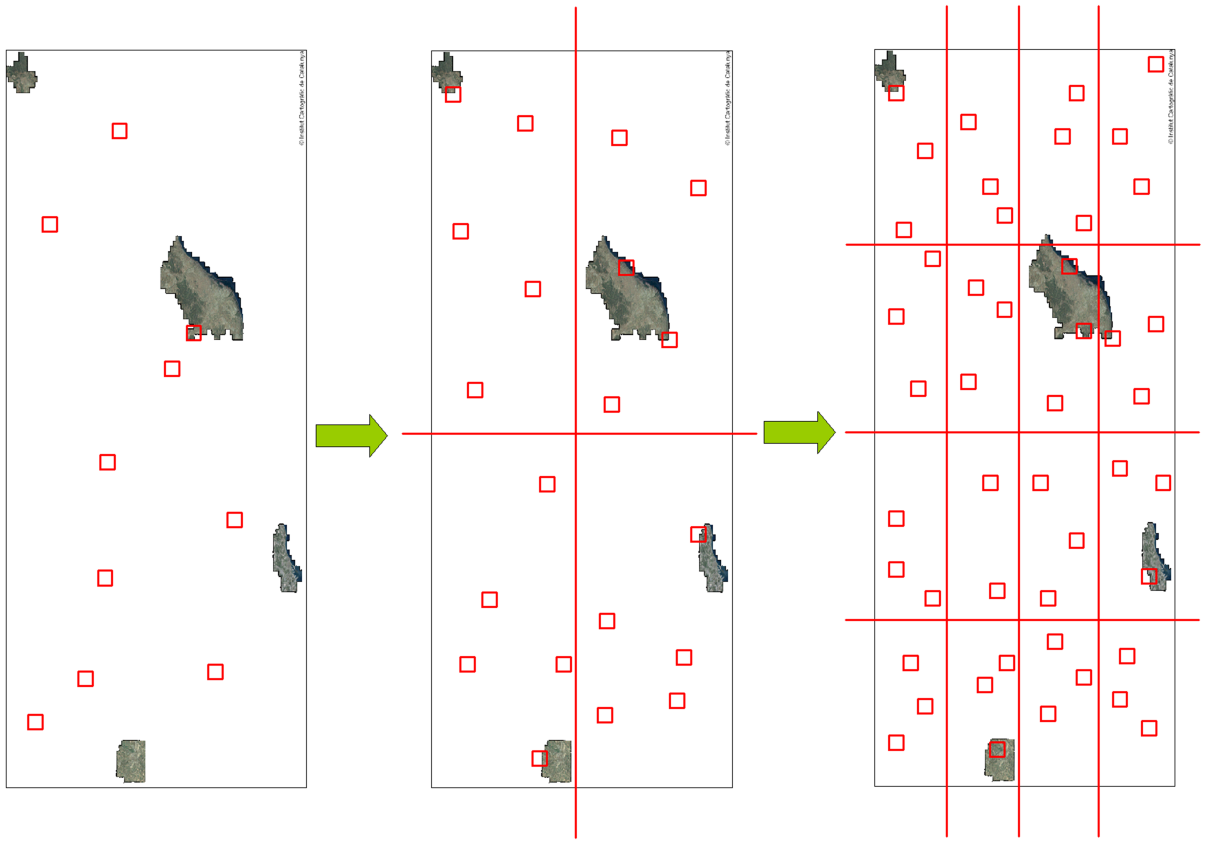

Figure 3.9: Example of image division in the *collection procedure*.

$$N_{Rw}(i) = \begin{cases} initialise, & \text{if } i = 1, \text{ for example } 100 \\ N_{Rw}(i-1)/2, & \text{if } i > 1; \end{cases}$$

- $N_{Img}(i)$ – a function that defines the expected number of images in the collection after each iteration (if no error occurred):

$$N_{Img}(i) = \begin{cases} N_D(i) * N_{Rw}(i), & \text{if } i = 1 \\ N_D(i) * N_{Rw}(i) + N_{Img}(i-1), & \text{if } i > 1; \end{cases}$$

- $T_{NvImg}(i)$ – a function that defines the threshold to discriminate a valuable image collection (at iteration i), i. e.:

Algorithm 1 Algorithm for collecting image fragments of a WMS layer.

1: **function** ImageCollection($T_{NImgC}, Rq_{Tmpl}, iter$) ▷ The threshold to determine a valuable image ($T_{NImgC} = 1$); the GetMap request template (Rq_{Tmpl}); the number of iterations ($iter \geq 1$).
2: $BBox_{Layer} \leftarrow Rq_{Tmpl}.bbox()$ ▷ The *layer BBox*.
3: $N_{vImg} \leftarrow 0$ ▷ The number of valuable GetMap responses.
4: $error \leftarrow FALSE$ ▷ The GetMap response error.
5: $S_r \leftarrow S(1)$ ▷ Initialising the request scale for the first iteration.
6: $changeScale \leftarrow FALSE$
7: $i \leftarrow 1$ ▷ The counter of iterations.
8: **repeat**
9: $Set_{BBoxw} \leftarrow split(BBox_{Layer}, N_D(i))$ ▷ The $BBox_{Layer}$ division
10: **if** *changeScale* **then** ▷ Scale change.
11: $S_r \leftarrow S(i)$
12: **end if**
13: **for all** $BBox_w$ in Set_{BBoxw} **do**
14: $r \leftarrow 1$ ▷ The counter of requests per work area.
15: **repeat**
16: $BBox_r \leftarrow rand(BBox_w, S_r, Rq_{Tmpl})$ ▷ The request bounding box
17: $res \leftarrow getMap(Rq_{Tmpl}, BBox_r)$ ▷ The response of GetMap request.
18: $error \leftarrow chkError(res)$ ▷ The response error.
19: **if** $\neg error$ **then**
20: $img \leftarrow getImage(res)$ ▷ The response image fragment.
21: $imgCollection.add(img)$ ▷ Adding *img* to the image collection.
22: $N_{ImgC} \leftarrow colourNumber(img)$ ▷ The number of colours of *img*.
23: **if** $N_{ImgC} > T_{NImgC}$ **then**
24: $N_{vImg} \leftarrow N_{vImg} + 1$ ▷ Counting valuable image fragments.
25: **end if**
26: **end if**
27: $r \leftarrow r + 1$
28: **until** $(error) \vee (r > N_{Rw}(i))$ ▷ It ends if there is any WMS error or all requests per the $BBox_w$ have been performed.
29: **end for**
30: $changeScale \leftarrow (N_{vImg} < T_{NvImgS}(i))$ ▷ The request scale should be changed in the next iteration, if the number of valuable image fragments is too small.
31: $i \leftarrow i + 1$
32: **until** $(error) \vee (N_{vImg} > T_{NvImg}(i)) \vee (i > iter)$ ▷ It ends if there is any WMS error or a valuable collection has been collected, or all permited iteration have been performed already.
33: **return** *imgCollection* ▷ The image collection.
34: **end function**

$$N_{vImg} > T_{NvImg}(i), \text{ where}$$
$$T_{NvImg}(i) = 10\% * N_{Img}(i) - 1, \text{ and}$$
$$N_{vImg} \text{ – the number of valuable images gathered so far;}$$

- $T_{NvImgS}(i)$ – a function that defines the threshold to decide a scale change (at iteration i), i. e.:

$$N_{vImg} < T_{NvImgS}(i), \text{ where}$$
$$T_{NvImgS}(i) = 2\% * N_{Img}(i) + 1, \text{ and}$$
$$N_{vImg} \text{ – the number of valuable images gathered so far;}$$

- $S(i)$ – a function that defines the appropriate scale for each iteration (i); this function must be customized for each experiment, using a sample set of inputs (Section 3.3.3).

First, the variables are initialised and almost all have value *0* or *false*. The *work BBox set* (Set_{BBoxw}) is initialised with one *work BBox* ($BBox_w$) equal to the *layer BBox* (i. e., $N_D(1) = 1$). Then, a limited number of requests (N_{Rw}) are performed with a random *request BBox* ($BBox_r$) within the *work BBox*. If the WMS response is an error or an image cannot be opened (e.g. it gives some errors), the algorithm finishes and the image collection is returned. Otherwise, an image is added to the image collection. Additionally, a colour histogram is created per each image and if the image is valuable, the counter of valuable images in the collection (N_{vImg}) increases. When an image collection is gathered at first iteration, the conditions on valuable collection and scale change (*changeScale*) are checked for the next iteration and the iteration counter increases. If the collection is not valuable and there is no error response, the next iteration starts. The $BBox_{Layer}$ is split into $N_D(i)$ elements (i. e. a new *work BBox set* is defined). A new *request scale* is set (S_r) if the *changeScale* is true. Then, the gathering task is performed per each *work BBox* in the *work BBox set* again. The number of iteration is limited with the parameter *iter*. The output of the algorithm is an image fragment collection.

Image analysis procedure Two heuristics of image analysis have been investigated. The first one, the *colour test*, operates on the colour histogram of an image. The other one, the *pixel test*, observes changes in the colour characteristics of a selected area between images generated for two spatially overlapping requests. The final *image analysis procedure* combines both of them.

The *colour test* performs the analysis of the number of colours. High quality orthoimages are characterised by the use of a bigger number of colours compared to other types of datasets. It consists of an array of pixels that the sensor picked up and may contain hundreds, or even thousands, of different colours depending on the capture conditions (light, sensor quality, etc.). Based on the above assumption, the collecting procedure has to gather images in a format that does not employ a lossy data compression to prevent loss of quality. Therefore, PNG format has been chosen. Then, the average number of colours of valuable images in the total image collection (N_{ct}) is compared to a threshold value (i. e. the *colour test threshold*, T_{ct}). The proper value of T_{ct} will be estimated from examples of orthoimages during an evaluation step previous to the experiment (Section 3.3.3).

The second heuristic, the *pixel test*, is a supporting procedure allowing detection of the lossy compression of raster format (e. g., ECW, JPEG2000), which is typically used for storing images. It assumes that the images portrayed by a WMS are stored with this format, and that it is possible to detect the pixel colour variance in images with lossy formats, which are rendered by a WMS in response to overlapping spatial requests. In the first step, a valuable image fragment is selected from the image collection. It is requested again with a lossy format (i. e. JPEG). A colourful pixel (i. e. the *base pixel*) is chosen from the obtained *base image*. Then, a *test image* is retrieved with a request that overlaps the *base image* in the area represented by the selected pixel (Figure 3.10). The match pixel is identified in the *test image* and compared with the *base pixel*. If a number of such tests determine that the colour varies in any of the pixel pairs, it is probably an image using

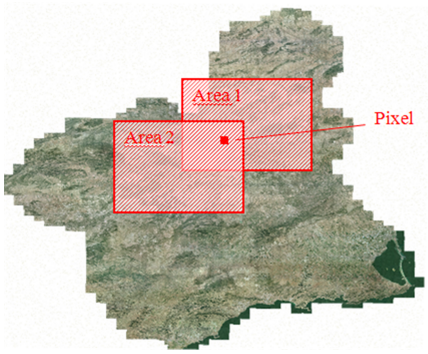

Figure 3.10: *Pixel test* heuristic. The *Area 1* represents the request BBox of the *base image* and the *Area 2* shows the request BBox of the *test image*.

internally a raster representation. Layers using internally a vector representation format are always rendered in the same way, assuming that the request parameters differ only in the request BBox.

Figure 3.8 presents the workflow of the developed *image analysis procedure*. If it has been possible to retrieve a valuable collection (i. e. $N_{vImg} > T_{NvImg}$), the image analysis is performed. First, the valuable image collection is analysed applying the *colour test*. If the average colour number (N_{ct}) is equal or bigger than the *colour test threshold* (T_{ct}), the layer is assumed to be an orthoimage layer (`Orthoimage`). If the number of colours is slightly lower (i. e. $N_{ct} >= T_{pt}$), the *pixel test* is performed. For example, the *pixel test threshold* value might be 10% lower than the *colour test threshold*. A valuable image fragment is selected from the collection and the request that has generated it is identified. Then, two image fragments are requested (*CollectImages)*. The first one is requested using the identified request but with image format necessary to proceed the *pixel test*. It is the *base image* (Img_1). Then, the *test image* (Img_2) is requested and the test performed as described previously. As result of the *image analysis procedure*, the layer is assessed to be an `Orthoimage` or `NonOrthoimage` layer.

3.3.3 Experiment

This section describes the experiment performed over a collection of WMS layers (i. e. *layer collection*) to filter orthoimage layers.

Data corpus and tuning of parameters

An OWS crawler has been used to exploit existing search engines to identify WMS instances on the Web (see López-Pellicer et al. (2011e) for more details) from the Spain and Portugal region, and the discovered WMS *Capabilities* documents have been stored in a repository. The *layer collection* has been extracted from the repository applying the following restrictions: (1) the service must be accessible and respond correctly to the standardised operations, and (2) the layer must be requestable

Layer collection (N_{LTotal})	**Vector** (N_{LVec})	**Coverage** (N_{LCov})	**Toposheet** (N_{LTopo})	**Satellite (non–orthoimage)** ($N_{LSatNOrt}$)	**Orthoimage** ($N_{LSatOrt}$)
5848	5259	241	108	117	123
100%	89.93%	4.12%	1.85%	2%	2.1%

Table 3.2: Number of layers per each layer type for the *layer collection* used in the experiment.

and a single layer. As a result, the *layer collection* consisted of 5848 layers offered by 708 different WMS services.

First, the *layer collection* has been manually analysed and annotated: (1) the associated capabilities information has been analysed, (2) the offered content has been visually examined, and (3) the features of `GetMap` responses have been identified. Table 3.2 presents the number of layers in the *layer collection* per each layer type.

Additionally, per each layer type a set of ten random layers (i. e. a *type layer set*) has been selected for tuning the parameters of the algorithm. An evaluation of these groups has allowed discovering the characteristics of a layer type, which have helped to establish the algorithm parameters used in the performed experiment.

For the selection of typical keywords describing orthoimages, the *Capabilities* documents of the *type layer set* have been processed to extract the $K_{oi}/K_{noi}(g)$ per each metadata group (Section 3.3.2). The *Layer Metadata Document* group has not been considered, as the capabilities do not provide this information (hardly 1% of the *layer collection*). Application of this method to the *General Service Metadata* and *Parental Layer Metadata* groups has resulted in empty lists. Only the *Layer Metadata* group has produced the deterministic set of reference keywords (i. e. $K_{oi}/K_{noi}(LayerMetadata)$): “orto”, “ortho”, “photo”, “ortofoto”, “ortofotografias”, “ortofotografía”, “pseudoorto”, “vuelo”, “LIDAR”, “PNOA”. Due to the characteristics of the *layer collection* that has been used, the result list contains mainly Spanish vocabulary.

For the construction of the `GetMap` operation requests used in the content–based analysis, the *Capabilities* document must be also analysed to derive the appropriate parameters of the request. The template of a `GetMap` request is defined as follows:

<$RequestPoint$>SERVICE=WMS&VERSION=<$Version$>&REQUEST=GetMap&LAYER=<$IDLayer$>&STYLE=<$Style$>
&BBOX=<$Bbox$>&[SRS=<Crs>|CRS=<Crs>] &HEIGHT=<H_{Img}>&WIDTH=<W_{Img}>&FORMAT=<F_{Img}>.

The <$Version$> and <Crs> values are important because they influence the formatting of spatial restriction parameters (Table 3.3). The WGS 84 coordinate system is used by default (i. e. $Crs = EPSG : 4326$). If the layer does not support it, one of the supported systems is selected. If the <$Style$> value is “Default”, it remains empty. The height (i. e. <H_{Img}>) and width (i. e. <W_{Img}>) of requested images have been defined as 500 and 400, respectively. For the *colour test* heuristic the images are requested in PNG format (a format with lossless data compression as explained in Section 3.3.2). An example of `GetMap` request might be as follows:

WMS Version	Coordinates System Definition	Coordinates System Example	Request Bounding Box Definition	Request Bounding Box Example
before 1.3.0	SRS= namespace:identifier	SRS=EPSG:4326	BBOX=minx,miny, maxx,maxy (min longitude, min latitude, max longitude, max latitude)	BBOX=-180,-90, 180,90
1.3.0	CRS= namespace:identifier	CRS=EPSG:4326	BBOX=minx,miny, maxx,maxy (min latitude, min longitude, max latitude max longitude)	BBOX=-90,-180, 90,180
		CRS= CRS:84	BBOX=minx,miny, maxx,maxy (min longitude, min latitude, max longitude max latitude)	BBOX=-180,-90, 180,90

Table 3.3: Parameters of `GetMap` request for definition of the spatial constraint according to the WMS service version and CRS.

Iteration (i)	$N_D(i)$	$N_{Rw}(i)$	$N_{Img}(i)$	$S_r(i)$	$T_{NvImg}(i)$	$T_{NvImgS}(i)$
1	1	100	100	1:5000	9	3
2	4	50	300	1:2500 (if not in metadata)	29	7
3	16	25	700	1:2500 (if not in metadata)	69	15

Table 3.4: Values of functions of the content collecting procedure per iteration ($iter = 3$, $N_{Rw}(1) = 100$).

```
http://ovc.catastro.meh.es/Cartografia/WMS/ServidorWMS.aspx?SERVICE=WMS&VERSION=1.1.1&REQUEST=GetMap
&LAYER=ELEMLIN&STYLE=&BBOX=-3.71,40.40,-3.70,40.41&SRS=EPSG:4326&HEIGHT=500&WIDTH=400&FORMAT=image/png.
```

Table 3.4 summarises the configuration of the collection algorithm used in this experiment. The number of iterations has been set to 3 (i. e., $iter = 3$) and the number of requests for the first iteration has been set to 100 (i. e., $N_{Rw}(1) = 100$). The algorithm has required establishing a starting *request scale*. For this reason, the range of presentation scales of the orthoimage layer group from the *type layer set* has been manually analysed. Only 60% provide data for scales lower than 1:125 000, 80% provide data for scales higher than 1:125 000, 90% provide data for scales higher than 1:29 000, and 100% services provide data for scales higher than 1:15 000. Then, the colour variability between orthoimage layers and other layers at different scales has been compared. Image fragment collections have been gathered for 1:25000, 1:10000 and 1:5000 scales per each layer group of the *type layer set* (i. e., fifteen image collections in total). The results have shown that at 1:5000 the variability of colours between orthoimage layers and others layers was bigger than using lower resolutions. Taking into account these results, the 1:5000 scale has been selected as a start scale.

Test	**Perform Condition**	**Pass Condition**	**Threshold Value**	F_{Img}
Colour test	$N_{vImg} \geq T_{NvImg}$	$N_{ct} \geq T_{ct}$	$T_{ct} = 50\,000$	*image/png*
Pixel test	$N_{vImg} \geq T_{NvImg}$, $N_{ct} \in < T_{pt}; T_{ct})$ where $T_{pt} = T_{ct} - 10\% * T_{ct}$	$PxelColour_{Img1} = PxelColour_{Img2}$	$T_{pt} = 45\,000$	*image/jpeg*

Table 3.5: Image analysis parameters.

	Description–based Analysis	**Hybrid Analysis**
Identified as Orthoimage Layer (R_{LTotal})	803	85
Possitive Results ($R_{LSatOrt}$)	120	74
Precision ($R_{LSatOrt}/R_{LTotal}$ %)	14.94%	87.06%
Recall ($R_{LSatOrt}/N_{LSatOrt}$ %)	97.56%	60.16%

Table 3.6: The results of the performed experiment.

The parameters of the image analysis algorithms applied in the experiment are summarised in Table 3.5. The PNG and JPEG formats have been requested during the *colour test* and the *pixel test* respectively. In order to estimate the *colour test threshold* (T_{ct}), the image features of each group from the *type layer set* have been manually analysed. In general, vector data usually returns images from 2 to 100 different colours; coverage data from 100 to 500 colours; non–orthoimage satellite data, up to 10 000 colours; and orthoimage, more than 50 000. Some noise has been introduced by toposheets, because some scanned maps are quite colourful. Considering these outcomes, the *colour test threshold* has been set at 50 000 colours. The *pixel test* is performed if the layer does not pass the *colour test* and the average colour number is only 10% lower than the *colour test threshold*.

Tests

Once the data corpus was fixed and the parameters were computed, two tests were performed: a first test using only the description–based heuristics, and a second test using both the description–based and the content–based heuristics. The test results are summarised in Table 3.6.

During the first test, the content–based analysis was switched off. The performed test has indicated that the description–based analysis is inefficient. Within 803 layers returned as orthoimage layers only 120 are correctly identified, which gives a precision of 14.94%. The inappropriate service description causes the low precision of this test result. On the other hand, only 3 orthoimage layers have not been identified because of lack of description, which gives a recall of 97.56%.

In the second test, the complete algorithm including both approaches has been executed. The method has identified 85 layers as orthoimage layers and 74 layers are true positive. Therefore, this gives a precision of 87.06 % and a recall of 60.16%.

The false positives are 5 colourful toposheets (5.88%) and 6 vector layers (7.06%). The vector layers that have been identified as orthoimage layer are characterised by the use of colourful icons, or

the presence of more than one geographic feature type. The vector layers that present more than one geographic information logic unit correspond with compound layers, and in practice they should have been split into various related layers. Compound layers have not been considered in this method. Therefore, this kind of single layers are interpreted as noise. No coverage or non–orthoimage satellite layers have been identified within the results.

The complete algorithm is characterised by improving the precision from 14.94% to 87.06%, which indicates that the applied *image analysis procedure* is suitable for the required functionality. On the other hand, the decrease of the recall (from 97.56% to 60.16%) seems to be influenced by the image fragment collection algorithm. Within the negative results there are layers where the algorithm was not able to gather a valuable collection. One of the questions of the collecting process is the dispersion of the requested bounding box within the layer bounding box. The requested bounding boxes are spread more homogeneously during the second and third collecting iterations. However, it has been observed that the number of collected valuable responses increases only in the case that the first iteration retrieves at least one valuable image fragment.

3.3.4 Implementation of the method as a WPS

The WMS layer analysis functionality has been published as a Web service in compliance with the OGC WPS specification (Schut, 2007). When a WPS offers a time consuming process, it is necessary to offer an asynchronous communication protocol with the client. Such a WPS usually supports two operations: one operation to invoke the procedure, and another to obtain information on process state and retrieve the results. The analysis of the content of a layer is one of these cases that might take some time. For this reason, the service prototype supports two processes whose execution might be invoked: “analyseLayer” and “getAnalysisResults”.

Figure 3.11 shows the class diagram of the implemented prototype. *OGCService* is an interface that must be implemented by all kinds of OGC services. The *OGCWPS* interface extends the *OGCService* interface and offers two methods, the `DescribeProcess` and `Execute`, in conformance with the OGC WPS specification. The *OGCLayerAnalysisWPS* interface extends the *OGCWPS* functionality to permit users to explicitly invoke the processes offered (“analyseLayer” and “getAnalysisResults”). This interface is implemented by the *LayerAnalysisWPS* class and the prototype is an instance of this class. As the service is multithread, various analysis might be performed at the same time. The *AnalysisProcess* class is responsible for the analysis process. The *ProcessInfo* and *RequestInfo* classes gather information on the process and the user request, respectively. The *AnalysisResult* class contains the results of the analysis and information on errors if they have occurred. Some errors may occur while initialising the user request (e.g. “Capabilities document could not be processed.”; “The requested layer $< IDLayer >$ does not exist.”) or during the analysis process (e.g. “Service connection error.”).

Figure 3.12 presents the sequence diagram of the communication between a client (the *User*) and

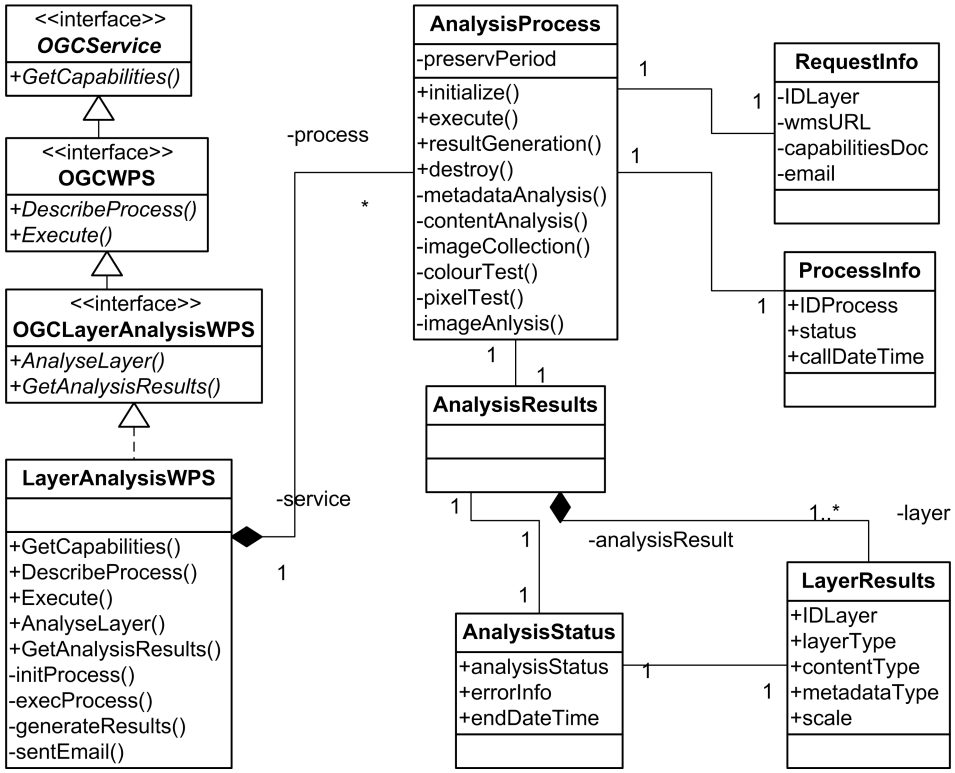

Figure 3.11: Class diagram of the implemented WPS.

the prototype service (the *OGCLayerAnalysisWPS*) via the *OGCWPS* interface. The execution of the "analyseLayer" process invokes the WMS layer analysis (the `Execute` operation with *AnalyseLayer* parameter). The process takes a `URL` of the `GetCapabilities` request of a WMS service (the *URL*) as an input parameter. A layer name (the *IDLayer*) and a user email (the *eMail*) are optional parameters. If the layer name is provided, only this layer will be analysed. Otherwise, all requestable and single layers found in the *Capabilities* document will be analysed. The response of this operation is a process identifier (the *IDProcess*) that the client has to use as a parameter when executing the "getAnalysisResults" process (the `Execute` operation with *GetAnalysisResult* parameter). When the analysis process finishes (*STATUS=Finished*), the results are stored on the server for some period of time (see the note *PreserveResults* in sequence diagram). The information on this period is provided in the *Abstract* metadata of the "getAnalysisResults" process description (it may be obtained via the `GetCapabilities` or `DescribeProcess` method). It may be configured by the service publisher. By default, it is one week from the time the process finished. If the user email has been provided, an email with the results obtained is sent as well (see *sendEmail* invocation in sequence diagram).

The "getAnalysisResults" response message (in an `XML` format) has four parts:

- *Status.* The *Status* block provides information about the current state of the process. The

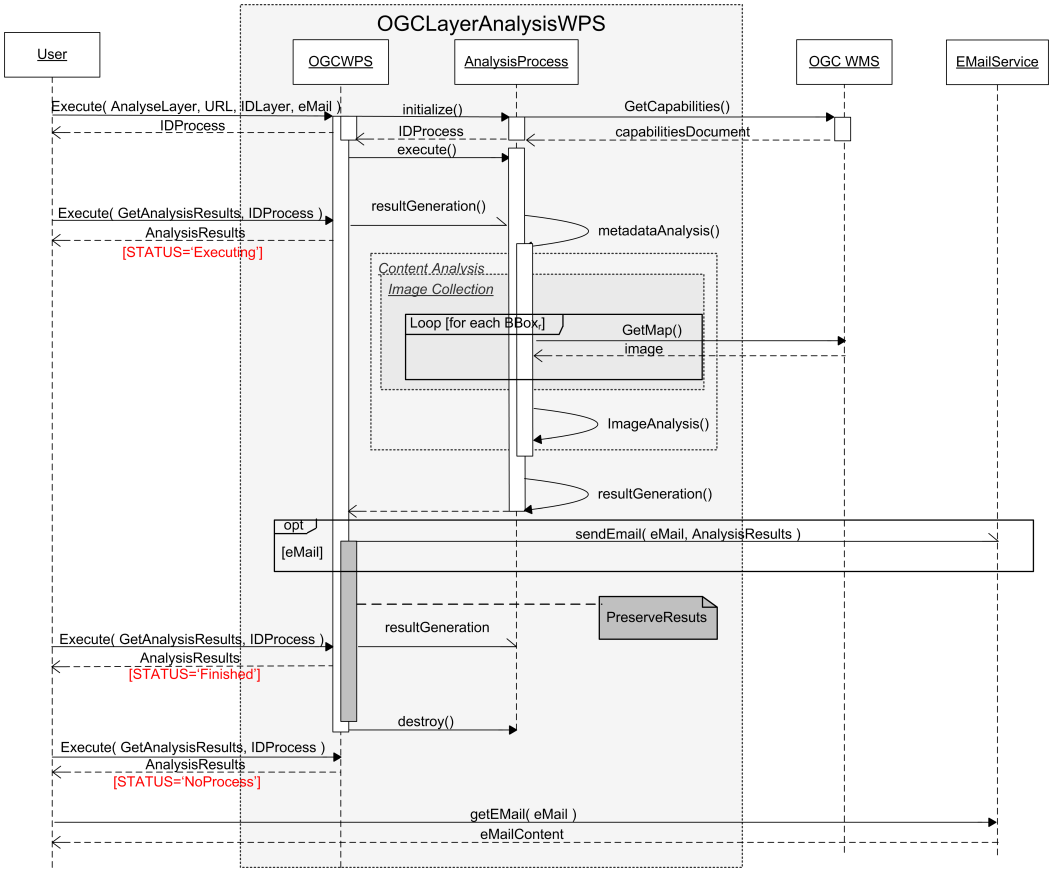

Figure 3.12: Sequence diagram of the communication between a client and the OGC WPS for the identification of the orthoimage WMS layer.

"Executing" value means that the process is still being performed. The "Finished" status informs that the analysis process has already finished. If 'NoProcess" value appears, it means that the process of the provided identifier does not exist: either it has been already destroyed or it has never been run.

- *ProcessInfo.* The *ProcessInfo* element contains the *IDProcess* element and the *CallDateTime* element with information on the process invocation moment.

- *RequestInfo.* This block contains information on the user request, i. e. the `GetCapabilities` request (the *wmsURL*) and the retrieved response (the *CapabilitiesDoc*), the layer name (the *IDLayer*), and the user email (the *eMail*) if provided.

- *Response.* The *Response* element gathers the analysis results at the moment of the "getAnalysisResults" request. If the user request has been initialised successfully, the analysis status (the *analysisStatus*) has "OK" value and the *AnalysisResult* element summarises the analysis

results per each analysed layer (the *LayerResult*). If the analysis status has the "Error" value, the *ErrorInfo* element is provided instead. The *LayerResult* element contains the layer name (the *IDLayer*) and estimated type (the *LayerType*), and the scale (the *Scale*) at which it has been analysed.

With respect to the process state (*Status*), the system generates different responses to the "getAnalysisResults" request.

This WPS service can be useful when the input of other processes or a potential WMS client needs to access information presented with a certain precision and form. For instance, this WPS service might be a component of an automated catalogue service able to analyse content and extract additional meta–information.

3.4 Application of the methods developed

The catalogs that use precise spatial representation instead of MBBOX approximation offer better functionality for the applications based on on–the–fly data integration. An example might be an application which allows displaying geographic information from different OGC services found in a services catalog (see Figure 3.13). The prototype of such enhanced CSW catalogue has been used as the core component of Web application[7], which displays spatial data provided from different OGC services. Only services that return responses encoded in GML (i.e. WFS and WCS services) has been considered for assignation of geoindentifiers. GML–encoded response allows unambiguous interpretation in terms of existence of geographic information, which can be confusing in case of an image (e.g. a one colour image might be empty or not). The Web client retrieves OWS services via MBBOX but only enables those which spatial objects overlaps with current bounding box (i.e. the displayed area). Since the list of selectable layers depends on area displayed, it improves user experience. The main disadvantage of this proposal has been the response time of the reference WFS service. To solve this problem we have created a local repository of spatial objects retrieved previously from the reference service. In this way, each instance of service is annotated with its precise spatial object as well. The cache techniques have also improved the services catalog response time and the behaviour of the end application.

The method for image layer identification has been applied in the *Virtual Spain* project, an R&D project supported by the Spanish Government through the Centre for the Development of Industrial Technology whose objective is to define architectures, protocols and standards for an envisioned 3D Internet focusing specially in 3D visualisation, virtual worlds, user interactions and the introduction of semantic capabilities. In particular, one of the experiments proposed in this project is focused on the crawling of services for access and exploitation of images[8]. The method presented

[7] http://www.idee.es/IDEE-ServicesSearch/ServicesSearch.html
[8] http://ev.unizar.es/EV42/

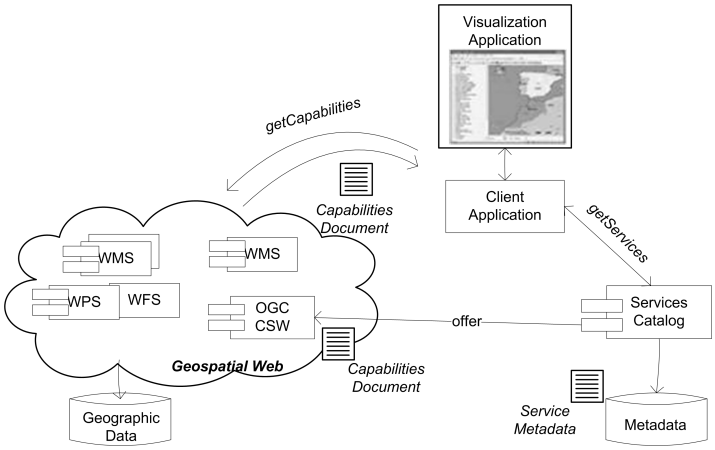

Figure 3.13: Services catalog as the support component for application based on–the–fly data integration.

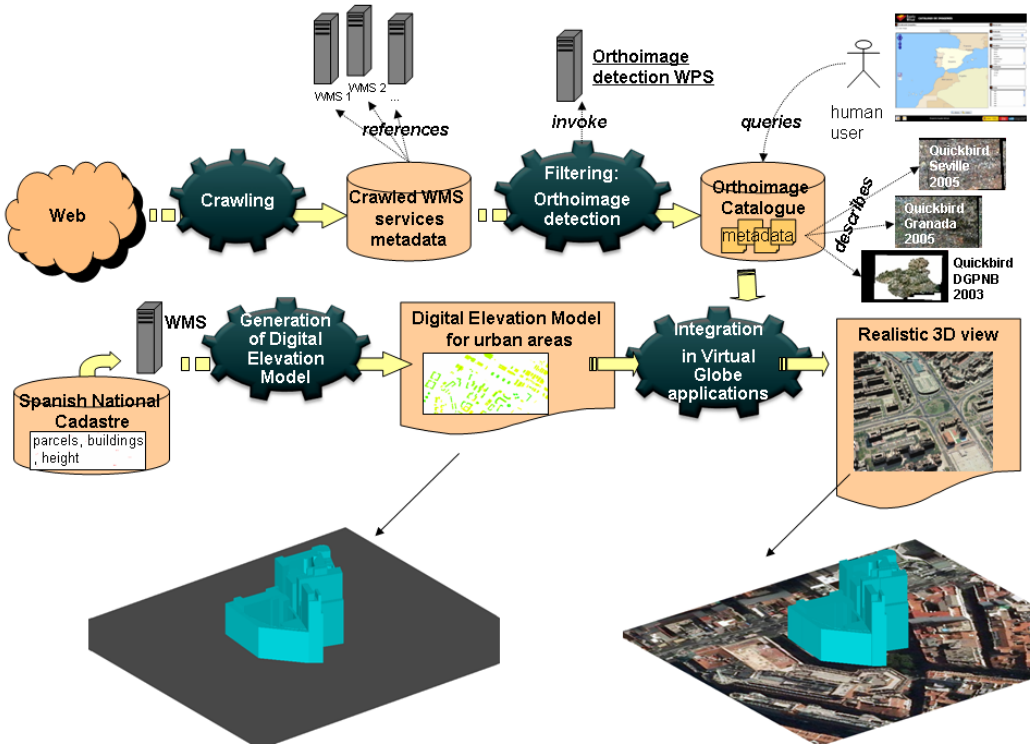

Figure 3.14: Integration of the WPS for orthoimage detection within the Virtual Spain project, an example of generated 3D representations for an urban area (*Gran Vía Street, Madrid*).

has been used in the development of a catalogue of orthoimages[9], which are compiled as a result of

[9]http://ev.unizar.es/EV-ResourcesCatalogueH7/ResourcesCatalogue.html?locale=en

crawling WMS services on the Web. The method, implemented through the *OGCLayerAnalysisWPS* class as explained in Section 3.3.4, contributes to the filtering of crawled services to detect if they contain image layers. This project also explores how to link these image services with Virtual Globe applications in order to provide a realistic display of Digital Elevation Models (DEMs) for urban areas, which are produced as the result of another experiment in this project. Virtual Globes provide computer–based representations of the real world that are receiving an increasing interest by experts in the geoscience field (Bailey and Chen, 2011). Figure 3.14 depicts the architecture of this workflow. On the one hand, an orthoimage catalogue has been created by filtering orthoimage layers of WMS instances that have been previously gathered from the Web. On the other hand, the generation of DEMs in urban areas has been investigated. Figure 3.14 shows also an example of generation of a 3D model of a real building in the *Gran Vía* street in Madrid. On the left side, there is a basic model, and on the right side the same building model is combined with an orthoimage requested from the "pnoa" layer (offered by PNOA–IGN WMS service[10]), which has been selected from the orthoimage catalogue.

The image catalog resulted from filtering crawler outcomes can use the additional information about the scale at which the image layer has been identified. This information can be exploited by Web client for recommendation for human users.

3.5 Summary

This Chapter has presented two example applications which require additional characterisation of geospatial resources. The approach proposed uses content–based heuristics for dataset sampling which take advantage of standardised interfaces of the Web resources analysed.

The first example uses semantic Web technologies to describe OWS. First, the idea of abstraction of a geographic feature from its spatial definition is presented. Then, the current approaches in referencing and identification of geographic features in the Semantic Web and the Geospatial Web are revised. Following the supposition that applying the best practices from the Semantic Web might be useful for the Geospatial Web, here, an *administrative geography* is created in accordance with Linked Data principles. The principal advantage of using Linked Data technology in geospatial solutions is the possibility of explicit identification of features and abstraction of their spatial definition from footprint and computational representation. The different spatial representation might be accessible via linked instances and chosen according to the application requirements. Such ontology is used as source of geoidentifiers in a geospatial solution and its main advantage lies in using more precise spatial representation and spatial reasoning on the semantic level. A Geoidentifier and corresponding MBBOX used to represent the geographic extents may improve the recall of OGC services catalog. For instance, this improvement has been used in developing Web–based applications that facilitate

[10] http://www.idee.es/wms/PNOA/PNOA?Request=GetCapabilities&Service=WMS&Version=1.1.1

on–the–fly data integration. Currently, this early design of the applied ontology has evolved to the *Spanish jurisdictional application ontology* described in López-Pellicer et al. (2011f).

The second part of this Chapter has been dedicated to providing a geoprocessing service for the automatic identification of orthoimages offered through WMS services. In order to demonstrate the feasibility of this geoprocessing service, the implementation has been tested on a collection of WMS instances found on the Web. The human supervision of the experiment results has proved the efficiency of the proposed method (87 % precision and 60 % recall). The image catalog resulted from filtering crawler outcomes has been applied in the *Virtual Spain* project.

Chapter 4

Semantic characterisation of Geospatial Web resources

4.1 Introduction

The aim of this Chapter is to propose an architecture of a system dedicated to the automatic creation of geographic metadata of Web resources. Such architecture should be prepared to support various metadata models and different types of Web resources (i.e. to be easily extensible). A heuristic–based method for geographic scope estimation of Web pages has been proposed as well. A prototype, which is able to generate a geographic metadata (in DC profile) of an HTML Web page, has been developed and tested. The metadata model is tailored to the SDI requirements because that is the model used by the metadata catalogues compliant with the OGC Catalogue Service standard. The experiments have been run on a realistic corpus made of OWS *Capabilities* documents (i.e. generated by a Web crawler focused on OWS (López-Pellicer et al., 2011e)), which has placed the developed prototype in the actual Web environment. These Web pages belong to sites of publishers of Geospatial Web resources; therefore the generated metadata could be used in a catalog.

This Chapter is organised as follows. Section 4.2 summarises the existing approaches in automatic creation of metadata from the Web community, with special focus on Web pages and geospatial domain. Section 4.3 introduces an architecture for automatic metadata generation for Web resources. First an abstract workflow of the process necessary to generate metadata is outlined and then, the architecture is briefly described. Section 4.4 presents application of the architecture proposed to generate geographic metadata for the Web pages which are part of Web site of publishers of Geospatial Web resources. Considering that the geographic metadata are hardly provided within Web pages, including those from the geospatial domain, a coverage estimation method has been developed. Then, the prototype implemented and results of performed experiments are described.

In the end, some conclusions are presented.

4.2 Web community and geographic metadata

The approaches to the automatic generation of metadata from the geospatial community are appropriate for Geospatial resources; however, they cannot be applied successfully to Web resources such as geoportal Web pages. Therefore, the usage of metadata in Web pages is examined and the main research work from the Web community related to this issue is studied.

There is much work done in the field of research on the development and maintenance of metadata of digital Web resources (Ossenbruggen et al., 2004; Nack et al., 2005; Foulonneau and Riley, 2008). Greenberg et al. (2001) shows that non–professionals equal professionals in the creation of metadata for Web resources. However, Web content publishers do not pay attention to asssure proper description of the resources or even deliberately distort it (Golliher, 2008). For example, the content publishers keep on using metadata to try to gain visibility within search engines because the metadata contained within HTML Web pages used to be the base of ranking method. Today, this assumption is erroneous because search engines rank resources mainly with graph–based algorithms (Brin and Page, 1998).

The geospatial–based solutions that use Web resources (e.g. HTML Web pages) as part of the searchable content are mainly LBS systems, which are popular in mobile environments. They require generation of some descriptions of the resources and then, indexing them for their further retrieval, and in these terms, they are similar to Web search engines. Although the metadata associated with a Web resource may not be reliable, it may be still be used as the base for the automatic creation of metadata. The header section of an HTML document compliant with the HTML 4 W3C Recommendation (Raggett et al., 1999) may contain metadata via the *meta* elements (<META>), which contain a property–value pair, i.e. the *name* (the property name) or *http–equiv* (the value of header of the HTTP response) and the *content* (the property value). A schema attribute may be added to specify how to interpret the property value. These values can be eventually described via a metadata profile declared in the *head* element via a URI. For example, the Dublin Core Metadata Initiative (DCMI) (DCMI, 1995–2012) recommends a DC metadata profile (Johnston and Powell, 2008) that can be encoded using HTML elements and attributes. However, there is no specification that enumerates legal values of the *name* attribute. The mapping of the metadata used popularly in the Web to a target metadata model might be developed by analysing the W3C and WHATWG recommendations (Hick, 2011; Hickson, 2011) and lists of the meta elements frequently used in Web pages gathered by initiatives such as *Metatags.org* (Metatags Company Inc., 2012). As for metadata for describing the geographic scope of a Web resource, it can be represented as a disambiguated textual description of a location, a spatial object (e.g. a point or a bounding box) or both. Apart from *coverage* of the DCMI (and its extensions, e.g. *DC.coverage.x*) there are other geographic

Meta Element	Format	Note	Source
DC.coverage (.x/y/z/ placeName/ longitude/ latitude)	x/y/height/ placename/ longitud/ latitud	The coordinate system must be defined by the additional scheme attribute when x or y is used. The WGS 84 is default system for *latitude* and *longitude* (e.g., "World", "51.66, 6.88")	DCMI
geographic–coverage	place–class, lower–case/ code	Region definition (e.g., "city, Sao Paulo, Sao Paulo, Brazil")	WHATWG
ICBM	latitude, longitude	WGS 84, (e.g., "51.66,6.88")	GeoURL
geo.position	latitude, longitude	WGS 84, (e.g., "51.66;6.88")	GeoTags
geo.placename	free text placename	Placename (e.g., "Steinbergweg, 46514 Schermbeck, Germany")	GeoTags
geo.region	ISO 3166–2 code (ISO, 2007b)	Code of country subdivision (e.g., "DE–Nordrhein–Westfalen")	GeoTags

Table 4.1: Geospatial *meta* elements used in Web pages.

meta elements (i.e. geotags) that might be found within the *header* of a Web page. They have been proposed to support geographic search engines. For example, the GeoURL ICBM Address Server (Hansen, 2008) uses geotags to create a location–to–URL reverse directory for finding URLs by their proximity to a given location. Other example might be GeoSearch[1] which uses geotags for HTML resource discovery (Daviel, 2007). However, this approach seems not to be successfully adopted within the Web community, because the Internet Drafts (Daviel and Kaegi, 2007; Daviel et al., 2007) on the scheme proposed for embedding geographic information in HTML pages are obsolete. Nevertheless, the geotags are still used due to the popularity of online free generators (e.g. *geo–tag.de*[2], *MyGeoPosition.com*[3]), which help Web page publishers to add the geotags to a Web page. Table 4.1 summarises the geographic metadata used in Web pages (*gMeta*).

Some approaches have emerged in the context of the automatic metadata generation for Web resources, including the metadata collection (Tika Apache (ASF, 2011)), the content extraction (DC–dot (Medeiros, 2001)), the automatic classification and indexing (Data Fountains (Mitchell, 2006)), the text and data mining (Humphreys, 2002), the social tagging, and the metadata generation from contextual information provided by related or associated resources (Polfreman and Rajbhandari, 2008).

Geospatial information is hardly provided in Web page metadata, even within those published by the Geospatial Web community (see Section 4.4.3). In such case, a page has to be georeferenced. The georeferencing of documents is a task intensively studied in the context of the Geographic Information Retrieval (GIR) (Leidner, 2007; Jones and Purves, 2008) and Web search (Silva et al., 2006; Campelo

[1] http://geotags.com/frameset.html
[2] http://www.geo-tag.de/generator/en.html
[3] http://www.mygeoposition.com/

and Souza Baptista, 2009). A text may contain references to multiple locations (i.e., toponyms). The Named Entity Recognition (NER) tools apply natural language processing techniques to identify place names in a text. The results may contain false positives, i.e. words or phrases that are not toponyms in the context used within the analysed text. A geocoder georeferences a toponym and returns a ranked list of matching locations (Goldberg, 2008). Research on the toponym resolution focuses on georeferencing toponyms in a text (Zong et al., 2005; Jones and Purves, 2008). The effectiveness of this task depends on the reference dataset and the algorithm used. A place may have several names (e.g., endonyms and exonyms) that may change over time. Its footprint may also change over time. These changes can result in an incomplete datasets. The algorithm must take into account the ambiguity: (1) common words should be distinguished from proper names (geo ambiguity/ non–geo) (Amitay et al., 2004), and (2) the mapping between toponyms and locations can be ambiguous (e.g. there are about 40 inhabited places named "London" in the world). There is a variety of approaches, such as using other place names found in the text to improve the place name disambiguation (Overell and Rüger, 2008), and the usage of simple taxonomies based on gazetteers (Amitay et al., 2004) or more complex ontologies (Jones et al., 2001) which might be transformed to a graph for the computation of an "importance" score (Silva et al., 2006).

The next section presents the architecture proposed for metadata generation. In the context of this work, it has been necessary to develop a heuristic–based method for geographic scope estimation of Web pages.

4.3 Knowledge Generator

This section introduces an architecture for knowledge generation, i.e. generation, validation and/ or improvement of semantic description of retrievable resources. Although the scope of this work is limited to the generation of geographic metadata for Web pages, the architecture proposed can be extended in the future, in terms of functionality and resource type supported. Therefore, the following elements have to be considered: (1) the Web resource type which is analysed, (2) the metadata model read/ produced, and (3) the logic of metadata generation. First, an overview of a general functionality of a system is provided. Then, the architecture proposed for generating metadata for Web pages is outlined.

4.3.1 Workflow

Figure 4.1 presents the high–level workflow performed by the architecture proposed. The system receives some resource metadata (*InMeta*) which enclose information on how to reach the Web resource (*WR*), for example via a `URI`. A metadata model ($Model_{MD}$) is initialised and characteristics of the *WR* are analysed. According to the resource characteristics (for example, the identified Media Type) several metadata extractors are called. Each extractor is specialised in extracting some pieces

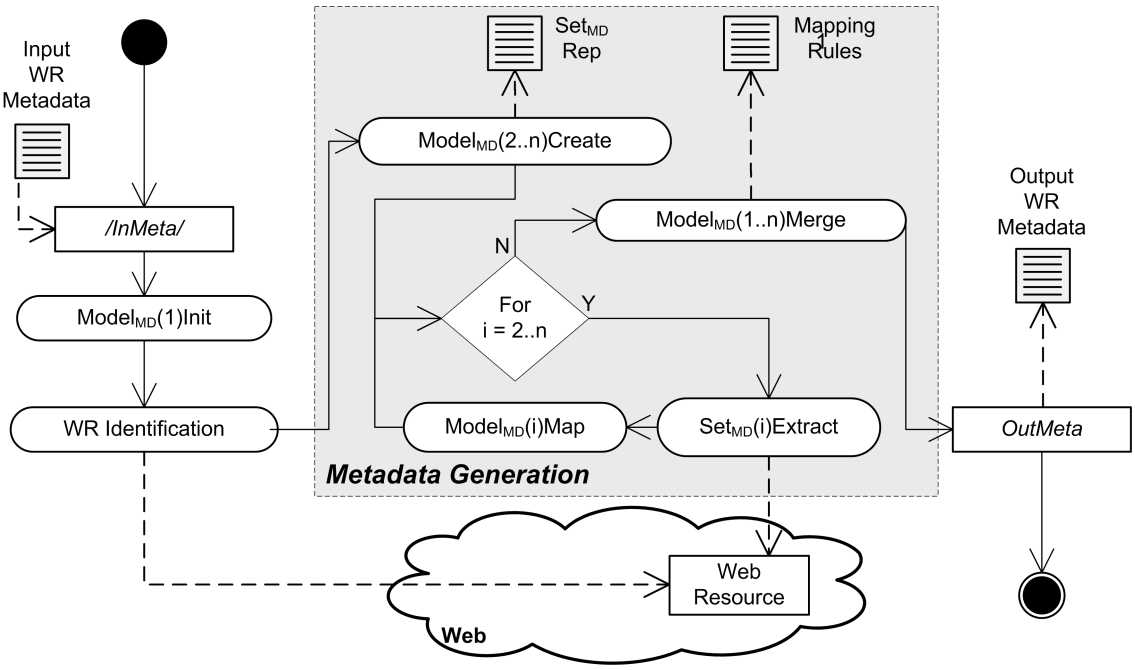

Figure 4.1: Overview of the system functionality.

of information (i.e. metadata set, Set_{MD}) from the *WR*. Then, a mapping is applied to transform the Set_{MD} into a $Model_{MD}$. In the end, there is a $Model_{MD}$ for each extractor that has been run ($Model_{MD}(2..n)$). In this way, the $Model_{MD}$ set is composed. Finally, the set is merged. The system generates the resource metadata (*OutMeta*) which model is extended with additional information to include some provenance information.

In this work, some mapping rules are provided to control the metadata transformation and the final merging. In the future, the transformation might be done by applying model–transformation tools instead of the simple mappings used in this work. For example, the importance of one element can depend on the extractor type or existence/ lack of other elements. The creation of final metadata could be more sophisticated as well in future, for example, it may include some recursive calls for the metadata generation of the Web resources which are linked from the resource that is being analysed. This approach also enables adding validation information in the future.

4.3.2 Architecture

The proposed architecture for knowledge generation that realises the previously outlined workflow is presented in Figure 4.2. The *InputDoc* contains the resource metadata. The metadata are encoded in a way that it enables automatic metadata processing. In this work, the system receives an `XML` metadata document (*InputDoc*) with declared Schema (however, it might be an `RDF`/`XML` as

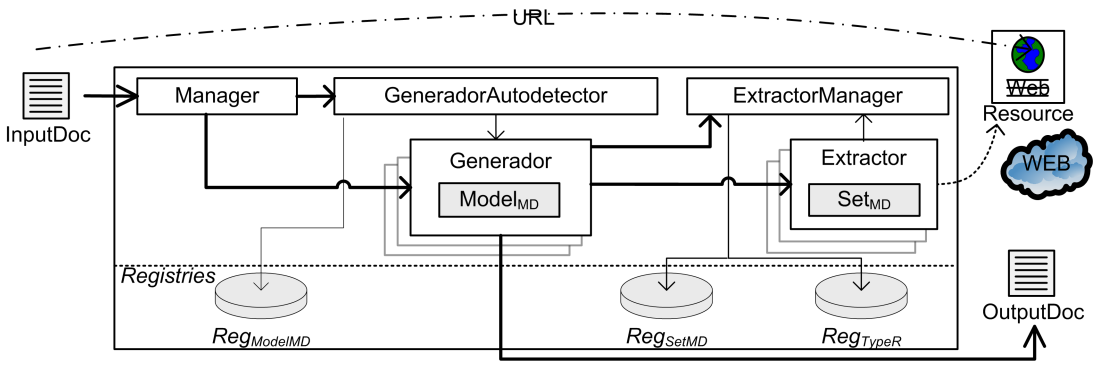

Figure 4.2: Overview of the system architecture.

well). The schema allows the metadata model ($Model_{MD}$) to be identified by the system. All models which can be processed by the system should be registered a priori in the $Reg_{ModelMD}$. Additionally, the supported types of Web resource and metadata sets are registered in the Reg_{TypeR} and the Reg_{SetMD}, respectively. The *Manager* calls the *GeneratorAutodetector*, and it returns an appropriate *Generator* according to the $Model_{MD}$ and the *Web Resource* type. Then, the *Generator* calls the *ExtractorManager*, which initialises the required *Extractors* due to detected Media Type and desired functionality. An *Extractor* gathers the Set_{MD} via simple parsing of the Web resource (e.g. a list of links within body, metadata from the *header* or the content of the *title* in the case of a Web page), or it can apply other extractors. The *Generator* can implement a variety of logics and it is responsible for the model transformation and final metadata generation. In the end, the resultant metadata are returned (*OutputDoc*).

As the work is dedicated to the geospatial domain, the system works with a metadata model that holds geographic information. Several extractors capable of identifying geographic information are necessary, for example, those which can extract the *gMeta*. A coverage estimation method has been proposed in Section 4.4.1. It has been provided as an extractor, the *CoverageExtractor*. This compound extractor calls other *Extractors* to get the L_{ner} *lists*. The prototype description includes implementation details and presents the developed components.

4.4 Applying the Knowledge Generator

Web pages have not caught the attention of the geospatial community, except for LBS systems that use Web technologies. However, existing approaches from the Web community allow development of a repository of Web resources that might be seen as alternatives sources of geospatial information. Crawler–based search engines proved to be successful in the dynamic Web community. Therefore, an architecture which enables the Knowledge Generator workflow for creation of the semantic description of Web pages has been proposed. The prototype developed is dedicated to the metadata

generation of Web pages for their deployment in a geospatial catalogue. The experiments show that Web pages, even these from the geospatial domain, lack geographic metadata. Therefore, one of the important issues of the prototype developed has been a method for the coverage estimation.

4.4.1 Method for coverage estimation

The goal of the coverage estimation method is assigning the MBBOX to a Web page. First, the heuristics used are described, and then, some details are presented. Finally, some disadvantages of the method developed are discussed.

The coverage estimation method consists in two heuristics: a content–based heuristic (H_3) and a host–based heuristic (H_{hip}). The heuristic named H_3 estimates the coverage by analysing geographic information found within different elements of Web pages (mainly the geocoded place names). The heuristic named H_{hip} is used when H_3 has not been successful. H_{hip} infers a country code (ISO 3166–1 alpha–2 codes (ISO, 2007a)) from the host (i.e. host name or IP), and then the code is geocoded to the MBBOX. Apart from MBBOX, the final coverage estimation method ($H_3 + H_{hip}$) returns a textual representation of the geographic scope, a *code* and some provenance information (see Section 4.4.2).

Figure 4.3 shows an overview of the coverage estimation method giving emphasis to the *code* attribute. It can be observed that the final value of the *code* might be `POINT` or `ESTIMATED`. First, the content–based heuristic tries to identify the *gMeta* within header metadata that provide latitude and longitude. However, this information is hardly provided (see Section 4.4.3). Therefore, the content–based heuristic focuses on the toponyms found within the Web page. The task for the coverage estimation from a text is comprised of three general steps:

1. *Toponym recognition.* This step produces a candidate place names list (L_{ner}).

2. *Toponym resolution.* This step identifies the geographic entity ($entity_g$, an element of a simple territorial ontology) to which refer each place name in the L_{ner}, and it produces a set of geographic entities (L_{ge}).

3. *Geo–scope estimation.* This step tries to estimate the MBBOX that best represents the extracted set of geographic entities.

Here, the task of the estimation of the representative geographic entity from a set of toponyms found in a Web page is called $Entity_gEstimation$. Two external tools are used in this task, a NER tool and a geocoder tool. The first one is used to create a L_{ner} *list*. The developed heuristic treats separately place names recognised in different elements of the Web page separately. According to the processed element, the following L_{ner} *lists* can be created:

1. Pn_{gMeta}, that is a L_{ner} of *gMeta* identified within the *header* element of Web page,

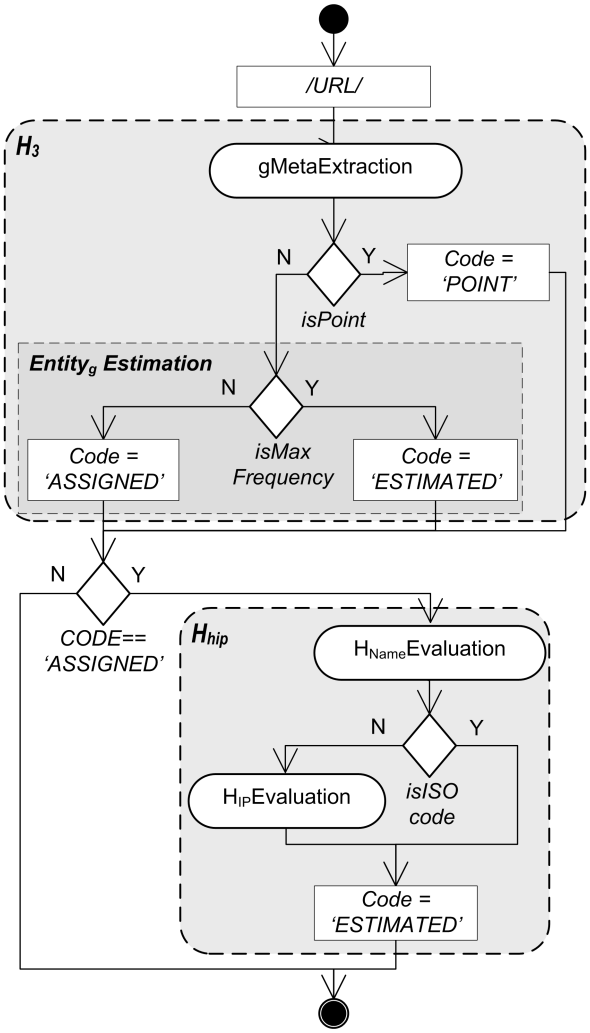

Figure 4.3: Overview of the coverage estimation method.

2. Pn_{meta}, that is a L_{ner} extracted from the *header* element (other than *gMeta*) and *title* element of Web page,

3. Pn_{body}, that is a L_{ner} of the Web page *body*, i.e. the visible text (including links) and the invisible tags of images.

The geocoder module is used to create the L_{ge} from a L_{ner}. The geocoder produces a ranked list of geographic entity proposals for each item of the L_{ner}. The geographic entity returned is encoded in an XML and the data model used allows the identification of the related concept from a territorial ontology. A simple territorial ontology has been used in this work, which is result of the analysis of three existing standard models: the FIPS 10–4 standard for countries, dependencies, areas of

special sovereignty and their principal administrative divisions developed by the United States Federal Government (National Institute of Standards and Technology, 1995); the ISO 3166 Codes for representation of names of countries and their subdivisions (ISO, 2007b); and the Nomenclature of Territorial Units for Statistics (NUTS) developed by the EU (EC, 2003). In this simple ontology, geographic entities are the concepts, and the only relationships of interest are the spatial aggregations, i.e., *has–part* or *part–of*. It is a modification of the Administrative Unit domain ontology proposed in López-Pellicer et al. (2008). Additionally, natural phenomena and towns have been considered as well. The resultant ontology gathers geographic entities of the following types:

1. *Feature* (FT) that represents a natural phenomenon, for example "Danube" (river) and "Alps" (mountains range),
2. *Earth region* (ERT) that defines international organisations, for example "European Union" and "United Nations",
3. *Country* (CT) that represents countries in the world,
4. *Region* (RT) that represents the top level administrative divisions of a country,
5. *Sub–region* (SRT) that represents the administrative divisions of a country lower than the top ones,
6. *Town* (TT) that refers to cities.

For example, in case of "Barcelona" toponym, the expected $entity_g$ is the "Barcelona" (TT) in "province of Barcelona" (SRT) in "Catalonia" (RT) of "Spain" (CT). The ERT entities are related to countries they gather (*has–part*), and FT entities are related to countries they belong to (*part–of*). The L_{ge} is created by assigning to each item in the L_{ner} the first $entity_g$ from the ranked list. The *geo–scope estimation procedure* uses a L_{ge} to calculate frequencies of the geographic entities for different levels of accuracy in the following order: TT, SRT, RT, CT, FT, ERT and EARTH. Each L_{ge} *item* is represented via the $entity_g$ to which it is related at the accuracy level that is being calculated (e.g. "Barcelona" (TT) will be represented by "Catalonia" at RT level of accuracy). The method returns an $entity_g$ of maximum frequency and the ESTIMATED *code*. If the method could not have estimated the coverage (e.g. it fails if the L_{ge} is empty), the "Global" $entity_g$ and the ASSIGNED *code* are returned.

The final heuristic ($H_3 + H_{hip}$) is performed as follows. First, the content–based heuristic is run (see Figure 4.4). The *gMeta* are checked and if the *gMeta* provide a point, it is used to create the MBBOX and the result *code* has POINT value. If no spatial object has been distinguished, the text values are analysed to create Pn_{gMeta} and then the corresponding L_{ge}. If the *geo–scope estimation procedure* fails (*code* has ASSIGNED value), a weighted list is created by joining the Pn_{gMeta} and the Pn_{meta} ($w_2(Pn_{gMeta}), w_1(Pn_{meta})$, where $w_i = i$). If the *geo–scope estimation procedure* fails

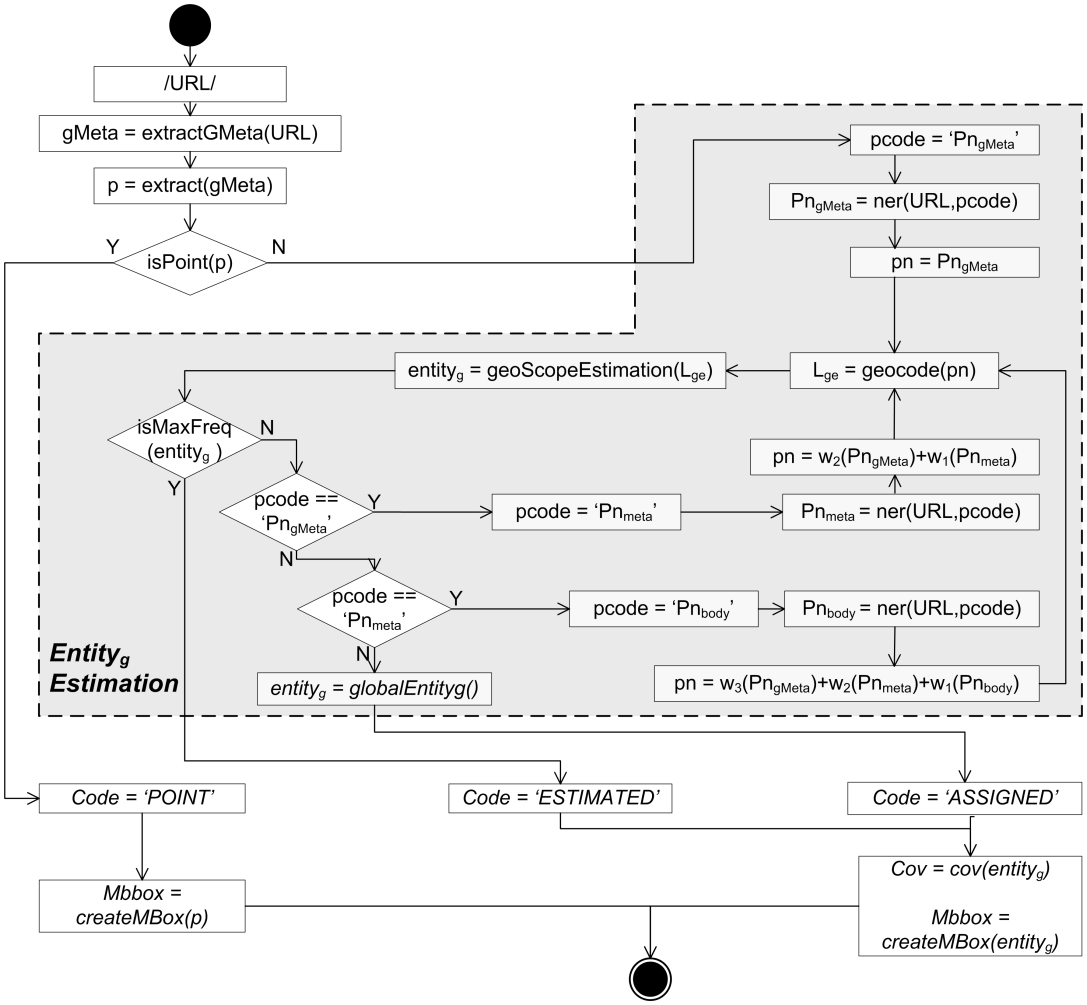

Figure 4.4: Overview of the content–based heuristic.

again, the Pn_{body} is added, new weights are assigned ($w_3(Pn_{gMeta}), w_2(Pn_{meta}), w_1(Pn_{body})$, where $w_i = i$) and the *geo–scope estimation procedure* is run again. The host–based heuristic is used only when the heuristic H_3 fails to estimate the coverage (i.e. it returns `ASSIGNED` *code*), which happens usually due to the lack of metadata and poor NER results. The heuristic H_{hip} tries to extract the ISO country code from host name of the analysed Web page (H_{Name}) and if it is not successful, its IP is georeferenced to an ISO country code (H_{IP}). Then, the ISO code is geocoded to a MBBOX and `ESTIMATED` *code* are returned. Table 4.2 shows some examples of Web pages whose coverage has been estimated by the host–based heuristic.

The developed content–based heuristic is simple and has several problems. First, the candidate place names are trimmed from their context when using the geocoder. For example, it does not consider other place names from the same L_{ner}, which have been identified near the searched place

URL	**Manual estimation**	H_3 **Code**	H_{Name} (H_{hip})	H_{IP} (H_{hip})	H_{hip}	$H_3 + H_{hip}$ **Code**
bnhelp.cz	*CZ*	ASSIGNED	CZ	–	CZ	ESTIMATED
b5m.gipuzkoa .net	*Gipuzkoa, Basque Country, ES*	ASSIGNED	–	ES	ES	ESTIMATED

Table 4.2: Example of the H_{hip} heuristic results.

name within the text. The procedure for the creation of L_{ge} delegates the ranking to the geocoder as well. The algorithm that creates L_{ge} could consider, for example, the re–ranking of geocoding list according to other items within the L_{ge}. The results of the experiments performed shows that this straightforward approach can be satisfactory in the context of this work.

4.4.2 Prototype

This section presents the prototype implementation and describes the experiments performed and their results. First, the supported metadata model is introduced with some details on mappings applied. Then, the main characteristics of the implemented prototype are presented. Next, the corpus used and the performed experiments are described, and then the results are discussed.

The implemented prototype is dedicated to the generation of geographic metadata of Web pages of Geospatial Web resource providers. The generated metadata will be deployed in an OGC catalogue service. Therefore, the metadata model consists of the core returnable properties supported by the OGC catalogue services (Nebert et al., 2007). Table 4.3 presents a metadata model and the main mappings applied by the system. The system receives an XML document whose schema conforms to this metadata model. The source element contains the URL of the Web page for which the metadata should be generated. The *modified* (i.e. the metadata creation date), *identifier* (i.e. a unique identifier of the metadata) and *type* (i.e. "Web page of a Geospatial Web resource provider." by default) are filled first. Additionally, the HTML *title* element is mapped to *title* and a procedure to extract copyright information from *body* element is implemented as well.

The prototype has been implemented in Java. It supports redirection, i.e. the HTTP redirections and a simple JavaScript declaration to follow a link. The Tika Apache project has been selected as the base for the implementation of the architecture. A generator, which works with the metadata model described above, uses several extractors (e.g. the *gMeta* extractor and metadata extractor which work on the *header* element, the *title* element extractor, coverage estimator). The extractors that extract the L_{ner} lists use the Stanford NER (Finkel et al., 2005). The NER tool is configured in function of the Web page language. Various classifiers have been tested (Manning and Klein, 2003; Faruqui and Padó, 2010). Although the NER tool used in this prototype only handles English or German text properly, it covers 58.3% of the analysed corpus as showed later. By default, the

CSW Record Model	Min..Max	HTML Element Mapping
contributor	0..*	DC.contributor
coverage	1..*	DC.coverage, coverage, ICBM, geo.position, geo.placename, geo.region, geographic–coverage, DC.coverage.x/ DC.coverage.y, DC.coverage.longitude/ DC.coverage.latitude
creator	0..*	DC.creator, author,
	0..*	webauthor
modified	1	–
description	0..*	DC.description, description
format	0..*	DC.format, content–type (*http–equiv*)
identifier	1	-
language	0..*	DC.language, language
publisher	0..*	DC.publisher, publisher
relation	0..*	DC.relation
rights	0..*	DC.rights, rights, copyrights
source	1	–
subject	0..*	DC.subject, keywords
title	0..*	DC.title, application–name (*http–equiv*)
type	1	–

Table 4.3: Metadata model and mapping to the HTML *meta* elements.

classifier trained with an English language corpus is selected. The coverage estimator implements the method presented in Section 4.4.1. A geocoding module has been implemented to support the required functionality. It uses the Google Maps API as an external geocoder, which has been chosen due to its global coverage, a rich data model (i.e. it includes administrative divisions up to CT), a multilanguage support and a relevance ranking. It was necessary to provide a geographic ontology with parent–child relations (*part–of* and *has–part*) for the other types of geographic features, i.e. feature and earth region. Additionally, an extension for the ISO codes of countries and their subdivisions (ISO, 2007b) has been added because the external geocoder does not support them adequately. Appendix A contains some details of the implemented prototype.

Listing 4.1 shows an example of the metadata generated by the prototype. It might be observed that apart from the *dc:coverage* element some provenance information is offered as well, i.e. the coverage code (*gse:code*), the coverage in textual format (*gse:coverage*), the extracted *gMeta* (*gse:covmeta*) and the heuristic code (*gse:hcode*) that informs which heuristic has produced the resultant coverage. The provenance information on metadata generation process is not required by the OGC CSW specification, however it may help in the evaluation of the accuracy of the generated metadata later.

4.4.3 Experiment

Some experiments have been run to evaluate the functionality of the implemented prototype. The corpus used permits the examination of the system in the conditions of the actual Web environment.

Corpus

The corpus is a realistic set of resources retrieved from Geospatial Web resources publishers. A set of OWS URLs returned by an OWS crawler (López-Pellicer et al., 2011e) was used to identify the publishers (each URL was trimmed to its host). In this work, it is assumed that it is highly probably that the Web pages reached via this list are pages published by the service publishers. In other words, it is assumed that OWSs are related in some way with the Web pages served by the same host.

Listing 4.1: Metadata generated when applying the coverage estimation method.

```
<Record>
        <dc:coverage>51.545027, -0.056262, 51.545027, -0.056262</dc:coverage>
        <dc:creator>1</dc:creator>
        <dc:creator>London Borough of Hackney</dc:creator>
        <dct:modified>2011-12-14T22:38:52Z</dct:modified>
        <dct:abstract>This site has been created by the London Borough of
            Hackney. The site provides access to information and online requests
             about the priority services delivered by the council and its
            partners.</dct:abstract>
        <dc:format>text/html; charset=iso-8859-1</dc:format>
        <dc:language>eng</dc:language>
        <dc:publisher>London Borough of Hackney, Town Hall, Hackney, London E8 1
            EA, Tel 020 8 356 5000, http://www.hackney.gov.uk</dc:publisher>
        <dc:source>http://www.map.hackney.gov.uk/LBHackneymap/</dc:source>
        <dc:subject>Hackney Map, Hackney, Hackney Council, Borough of Hackney,
            Hackney Where, HackneyWhere, Property, Map, LLPG, UPRN</dc:subject>
        <dc:rights>Copyright  London Borough of Hackney, Town Hall, Hackney,
            London E8 1EA, Tel 020 8 356 5000, http://www.hackney.gov.uk [Or
            this could be a link to a copyright declaration page]</dc:rights>
        <dc:title>Map.Hackney 2.0</dc:title>
        <dc:type>Geospatial Web Resource Provider</dc:type>
        <dc:identifier>beb14402-0018-4d20-997b-ba7e5176a019</dc:identifier>
<!-- Provenance Information -->
        <gse:coverage></gse:coverage>
        <gse:code>POINT</gse:code>
        <gse:hcode>H3</gse:hcode>
        <gse:covmeta>London Borough of Hackney, London, UK, Global</gse:covmeta>
        <gse:covmeta>51.545027, -0.056262</gse:covmeta>
        <gse:covmeta>Hackney</gse:covmeta>
        <gse:covmeta>GB-HCK</gse:covmeta>
        <gse:covmeta>51.545027;-0.056262</gse:covmeta>
</Record>
```

Type	Code	%	Note
geoportal	Gp	48.09%	Geoportal main pages
portal	P	13.45%	Portal main pages
resource	P/Gp	2.36%	A logical part of a portal dedicated to geographic information
map visor	AV	12.73%	Map visor based Web pages or even geoportals
other	S	23.46%	Web pages which usually provide demo services or it was difficult to analyse them. They include:
		6.18%	Companies Web pages.
		4.36%	Research group Web pages.
		4%	Entrance point, i. e. pages that require some kind of interaction with user to proceed (e.g. logging or language selection),
		2%	Personal Web pages
		6.73%	Other Web pages (mainly community or software pages).

Table 4.4: Classification of Web pages in the corpus according to Web site characteristics.

Type	Code	%	Coverage
local	L	49%	Coverage refers to a part of a country
national	N	30.97%	Coverage of a country
regional	R	2.91%	Coverage crosses country boundaries
global	G	10.38%	Coverage of the Earth
out–of–Earth	O	0.36%	Coverage does not refer to the Earth (two examples)
notKnown	NN	6.38%	It was impossible to determine the coverage manually (mainly pages classified as *S*)

Table 4.5: Classification of Web pages in the corpus according to the coverage estimated.

The OWS host list (1122 elements) was analysed manually in October 2011 using Chrome browser (version 14). More than half of them (51.1%) were not considered due to some errors (e.g. duplication, connection and page loading errors) or did not provide information that might be processed (e.g. "Under construction", an empty page, a server test page). The rest of the OWS hosts (549 elements) were analysed in order to identify the geographic scope and the language.

Table 4.4 shows classification of the pages according to their origins. It can be observed that 63.9% of Web pages are portal (or geoportal) pages. Some Web sites of map visor–based sites can be classified as geoportals, however, they are treated separately due to the technology used that makes it difficult to analyse their content automatically.

The manual estimation of coverage treats a Web site as a whole and considers its published geospatial resources in the estimation. Table 4.5 shows the classification of the corpus according to the coverage estimated.

Most pages are in English (43.9%), German (14.4%), Spanish (12%), Polish and Italian (about 4% each), and Czech, French and Catalan (about 3% each). The rest of the examined pages are

Metadata element	% filled
coverage	3.21%
title	97.54%
subject	43.67%
description	42.72%
creator	25.71%
contributor	1.32%
publisher	7.56%
rights	10.96%
format	80.91%
language	34.22%

Table 4.6: Summary of metadata extraction.

mainly in one of the official languages of Europe but there are examples of Web pages in languages of Asia (e.g. Thai or Chinese). Additionally, there are six pages whose content is in two languages (e.g. atlastenerife.es).

The corpus analysis indicates that 87.7% of the sites of the *regional* or *global* geographic scope (14 and 50 pages respectively) use the English language. This feature is in accordance with common sense when targeting an international audience, a language used at global–level is preferable (or at least language used in the geographic scope, for example, Spanish language seems to be adequate to the region of Latin America). On the other hand, the internationalisation of Web pages (i.e. availability in other languages) has not been evaluated. Therefore it should be understood that the language attribute of a page refers to the language in which the Web page is offered by default.

Experiments and result

The experiment consisted in generating the geographic metadata for each Web page from the corpus. Due to the dynamic characteristics of the Web (temporal unavailability of the Web resources), several test runs have been performed during the period of November and December of 2011. In general, 3.1% of the elements of the corpus were not processed due to some errors (e.g. data format errors, frequently repeated connexion problem). Table 4.6 summarises the extracted metadata by applying the defined mapping. The generated metadata will be exploited in a catalogue. Therefore, the faulty metadata have been filtered out (i.e. metadata without title, description or subject fields). After removing metadata that do not conform to this restriction (2.4%), all remaining elements do have at least title, description and subject fields. It can be observed that the information about the geographic scope (coverage) is rare in the examined corpus, and it varies in format (i.e. textual information in different format, lat/ long point). Therefore, the next experiment consisted in applying the coverage estimator to obtain a geographic scope (i.e. MBBOX). The Web pages for which it was not possible to estimate coverage manually (6.4%) have been removed from the corpus as well. Table 4.7 summarises the percentage of removed elements of corpus.

The results of the coverage estimation experiment (see Table 4.8) shows that H_3 produces *equal*

Lang	**Total Analysed**	**Process error**	**Not valuable metadata**	**Coverage NN**	**Processed**
EN	241	11	6	29	**195**
DE	79	3	2	1	**73**
Other	229	3	5	5	**216**
Total	549	17	13	35	**484**
	(100%)	(3.10%)	(2.37%)	(6.38%)	**(88.16%)**

Table 4.7: Trimmed corpus (Lang–language, NN – coverage not estimated).

results to the manual estimation of the geographic scope of Web pages in almost half of cases. 73.4% of results are correct with the country accuracy or better (i. e. the country of the computed $entity_g$ is equal to the country of the manually estimated $entity_g$). In other words, the H_3 procedure yields *acceptable* results in 73%, and the erroneous results in 26.6%.

After applying the H_{hip} the coverage estimator produced acceptable results in 78.9%. The result is poor due to several problems. The NER tool is not properly configured for almost half of the corpus. Therefore, the H_3 produces poor results for them. Surprisingly, the H_3 produces a similar percentage of errors for the EN corpus. Closer analysis of the corpus and the H_3 results have shown that H_3 behaves better for pages classified as *national* or *local* than for those classified as *global*, *regional* or *out–of–Earth*. Since 87.7% pages of those classified as *global*, *regional* or *out–of–Earth* consist of Web pages in English, therefore, the results are worse than expected for the EN corpus. An improvement of the *equal* results after applying the H_{hip} should not be expected because most of the elements of the corpus are classified as *local* and the H_{hip} handles only the country level coverage or higher. Nevertheless, the result shows in fact an improvement, i.e. the number of errors decreases and the percentage of the acceptable results increases. This tendency is not shown in the EN language part of the corpus. After meticulous analysis of the results it has been observed that this effect is produced by the fact that 72.3% of the elements evaluated by the H_{hip} do not permit the estimation of the ISO code from the host name. In such a case, the H_{hip} georeferences IP and introduces an error. The coverage estimator is quite good for the DE language part of the corpus. Therefore, it might be suspected that if the NER tool is properly configured, the coverage estimator is efficient (for Web pages classified as local and national) for the DE language part of the corpus at least. Nevertheless, the estimator should be improved for the Web pages classified as global, regional and other.

4.4.4 Evaluation of the content–based heuristic

As mentioned before, the functionality of NER tool and its configuration influence the results of the coverage estimation method. In this work, the method uses a geocoder module to geocode the NER results. It removes items that do not match any geographic entity of a valuable type (e.g., a street should be removed because it is not considered). In practice, a NER item can be:

	EN	**DE**	**Other**	***Total***
Total	195	73	216	484
H_3	**137**	**45**	**89**	**55.99%**
(% of Total)	(70.26%)	(61.64%)	(41.20%)	
H_3 Acceptable	97	41	61	73.43%
(% of H_3)	(70.80%)	(91.11%)	(68.54%)	
H_3 *Equal*	*66*	*32*	*33*	*48.34%*
(% of H_3)	*(48.18%)*	*(71.11%)*	*(37.08%)*	
H_3 Error	40	4	28	26.57%
(% of H_3)	(29.20%)	(8.89%)	(31.46%)	
$H_3 + H_{hip}$	**195**	**73**	**216**	**100%**
(% of Total)	(100%)	(100%)	(100%)	
$H_3 + H_{hip}$ Acceptable	134	67	181	78.93%
(% of Total)	(68.72%)	(91.78%)	(83.80%)	
$H_3 + H_{hip}$ *Equal*	*81*	*39*	*81*	*41.53%*
(% of Total)	*(41.54%)*	*(53.43%)*	*(37.50%)*	
$H_3 + H_{hip}$ Error	61	6	35	21.07%
(% of Total)	(31.28%)	(8.22%)	(16.20%)	

Table 4.8: Results of the experiment on coverage estimation.

- *Error.* It is a false positive that is not a place name at all (i.e., there is no place of such a name). The NLP techniques analyses language structure and may produce some errors. These values are usually geocoded to a street and, in this way, can be removed.

- *Mistake.* It is a false positive which is recognised as a toponym due to ambiguity problem (i.e., it is a toponym in another context). The elimination of these false positives depends on the type of geocoded entity.

- *Hit.* It is a positive result which is a toponym in the used context. Such toponym might be eliminated or eventually geocoded wrongly, if its type is ignored by the system (e.g. it is a street name). In the case of toponym elimination, it is an expected behaviour. However, the geocoding errors are not desired.

The result of the method can also be affected by the relevance ranking of the geocoder applied. If a text is geocoded successfully, it produces a ranked list of geocoded toponyms. In the case of the geocoder used in this work, the list size depends on ambiguity of the toponym, and it varies between 1 and 15 items usually. The first item, whose type is valuable, is returned. The geocoder may introduce false positives, for example, a street name can be geocoded to a town. Additionally, any errors and lacks in a reference dataset of the used geocoder may influence the final results.

The efficiency of a GIR system is usually measured via *precision* and *recall*. In the case of a NER tool, for example, the first one is measured as the number of toponyms returned by the NER tool divided by the total number of items in the NER list. As for the *recall*, it is measured as the number of toponyms returned by the NER tool divided by the total number of toponyms within the

Text Language	**NER Tool Configuration**	P_{G-R}	P_{CEP}
EN	EN	86.10%	78.31%
DE	DE	74.49%	69.78%
ES	EN	81.52%	67.20%
SE	EN	63.64%	66.87%
PL	EN	90.91%	72.73%
DE	EN	50.90%	48.34%

Table 4.9: Evaluation of the coverage estimation method.

analysed text. In this work the precision of the coverage estimation method (P_{CEP}) is evaluated. The *geocoder–relative precision* (*G–R precision*, P_{G-R}), a modified precision metric, is measured as the number of items of the NER list that have been geocoded successfully divided by the total size of the list, i.e.:

$$P_{G-R} = \frac{G_T(L_{ner})}{S(L_{ner})}$$

$G_T(list)$ is a geocoding function that considers only the $entity_g$ of type T;
$T \in (ET, FT, ERT, CT, RT, SRT, TT)$;
$S(list)$ is a function to calculate the size of a $list$.

This metric is adjusted to the proposed approach and it indicates the procentage of the *NER list* that will be used in the coverage estimation method. Although it may introduce some errors due to ambiguity problems of toponyms and the ranking applied, it can be calculated automatically for any text. Evaluation of the system precision requires manual analysis of the parsed text. Therefore, it has been done for a sample of documents from the corpus. Each toponym generated by the method has been validated with text in order to check if it has been geocoded to the right entity.

In general, the better NER and its configuration, the better functionality of the coverage estimator will be achieved. As expected, the improvement of NER tool functionality is observed when using the language dedicated classifier. Table 4.9 presents the experiment results performed over the corpus when considering text language and the configuration of the Stanford NER tool (i.e. language dedicated classifiers). It can be observed that the NER results are better when it is configured adequately to the German language. Additionally, the NER trained for English language performs with similar *G–R precision* for English and Spanish text. For example, in the case of the Web pages from the Polish language corpus, surprisingly the value of P_{G-R} is even higher than for English (90.9%). The P_{CEP} decreases in almost all cases but it maintains around 70% (except EN language configuration for German text).

4.4.5 Improvement discussion

First of all, it is important to stress that the developed coverage estimation method works efficiently for *national* and *local* Websites, and these sites state 79.97% of the corpus used. Since the content–based heuristic fails mainly for *global* and *regional* sites, this aspect should be considered in the first place. The results of the corpus analysis indicate that the sites of the *regional* or *global* geographic scope usually use the English language (almost 90%). Therefore, the language of the text could be the first important hint. An additional analysis of the Web pages of geoportals has been performed to provide deterministic features of Web pages which can be used to enhance the coverage estimation results. In the case of *global* geoportals, the main source of the toponyms is the visible text of the *body* element without links, and the majority of toponyms are names of countries which are spread around the World. As for *regional* geoportals, the visible text of the *body* is the principal source of toponyms and (almost) all countries belong to the same geographic region. The *national* geoportals characterise the lack of toponyms in general, and if any appears it is the country name or its capital. In the case of the *local* geoportals, the *body* element (including the invisible text of the images, i.e. the value of the *alt* attribute) is the source of valuable toponyms, and the toponym of the highest frequency represents the administrative or geographic region of country. The names of countries hardly appear, and if they appear, one of them is the referred country. These hints can be used to enhance the current content–based heuristic.

4.5 Summary

This Chapter presents an architecture of the Knowledge Generator, a system for automatic generation of metadata for Web resources. In the context of this work, capturing geographic scope of Web resources is the centre of interest. The prototype developed is able to generate automatically geographic metadata for Web pages. This tool has been used to generate metadata for Web pages that belong to the Websites that publish Geospatial Web resources. The empirical study shows that straightforward heuristics for geographic coverage estimation can automatically supply this information when a publisher does not provide it. In general, the coverage estimator method produces acceptable results in almost 80% of cases, and its precision might be even higher for non–English Web sites if the system is configured appropriately (e.g. 91.8% for German Web sites).

The empirical study provides a brief overview of characteristics of Geospatial Web publishers as well. Geospatial Web resources can be found mainly in geoportals and general portals (63.9%) but also in the Websites of companies, research group, communities and personal Web pages (frequently as demo resources). Most of the Websites have been classified as *local* (49%) or *national* (31%). The Web sites of *regional* and *global* geographic scope usually use the English language (i. e. 87.7%). Also, this study uncovers current practices in the geospatial community in providing metadata for Web pages, the lack of geographic metadata in particular.

Chapter 5

Geospatial Web Search Engine

5.1 Introduction

This Chapter presents a Geospatial Web Search Engine and Web search client that support non–expert users in searching for Geospatial Web resources. In the context of this work, searching for Geospatial Web resources on the Web is equivalent to search for domain–specific Web resources. Therefore the approaches to search for such Web resources are examined (Section 5.2). Section 5.3 introduces the user search goals and search strategies which might be developed when searching for Geospatial Web resources by means of a general search engine. Then, Section 5.4 discusses faceted search and browsing as support for non–expert users. Section 5.5 presents some relevant examples of existing searching applications. The important issues of the proposed system and its architecture are presented in Section 5.6. Then, the results of evaluation studies are presented (Section 5.7). Finally, some conclusions are outlined (Section 5.8).

5.2 Domain searches on the Web

In the context of this work, searching for Geospatial Web resources on the Web is equivalent to search for domain–specific Web resources. More precisely, it refers to search for Web services.

5.2.1 Searching for Web services

Research work on service discovery is focused on a variety of aspects of service discovery and selection, such as semantic matching and automatic composition (Bussler, 2002; ShaikhAli et al., 2003; Mandell and McIlraith, 2003; Benatallah et al., 2003; Sycara et al., 2003; Kwon, 2003; Verma et al., 2004; Broens, 2004; Gooneratne and Tari, 2008), semantic description (Akkiraju et al., 2005; Martin et al., 2007a), semantic annotations (Martin et al., 2007b; Vitvar et al., 2008; Talantikite

et al., 2009) and ontology–based discovery and selection (Keller et al., 2005; Sriharee, 2006; Stollberg and Norton, 2007; Kopecký and Simperl, 2008; Chitra et al., 2010); requirements matching (Hausmann et al., 2004) and contract–based discovery (Luca and Padovani, 2010); QoS–based discovery, selection (Wu and Wu, 2010) and composition (Wang et al., 2010b) personalised search and selection (Wolf-Tilo Balke, 2003; Balke and Wagner, 2004), context–aware (Doulkeridis et al., 2006; Dietze et al., 2008) and recommendation–based discovery (Kokash et al., 2007; Chukmol, 2008); ranking (Palmonari et al., 2009; Hao et al., 2010) and classification (Wang et al., 2010a) algorithms; IR–based discovery (Chen and Wu, 2011), or description anti–patterns (Rodriguez et al., 2010). All these works on service discovery are dedicated to a service infrastructure based on one common repository. This approach has been motivated by the fact that in the beginnings of the Web Service era, Universal Description, Discovery and Integration (UDDI) specification (Clement et al., 2004) was proposed as a solution to publish and search services. As accurately pointed by Sreenath and Singh, in such a situation *"the key challenge is not discovery but selection: ultimately, the service user must select one good provider."* (Sreenath and Singh, 2004). In practise, the decentralised character of Web service discovery should be considered. Garofalakis et al. (2006) examines two different approaches to service discovery mechanism, (1) the traditional catalogue–based (i.e. UDDI), and (2) the decentralised discovery, i.e. peer-to–peer (P2P). Other work dedicated to Web service discovery in P2P networks examines requirements for semantic–enablement (Antonellis et al., 2006) and another one presents empirical evaluation of P2P infrastructures for large scale Web service discovery (Sioutas et al., 2009). Even these works, however, follow the SOA paradigm, i.e. they assume the existence of registries/catalogues as the linking points between providers and users.

In practice, the distributional and liberated characteristics of the Web have eclipsed the SOA approach out of the business domain. This can be observed by the fact that the UDDI standard has not prevailed in the domain of publicly available Web Services. The last UDDI version 3.0.2 dates from 2004 (Clement et al., 2004) (approved in 2005[1]) and the OASIS UDDI Specification Technical Committee that defined it has been disbanded in late 2007 (Clark, 2008). Microsoft, IBM and SAP shut down their public UDDI (i.e. UDDI Business Registry (UBR)) back in 2006[2] as well.

Today, from the provider perspective, Web services are often registered on specialised portals (not necessarily UDDI–based) following SOA paradigm or are simply put on the Web together with some Web pages describing the features of the service. This leads to two main ways in which services are searched today by a user:

- *Searching over the specialised portals.* For example, *XMethods*[3] offer a list of publicly available Web services (i.e. 393) but most of them are obsolete.

- *Searching over general Web search engines.* Although searching for Web resource of a specific

[1] http://www.oasis-open.org/standards
[2] http://soa.sys-con.com/node/164624
[3] http://www.xmethods.com/ve2/index.po

type (e.g. a Web service) seems to be considered as a second–class category in the research about search engines (Rose and Levinson, 2004), the characteristics of the Web service might give hints on heuristics necessary for developing expert strategies (e.g. "sms inurl:WSDL filetype:asmx")(Al-Masri and Mahmoud, 2008).

Several works provide evaluation of approaches used to service discovery on the Web (Bachlechner et al., 2006b,a; Hagemann et al., 2007). Lausen and Steinmetz (2008) presents the last well–known survey on service findability in the Web that compares search using dedicated portals and general Web SE (Alexa and Google). In this work, the Web SEs had significantly better coverage than the discovery portals studied. Alexa[4] has been identified as the best one (significantly better than Google). However, the portals were much better in terms of the precision, achieving up to 83% where Google's precision was merely 15%. Additionally, the portals usually offer a browsing mechanism and Web search engines do not offer it.

5.2.2 Searching for Geospatial Web resources

In terms of visibility, the Geospatial Web shares the same problems as the Web in general (see (López-Pellicer, 2011)). Publishing geospatial datasets and services by the geospatial community has led to the situation in which the Geospatial Web is characterised by strong tendency towards deep Web. Raghavan and Garcia-Molina (2001) present approaches to indexing the content hidden behind forms. One of the indicated disadvantages was the necessity for human interaction. Although searching for a Web resource of a specific type (e.g. a Web service) seems to be considered a second–class category in the research about search engines (Rose and Levinson, 2004), the characteristics of the Web service might give hints on heuristics necessary for developing a focused–crawler (Al-Masri and Mahmoud, 2008).

The OWS specifications permit the development of dedicated solutions for service discovery and content indexing. Patterns can be used to identify service requests calls in the Web, and ensure automatic generation of parametrised request templates and processing of the OWS responses (Whiteside and Greenwood, 2010) (all OWS provide information on used data models via `XML` schemas). Therefore, the potential of search engines has started to be acknowledged for the discovery of geospatial resources. There are findings about the ability of SEs as a replacement for the DL system for the discovery of Geospatial Web services (Sample et al., 2006; López-Pellicer et al., 2010c, 2011e), especially the recent work on a geo–domain focused crawler bring significant advances in this area. They present the systems for discovery of OWSs in the Web, which exploit the knowledge on standards and specifications of Geospatial Web resources (Li et al., 2010) and behaviour of a general Web search engine as well (López-Pellicer et al., 2011e). White et al. (2008) indicates that it might also be necessary to explore more than one of the existing search engines to extend the Web cover, which has been proven to be true also for the Geospatial Web (López-Pellicer et al., 2011e).

[4] `http://www.alexa.com/`

To summarise, search for domain–specific Web resources presents some challenges for non–expert users. The existing domain–oriented searching techniques indicate that an effective search demands:

- exploring more than one source;
- expert knowledge on functionality of the used search engine;
- expert knowledge on the domain (Wukovitz, 2001; Chen et al., 2003; Leroy et al., 2006);
- it might be necessary to perform additional Web mining to obtain better results (Menczer, 2003; Almpanidis et al., 2007).

5.3 Searching for Geospatial Web resources via a general search engine

This section introduces some recent advances in searching for Geospatial Web resources by means of a general SE. In this context, the two aspects of a search task from user perspective should be considered according to López-Pellicer et al. (2011e): (1) the search goals of a potential user of a geo–resource and (2) the search strategies that are necessary to follow to gather relevant resources.

5.3.1 Search goals

A search engine supports users in their search task. Different users have different searching goals, and supporting them means returning a list of relevant resources. Broder (2002) presents a well–known study on user search in context of the Web. The authors classified Web queries into three general classes:

- *Navigational*. This query represents an intent to reach a particular Web site, for example a homepage of an institution,
- *Informational*. These kinds of queries are performed when a user gathers some information assumed to be found on one or more Web pages,
- *Transactional*. This query represents an intent to perform some Web activity such as downloading songs.

Rose and Levinson (2004) refine the presented classification but in general the three main classes remain: (1) *navigational*, (2) *informational* and (3) *resource* (i.e. *transactional*) queries.

These three different search goals can be also applied within the Geospatial Web considering the characteristics of the resources, i.e. OWS services and related resources. For example, an ornithologist who is looking for resources to study the effect of the climate change on a bird spieces may develop the three search goals.

Navigational goal

In the case of the *navigational* goal, the user needs to find a service or Web page that he has in mind. For example, the ornithologist might want to find the home page of a concrete provider to see if any other services are offered as well.

Informational goal

The user also may expect a ranked list of OWS resources and Web pages related to some topic (*informational*). For example, the ornithologist can make an exploratory search of desertification maps and related pages, or he might gather OWS for building a thematic collection of services about climate components of the Iberian Peninsula.

Transactional goal

If the user search goal is *transactional* the user wants to learn the technical details of an OWS service. For example, the ornithologist wants to be able to test a WMS service.

These domain–specific search goals might be achieved with help of a general search engine. However, knowledge about its functionality and resource characteristics is required to develop proper search strategies.

5.3.2 Search strategy

In the scenario of OWS discovery through a general search engine, users make use of the fact that geoportals often publish documents with hyper–links to them, usually through catalogue viewers (Maguire and Longley, 2005). Therefore search engine crawlers can access and index service descriptions. However, the query result varies in function of the strategy. Two main strategies might be developed in general: *basic* and *expert strategy*, which depend on the user's ability to work with search engines.

Basic strategy

In the *basic strategy* the terms in query are those found in the `URL` of OWS descriptions along with terms related to the required information. For example, the search terms '*getcapabilities coastal*' will return a list of Web resources that contains those terms, such as an `HTML` page (e.g. a service description) with a hyper–link that invokes the `getCapabilites` operation.

Expert strategy

An *expert strategy* refines the basic strategy (Bartley, 2005; Sample et al., 2006), as search engines usually allow users to define constraints on the `URL` of the target resource. In this way, the query

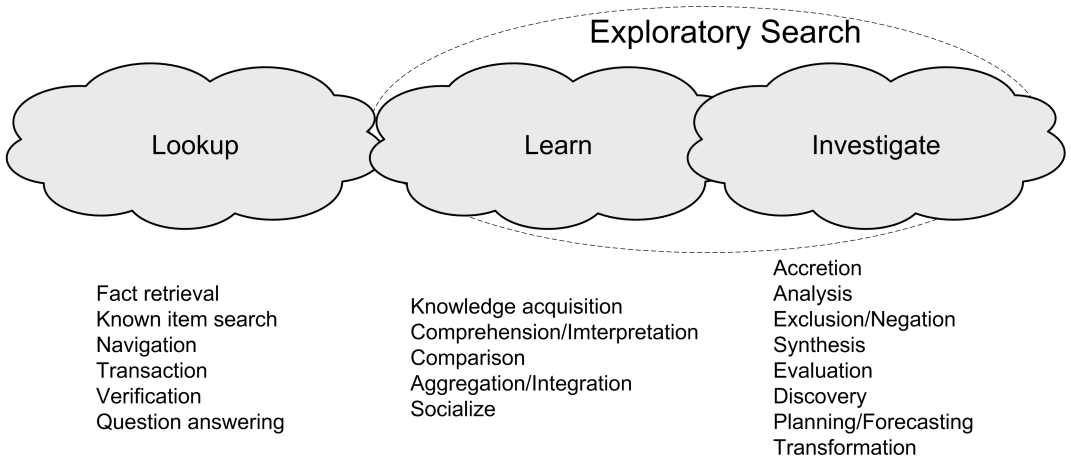

Figure 5.1: Three kinds of search activities and exploratory search (source: Marchionini (2006)).

may be restricted to the documents whose `URLs` match the pattern of requests for the OWS service description. For example, *'*inurl:getcapabilities coastal*'* request in case of Google SE returns the *Capabilities* documents as the response to the `getCapabilities` operation.

Users who use general search engines to search for geospatial resources might be practitioners of GIS tools but not necessarily specialists in geographic information and related technology. Therefore they will have to investigate the proper way to exploit general search engines for their purposes (i.e. to look thorough a SE documentation and OWS characteristics). However, having the possibility to use a tool that supports the non–expert users in searching for geospatial resources would save their time and effort.

5.4 Non–expert user support

As shown in previous section, discovery and searching for geospatial resources on the Web requires expert knowledge. This is because searching for OWS instances has a mainly *transactional* character. Therefore, a domain–oriented SE that applies special techniques may support the users out of the GI community. For example, a search layer over a general SE that applies the expert search strategy (i.e. via automatised extension of user requests with specialised operators) might be one of the possible solutions. Additionally, a search client should be designed in the manner needed to support user goals other than *transactional*. Humans have poor short–term memory, i.e. they have a limited ability to search and interpret textual and/or tabular data (Miller, 1956). Therefore, a crucial aspect of achieving success in implementing such systems for their reuse is its acceptance. In this scope, the system's user interface is critical, as even minor usability problems will demotivate users (Seaman et al., 2003).

There are many important theoretical models of information search (e.g. a stratified model

presented in Saracevic (1997) which summarises Belkin's and Ingrewsen's). In the context of this work, exploratory search seems to be the best choice because it has been identified as especially appropriate for users that lack domain–specific knowledge. Such a search system helps users to explore, overcome uncertainty, and learn, rather than just providing search results (White et al., 2006; Marchionini, 2006) (see Figure 5.1). Exploratory search can be supported in Web search engines by using categorised overviews of Web search results (Kules, 2006), based on meaningful and stable categories. This approach *"can provide substantial benefits when searchers need to explore, understand, and assess their results. When information needs are evolving or imprecise, categorized overviews can stimulate relevant ideas, provoke illuminating questions, and guide searchers to useful information they might not otherwise find. When searchers need to gather information from multiple perspectives or sources, categorized overviews can make those aspects visible for interactive filtering and exploration."* (Kules, 2006)

One of the existing, successful approaches to search interface that support exploratory search is faceted search and browsing. It is popularly used by search application for searching within DLs in diverse of domains, such as e–commerce (e.g., Ebay[5], Amazon[6]) or scientific results (e.g. ScienceDirect[7], PubMed[8]). Faceted search interface is successful as it usually supports and encourages two aspects of user search behaviour: browsing (i.e. exploration) and querying (in particular query refinement).

5.4.1 Faceted classification

Facets have been used by librarians and information scientists to structure information for a long time already. The Classification Research Group indicated faceted classification to be the basis for all IR in 1955 (Broughton, 2006), and then, also librarians started using facet analysis (Ranganathan, 2006).

Facets are defined as "*a set of meaningful labels organized in such a way as to reflect the concepts relevant to a domain*" (Hearst, 2006), and represent "*the categories, properties, attributes, characteristics, relations, functions or concepts that are central to the set of documents or entities being organized and which are of particular interest to the user group*" (La Barre, 2007). Ranganathan (1967) states that faceted classification "*of a particular universe* [of entities] *is made on the basis of characteristics*". Following La Barre (2010), "*(...) characteristic [of entities] is equivalent to a mathematical parameter presenting a range of possible factors, aspects, or elements that assist in the identification of a collection of distinct cases. In facet theory, each parameter creates a dimension, or small number of groupings, and each grouping represents a facet.*" Therefore, each facet has potentially multiple dimensions (Ranganathan, 1967).

[5] http://www.ebay.com/
[6] http://www.amazon.co.uk/
[7] http://www.sciencedirect.com/
[8] http://www.ncbi.nlm.nih.gov/

Not each classification system is a faceted classification (Ranganathan identifies eight kinds of classification systems). There are three important issues which distinguish faceted classification (Vickery, 1966):

- Faceted classification system is created by strict application of rules and facet analysis such that: "*every distinctive logical category should be isolated, every characteristic should be clearly formulated, and each new relation should be recognized.*"
- Facets are not locked into "*rigid enumerative schedules, but are left free to combine with each other in fullest freedom, so that every type of relation between terms and between subjects may be expressed.*"
- Faceted classification system "*breaks free from the restriction of traditional classification to the hierarchical, genus–species relations. By combining terms in compound subjects, it introduces new logical relations between them, thus better reflecting the complexity of knowledge.*"

5.4.2 Faceted search and browsing

Faceted classification provides a usable method which permits browsing information collections via multiple categories simultaneously (Hearst, 2000; Hearst et al., 2002). According to Hearst (2009) *"[i]n a properly designed faceted navigation interface, the user can browse the information collection from any of the different facets as a starting point, and after starting with one facet, can then navigate using any other facet."*

Fagan (2010) provides summary of empirical evidence related to faceted browsing found in the information science literature as follows:

- *Facets are useful for creating navigation structures.*
- *Faceted categorisation greatly facilitates efficient retrieval in database searching.*
- *Facets help avoid dead ends.*
- *Users are faster when using a faceted system.*
- *Success in finding relevant results is higher with a faceted system.*
- *Users find more results with a faceted system.*
- *Users also seem to like facets, although they do not always immediately have a positive reaction.*
- *Users prefer search results organised into predictable, multidimensional hierarchies.*
- *Participants' satisfaction is higher with a faceted system.*

Additionally, Fagan performs an empirical study on the usability of search systems; the study shows that faceted browsing improves user performance. Fagan (2010) argues that all these results build a solid case for including facets in search interfaces.

Kules proposes a *set of principles for the design of exploratory search interfaces* in his dissertation (Kules, 2006):

- *Provide overviews of large sets of results.*
- *Organise overviews around meaningful categories.*
- *Clarify and visualise category structure.*
- *Tightly couple category labels to result list.*
- *Ensure that the full category information is available.*
- *Support multiple types of categories and visual presentations.*
- *Use separate facets for each type of category.*
- *Arrange text for scanning/ skimming.*
- *Visually encode quantitative attributes on a stable visual structure.*

This set of principles is *"useful for digital library and Web search designers, information architects, and Web developers because they provide guidance for the appropriate integration of visual overviews with search result lists, and particularly for the textual surrogates embedded in result lists. These principles embed a strong call for the surfacing of structure – which is often used internally by search engines, but less often exposed at the user interface – without abandoning the tried and true value of text"* (Kules, 2006). Therefore, a search system that uses faceted classification should be designed and developed with this set of principles in mind, e.g. several meaningful categories should be used, the category structure should be clarified and visualised, and search result lists should be tightly coupled with category labels.

Other works in this context cover technical aspects of the faceted search. Koren et al. (2008), for example, indicates that a faceted search engine essentially performs two separate retrieval tasks: (1) user query matching documents, and (2) recommended query refinement. The first one has been intensively investigated within the IR research community. The faceted classification features of the category isolation and ability to combine facets enables query refinement in a straightforward manner. It is based on the usage of *refinement operators*. Values of such specialised operators are mapped to facet dimensions which allow grouping of resources via categories.

In general, two approaches to query refinement, (1) *exclusion* and (2) *limitation*, have been distinguished after analysing some examples of popular online applications that offer faceted search (e.g. Amazon, ScienceDirect). *Exclusion* allows removal from the result list of those elements whose

category belongs to one of the defined values. *Limitation* trims the result list to those elements which are categorised as one of the selected values. This system behaviour is usually offered via interaction with GUI, for example by series of check–boxes on a category values. The refinement operators can also be offered by *search interface*, i.e. they can be used as *search operators*. There are *free–text search operators* and *restricted search operators*. *Free–text search operators* are those *search operators* whose values can be explicitly specified by a user. Values of the *restricted search operators* have to belong to a controlled set. Such an operator might appear in an advanced search GUI with a controlled list or check–box list. In this way, only values from the controlled set are used as values of the *restricted search operators* within the search request. This restriction on the input values eliminates user errors and limits possible options which improve system performance. Some works indicate that, even providing the advanced search capacity, the simple keywords search should be still available for a user. If simple search is enabled (i.e. there is one search box to input a request), usually only the *free–text search operators* are processed as part of a user request. They might be allowed with some restrictions as well, for example, without the possibility of using multiple values or complex logical queries (i.e. "AND", "OR"). The *refinement operators* can also be part of the *navigation interface* to enable browsing of resources via a category. For example, a check–box can also be a link to a list of resources grouped via the specified category. These operators usually correspond to the *restricted search operators*. The coherent *navigation interface* and the resultant category–based lists of resources exposed for their further exploration (i.e. with associated links to description of related resources) are the basis of the exploratory search.

A search system which offers faceted search and browsing fulfils one of the principles for Web search interfaces provided by Rose (2006), i.e. "the interface should support the iterative nature of the search task. In particular, it should invite refinement and exploration." Additionally, according to Rose "different interfaces (or at least different forms of interaction) should be available to match different search goals". Therefore, the system proposed should be adjusted to search for OWS resource and related information. In particular faceted search and browsing supports the *informational* search goal because it helps with the gathering of information on a particular subject. However, the *transactional* character of search for OWS instances demands some specific facilities, for example, ability to interact with OWS instances.

5.5 Domain–specific search user interface

The design requirements of the Search User Interface (SUI) that exploits facets invlove important issues when designing a search system, in particular, coherence of search, browsing and navigational interaction model. Additionally, the domain–specific characteristics of searched resources should be considered when defining the set of facets and interactions with the system. Examples of faceted search and browsing systems can be found on the Web, including online portals that help users to

search for Web resources relevant to the scope of this work. Therefore, two main communities have been considered as those of major relevance to this work, the Web service and the Geospatial Web community. First, some relevant Web search clients from related areas are examined to identify their main characteristics. Then, Section 5.6.1 discusses the proposed facets, refinement operators and the implementation decisions which have been taken.

5.5.1 Web service community

The important elements that influence the usability of search system are the SE interface and the search client design. Mohammed et al. (2006) describe a UDDI–based SE offered as a desktop application. The SE search for medical Web services that appear on the Web. According to the authors, the developed SVG Search Engine client (i.e. a UDDI client) allows discovery and usage of the required Web services. The search criteria are (1) the service name, (2) the thematic classification (i.e. hierarchical classification schemes of Breast Cancer), and (3) the service groups. The work focuses on the security aspects of such system, and it has been developed as a desktop application. There is no reference to an existing tool; therefore, it is difficult to evaluate functionality of the developed interface.

A good reference on a Web service search portal might be seekda's Web Services portal[9], a result of several research projects in which seekda[10] has been involved. The main projects are the *Service–Finder* and the *Service Detective*. The *Service–Finder* project[11], (funded by the 7th Framework Programme for Research and Development of the European Commission and finalised in December 2009) has been dedicated to "*develop a platform for service discovery in which Web Services are embedded in a Web 2.0 environment*"[12]. Brockmans et al. (2008) and Brockmans et al. (2009) discuss the requirements and architecture of such a system. These requirements refer to system components (i.e. service search interface, service related information and user community interface), external resources (i.e. provider and service related information sources), and service description standards (i.e. `WSDL` documents in the first place and RESTful APIs as the next step). The proposed architecture distinguishes following components:

- Ontologies (i.e. Generic Service–Finder Ontology and Service Categories)
- Service Crawler
- Automatic Annotator
- Conceptual Indexer and Matcher
- Service–Finder Portal

[9] `http://webservices.seekda.com/`
[10] `http://www.seekda.com/`
[11] `http://www.service-finder.eu/`
[12] `http://www.seekda.com/en/research/projects/service-finder`

- Cluster Engine
- Service–Finder Interface (Search Interface, Navigation Interface)

Service Detective project[13], funded by FIT–IT[14], that "*create an architecture for a Web Service SE that automatically aggregates information from heterogeneous sources to facilitate discovery of both* WSDL *and RESTful services*"[15]. The project has been dedicated to the provision of a new approach to Web service ranking (Steinmetz and Lausen, 2009) (e.g. the description that is available on the Web, their hyperlink relations, monitoring information, etc.), a focused crawler that gathers SOAP Web Services (i.e. their WSDL documents), the RESTful Web Services (i.e. the online documents on their APIs, and a WSDL document if provided), and related documents. The system aggregates Web service annotations (Steinmetz et al., 2009) and stores meta–information in RDF triples. The ontologies used have partly been developed in the scope of the *Service–Finder* project (i.e. *Service–Finder Ontology*[16] and *Crawl Ontology*[17]).

The seekda's Web Services portal uses a Web service repository which contains 28.606 services offered by 7.739 providers[18]. The services have been gathered via a focused crawler that searches the Web for the valid WSDL documents or the links submitted manually by providers. This portal favours creation of user communities. It encourages users to create an account as it is the only way to add the free–form annotations (i.e. tags) or wiki–like comments about services and providers, and to rate services and provider. Each service is described via a country, its provider, a URL of its WSDL document and its cached version (i.e. an XML or HTML document), age, the type of server, a category according the documentation found within the WSDL document (i.e. "none", "partial" or "good"), a textual description (from the WSDL document, or added by the provider and/or users), and user rating and tags. The portal also offers some monitoring information about services (i.e. availability) and the possibility of invoking the operation declared in the associated WSDL document. Each provider is annotated with a country (via IP georeferencing), some textual descriptions, a list of tags and the homepage URL. A list of offered services is available as well. As for SE interface, it processes requests with restrictions on the country, the provider and free–form tags. The ranking might be modified according to relevance, available documentation, availability, provider, country and age. The first 250 results are presented for each query (there is information on the total number of matching results). The client search interface offers both kinds of search form, a simple form (i.e. the one field form) and an advanced search form (one field per each restriction operator allowed). Although the exploratory navigation is intuitive and coherent, only simple logic requests are supported. The system does not support full faceted search which requires combination of multiple logic restrictions

[13] http://service-detective.sti2.at/
[14] http://www.fit-it.at/
[15] http://www.seekda.com/en/research/projects/service-detective
[16] http://www.service-finder.eu/ontologies/ServiceOntology
[17] http://seekda.com/ontologies/CrawlOntology
[18] Last accessed: 09/02/2012

on declared operators. Currently, there is on–going work on a new version of the portal which includes support to RESTful Web Services and information on mashups as well.

5.5.2 Geospatial community

A catalogue service plays a fundamental role in an SDI as it allows the reusage of resources, and it is a central element of a system for the metadata management and discovery. A geoportal is an effective solution for such a system (Baldini et al., 2010), as it is accessible from a Web browser. Currently, there is a variety of configurable frameworks, and it is possible to develop geoportals in an easy and fast manner. For example, GeoNetwork[19] is open–source software designed to improve the accessibility of a wide variety of data and the associated metadata. It implements the Portal and Catalog SDI components as defined in the OGC Reference Architecture. GeoServer[20] is another example of such software. TerraViva! GeoServer[21] is a geoportal that helps in finding interactive maps, GIS datasets, satellite imagery and related applications.

Following the OGC GP–RA, geoportals can offer an interoperable search across different catalogues (Giuliani et al., 2011). Such system is a gateway that queries one or more registered catalogues. The GEOSS project is a good example of such an interoperable system that can query multiple catalogues registered in its system. GEOSS contributors use the *Component and Service Registry* to register their geospatial services for their further discovery through *Clearinghouse* and invocation by *GEO Portal*. The central element of the GEO Portal[22]is a map viewer which is used to show recent updates of resources. The main browsing menu allows users to explore resources according to the SBAs (two–level categorisation schema) and then, to refine the result by the location (a controlled list of regions). Simple search (i.e. free keywords search) and mentioned browsing return resources categorised via their type with indication of number of resources which fall into each category. The top–level categories are: Dataset, Monitoring and Observation Systems, Computational Model, Initiatives, Web sites and documents, Data Services, Software and Applications and Others. A resource in the result list is described via part of the abstract, links to related Web resources (documents, Web pages or Visors) and link to resource description tab. The resource description tab has two possible views, (1) "Summary", a default view which shows selected metadata (Contact Details, Abstract, Organisation Name, Distributor, Date Stamp and related documents) and (2) "Full Description" which returns full metadata record (encoded in `HTML`) from the system catalogue. The refinement of search results refers to (1) location (as (1.1) bbox, (1.2) region name from a list or (1.3) region from an interactive map), (2) additional keywords, and (3) social benefit area (one or more category from two–level classification schema). In the context of search for OWS services, identification of OWS services might be a bit confusing in the beginning. As for user search goals, the *informational*

[19] `http://geonetwork-opensource.org/`
[20] `http://geoserver.org/display/GEOS/Welcome`
[21] `http://geoserver.isciences.com:8080/geonetwork/srv/en/main.home`
[22] `http://www.geoportal.org/web/guest/geo_home`

goal is quite well supported. A topical query returns a variety of resources (not only OWS services) and the returned OWS services can have associated related documents. The *navigational* goal (e.g. looking for homepage of a service provider) might be a bit difficult due to an overwhelming number of results for some queries. Accomplishment of the *transactional* goal seems to be a bit demanding for a non–expert user. Although there is a link to the *Capabilities* document in the main result list it is automatically redirected to the map visor[23]. The request point of OWS service can be found after navigating to the "Full Description"

As described in Bai et al. (2009), the resources are manually registered. Resource incorporation into an SDI is characterised by a top–down control which involves metadata curation. Such approach is not effective in systems based on Web crawling. Growing interest in crawling the Web for the Geospatial Web resources (López-Pellicer et al., 2011e; Li et al., 2010) and non–SDI resources as an alternative source of valuable information (Craglia et al., 2008; Keßler and Bishr, 2009) give bases to expect that OWS service portals similar to the seekda's portal will become popular soon.

5.6 Geospatial Web Search Engine

The proposed GWSE exploits a general Web SE (i.e. a remote SE) to search for geospatial resources. Focused crawlers are used to identify OWS calls via patterns within the results of the remote SE. In this work, the crawlers are developed to detect the most popular OWS specifications. The OWS resources discovered are classified into *service description* (the *Capabilities* document), *item description* (the response to the *Describe** operations) and *item* (the content published by the service). Table 5.1 summarises the description of the main operations. As for the Search User Interface (SUI), the proposed search system offers to the user the ability to define a search strategy (*none*, *basic*, *expert*), and the search client has been developed in order to help support the three search goals (i.e. *navigational*, *informational* and *transactional* as introduced in Section 5.3.1) in the context of the search for OWS resources.

Focused crawlers applied are not levered here because this topic is out of scope of this work (please refer to López-Pellicer et al. (2011e) and Li et al. (2010) for details of those systems). The following Section presents the offered search, browsing and navigational interface. Next, the developed search strategies and process of indexing of OWS content (the *OWS Crawler*) are presented, and then the role of search operators in USI design is outlined. Finally, the architecture and implementation of a prototype is briefly described.

5.6.1 Facets for search, browsing and navigational interface

In a faceted search and browsing system, three types of interactions with USI have to be considered, i.e. search, browsing and navigation. In the context of this work, these concepts are understood as

[23]There are some usability issues which make it difficult to evaluate this approach.

Service Specification [Operation]	**Brief Description**
Web Catalogue Service (CSW)	It supports the ability to publish and search collections of descriptive information (metadata) of data, services, and related resources.
[DescribeRecord]	It allows a client to discover elements of the supported data model.
[GetRecords]	It allows discovering resources with possibility to apply spatio–temporal constraints.
Web Map Service (WMS)	It produces dynamically maps of spatially referenced data from geographic information.
[GetCapabilities]	It enumerates layers that might be rendered and supported parameters (e.g. graphic format).
[GetMap]	It produces maps.
Web Coverage Service (WCS)	It supports electronic interchange of coverages (values or properties of a set of geographic locations) that represents space–varying phenomena.
[GetCapabilities]	It enumerates coverages that might be rendered and supported parameters.
[DescribeCoverage]	It provides a full description of a coverage.
[GetCoverage]	It returns a coverage.
Web Feature Service (WFS)	It allows direct fine–grained access to geographic information at the feature and feature property level.
[GetCapabilities]	It lists the features that might be requested.
[DescribeFeatureType]	It returns a schema description of the requested feature.
[GetFeature]	It operation returns a document that contains selection of features (retrieved from a relatively static data store), which satisfy the query expressions specified in the request.
Web Processing Service (WPS)	It allows invoke processing functionality at the feature and feature property level.
[GetCapabilities]	It lists the processes that might be executed.
[DescribeProcess]	It returns the description of the requested process.
[Execute]	It executes requested process.

Table 5.1: The main operation of OGC Web Services.

follows:

- *Faceted browsing.* It is the act of reviewing the collection of resources grouped via a category.
- *Faceted search.* It refers to the usage of search operators that correspond with facet categories.
- *Faceted navigation.* It is the act of switching a facet category when navigating.

A coherent search interface should be designed to help the user in exploring search results. Therefore, the selection of faceted classifications, search operators and their usage to support facet–based interactions are critical to acquire appropriate behaviour of the system.

Facets

The characteristics of the aimed resources have been analysed to define faceted classification. According to the design principle about meaningful categories, the following facets have been chosen:

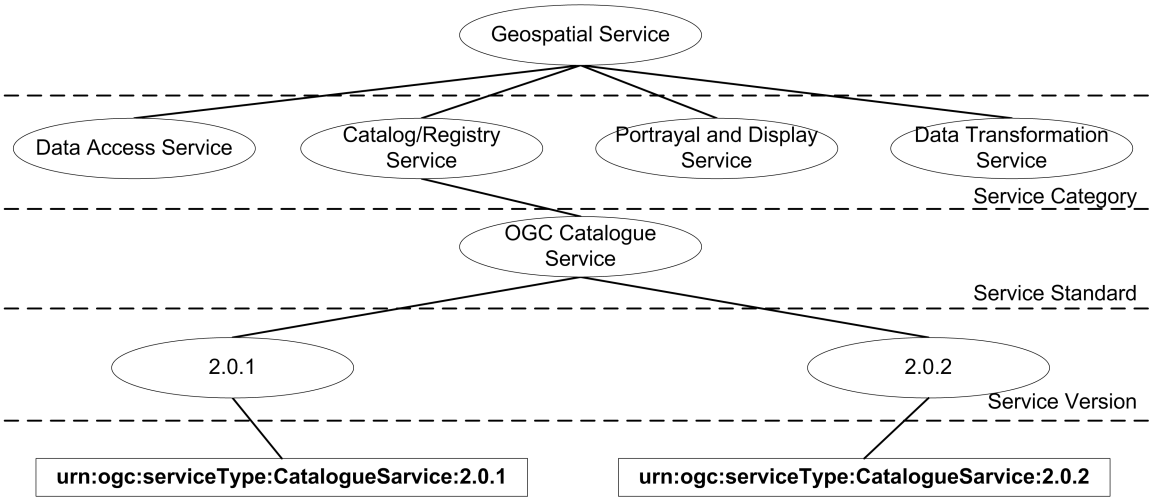

Figure 5.2: Geospatial service taxonomy proposed in Bai et al. (2009) (source: Bai et al. (2009)).

- *OWSResource*. This facet classifies resources from a list according to their type, i.e. *service description*, *item description* and *item*
- *OWSTaxonomy*. This facet classifies resources according to the taxonomy of service the resource is related to.
- *Reliability*. This facet informs on reliability status of services. It allows connecting to a service monitoring framework. The values are ranges of reliability (i.e. 100%–75%, 75%–50%, 50%–25%, 25%–0%).
- *Domain*. This facet classifies resource according to their domain (i.e. the categories are extracted via patterns from the resource `URL`).
- *Provider*. This facet classifies resource according to providers defined within the *Capabilities* document of service the resource is related to (extracted via NER methods).

Two different taxonomies for geospatial services have been studied and evaluated (Bai et al., 2009; Zhang et al., 2009). The service taxonomy proposed in Bai et al. (2009) has been selected because it is lightweight service taxonomy especially useful to capture knowledge around services characteristics, so that geospatial services can be classified according to their service category, particularly what standards are followed. This classification scheme is used in the GEOSS *Component and Service Registry*, one of the main elements of the GEOSS architecture. Figure 5.2 shows the taxonomy proposed in Bai et al. (2009)). The service taxonomy has been restricted to "Service Version" (i.e. only the `HTTP` binding is supported in this application). The taxonomy is developed in accordance with technical standards. In this way, the extension of *OWSTaxonomy* vocabulary and resource annotation can be automatised because the *Capabilities* document provides all information on service

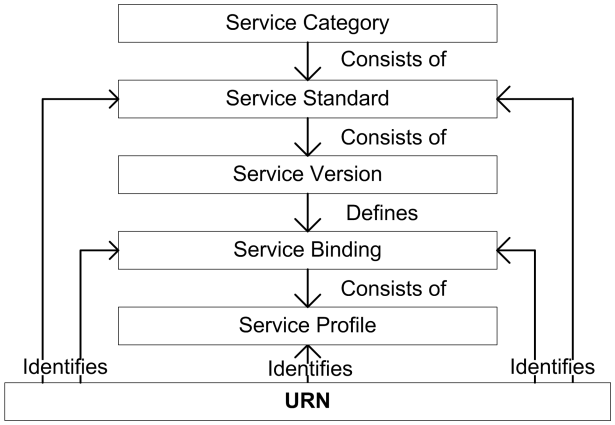

Figure 5.3: Multi–layer logical structure of the URN taxonomy (source: Bai et al. (2009)).

type that is needed. *Domain* and *Provider* vocabularies have to be automatically extensible because the system has to support automatised new resources added to the repository.

The faceted classification used in this work have been developed in Simple Knowledge Organization System (SKOS) (Isaac and Summers, 2009), a W3C standard for porting knowledge organisation systems to the Semantic Web. The Web Ontology Language (W3C OWL Working Group, 2009) (OWL) has been considered as well. It offers a general and powerful framework for knowledge representation. This application does not demand such advanced capabilities to define vocabularies required. SKOS is a simple language with just a few features, tuned for sharing and linking knowledge organisation systems, and it can be used for this purpose. The vocabularies created are used to annotate the gathered Web resources. Appendix B contains the used SKOS vocabularies. Only the *OWSTaxonomy* faceted classification has hierarchical structure. Putkey (2011) creates a SKOS–compliant faceted taxonomy that preserves the required hierarchical structure. Here, the usage of `URIs` helps to preserve the structure of the *OWSTaxonomy* taxonomy (see Figure 5.3).

Search operators

Table 5.2 summarises the search operators supported by the system. The *"service"* and the *"resource"* refinement operators refer to the *OWSTaxonomy* and *OWSResource*, respectively. If they are used as free–text operators, the combination of both (e.g. "service:WMS resource:item") might produce an empty response. The *"site"* and the *"inurl"* operators have similar functionality to the corresponding operators supported by existing SEs (e.g. Google). Also the *"+"* and *"–"* modifiers are supported to extend or restrict the search (e.g. "inurl:tata –inurl:en –site:com").

For more effective search of geospatial resources, the proposed system extends text–based searching with spatial search capacity. In this work, spatial search is understood as: *"give me all resources which are related with this spatial location"*. The system supports two forms of spatial constraints.

Operator	**Classification Scheme**	**Description**	**Free–text example**
`service`	*OWSTaxonomy*	restriction on service specification	"service:urn:ogc:serviceType:WMS"
`resource`	*OWSResource*	restriction on OWS resource type	"resource:service"
`provider`	*Provider*	restriction on OWS provider	"provider:'ING"'
`domain`	*Domain*	restriction on OWS domain	"domain:'www.idee.es"'
`inurl`		the text has to appear in the resource `URL`	"inurl:es" "inurl:es"
`location`		spatial restriction via a toponym	"location:'Washington DC"'
`point`		spatial restrinction via a coordinate pair (might be omitted)	"point:'36.533333, -6.283333"' or "point:'36,533333, -6,283333"'

Table 5.2: Searching interface supported by the Geospatial Web Search Engine.

A latitude/ longitude pair coordinates might be used to define the location of interest explicitly. The point coordinates are assumed to be of WGS 84 (NIMA, 2004), a reference system which is commonly adopted in the Web community (e.g. GeoRSS[24]). The *"point"* spatial operator can be omitted because the parser of the free–text query tries to extract a coordinate pair as well. For example "36.533333, -6.283333" as a query will return any geospatial resources that offers data of the area that contains the point defined. The "," is not necessary as it will be removed during the query pre–processing task. It is also possible to define spatial restriction via the *"location"* operator because the system is dedicated to support non–expert users as well. However, this kind of the location definition has a disadvantage when comparing to an explicit point. It inherits ambiguity of toponyms, as different places may have the same name (e.g. "Madrid" in Spain or "Madrid" in Iowa, USA). To offer a proper support to the spatial restrictions, it should be translated into explicit coordinates. In the GWSE presented in this work, a toponym is translated into a list of candidate points by means of a gazetteer (Hill et al., 1999), and a user is asked to select the desired location from the list.

While querying the remote SE, the point identified is removed from the request because a general SE treats coordinates as a pure text usually, and such a search produces an empty result frequently. If an SE does not offer any operator with a similar semantic to that of the *"location"* operator, the place name is used as a free–text in the remote SE query.

[24] `http://www.georss.org/`

5.6.2 Search User Interface design

As for the SUI design, Table 5.3 summarises the search operators and the way they have been used in the Simple Search Interface (SSI), the Advances Search Interface (ASI), the Refinement Interface (RI) and the Navigation Interface (NI).

Operator **Interface**	**Description**
service	
SSI–R	The SSI only accepts this operator if its value is correct.
ASI–R	The search operator is associated with a controlled list available via in the advanced search form.
RI–R	The *OWSTaxonomy* is exposed as check–boxes in the refinement menu to manipulate the search results.
NI	The browsing via type of related service is enabled. Users can browser through whole content using a link in the main menu ("Browse via service type") and the service description form ("service type" field). The check–box links in the refinement menu permit to access to the subset of resources according to user search query.
resource	
SSI–R	The SSI only accepts this operator if its value is correct.
ASI–R	The search operator is associated with a controlled list available via the advanced search interface, in the advanced search form.
RI–R	The *OWSResource* is exposed as check–boxes in the refinement menu to manipulate the search results.
NI	The browsing via OWS resource type is enabled. Users can browser through whole content using a link in the main menu ("Browse via resource type") and the service description form ("resource type" field). The check–box links in the refinement menu permit to access to the subset of resources according to user search query.
domain	
SSI–FT	No restriction on value.
ASI	The search field accepts free text.
RI–R	The *Domain* is exposed as check–boxes in the refinement menu to manipulate the search results.
NI	The browsing via OWS domain is enabled. Users can browser through whole content using a link in the main menu ("Browse via domain") and the service description form ("domain" filed). The check–box links in the refinement menu permits to access to the subset of resources according to user search query.
provider	
SSI–FT	No restriction on value.
ASI	The search field accepts free text.
RI–R	The *Provider* is exposed as check–boxes in the refinement menu to manipulate the search results.
NI	The browsing via OWS provider is enabled. Users can browser through whole content using a link in the main menu ("Browse via provider") and the service description form ("provider" field). The check–box links in the refinement menu permit to access to the subset of resources according to user search query.

Table 5.3: Search operators and user interface design (SSI–FT – free–text operator of simple search interface; SSI–R – restricted operator of simple search interface; ASI–FT – free–text operator of advanced search interface; ASI–R – restricted operator of advanced search interface; RI–R – operator used in the refinement interface of search results; NI – operator used in navigation interface).

A Web client has been developed following Jacob' Laws (Nielsen, 2008) on Web application

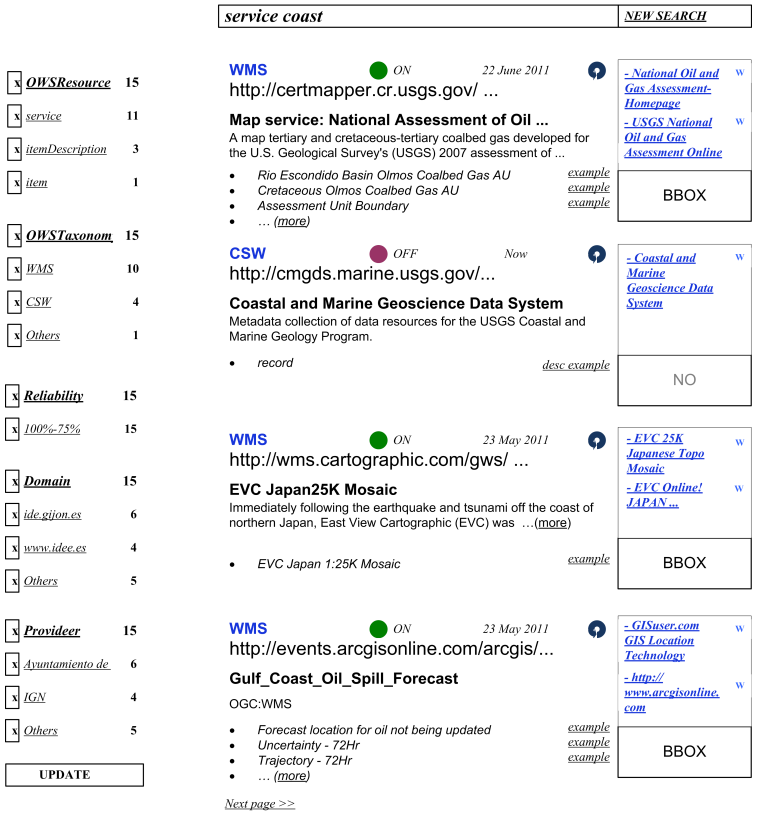

Figure 5.4: Prototyped GUI of the search result offered by the Geospatial Web Search Engine.

usability (i.e. "*users spend most of their time on other websites*"; "*Users have several thousand times more experience with standard GUI controls than with any individual new design.*"). Therefore its GUI design is similar to the faceted search offered by the typical online e–commerce applications (e.g. Amazon) Figure 5.4 presents the pattern of representation of the search result implemented by the GWSE client. The result list is in the centre. Each resource is described via a summary description that contains snippets of information from a full description, and some examples of operations that might be used to the retrieve the descriptive information or an example of item. On the left, the refinement menu is shown. The check–boxes allow refinement of the results which is performed after pressing "Update". In the case of the *Domain* and *Provider* facets, the number of possible values in the refinement menu might overwhelm a user. Therefore, only the first five values of the highest frequency are displayed and the sixth one ("Others") gathers the remaining resources. The check–box terms are also links, that offer explanation of the applied classification and expose a list of navigable categories to browse the related OWS resources. On the right, there is the related resources' menu to expose the related resources to users. They are the Web pages ("W") or services ("S") from which the resource might be accessed. Most of them are results of the crawling process

(i.e. they have been used by a crawler in order to identify the OWS resource). The others are Web pages of the OWS publisher. These Web pages have been generated from the domain URL as described in Chapter 4. There is also a link to visualise the resource in a map.

The full description of a resource is accessible via the "title" link in the summary description. Each resource is described via information from the *Capabilities* document which has been extracted using a metadata crosswalk (Nogueras-Iso et al., 2004). In the case of a service, user has the ability to invoke the descriptive operations or to retrieve some items. The requestable elements are described via information from the *Capabilities* document and the result of descriptive operation is encoded in HTML. Also, some summaries of dataset are shown if it is possible to create them. The *discriminative elements* are identified. The values from *low variable fields* are listed with quantitative attributes (i.e. the number of items in a dataset which satisfy the restriction), and an appropriate request is generated and enabled via a link.

Theoretical evaluation

The design principles proposed in Kules (2006) help in development of systems that support non–expert users in exploring the search results. Table 5.4 describes the manner in which the design principles are accomplished by the GWSE.

Id	Design Principle	Application
1	Provide overviews of large sets of results.	The first 50 results are available. The full number of matching resources is showed also.
2	Organise overviews around meaningful categories.	Several stables classification schemes are used.
3	Visualise and clarify category structure.	The check–box and browsing tabs are organised according to the schemes' structure.
4	Tightly couple category labels to result list.	The check–boxes indicate which categories are currently applied.
5	Ensure that full category information is available.	Not applicable.
6	Support multiple types of categories and visual presentations.	Not supported.
7	Use separate facets for each type of category.	Supported.
8	Arrange text for scanning/skimming.	The result list contains descriptional snippets.
9	Visually encode quantitative attributes on a stable visual structure.	The refinement menu and browsing tabs provide quantitative attributes.

Table 5.4: Accoplishment of the design principles for exploratory search interfaces.

Seven of the nine principles are fully supported. The fifth design principle is not applicable in this work because no taxonomy of deep hierarchy has been used. In this context, the wide vocabularies can be the obstacle. *Domain* and *Provider* are horizontally extensive. Anyway, the browsing menu allows users to browse through them. The sixth design principle is not accomplished because the user cannot define their own categories and no only single presentation style of search results is enabled.

As for user search goals, faceted search interface, along with the information model, can support the *informational* goal especially well. The *informational* goal can be reached via the exploratory search, as it enables search and browsing collections of resources using multiple categories. Also, the association of non–OWS Web resources (e.g. provider Web site) extends the information scope that can be explored by users. The *navigational* goal can be achieved, for example, via the "site" operator or by the navigational interface (e.g. by navigating to the provider Web site). The enablement of interaction with OWS services offers extended assistance to the *transactional* goal. A *test module* has been created, similar to that one which is offered by seekda' portal for invoking the operations declared within a `WSDL` document. It generates a set of modifiable requests in order to allow users to interact with the resource. In this work, the set of request–response fields are only generated for WFS, WCS and WMS services.

5.6.3 OWS search and indexing procedure

When the user invokes a search query, two parallel search processes are run, i.e. the task of querying the remote SE (*RemoteSE*) and the task of querying the local repositories (*Repository*). The *Fast-WebCrawling* is performed over the results from a remote SE when the user query is being handled, while the long–time *MainOWSCrawling* is performed in the idle time of the system. Figure 5.5 shows an overview of the searching and indexing tasks that are performed by the developed system.

The following Sections present some details of these processes.

Search procedure

When search query is invoked, two parallel search processes are run, i.e. the remote SE is queried and the local repository is exploited. According to the selected search strategy, the user query may be extended automatically with additional tags to improve remote SE efficiency. Then, the *fast Web crawling* is performed over the SE response. The *fast Web crawling* consists in the identification of OWS operation calls among resources obtained from the remote SE. The service descriptions of identified OWS instances are indexed and integrated with results from the local repository to produce the user response.

The remote SE response is saved for the *main OWS crawling* process which is invoked apart for performance reason. This process applies the *Web Crawler* and the *OWS Crawler* (detailed in Section 5.6.3) to generate the local repository content, i.e. to identify OWS resources in the Web,

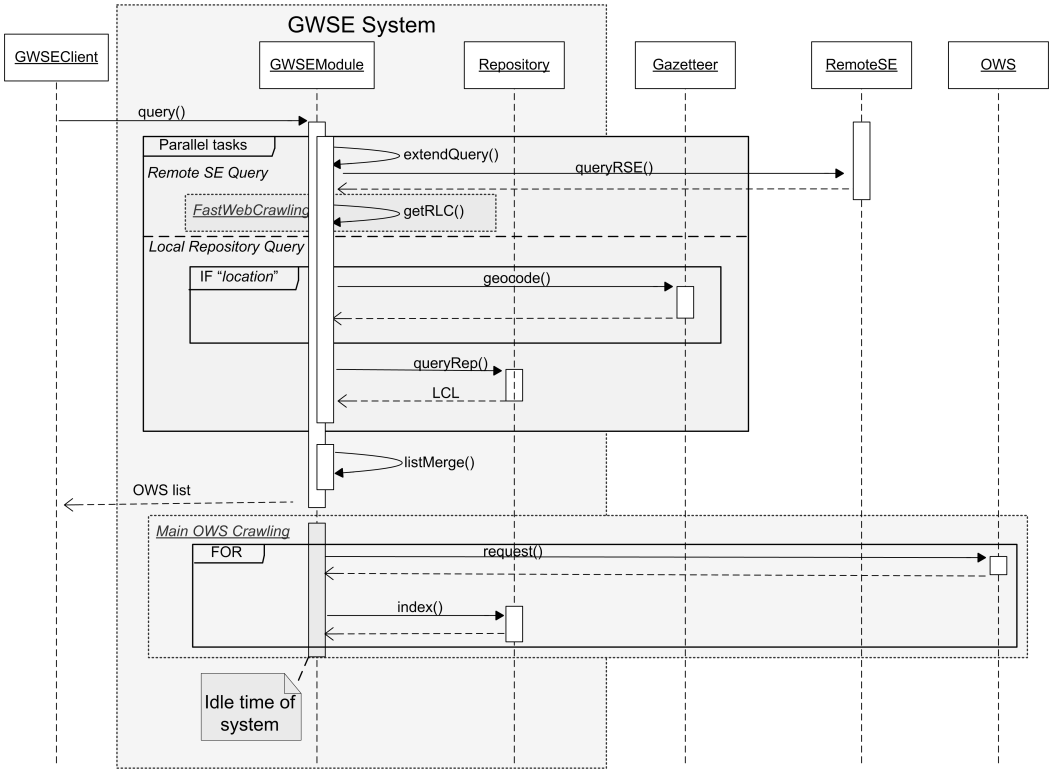

Figure 5.5: Overview of the searching and indexing tasks performed by Geospatial Web Search Engine.

retrieve and index OWS content. The *Web Crawler* returns OWS resource `URLs` with a "parent" link associated (i.e., a `URL` of the Web document where the OWS resource has been identified). This contextual information is necessary to supply the *informational* search goal. If the OWS resource `URL` is the *item description* or *item*, for example a `GetFeature` operation call, the automatically created *service description* `URL` of the identified service instance will be also associated with the parent resource with the `GetFeature` operation `URL`. In this way, the local repository gathers resources obtained from the past requests and the system searchable content is growing continuously.

The search strategy selected influences the search task. The *none* strategy deactivates the search in the local repository (OWS descriptions, OWS content and parent information), and the returned OWS resources are those identified within the remote SE response. The only difference between the *basic* and *expert* strategies is the manner of extending the user query before passing it to the remote SE, i.e. with "getcapabilities" or "inurl:getcapabilities", respectively. Additionally, when a spatial search is used, it is removed from the request that is passed to the remote SE.

The responses from the remote SE are crawled by Web robots (with a depth factor of 1) which identify and extract OWS resource candidate list for future content retrieval and indexing. The

service description links of the identified OWS resources state the *Remote Candidate List* (RCL). If a spatial restriction is used, the geographical extent defined within the *Capabilities* documents is checked.

At the same time, the search through local repositories produces the *Local Candidate List* of OWS resources (LCL). While searching without any spatial restrictions, only the text indexes are exploited. First, with the consideration of *"site"* and *"inurl"* operators if defined, the text index is used to identify those OWS resource links that match the searching terms. It produces a scored list of OWS candidates. The final OWS list is ordered according to the OWS relevance calculated from the matching score.

In the case of a spatial search, the spatial restriction is applied first: the OWS description repository is searched for all candidates that provide data from the requested area (with consideration of *"site"* and *"inurl"* operators if defined). Then, the term–based search described above is applied for the obtained list of OWS candidates.

The type of the spatial restriction used determines how the OWS candidate list is retrieved. When a point is defined, only this geometry is used in the spatial search. In the case of the *"location"* operation, the toponym is transformed into a list of geometry candidates using a gazetteer, and the user has to chose one of them. The geometry is applied for OWS search. When both spatial restrictions are used, first the *"location"* operator is considered.

The results from the RCL and LCL are merged. The RCL resources that appear in the LCL are removed and the rest of the elements are indexed and scored (the content retrieval is postponed). Then, one result list is created from RCL and LCL according to the obtained score and the links to the OWS resources are annotated with links to the Web pages, where the resources have been found.

Indexing OWS resources

OWS service can be seen as an entrance point to the content of the invisible Web. Ru and Horowitz (2005) have identified challenges which are raised when such content is being indexed:

1. The lack of knowledge of the underlying database schema. Therefore, it is difficult to generate the form assignments that generate information–rich resulting pages.

2. A variety of interfaces of the invisible Web sites (even in the same domain). Therefore it is hard to design a universal form–filling method.

3. The volume of the information in the invisible Web.

In the case of OWS resources, there is a set of known interfaces guided by OGC specifications which focus on technical aspects and detail interfaces and encodings. These interfaces are self–describing by providing the *Capabilities* documents. Additionally, the services which expose content of potential

Service/Resource	Service Description	Element Description	Item
WMS	GetCapabilities		
WCS	GetCapabilities	DescribeCoverage	
WPS	GetCapabilities	DescribeProcess	
WFS	GetCapabilities	DescribeFeatureType	GetFeature
CSW	GetCapabilities		GetRecords

Table 5.5: OWS resources of potential interest to be indexed by the Geospatial Web Search Engine.

interest to be indexed (e.g. WFS), provide data model scheme, and it allows the development of heuristics for automatic procedures.

A specialised module called *OWS crawler* has been developed, which is responsible for indexing the OWS resources. It requests OWS services whose *Capabilities* documents (i.e., the *Service Description*) have been stored in the temporary repository (i.e. a repository which gathers the *Capabilities* documents identified within response of the remote SE). A type of OWS service provides some hints on operations that return the valuable content for indexing. Table 5.5 shows the OWS resources (per an OWS type) that may be considered by the indexing system, and the operation that retrieves them. For example, the *Capabilities* document of any OWS service should be indexed. In the case of WCS, WPS or WFS, the descriptive documents of the requestable elements (i.e. the *Element Description*) should be also retrieved and indexed, and in the case of the WFS service, the `GetFature` operation allows the system to index the served content (i.e. the *Items*).

It is possible to determine the number of requests necessary to be performed in order to index the descriptive documents published by an OWS service (i.e. one `GetCapabilities` call per each service and one `Describe*` call per each requestable element defined within the *Capabilities* document). The text–indexing function ignores all `XML` tags, and all nodes that define geometries or contain numeric data. OWS resources are also indexed spatially via a bounding box defined within the descriptive documents.

Indexing the whole content offered by a WFS service is not efficient. A more appropriate approach can be the creation and indexing of some data summaries, (e.g. a data distribution histogram used in the database community (Ioannidis, 2003)). In this work, the data summaries are created by identifying the *discriminative elements* within the data model, i.e. the *low variable fields* and *high variable fields*. For example, in the case of a gazetteer service, the field that contains an entity name will be a *high variable field* and the field which contains the entity category will be a *low variable field*. The *low variable fields* are the classifying fields. The unique values of these fields can be extracted and indexed. Then, the histograms of these fields can be created. The *high variable fields* are used by a content retrieval procedure to create histograms.

The content retrieval procedure is developed by exploiting the OGC Filter Encoding Standard (Vretanos, 2010a), that allows restrictions to be added on the returned items (or features).

It defines an XML encoding for filter expressions that logically combines the constraints on the properties of an item in order to identify a particular subset of items to be operated upon. For example, it is possible to identify a subset of items by using constraints specified on values of spatial, temporal and/or scalar properties.

First, several test requests are performed and textual fields are distinguished. The fields with the *discriminative elements* are identified via additional validation requests. For example, the values of a *low variable field* change less frequently than the values of the other fields. Therefore, restrictions on this *discriminative element* in the future retrieval requests (i.e. with values already gathered) should return items distinct from those already retrieved. Appendix C presents some details of the method developed for creating and indexing data summaries of the content published by WFS services.

An additional issue is the periodical update of the remote content which is considered by any centralised system based on indexing remote content. In this work, it is assumed that the content published by the remote services is relatively static. The procedures for periodical update should focus on the *Capabilities* documents, and the indexing of the service resources should be repeated only if the service description changes.

5.6.4 Architecture and Implementation

The overview of the architecture proposed for the GWSE system is shown in Figure 5.6. The *RequestManager* is responsible for the support of the user search. The input user request is parsed, which consists of the query validation, cleaning (i.e. trimming and steaming) if it is necessary, the query extension (according to chosen strategy), and then the internal query model is initialised. The SE adapter (the *RemoteSEAdapter*) maps this initialised model into the first–level SE interface and dispatches the query created. The response is stored in the local repository (the *TMPRep*) for future crawling. In a parallel manner, the fast crawling is performed by the *Web Crawler* to generate the RCL as described in Section 5.6.3. The *OWSrSearcher* is responsible for generation of the LCL, which then is integrated with the RCL by the *RequestManager*. Finally, the *FacetController* manages the presentation of the results according to selected facets. In the idle system time, the responses gathered in the *TMPRep* are crawled (the *Web Crawler*) and the OWS resource retrieval and indexing is performed (the *OWS Crawler*).

The system prototype uses Google WebSearch API[25] as the remote general SE. The *"location:<LOC>"* operator is removed from the search request and the toponym (i.e. *"<LOC>"*) is added as an additional term. The libraries under open source licenses have been used for implementation: indexing and textual searches are supported by using Lucene[26], and Hibernate[27] is used to communicate with

[25] http://code.google.com/apis/websearch
[26] http://lucene.apache.org/
[27] http://www.hibernate.org/

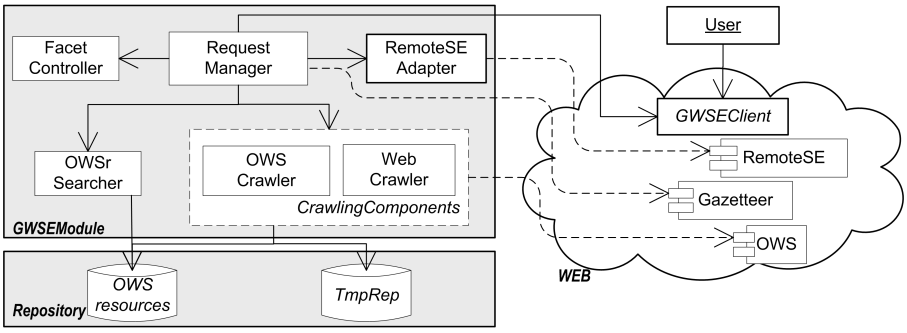

Figure 5.6: Overview of the architecture of the Geospatial Web Search Engine.

PostgreSQL[28] database. These technologies provide support to the required spatial functionality.

5.7 System Evaluation

This Section presents two experiments performed in order to evaluate the prototype developed in this work. The first one is an automated test that compares the result precision of the GWSE prototype with the precision of a selected general SE (i.e. Google Web Search API). The second experiment is dedicated to a human evaluation of the SUI design proposed. The participants have been asked to perform a search task and fill an online questionnaire.

5.7.1 Precision testing

Several test have been performed to compare the response precision of different search strategies for different search goals. The search goals have been defined for purpose of this work according to the characteristics of the searched resources, and Google Web Search API has been used as the remote SE. The NOMGEO database[29] of georeferenced toponyms of Spain has been used to create the local gazetteer for the experiment purpose. This restriction conditioned the way of "location" operation testing, i.e. only place names from Spain could be used. As the *none* strategy deactivates geospatial functionality, the results of this strategy have not been considered. The system gathers information during its work; therefore the evaluation has been performed at the moment when there were about one hundred services indexed in the local repository. The top ten results have been considered for precision estimation, following the common approach from studies on the SE evaluation which usually uses ten or twenty first results (Griesbaum, 2004; Kumar et al., 2005; MacFarlane, 2007; Tawileh et al., 2010).

First, the precision of the Google Web Search Engine has been estimated by evaluating the results from the RCL. The precision has been calculated separately for each search goal. In the case of the

[28] http://www.postgresql.org/
[29] http://www.idee.es/IDEE-WFS-Nomenclator-NG/services

Strategy/Goal	***list***	***interact***	***discovery***
expert	98%	94.6%	94.7%
basic	68%	18%	76%

Table 5.6: Geospatial search precision for different combination of search strategies with search goals.

interaction search goal a hit is the link to a service *Capabilities* document. The *list* search goal is satisfied by a service *Capabilities* document or a page with a link to it. In the case of the *discovery* search goal, the hits from the *list* search goal are extended with all pages which allow a client to find a service *Capabilities* document when navigating from them with a maximum of three hops.

The evaluation queries have been divided into non–spatial queries, spatial queries with a point, and spatial queries with the *"location"* operator. Each of the sets consists of 15 queries. Six test runs have been performed (October 2011) to evaluate precision results when combining search strategies with search goals.

While searching without spatial restriction, the expert strategy precision was high for all search goals, and the basic strategy precision was the worst for the *interaction* search goal (see Table 5.6). Independently from the search goal, the spatial search with a point usually produced an empty result. Analysis of results has shown that the point is treated as text and the engine returns only the textual matches. When using the *"location"* operator, the results produced have the same precision as those obtained when searching without spatial restriction. This behaviour is correct because the value of the spatial operator is added as an additional term when calling the remote SE.

In the second stage of the experiment, the system was always successful (100%) as the system operated on an initialised service repository. As for system performance, the response time increased significantly in the second stage of the experiment, however, this experiment was not focused on this aspect of the system.

5.7.2 Interface Evaluation

Utility and *usability* of a system design are the key quality attributes which determine whether the system is *useful* (Nielsen, 2003). Design utility indicates whether it provides the features aimed, and the second one refers to those properties of the interface that determine how easy it is to use. Nielsen (2003) identifies five components of usability:

- *Learnability.* This quality component informs how easy is it for users to accomplish basic tasks the first time they encounter the interface.
- *Efficiency.* This quality component assesses how quickly users can accomplish their tasks after they learn how to use the interface.

- *Memorability.* This quality component assesses how long it takes users to reestablish proficiency after a period of non–use.
- *Errors.* This quality component assesses how many errors users make, how severe are these errors, and how easy is it for users to recover from these errors.
- *Satisfaction.* This quality component assesses how pleasant or satisfying is it to use the interface.

One of the popular approaches to the evaluation of UI usability is the discount usability testing (Nielsen, 1989), which involves having a small set of evaluators who examines the interface and judges its compliance with recognised usability principles. The SUI design presented in this work focuses on the information model for search and browsing of Geospatial Web resources. Therefore, design utility should be the target of this evaluation.

The principal goal of the SUI design developed is the support to non–expert users, i.e. the users that lack specific knowledge on geographic resources. Therefore, experts in the geospatial domain and non–expert users should be able to perform a search task successfully. Additionally, the interface design applied favours exploratory search which should breed an effect of learning about geospatial resources. As for evaluation of usability, it might be expected that following known and successful approaches from the Web should assure system usability. Here, the successful execution of a search task by a user will be considered as a positive rate of learnability and efficiency in some degree. User satisfaction will be considered in this work as well.

An online questionnaire has been created to evaluate the GWSE prototype, and a group of Web users have been invited via emails to participate in it. Since the target participants should consist of expert and non–expert users, some efforts have been made to encourage researchers in the Geospatial domain and out of it to participate. The survey has three parts as follows:

- *Pre–search questionnaire.* This part of survey gathers some general information on participants (e.g., age, gender, country, education level, etc.). It also allows the evaluation the user's prior knowledge on geographic information resources (in order to classify the user as an *expert* or *non–expert*), and a potential interest in these resources.
- *Search–task questionnaire.* The participants are asked to perform a search for geographic information resources. The required task involves the three potential search goals (i.e., *informational*, *transactional*, *navigational*). The search task has to be performed by using the prototype application (*GWSE*) or the Google Search Engine (*GoogleSE*). The type of the SE is assigned in order to maintain the proportion between the SE used (i.e. *GWSE* and *GoogleSE* users) and also considering the classification of the user.
- *Post–search questionnaire.* This part of survey allows participants to express their opinion on the tool used and satisfaction with the task result. It also attempts to estimate the learning

User Perfil/Search Engine	**Expert**		**Non–expert**	
	Completion	*Error*	*Completion*	*Error*
GoogleSE	**6**	4	**7**	4
GWSE	**7**	5	**7**	3

Table 5.7: Classification of the survey participants. *Completion* and *Error* columns indicate the number of questionnaires which have been completed properly or not, respectively.

Search Task Result/Search Engine	**Success**		**Failure**	
	Expert	*Non–expert*	*Expert*	*Non–expert*
GoogleSE	5 (83.3%)	0 (0%)	1 (16.7%)	7 (100%)
GWSE	7 (100%)	5 (71.4%)	0 (0%)	2 (28.6%)

Table 5.8: The survey results.

effect of the exploratory search.

The search system developed is characterised by a focus on the Geospatial domain with strong bias towards OWS resources. Therefore, only some aspects of the recommendations on exploratory search task design provided in Kules and Capra (2008) have been employed. A participant has been asked to imagine himself/herself to be a researcher who is going to write a report on the fresh water supplies in the Mediterranean region, including information on droughts and floods if it is possible. This introduction also includes a brief overview of targeted resources (i.e. what is a Geospatial Web service and types of OWS services). The search task which should be performed consists in gathering necessary information for the report: (1) some Web sites which publish geospatial resources on the subject should be found, (2) the information resources should not be in a form of a map only (i.e. some examples of resources of processable data should be given), and published via services are preferable. The services which offer the most useful information should be identified among the found services and the reasons for the selection should be given. The homepages of providers of the selected services should be identified if possible as well. The task has been restricted by time limit (i.e. 20 minutes). At the end, the test participants have been asked to rate each search engine in terms of the intervals: Terrible–Wonderful, Frustrating–Satisfying, Dull–Stimulating, Confusing–Clear and Rigid–Flexible, which are based on the Questionnaire for User Interface Satisfaction (QUIS) for the subjective evaluation measures (Chin et al., 1988). Additionally, the participants were asked to evaluate the increase of their knowledge about Geospatial Web resources.

The survey ran in January 2012. In total, 43 participants have been invited to participate. Only the well–completed questionnaires have been considered and the questionnaires which have not been finished (e.g. abandoned) have been removed from the further analysis. Table 5.7 summarises characteristics of the participants according to the declared knowledge on geospatial resources and

the type of search engine assigned to carry out the search task. Additionally, the distribution of the questionnaires which were filled properly and the incomplete ones is shown.

The search task has been accomplished successfully if the participant has found (1) several Web pages related to the search task topic that publish geospatial services or datasets, (2) at least one appropriate and working service for each service type required (i.e., a map and text–based geospatial information), and (3) homepages of providers of the selected services. The survey results are summarised in Table 5.8. The analysis has shown that only experts in the geospatial domain were able to realise the task favourably by means of the *GoogleSE*. The main problems found by the non–experts which used the *GoogleSE* were the identification of OWS services and the interaction with them (e.g. to test them). The *GWSE* users completed their task successfully in most cases. Only two non–expert users had some problems finding an example of service for each required type. When it comes to reasons for choice of OWS services, some of the non–experts from the *GoogleSE* user group provided information found on Web pages and even information offered by the *Capabilities* documents (only few examples). The experts from the *GoogleSE* user group focused mainly on the *Capabilities* documents, and some of them also interacted with services in order to gather more information. But they acknowledged that the interaction with a service was time–consuming and even annoying. The *GWSE* users took advantage of a variety of information offered by the SE, i.e. the *Capabilities* documents, the other services' responses, and summaries of service datasets (if any). Overall, the *GWSE* users were more satisfied with the results, and the non–expert users felt that they learned about geospatial resources and how to use them. In terms of QUIS, *GWSE* was rated higher than *GoogleSE*.

The survey results show that even non–expert users can perform a satisfactory search and assessment of geospatial resources using *GWSE*. Experts still can be successful with the use of *GoogleSE* but as some of them pointed out, it is sometimes annoying to interact with a service without a tool that conceals the painful creation of requests. The main drawback of *GWSE* identified by participants is the lack of a map to search and browse resources. There are some works which offer faceted data browsers combined with maps (see (Auer et al., 2009b) for instance), but in the present work the spatial representation of the resource may vary (e.g. a point, a bounding box, a multi–polygon). In this scenario, the creation of a proper faceted search engine with spatial map–based search capacity is not trivial. This requires a combination of faceted search with map–based visualisation in the manner which ensures a consistent presentation of results and coherent user interaction with the application.

5.8 Summary

The shift to the Web platform permits support for new requirements coming from future community–based GCIs that need to perform cross–domain e–science (Pierce et al., 2009). One of the important

issues in promoting new e–science is information and knowledge sharing (Longueville, 2010). However, searching for Geospatial Web resources is not trivial for non–expert users. This Chapter has presented a Geospatial Web Search Engine that allows non–expert users to search for geospatial resources in the Web. The system can operate over any existing general search engine. It uses the results of a remote search engine to collect these links that lead to resources published by the OGC Web services. The Web cover of the system increases with time because the found OWS resources are stored locally for use in future queries. The system enables spatial search using a point or a toponym, which is converted to the corresponding footprint. Empirical study has shown that, in the beginning, the search precision has been at least as high as the precision of a remote search engine. When the system repository was initialised, the precision of the system improved. In addition, a survey has been conducted to assess the utility of the interface design proposed for GWSE. This study has indicated that even non–expert users can perform a search task satisfactorily.

The presented system can be seen as an integration point in the Geospatial Web, because it assists in searching for geospatial resources distributed on the Web. In addition, it allows the creation a virtual meeting place for publishers of geospatial resources and potential users (and customers). If the found resource does not meet user requirements, the user may contact the supplier in order to obtain information on alternative resources.

Chapter 6

Conclusions

6.1 Summary of Contributions

The work developed in this thesis has shown that it is possible to improve search in the context of SDI by applying practices from other communities, especially the Web and Semantic communities. The semantics and content–based approaches can help in searching for information about geographic features, and in searching for geospatial resources in general. This fact has some important implications for SDI. First of all, the principle of meaningfulness should be applied at all levels of SDI development and implementation. For example, the vocabulary used in schema of a geographic feature published by WFS should catch the model followed. Additionally, semantic technologies (e.g. Linked Data) could be relevant in cost reduction. As for content analysis, resource providers should consider the need for test and / or sampling procedures. This work shows also that good practice for metadata creation should be cultivated when creating Web geoportals. It might be recommended to the SDI community to consider Search engine optimisation guidelines to improve the discoverability of published resources on the Web.

Following the structure of this memory, the most relevant contributions are presented below.

- Chapter 2 has described an approach to **an enhanced search for geospatial entities** from the perspective of traditional geocoding. The *compound geocoding architecture* proposed in this work ensures the improvement of geocoding results thanks to the use of different geographic information suppliers. In this approach, structural design patterns have been used for service integration. The usage of ontologies has yielded an advanced architecture in terms of extensibility, flexibility and adaptability. The framework for geocoding service selection allows the development of a methodology to geocode diverse categories of geospatial data (e.g. geographic features, points of interest), which is an essential functionality of a geolocating service. The results of this study have been the subject of several research publications (Florczyk et al.,

2008, 2009b,a,c; Pérez-Pérez et al., 2009; Florczyk et al., 2010c).

- Chapter 3 has presented two representative applications which require an **additional semantic characterisation of geospatial resources**. The approach proposed in this work uses content–based heuristics for dataset sampling. The first part has introduced the idea of abstraction of a geographic feature from its spatial definition. It has shown how best practices from the Semantic Web can be used to describe OWS services by means of **geoidentifiers** (i.e., entities from a geographic ontology), which has allowed access to different representations of a geographic feature in a flexible manner. This approach has required the development of a content–based heuristic in order to extract and create the additional semantics. This research and related contributions have resulted in several publications (López-Pellicer et al., 2008; Florczyk et al., 2010a,b; López-Pellicer et al., 2011f).

  The second part of Chapter 3 is dedicated to providing a geoprocessing service for the automatic identification of orthoimages offered through a WMS service. A method for the identification of a **WMS orthoimage layer** has been proposed. The experiment running on a realistic corpus has proved the efficiency of the proposed method (87 % precision and 60 % recall). The image catalogue resulted from filtering crawler outcomes has been applied in the *Virtual Spain* project. The results of this study have been published in Florczyk et al. (2011).

- Chapter 4 has presented analysis of issues related to the creation of metadata for Web resources in the context of the Geographic domain. An architecture for **automatic generation of geographic knowledge of Web resources** is proposed. Since geographic metadata are hardly used in Web pages, including those from the Geospatial domain, content–based heuristics for geographic coverage estimation of Web pages has been proposed. The prototype developed generates metadata whose model holds the recommended minimal set of elements required by an OGC catalogue. Additionally, the model encompasses some provenance information regarding the estimated coverage which can be useful for accuracy evaluation. An experiment proves the applicability of the system in the realistic Web environment. This study determines some **characteristics of the current Geospatial Web**. First of all, it offers some characteristics of the publishing market. Geospatial Web resources can be found mainly in geoportals and general portals (63.9%). Also, they can be found in the Web sites of companies, research groups, communities and personal Web pages (frequently as demo resources). Most of the Web sites have been classified as *local* (49%) or *national* (31%). The usage of English language dominates in the Web sites of *regional* and *global* geographic scope (i.e. 87.7%) . Also, this study uncovers some practices in the Geospatial community in providing metadata for Web pages, i.e. the lack of geographic metadata in particular. The results of this study have been published in Borjas et al. (2011a) and Borjas et al. (2011b).

- One of the important issues in promoting new e–science is information and knowledge sharing (Longueville, 2010). The shift to the Web platform permits support for new requirements coming from future community–based Geospatial Cyberinfrastructures that need to perform cross–domain e–science (Pierce et al., 2009). However, searching for resources of the Geospatial Web is not trivial for non–expert users. Chapter 5 has examined the issue of **supporting non–expert users** in searching for Geospatial Web resources. The Geospatial Web Search Engine proposed in this work can use an existing search engine and supports the exploratory search for geospatial resources in the Web. The experiment on precision and recall has shown that the prototype developed in this work is at least as good as the remote search engine. Additionally, a survey, dedicated to the system utility, indicates that even non–expert users can perform a search task with satisfactory results. This study has contributed in several research lines, i.e. the discovery of OWS services in the Web (López-Pellicer et al., 2011e,b, 2010a, 2011a,b, 2010b), the semantic description of OWS services and republishing OWS services on the Web (Florczyk et al., 2010d; López-Pellicer et al., 2010c, 2011d,c).

6.2 Future Work

Each research line presented in this work gives an opportunity for future research. Following, a set of the most immediate research goals are presented.

1. **Compound Geocoding**. One improvement of the presented system should be the incorporation of advances in Web service interoperability in order to help in the automatic discovery and use of the geographic data providers. The Semantic Web Service community offers a variety of technological standards for semantic description of services. An effort should be made to create the formal definition of the ontology for geo–service description that includes the features enumerated in this work. This will give bases to apply the ontology reasoning for improvement of the service selection.

2. **Geoidentifiers**. One of the future tasks should be the application of the *administrative geography* as a guideline to map instances of the same geo–concept entities between two different gazetteers for the purpose of merging. Another research issue is an adaptable framework to support complex spatial requests applying different spatial representations of features.

3. **Image layers**. In the case of the content–based heuristics for image layer estimation the research should be focused on improving the presented algorithm, especially in terms of recall. The algorithm for collecting a representative set of fragments of the image layer should be improved to reduce the number of layers for which an invalid collection has been obtained. For example, the analysis of other layers offered by the same service might help estimate the minimal presentation scale and bounding box. The support for other languages of the *Capabilities*

documents should be added as till now only the Spanish language has been considered. This will require a training set of *Capabilities* documents in other languages. Then, the detection of other layer types should be considered as well: for example, vector layers and coverage data layers. The lessons learned can help in the research of the same kind of analysis that might be applied to WCS services.

4. **Knowledge generator**. The system developed in this work is the first step to the creation of a tool which will be capable to geospatially characterise Web resources by analysing contextual information provided by related Web pages (for example, KML (Wilson et al., 2008)). This context–based approach may be also useful to improve existing SDI resource metadata, for example, the metadata offered by an OGC Web Service *Capabilities* document. The prototype can be a base for the development of a tool for the validation and the improvement of OWS *Capabilities* documents. What is more, this work might be a starting point for the development of a heuristic framework for the automatic classification of Web sites which publish Geospatial Web resources. First of all, however, the improvement of the coverage estimation method should be investigated, especially in terms of the identification of a *regional* or *global* scope of a Web page.

5. **Geospatial Web search engine**. The main work for the future is identification of the proper approach to offer spatial search using a map from the Human–Computer Interaction perspective. It will require some research in combining faceted search and map–based visualisation to offer consistent views of the results. In this way, other presentation styles of search results will be enabled (as required by the sixth design principle proposed in Kules (2006): "Support multiple types of categories and visual presentations"). Additionally, the other spatial dataset summaries should be investigated to improve ranking function, for example: those proposed in Hariharan et al. (2008). Another issue under consideration should be multilingual support. Semantic support can improve the system effectiveness in case of thematic search. Also, other popular geospatial resources should be under consideration, for example KML or shape files, which can be found via a general search engine. Adding a new resource type requires the development of a new SE strategy for the resource type recognition instead of the link pattern approach applied in this work, and the adjustment of the search interface and facets. Another aspect worthy of investigation is the integration of the Web 2.0 approaches into the current version of the system. Also the service monitoring framework should be considered. In this work, some monitoring information has been used, however, such a system should be disconnected from the search engine, and the interaction between those two systems should be analysed.

Appendix A

KnowledgeGenerator: Prototype details

This appendix contains details of the "Knowledge Generator" prototype implemented. Table A.1 summarises the possible extraction methods which have been described in the revised literature or proposed in this work, and it points out the metadata elements which could be filled be means of these methods. The extractors are identified via codes, and some rules on the extractor chains are proposed. The extractors are identified via aliases in the table, and their names are as follows:

- *A* (*<META>*). This extraction method analyses metadata of the *header* element. It extracts the *name*, *http-equiv* and *content* attributes. Additionally, it extracts value of the *title* element (Mattmann and Zitting, 2011).

- *B* (*N–Gram*). This method is used to identify the language of text. The algorithm is based on the frequency of words and characters (Mattmann and Zitting, 2011; Cavnar and Trenkle, 1994).

- *C* (*<H1>*). This method extracts the content of the *H1* elements of an HTML document (Paynter, 2005).

- *D* (*D. Text50*). This method extracts the first 50 characters of the *body* element of an HTML document (Paynter, 2005).

- *E* (*PhraseRate*). This method extracts the keywords (two or five) from an HTML document, and gives good results for domain–oriented and well written text (Humphreys, 2002).

- *F* (*AutoAnnotator*). This method extracts a simple paragraph which summarises the content of an HTML document (Kedzierski, 2002).

- *G* (*NER*). This method extracts toponyms from the content of an HTML document, or from a free text (Finkel et al., 2005).
- *H* (<*IMG*>*NER*). This method extracts toponyms from the alternative text (i.e. *alt*) of the *IMG* element of an HTML document.
- *I* (<*A*>*NER*). This method extracts toponyms from the visible text of links found within the content of an HTML document.
- *J* (*License*). This method extracts the copyright links which are explicitly declared (i.e. the attribute *rel* has "Copyright" or "Licence" value) (Abelson et al., 2008).
- *L* (*LCSH*). This method extracts the LCSH (Library Congress Subject Heading) classification which is based on the Naive Bayes algorithm (Mitchell et al., 2003).
- *L* (<*BODY*>*NER*). This method extracts toponyms from the content of an HTML document from which all links and images have been removed previously.
- *M* (<*META*>*NER*). This method extracts toponyms from the metadata of an HTML document.

Metadata/ Extractor	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**J**	**K**	**L**	**M**	**Rule Used**
contributor	x													
coverage	x						x	x	x			x	x	$\neg A \rightarrow M$ $\neg M \rightarrow I$ $\neg I \rightarrow H$ $\neg H \rightarrow L$
creator	x													
modified	x													*Default value*
descritpion	x					x								
format	x													
identifier	x													*Default value*
format	x													
language	x													$\neg A \rightarrow B$ (*Tika*: $B \subset A$
publisher	x													$\neg A \rightarrow G$
relation	x													*Default value*
rights	x									x				$\neg A \rightarrow J$
source	x													*Default value*
subject	x				x						x			$\neg K \rightarrow E$ $\neg E \rightarrow A$
title	x		x	x										$\neg A \rightarrow C$ $\neg C \rightarrow D$
type	x													*Default value*

Table A.1: Summary of the existing methods for the metadata extraction and proposed extracting rules.

Appendix B

SKOS classification schemes for faceted–search.

This appendix contains SKOS vocabularies of the developed classification schemes used to support the faceted search and browsing interface. The vocabularies presented here have been generated by the repository of the GWSE system. The vocabulary for the OWS service Taxonomy and the OWS resources have been initially created in ThManager (Lacasta et al., 2007) and imported into the system repository. The vocabulary for the OWS service Taxonomy has been extended automatically by the GWSE with new interfaces identified in the *Capabilities* documents of crawled OWS instances. The vocabularies for the domains of OWS services and the providers of OWS services have also been created by the GWSE as a result of the discovery of new resources.

Listing B.1: SKOS vocabulary for the OWS service Taxonomy.

```
1  @prefix skos: <http://www.w3c.org/2004/02/skos/core#> .
2  @prefix rdf: <http://www.w3c.org/1999/02/22-rdf-syntax-ns#> .
3  @prefix dct: <http://purl.org/dc/terms/> .
4  @prefix ows: <http://example.com/gwse/ows-taxonomy/> .
5
6  ows:owsTaxonomy rdf:type skos:ConceptScheme;
7   dct:title ‘‘Taxonomy of OWS services’’;
8
9  ows:DataAccessService rdf:type skos:Concept;
10  skos:prefLabel ‘‘Data Access Service’’@en;
11  skos:inScheme ows:owsTaxonomy.
12
13 ows:CatalogService rdf:type skos:Concept;
14  skos:prefLabel ‘‘Catalog/Registry Service’’@en;
```

```
 skos:inScheme ows:owsTaxonomy.

ows:PortayalService rdf:type skos:Concept;
 skos:prefLabel ‘‘Portayal and Display Services’’@en;
 skos:inScheme ows:owsTaxonomy.

ows:DataTransformationService rdf:type skos:Concept;
 skos:prefLabel ‘‘Data Transformation Service’’@en;
 skos:inScheme ows:owsTaxonomy.


ows:WFSS rdf:type skos:Concept;
 skos:prefLabel ‘‘OGC Web Feature Simple Service’’@en;
 skos:inScheme ows:owsTaxonomy;
 skos:broader ows:DataAccessService;
 dct:identifier ‘‘urn:ogc:serviceType:WFSS’’.

ows:WFSS100 rdf:type skos:Concept;
 skos:prefLabel ‘‘OGC Web Feature Simple Service, version 1.0.0’’@en;
 skos:inScheme ows:owsTaxonomy;
 skos:broader ows:WFSS;
 dct:identifier ‘‘urn:ogc:serviceType:WFSS:1.0.0’’.

ows:WFS rdf:type skos:Concept;
 skos:prefLabel ‘‘OGC Web Feature Service’’@en;
 skos:inScheme ows:owsTaxonomy;
 skos:broader ows:DataAccessService;
 dct:identifier ‘‘urn:ogc:serviceType:WFS’’.

ows:WFS100 rdf:type skos:Concept;
 skos:prefLabel ‘‘OGC Web Feature Service, version 1.0.0’’@en;
 skos:inScheme ows:owsTaxonomy;
 skos:broader ows:WFS;
 dct:identifier ‘‘urn:ogc:serviceType:WFS:1.0.0’’.

ows:WFS110 rdf:type skos:Concept;
 skos:prefLabel ‘‘OGC Web Feature Service, version 1.1.0’’@en;
 skos:inScheme ows:owsTaxonomy;
 skos:broader ows:WFS;
 dct:identifier ‘‘urn:ogc:serviceType:WFS:1.1.0’’.

```

```
ows:WFS200 rdf:type skos:Concept;
 skos:prefLabel ‘‘OGC Web Feature Service, version 2.0.0’’@en;
 skos:inScheme ows:owsTaxonomy;
 skos:broader ows:WFS;
 dct:identifier ‘‘urn:ogc:serviceType:WFS:2.0.0’’.

ows:WCS rdf:type skos:Concept;
 skos:prefLabel ‘‘OGC Web Coverage Service’’@en;
 skos:inScheme ows:owsTaxonomy;
 skos:broader ows:DataAccessService;
 dct:identifier ‘‘urn:ogc:serviceType:WCS’’.

ows:WCS100 rdf:type skos:Concept;
 skos:prefLabel ‘‘OGC Web Coverage Service, version 1.0.0’’@en;
 skos:inScheme ows:owsTaxonomy;
 skos:broader ows:WCS;
 dct:identifier ‘‘urn:ogc:serviceType:WCS:1.0.0’’.

ows:WCS102 rdf:type skos:Concept;
 skos:prefLabel ‘‘OGC Web Coverage Service, version 1.0.2’’@en;
 skos:inScheme ows:owsTaxonomy;
 skos:broader ows:WCS;
 dct:identifier ‘‘urn:ogc:serviceType:WCS:1.0.2’’.

ows:WCS110 rdf:type skos:Concept;
 skos:prefLabel ‘‘OGC Web Coverage Service, version 1.1.0’’@en;
 skos:inScheme ows:owsTaxonomy;
 skos:broader ows:WCS;
 dct:identifier ‘‘urn:ogc:serviceType:WCS:1.1.0’’.

ows:WCS111 rdf:type skos:Concept;
 skos:prefLabel ‘‘OGC Web Coverage Service, version 1.1.1’’@en;
 skos:inScheme ows:owsTaxonomy;
 skos:broader ows:WCS;
 dct:identifier ‘‘urn:ogc:serviceType:WCS:1.1.1’’.

ows:WCS112 rdf:type skos:Concept;
 skos:prefLabel ‘‘OGC Web Coverage Service, version 1.1.2’’@en;
 skos:inScheme ows:owsTaxonomy;
 skos:broader ows:WCS;
 dct:identifier ‘‘urn:ogc:serviceType:WCS:1.1.2’’.
```

```
97
98  ows:WMS rdf:type skos:Concept;
99   skos:prefLabel ‘‘OGC Web Map Service’’@en;
100  skos:inScheme ows:owsTaxonomy;
101  skos:broader ows:PortayalService;
102  dct:identifier ‘‘urn:ogc:serviceType:WMS’’.
103
104 ows:WMS100 rdf:type skos:Concept;
105  skos:prefLabel ‘‘OGC Web Map Service, version 1.0.0’’@en;
106  skos:inScheme ows:owsTaxonomy;
107  skos:broader ows:WMS;
108  dct:identifier ‘‘urn:ogc:serviceType:WMS:1.0.0’’.
109
110 ows:WMS101 rdf:type skos:Concept;
111  skos:prefLabel ‘‘OGC Web Map Service, version 1.0.1’’@en;
112  skos:broader ows:WMS;
113  dct:identifier ‘‘urn:ogc:serviceType:WMS:1.0.1’’.
114
115 ows:WMS107 rdf:type skos:Concept;
116  skos:prefLabel ‘‘OGC Web Map Service, version 1.0.7’’@en;
117  skos:inScheme ows:owsTaxonomy;
118  skos:broader ows:WMS;
119  dct:identifier ‘‘urn:ogc:serviceType:WMS:1.0.7’’.
120
121 ows:WMS110 rdf:type skos:Concept;
122  skos:prefLabel ‘‘OGC Web Map Service, version 1.1.0’’@en;
123  skos:inScheme ows:owsTaxonomy;
124  skos:broader ows:WMS;
125  dct:identifier ‘‘urn:ogc:serviceType:WMS:1.1.0’’.
126
127 ows:WMS111 rdf:type skos:Concept;
128  skos:prefLabel ‘‘OGC Web Map Service, version 1.1.1’’@en;
129  skos:inScheme ows:owsTaxonomy;
130  skos:broader ows:WMS;
131  dct:identifier ‘‘urn:ogc:serviceType:WMS:1.1.1’’.
132
133 ows:WMS113 rdf:type skos:Concept;
134  skos:prefLabel ‘‘OGC Web Map Service, version 1.1.3’’@en;
135  skos:inScheme ows:owsTaxonomy;
136  skos:broader ows:WMS;
137  dct:identifier ‘‘urn:ogc:serviceType:WMS:1.1.3’’.
```

```
138
139 ows:WMS130 rdf:type skos:Concept;
140  skos:prefLabel ‘‘OGC Web Map Service, version 1.3.0’’@en;
141  skos:inScheme ows:owsTaxonomy;
142  skos:broader ows:WMS;
143  dct:identifier ‘‘urn:ogc:serviceType:WMS:1.3.0’’.
144
145 ows:WMS132 rdf:type skos:Concept;
146  skos:prefLabel ‘‘OGC Web Map Service, version 1.3.2’’@en;
147  skos:inScheme ows:owsTaxonomy;
148  skos:broader ows:WMS;
149  dct:identifier ‘‘urn:ogc:serviceType:WMS:1.3.2’’.
150
151 ows:WMTS rdf:type skos:Concept;
152  skos:prefLabel ‘‘OGC Web Map Transactionl Service’’@en;
153  skos:inScheme ows:owsTaxonomy;
154  skos:broader ows:PortayalService;
155  dct:identifier ‘‘urn:ogc:serviceType:WMTS’’.
156
157 ows:WMTS100 rdf:type skos:Concept;
158  skos:prefLabel ‘‘OGC Web Map Transactional Service, version 1.0.0’’@en;
159  skos:inScheme ows:owsTaxonomy;
160  skos:broader ows:WMTS;
161  dct:identifier ‘‘urn:ogc:serviceType:WMTS:1.0.0’’.
162
163 ows:WCTS rdf:type skos:Concept;
164  skos:prefLabel ‘‘Web Coordinate Transformation Service’’@en;
165  skos:inScheme ows:owsTaxonomy;
166  skos:broader ows:DataTransformationService;
167  dct:identifier ‘‘urn:ogc:serviceType:WCTS’’.
168
169 ows:WCTS002 rdf:type skos:Concept;
170  skos:prefLabel ‘‘Web Coordinate Transformation Service, version 0.2.2’’@en;
171  skos:inScheme ows:owsTaxonomy;
172  skos:broader ows:WCTS;
173  dct:identifier ‘‘urn:ogc:serviceType:WCTS:0.2.2’’.
174
175 ows:WCTS030 rdf:type skos:Concept;
176  skos:prefLabel ‘‘Web Coordinate Transformation Service, version 0.3.0’’@en;
177  skos:inScheme ows:owsTaxonomy;
178  skos:broader ows:WCTS;
```

```
 dct:identifier ‘‘urn:ogc:serviceType:WCTS:0.3.0’’.

ows:WCTS100 rdf:type skos:Concept;
 skos:prefLabel ‘‘Web Coordinate Transformation Service, version 1.0.0’’@en;
 skos:inScheme ows:owsTaxonomy;
 skos:broader ows:WCTS;
 dct:identifier ‘‘urn:ogc:serviceType:WCTS:1.0.0’’.

ows:CSW rdf:type skos:Concept;
 skos:prefLabel ‘‘OGC Catalogue Service’’@en;
 skos:inScheme ows:owsTaxonomy;
 skos:broader ows:CatalogService;
 dct:identifier ‘‘urn:ogc:serviceType:CSW’’.

ows:CSW100 rdf:type skos:Concept;
 skos:prefLabel ‘‘OGC Catalogue Service, version 1.0.0’’@en;
 skos:inScheme ows:owsTaxonomy;
 skos:broader ows:CSW;
 dct:identifier ‘‘urn:ogc:serviceType:CSW:1.0.0’’.

ows:CSW101 rdf:type skos:Concept;
 skos:prefLabel ‘‘OGC Catalogue Service, version 1.0.1’’@en;
 skos:inScheme ows:owsTaxonomy;
 skos:broader ows:CSW;
 dct:identifier ‘‘urn:ogc:serviceType:CSW:1.0.1’’.

ows:CSW200 rdf:type skos:Concept;
 skos:prefLabel ‘‘OGC Catalogue Service, version 2.0.0’’@en;
 skos:inScheme ows:owsTaxonomy;
 skos:broader ows:CSW;
 dct:identifier ‘‘urn:ogc:serviceType:CSW:2.0.0’’.

ows:CSW201 rdf:type skos:Concept;
 skos:prefLabel ‘‘OGC Catalogue Service, version 2.0.1’’@en;
 skos:inScheme ows:owsTaxonomy;
 skos:broader ows:CSW;
 dct:identifier ‘‘urn:ogc:serviceType:CSW:2.0.1’’.

ows:CSW202 rdf:type skos:Concept;
 skos:prefLabel ‘‘OGC Catalogue Service, version 2.0.2’’@en;
 skos:inScheme ows:owsTaxonomy;
```

```
220  skos:broader ows:CSW;
221  dct:identifier ‘‘urn:ogc:serviceType:CSW:2.0.2’’.
```

Listing B.2: SKOS vocabulary for the OWS resources.

```
1  @prefix skos: <http://www.w3c.org/2004/02/skos/core#> .
2  @prefix rdf: <http://www.w3c.org/1999/02/22-rdf-syntax-ns#> .
3  @prefix dct: <http://purl.org/dc/terms/> .
4  @prefix ows: <http://example.com/gwse/ows-resource/> .
5
6  ows:owsResource rdf:type skos:ConceptScheme;
7   dct:title ‘‘OWS resources’’;
8
9  ows:service rdf:type skos:Concept;
10  skos:prefLabel ‘‘OWS Service’’@en;
11  skos:inScheme ows:owsResources;
12  dct:identifier ‘‘urn:gwse:resource:service’’.
13
14 ows:element rdf:type skos:Concept;
15  skos:prefLabel ‘‘Requestable element of an OWS Service’’@en;
16  skos:inScheme ows:owsResources;
17  dct:identifier ‘‘urn:gwse:resource:element’’.
18
19 ows:item rdf:type skos:Concept;
20  skos:prefLabel ‘‘Information item retrieved from an OWS Service’’@en;
21  skos:inScheme ows:owsResources;
22  dct:identifier ‘‘urn:gwse:resource:item’’.
```

Listing B.3: SKOS vocabulary for the domains of OWS services (an example).

```
1  @prefix skos: <http://www.w3c.org/2004/02/skos/core#> .
2  @prefix rdf: <http://www.w3c.org/1999/02/22-rdf-syntax-ns#> .
3  @prefix dct: <http://purl.org/dc/terms/> .
4  @prefix foaf: <http://xmlns.com/foaf/0.1/> .
5  @prefix ows: <http://example.com/gwse/ows-domain/> .
6
7  ows:owsDomain rdf:type skos:ConceptScheme;
8   dct:title ‘‘Domain of OWS services’’;
9
10 ows:sgiisprambienteit rdf:type skos:Concept;
11  skos:prefLabel ‘‘Domain: sgi.isprambiente.it’’;
12  dct:source ‘‘http://sgi.isprambiente.it/geoportal/csw/discovery?service=csw’’;
```

```
 skos:inScheme ows:owsDomain;
 dct:date ‘‘2011-10-05’’;
 dct:identifier ‘‘sgi.isprambiente.it’’.

ows:mesonetagroniastateedu rdf:type skos:Concept;
 skos:prefLabel ‘‘Domain: mesonet.agron.iastate.edu’’;
 dct:source ‘‘http://mesonet.agron.iastate.edu/cgi-bin/wms/nexrad/ntp.cgi?
     service=wfs’’;
 dct:date ‘‘2011-09-07’’;
 skos:inScheme ows:owsDomain;
 dct:identifier ‘‘mesonet.agron.iastate.edu’’.
```

Listing B.4: SKOS vocabulary for the providers of OWS services (an example).

```
@prefix skos: <http://www.w3c.org/2004/02/skos/core#> .
@prefix rdf: <http://www.w3c.org/1999/02/22-rdf-syntax-ns#> .
@prefix dct: <http://purl.org/dc/terms/> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix ows: <http://example.com/gwse/ows-provider/> .

ows:owsProvider rdf:type skos:ConceptScheme;
 dct:title ‘‘Providers of OWS services’’;

ows:USGeologicalSurveyEasternMineralResourcesTeam rdf:type skos:Concept;
 skos:inScheme ows:owsProvider;
 skos:prefLabel ‘‘U.S. Geological Survey Eastern Mineral Resources Team’’;
 dct:source ‘‘http://mrdata.usgs.gov/services/nuresed?service=wfs’’;
 dct:date ‘‘2012-01-09’’;
 dct:identifier ‘‘U.S. Geological Survey Eastern Mineral Resources Team’’.

<http://minerals.usgs.gov/east/> rdf:type foaf:Document;
 dct:subject ows:USGeologicalSurveyEasternMineralResourcesTeam;
 skos:note ‘‘It is result of an automatised process and the association might be
      erroneous.’’;
 dct:date ‘‘2012-01-14’’;
 dct:source ‘‘Google Search Engine’’.
```

Appendix C

Generation of OWS dataset summaries.

This appendix contains some details on the retrieval–indexing procedure related to the creation of the dataset summaries. In general, the procedure varies in function of the service type. Here, the heuristic which handles WFS service is explained because it is the only service whose content is considered.

The procedure receives a `URL` of the `GetCapabilities` request. The *Capabilities* document details the implemented interface and lists the published features. In function of the service type some additional queries are created. Only the requestable features are considered further. The MBBOX which is declared within the *Capabilities* document is verified, and the corresponding administrative unit is identified if possible (i.e. geoidentifier). The `DescribeFeatureType` query is created for each feature to get the `XML` Schema of the used data model. The information on element cardinality is especially useful later. Then, a sample set of items is retrieved using the `GetFeature` query with a spatial restriction. This heuristic aims to sample items within the whole geographic extent declared in the *Capabilities* document. The number of items which shall be gathered should depend on the total number of items in the dataset. This information can be obtained by means of the `resultType=hits` parameter of the `GetFeature` operation which only returns the number of items which satisfy the query. A limited number of items are gathered per request using the `Maxfeatures` parameter within a query. Duplications are detected and removed. Additional queries may be performed to gather new items if it is necessary. The gathered examples are analysed. First, a list of the requestable properties is inferred from the examples and the `XML` Schema. Then, the properties are tested. The total number of items which use them is requested (i.e. the number of items that provide non empty value for these properties). In this way, the heuristic identifies the properties provided typically. Those never used, or not frequently provided are not considered further. Similarly, the properties

that contain spatial object or a numeric value are not considered either. The cardinality of properties is estimated within the data model of the requested feature. Next, unique values from the example set are assigned to the properties, and for each the number of the item is requested. In this way, the *low variable fields* and the *high variable fields* are distinguished. Next, series of data retrieval queries are performed to identify other values of the *low variable fields*. The filter encoding is used to restrict values which have been already found. The new items are also parsed in order to find new values for the other properties. If a new value of an analysed property is found, it is necessary to perform additional analysis. It should be noted that not all values might be identified. However, it might be assumed that those of the highest frequency will be found.

During the analysis, additional information is gathered which characterise how geospatial information is encoded, i.e. the content bounding box (which is shown on a map), the type of spatial object which is offered within the model (e.g. "Point"), and the spatial reference used by default (i.e. EPSG codes).

The method developed has been applied to OGC WFS services found by the GWSE. Here, an example of the performed analysis is presented. It is a result of the analysis and indexing of a gazetteer service deployed within the Spanish SDI, named *Nomenclátor Geográfico Conciso de España*[1] (NGCE). The NGCE service conforms to the OGC WFS specification. Listing C.1 shows some requests performed by the method. The first request (1) retrieves the *Capabilities* document of the NGCE service. The document details the implemented interface and lists the publisher features. The NGCE service publishes only one feature type, the *mne:Entidad* (of *mne=http://www.idee.es/mne* namespace). The requests 2 and 3 retrieve the XML Schema and an example feature item (Listing C.2), respectively. The dataset of the analysed service contains 3667 items (request 4) .

Listing C.1: Examples of requests.

```
1 http://www.idee.es/IDEE-WFS-Nomenclator-NGC/services?SERVICE=WFS&VERSION=1.1.0&
      REQUEST=GetCapabilities
2 http://www.idee.es/IDEE-WFS-Nomenclator-NGC/services?SERVICE=WFS&VERSION=1.1.0&
      REQUEST=DescribeFeatureType&NAMESPACE=xmlns(mne=http://www.idee.es/mne)&
      TYPENAME=mne:Entidad
3 http://www.idee.es/IDEE-WFS-Nomenclator-NGC/services?SERVICE=WFS&VERSION=1.1.0&
      REQUEST=GetFeature&MAXFEATURES=1&NAMESPACE=xmlns(mne=http://www.idee.es/mne)
      &TYPENAME=mne:Entidad
4 http://www.idee.es/IDEE-WFS-Nomenclator-NGC/services?SERVICE=WFS&VERSION=1.1.0&
      REQUEST=GetFeature&NAMESPACE=xmlns(mne=http://www.idee.es/mne)&TYPENAME=
      mne:Entidad&resultType=hits
```

[1] http://www.idee.es/IDEE-WFS-Nomenclator-NGC/services?

Listing C.2: Example of the GetFeature response which retrieves only one instance.

```
<?xml version="1.0" encoding="UTF-8"?>
<wfs:FeatureCollection numberOfFeatures='1' xmlns:gml="http://www.opengis.net/
    gml" xmlns:wfs="http://www.opengis.net/wfs" xmlns:xsi="http://www.w3.org
    /2001/XMLSchema-instance" xmlns:mne="http://www.idee.es/mne" xmlns:xlink="
    http://www.w3.org/1999/xlink" xsi:schemaLocation="http://www.idee.es/mne␣
    http://www.idee.es:80/IDEE-WFS-Nomenclator-NGC/services?SERVICE=WFS&amp;
    VERSION=1.1.0&amp;REQUEST=DescribeFeatureType&amp;TYPENAME=mne:Entidad&amp;
    NAMESPACE=xmlns(mne=http://www.idee.es/mne)␣http://www.opengis.net/wfs␣http:
    //schemas.opengis.net/wfs/1.1.0/wfs.xsd"><gml:boundedBy><gml:Envelope
    srsName='EPSG:4230'><gml:pos srsDimension='2'>-17.8833333333333 27.8</
    gml:pos><gml:pos srsDimension='2'>-17.8833333333333 27.8</gml:pos></
    gml:Envelope></gml:boundedBy><gml:featureMember>
<mne:Entidad gml:id="ES.IGN.NGC_0761"><gml:boundedBy><gml:Envelope srsName='
    EPSG:4230'><gml:pos srsDimension='2'>-17.8833333333333 27.8</gml:pos><
    gml:pos srsDimension="2">-17.8833333333333 27.8</gml:pos></gml:Envelope></
    gml:boundedBy>
<mne:nombreEntidad><mne:NombreEntidad gml:id="ES.IGN.NGC_0761.GMZDMNY171400">
<mne:nombre>Punta de la Caleta</mne:nombre>
<mne:idioma>spa</mne:idioma>
<mne:claseNombre>preferente</mne:claseNombre>
<mne:estatus>oficial</mne:estatus>
<mne:fuente>1:1 mill n: Mapa de la Pen nsula Ib rica, Baleares y Canarias,
    escala 1:1.000.000. Instituto Geogr fico Nacional, 2000</mne:fuente></
    mne:NombreEntidad></mne:nombreEntidad>
<mne:nombreEntidad><mne:NombreEntidad gml:id="ES.IGN.NGC_0761.GMZDMOA171400">
<mne:nombre>Punta de M rquez</mne:nombre>
<mne:idioma>spa</mne:idioma>
<mne:claseNombre>variante</mne:claseNombre>
<mne:estatus>normalizado</mne:estatus>
<mne:fuente>Canarias: Gobierno de Canarias</mne:fuente></mne:NombreEntidad></
    mne:nombreEntidad>
<mne:tipoEntidad><mne:TipoEntidad gml:id="ES.IGN.NGC.TP.INHVGVCB171400">
<mne:tipo>COSTA</mne:tipo>
<mne:catalogoEntidades>URL:http://www.idee.es/show.do?to=pideep_conciso.ES</
    mne:catalogoEntidades></mne:TipoEntidad></mne:tipoEntidad>
<mne:tipoEntidad><mne:TipoEntidad gml:id="ES.IGN.NGC.TP.INQWE3Y171400">
<mne:tipo>Cabo</mne:tipo>
<mne:catalogoEntidades>URL:http://www.idee.es/show.do?to=pideep_conciso.ES</
    mne:catalogoEntidades></mne:TipoEntidad></mne:tipoEntidad>
```

```
22 <mne:posicionEspacial><mne:PosicionEspacial gml:id="ES.IGN.NGC.PS.HE3TQNJX171400
       "><gml:boundedBy><gml:Envelope srsName='EPSG:4230'><gml:pos srsDimension='2'
       >-17.8833333333333 27.8</gml:pos><gml:pos srsDimension="2">-17.8833333333333
        27.8</gml:pos></gml:Envelope></gml:boundedBy><mne:geometria><gml:Point
       srsName="EPSG:4230"><gml:pos srsDimension="2">-17.8833333333333 27.8</
       gml:pos></gml:Point></mne:geometria></mne:PosicionEspacial></
       mne:posicionEspacial>
23 <mne:entidadLocal><mne:EntidadLocal gml:id="ES.IGN.NGC.EL.
       KNQW45DBEBBXE5L2EBSGKICUMVXGK4TJMZSQ171400">
24 <mne:comunidadAutonoma>Canarias</mne:comunidadAutonoma>
25 <mne:provincia>Santa Cruz de Tenerife</mne:provincia></mne:EntidadLocal></
       mne:entidadLocal>
26 <mne:codificacion><mne:Codificacion gml:id="ES.IGN.NGC.CO.NGC_0761171400">
27 <mne:codigo>NGC_0761</mne:codigo>
28 <mne:sistemaCodificacion>Nomencl tor Geogr fico Conciso 1.0. Instituto
       Geogr fico Nacional</mne:sistemaCodificacion></mne:Codificacion></
       mne:codificacion>
29 <mne:mapa><mne:Mapa gml:id="ES.IGN.NGC.MA.JVKE4NJQ.GEYTANI171400">
30 <mne:serie>MTN50</mne:serie>
31 <mne:hoja>1105</mne:hoja></mne:Mapa></mne:mapa>
32 <mne:mapa><mne:Mapa gml:id="ES.IGN.NGC.MA.JVKE4NJQ.GEYTAOA171400">
33 <mne:serie>MTN50</mne:serie>
34 <mne:hoja>1108</mne:hoja></mne:Mapa></mne:mapa></mne:Entidad></gml:featureMember
       ></wfs:FeatureCollection>
```

First, a list of the possible requestable properties is generated automatically. After retrieving several distinct features, the values from the corresponding fields are assigned (excluding spatial objects). The requestable fields that return numeric values are eliminated from the list.

The method has gathered 18 items during the sampling step (0.5% of the dataset). Table C.1 shows the generated list of properties and their values of two different items selected from the sample set. The unique values have been used to analyse characteristics of the properties as outlined above. Table C.2 and C.3 gather information on properties: number of items that use them, examples of frequency of their values according to the `resultType=hits` parameter. Table C.4 shows properties which have been excluded during the analysis.

The procedure developed supports the features that have "flat" models very well, i.e. the models that have no *multivalue properties*. For example, the `mne:nombreEntidad` property of the `mne:Entidad` feature is a complex property which can appear more than once per feature. The requestable properties of the `mne:nombreEntidad`, which are simple ones, are called *multivalue properties*. When requesting the *multivalue properties*, it is not easy to provide the required

filtering behaviour. An item is returned if at least one of the properties satisfies the restriction defined. For example, the request for items whose requestable property *mne:nombreEntidad/ mne:NombreEntidad/ mne:claseNombre* has value equal to '*variante*', both items from Table C.1 will be returned. If the filtering statement restricts values of this property in order to exclude the values equal to '*variante*' (i.e. `FILTER=<Filter xmlns:mne="http://www.idee.es/mne"> <Not> <PropertyIsEqualTo> <PropertyName> mne:nombreEntidad/mne:NombreEntidad/mne:claseNombre </PropertyName> <Literal>variante</Literal> </PropertyIsEqualTo> </Not> </Filter>`), at least one of the properties has to satisfy this restriction. As a result, both items will be returned again because one of the names of the item '*1*' is '*preferente*'.

In summary, the following aspects have been considered:

- The earlier versions of WFS service may not support the `resultType=hits` parameter of the `GetFeature` query (e.g. version 1.0.0). In this case the sampling task is limited to retrieve only ten items for each requestable feature.
- The publisher of the WFS service may restrict the number of items returned for each `GetFeature` request. Typically, the amount of returned items is limited to ten. This restriction does not affect the sampling method, since at most ten elements are retrieved in a single request.
- The encoding problems are handled.
- The time–out errors are handled.
- The non–flat feature model influences the construction of filter encoding statements.
- All *xLinks* are removed from the analysed items.
- If nested items are detected, they are removed from the analysed items.
- Some dependencies among requestable features can be identified but they are not considered.

The generated dataset summary is quite simplistic because it offers an overview based on simple textual properties. This approach does not inform on relations among values of complex properties, for example the `mne:nombreEntidad` property is composed of four simple properties which should be analysed together in order to understand its meaning properly. However, even such a simplistic approach is useful for indexing and quick overwiev of the content published.

Field	Id	Values
mne:nombreEntidad/mne:NombreEntidad /mne:nombre	1	Punta de la Caleta
	1	Punta de Márquez
	2	Canal del Henares
mne:nombreEntidad/mne:NombreEntidad /mne:idioma	1	spa
	1	spa
	2	spa
mne:nombreEntidad/mne:NombreEntidad /mne:claseNombre	1	preferente
	1	variante
	2	preferente
mne:nombreEntidad/mne:NombreEntidad /mne:estatus	1	oficial
	1	normalizado
	2	oficial
mne:nombreEntidad/mne:NombreEntidad /mne:fuente	1	1:1 millón: Mapa de la Península Ibérica, Baleares y Canarias, escala 1:1.000.000. Instituto Geográfico Nacional, 2000
	1	Canarias: Gobierno de Canarias
	2	1:1 millón: Mapa de la Península Ibérica, Baleares y Canarias, escala 1:1.000.000. Instituto Geográfico Nacional, 2000
mne:tipoEntidad/mne:TipoEntidad /mne:tipo	1	COSTA
	1	Cabo
	2	HIDRO
	2	Canal
mne:tipoEntidad/mne:TipoEntidad /mne:catalogoEntidades	1	URL:http://www.idee.es/ show.do?to=pideep_conciso.ES
	1	URL:http://www.idee.es/ show.do?to=pideep_conciso.ES
	2	URL:http://www.idee.es/ show.do?to=pideep_conciso.ES
	2	URL:http://www.idee.es/ show.do?to=pideep_conciso.ES
mne:entidadLocal/mne:EntidadLocal /mne:comunidadAutonoma	1	Canarias
	2	Castilla-La Mancha
mne:entidadLocal/mne:EntidadLocal /mne:provincia	1	Santa Cruz de Tenerife
	2	Guadalajara
mne:entidadLocal/mne:EntidadLocal /mne:municipio	1	
mne:entidadLocal/mne:EntidadLocal /mne:comarca	1	
mne:entidadLocal/mne:EntidadLocal /mne:isla	1	
mne:entidadLocal/mne:EntidadLocal /mne:EATIMNombre	1	
mne:codificacion/mne:Codificacion /mne:codigo	1	NGC_0761
	2	NGC_1633
mne:codificacion/mne:Codificacion /mne:sistemaCodificacion	1	Nomenclátor Geográfico Conciso 1.0. Instituto Geográfico Nacional
	2	Nomenclátor Geográfico Conciso 1.0. Instituto Geográfico Nacional
mne:mapa/mne:Mapa/mne:serie	1	MTN50
	1	MTN50
	2	MTN50
mne:mapa/mne:Mapa/mne:hoja	1	1105
	1	1108
	2	0486
mne:posicionEspacial /mne:PosicionEspacial/mne:geometria	1	-17.883333333333 27.8

Table C.1: Examples of values of analysed fields of two different features.

Field	Value	Hits
mne:nombreEntidad/mne:NombreEntidad /mne:idioma	*	3667
	spa	2596
	cat	879
	glg	211
	eus	131
	ast	45
mne:nombreEntidad/mne:NombreEntidad /mne:claseNombre	*	3667
	preferente	3667
	variante	110
	alternativo	92
	anterior	66
mne:nombreEntidad/mne:NombreEntidad /mne:estatus	*	3667
	oficial	3667
	normalizado	176
mne:nombreEntidad/mne:NombreEntidad /mne:fuente	*	3667
	1:1 millón: Mapa de la Península Ibérica, Baleares y Canarias, escala 1:1.000.000. Instituto Geográfico Nacional, 2000	920
	Atlas: Atlas Nacional de España. El Medio Físico 1. Instituto Geográfico Nacional, 2000	149
	1:200.000: Mapas Provinciales. Instituto Geográfico Nacional	115
	C. Valenciana: Generalitat Valenciana	108
	Cataluña: Generalitat de Catalunya y	82
mne:tipoEntidad/mne:TipoEntidad /mne:tipo	*	3667
	POBLA	1835
	Población 2	1729
	HIDRO	778
	Río	493
	OROGR	416
mne:tipoEntidad/mne:TipoEntidad /mne:catalogoEntidades	*	3667
	URL:http://www.idee.es/show.do?to=pideep_conciso.ES	3667
mne:entidadLocal/mne:EntidadLocal /mne:comunidadAutonoma	*	3667
	Canarias	166
	Castilla-La Mancha	295
	No aplica	9
	Aragón	191

Table C.2: List of fields and some examples of their possible values ("*" – any values). Part 1.

Field	Value	Hits
mne:entidadLocal/mne:EntidadLocal /mne:provincia	*	3667
	Santa Cruz de Tenerife	89
	Las Palmas	76
	No aplica	39
mne:codificacion/mne:Codificacion /mne:sistemaCodificacion	*	3667
	Nomenclátor Geográfico Conciso 1.0. Instituto Geográfico Nacional	3667
mne:mapa/mne:Mapa/mne:serie	*	3662
	MTN50	3662

Table C.3: List of fields some examples of their possible values ("*" – any values). Part 2.

Field	Value Example	Hits	Note
mne:nombreEntidad/mne:NombreEntidad /mne:nombre	*	3667	High variability
	Punta de la Caleta	1	
	Canal del Henares	1	
mne:codificacion/mne:Codificacion /mne:codigo	*	3667	High variability
	NGC_0761	1	
	NGC_1633	1	
mne:entidadLocal/mne:EntidadLocal /mne:municipio	*	68	Not typical
	Malpica de Bergantiños		
	Marratxí		
	Mieres		
	Parres		
	Piélagos		
	Derio		
	Galdakao		
	Camargo		
	Carreño		
mne:entidadLocal/mne:EntidadLocal /mne:comarca	*	0	Empty
mne:entidadLocal/mne:EntidadLocal /mne:isla	*	0	Empty
mne:entidadLocal/mne:EntidadLocal /mne:EATIMNombre	*	0	Empty
mne:mapa/mne:Mapa/mne:hoja	*	3662	Numeric value
	1105	10	
	1108	10	
	0486	5	
mne:posicionEspacial /mne:PosicionEspacial/mne:geometria	*	3667	Spatial Object

Table C.4: List of fields which have been dismissed from further evaluation ("*" – any values).

Bibliography

Abelson, H., Adida, B., Linksvayer, M., Yergler, N., 2008. ccREL: The Creative Commons Rights Expression Language. Technical report, Creative Commons.
URL http://www.w3.org/Submission/ccREL/

Akkiraju, R., Farell, J., Miller, J. A., Nagarajan, M., Sheth, A., Verma, K., 2005. Web Service Semantics – WSDL–S. Version 1.0. W3C Discussion Paper.
URL http://www.w3.org/Submission/WSDL-S/

Al-Masri, E., Mahmoud, Q., 2008. Investigating Web Services on the World Wide Web. In: Proceedings of the 17th international conference on World Wide Web. WWW'08. ACM, New York, NY, USA, pp. 795–804.
URL http://dx.doi.org/10.1145/1367497.1367605

Alarcón, R., Wilde, E., 2010. Linking Data from RESTful Services. In: Third Workshop on Linked Data on the Web, Raleigh, North Carolina, April 2010.

Almpanidis, G., Kotropoulos, C., Pitas, I., 2007. Combining text and link analysis for focused crawling – An application for vertical search engines. Information Systems 32 (6), 886–908.
URL http://dx.doi.org/10.1016/j.is.2006.09.004

Alonso, M., Malpica, J., 2008. Classification of Multispectral High-Resolution Satellite Imagery Using LIDAR Elevation Data. In: Proceedings of the 4th International Symposium on Advances in Visual Computing, Part II, Las Vegas, Nevada, USA. ISVC'08. Springer-Verlag, Berlin, Heidelberg, pp. 85–94.
URL http://dx.doi.org/10.1007/978-3-540-89646-3_9

Amitay, E., Har'El, N., Sivan, R., Soffer, A., 2004. Web–a–Where: Geotagging Web Content. In: SIGIR '04: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, New York, NY, USA, pp. 273–280.
URL http://dx.doi.org/10.1145/1008992.1009040

ANSI, 2009. North American Profile (NAP) of ISO 19115:2003, Geographic Information – Metadata, (NAP-Metadata).

Antonellis, V. D., Melchiori, M., Salvi, D., Bianchini, D., 2006. Peer-to-peer Semantic-based Web Service Discovery: State of the Art. Tech. rep., Dipartimento di Elettronica per l'Automazione Università di.

Apache Software Foundation, 2012. Apache Jena project.
URL http://incubator.apache.org/jena/index.html

ASF, 2011. Apache Tika – a content analysis toolkit.
URL http://tika.apache.org/

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z., 2008. DBpedia: A Nucleus for a Web of Open Data. In: Proceedings of 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference (ISWC+ASWC 2007). pp. 722–735.
URL http://dx.doi.org/10.1007/978-3-540-76298-0_52

Auer, S., Dietzold, S., Lehmann, J., Hellmann, S., Aumueller, D., 2009a. Triplify – Light–Weight Linked Data Publication from Relational Databases. In: Proceedings of the 18th international conference on World wide web. WWW '09. ACM, New York, NY, USA, pp. 621–630.
URL http://dx.doi.org/10.1145/1526709.1526793

Auer, S., Lehmann, J., Hellmann, S., 2009b. LinkedGeoData: Adding a Spatial Dimension to the Web of Data. In: Proceedings of the 8th International Semantic Web Conference. ISWC'09. Springer–Verlag, Berlin, Heidelberg, pp. 731–746.
URL http://dx.doi.org/10.1007/978-3-642-04930-9_46

Bachlechner, D., Siorpaes, K., Fensel, D., Toma, I., 2006a. Web Service Discovery – A Reality Check. Tech. rep., Digital Enterprise Research Institute.

Bachlechner, D., Siorpaes, K., Lausen, H., Fensel, D., 2006b. Web Service Discovery – A Reality Check. Demos and Posters of the 3rd European Semantic Web Conference (ESWC 2006), Budva, Montenegro, 11–14 June, 2006.

Baharudin, B., Qahwaji, R., Jiang, J., Rahman, P., 2007. Combining image features for image classification. In: International Conference on Intelligent and Advanced Systems 2007, Kuala Lumpur, Malaysia. pp. 268–272.

Bai, Y., Di, L., Wei, Y., 2009. A taxonomy of geospatial services for global service discovery and interoperability. Computers & Geosciences 35 (4), 783–790.

Bailey, J. E., Chen, A., 2011. The role of Virtual Globes in geoscience. Computers & Geosciences 37 (1), 1–2.

Bakshi, R., Knoblock, C. A., Thakkar, S., 2004. Exploiting Online Sources to Accurately Geocode Addresses. In: GIS '04: Proceedings of the 12th annual ACM international workshop on Geographic information systems. ACM Press, New York, NY, USA, pp. 194–203.
URL http://dx.doi.org/10.1145/1032222.1032251

Baldini, A., Boldrini, E., Santoro, M., Mazzetti, P., 2010. GeoNetwork powered GI–cat: a geoportal hybrid solution. Poster Session: Real Use of Standards and Technologies, European Geosciences Union, General Assembly, Vienna, Austria, 02–7 May 2010.

Balke, W.-T., Wagner, M., 2004. Through Different Eyes – Assessing Multiple Conceptual Views for Querying Web Services. In: Proceedings of the 13th international conference on World Wide Web, WWW 2004, New York, NY, USA, May 17–20, 2004. ACM Press, pp. 196–205.

Bartley, J. D., 2005. Mapdex: An Index of Web Mapping Services. Kansas Geological Survey at the University of Kansas.

Batcheller, J. K., 2008. Automating geospatial metadata generation – An integrated data management and documentation approach. Computers & Geosciences 34 (4), 387–398.
URL http://dx.doi.org/10.1016/j.cageo.2007.04.001

Baumann, P., 2010. OGC WCS 2.0 Interface Standard – Core. OGC 09-110r3.

Beal, J. R., 2003. Contextual Geolocation: A Specialized Application for Improving Indoor Location Awareness in Wireless Local Area Networks. In: MICS2003: The 36th Annual Midwest Instruction and Computing Symposium, Duluth, Minnesota, USA.

Beale, R., 2006. Improving Internet interaction: From theory to practice. Journal of the American Society for Information Science and Technology 57 (6), 829–833.
URL http://dx.doi.org/10.1002/asi.20302

Becker, C., Bizer, C., 2008. DBpedia Mobile: A Location-Enabled Linked Data Browser. In: Proceedings of the Linked Data on the Web Workshop, Beijing, China, April 22, 2008.

Becker, C., Bizer, C., 2009. Exploring the Geospatial Semantic Web with DBpedia Mobile. Web Semantics: Science, Services and Agents on the World Wide Web 7 (4), 278–286.
URL http://dx.doi.org/10.1016/j.websem.2009.09.004

Behr, F.-J., Rimayanti, A., 2008. OPENGEOCODING.ORG - A free, participatory, community oriented Geocoding Service. In: The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences. Vol. 36.

Béjar, R., 2009. Contributions to the Modelling of Spatial Data Infrastructures and their Portrayal Services. Ph.D. thesis, Universidad de Zaragoza.

Béjar, R., Nogueras-Iso, J., Ángel Latre, M., Muro-Medrano, P. R., Zarazaga-Soria, F. J., 2009. Digital Libraries as a Foundation of Spatial Data Infrastructures. In: Theng, Y. L., Foo, S., Goh, D. H.-L., Na, J.-C. (Eds.), Handbook of Research on Digital Libraries. IGI Global, pp. 382–389.
URL http://dx.doi.org/10.4018/978-1-59904-879-6.ch039

Benatallah, B., Hacid, S., Rey, C., Toumani, F., 2003. Request Rewriting–Based Web Service Discovery. In: Fensel, D., Sycara, K. P., Mylopoulos, J. (Eds.), The Semantic Web - ISWC 2003, Second International Semantic Web Conference, Sanibel Island, FL, USA, October 20-23, 2003, Proceedings. Lecture Notes in Computer Science. Springer, pp. 242–257.

Bergman, M. K., 2001. The Deep Web: Surfacing Hidden Value. Journal of Electronic Publishing 7 (1).
URL http://dx.doi.org/10.3998/3336451.0007.104

Bernard, L., Kanellopoulos, I., Annoni, A., Smits, P., 2005. The European geoportal–one step towards the establishment of a European Spatial Data Infrastructure. Computers, Environment and Urban Systems 29 (1), 15–31.
URL http://dx.doi.org/10.1016/j.compenvurbsys.2004.05.009

Bizer, C., Heath, T., Berners-Lee, T., 2009. Linked Data – The Story So Far. International Journal on Semantic Web and Information Systems, Special Issue on Linked Data. 5 (3), 1–22, linked Data - The Story So Far.

Bizer, C., Seaborne, A., 2004. D2RQ – Treating Non–RDF Databases as Virtual RDF Graphs. In: ISWC2004 (posters).

Boes, U., Pavlova, R., 2008. Is there a Future for Spatial Data Infrastructures? In: IfGIprints (Ed.), GI–Days 2008, Proceedings of the 6th Geographic Information Days. Vol. 32. pp. 305—-314.

Bonner, M. R., Han, D., Nie, J., Rogerson, P., Vena, J. E., Freudenheim, J. L., Jul 2003. Positional Accuracy of Geocoded Addresses in Epidemiologic Research. Epidemiology 14 (4), 408–412.

Borjas, B., Florczyk, A. J., López-Pellicer, F. J., Nogueras-Iso, J., Zarazaga-Soria, F. J., 2011a. Automatic metadata generation for the Web geo–resources. In: INSPIRE Conference 2011. European Commission Joint Research Centre.

Borjas, B., Florczyk, A. J., López–Pellicer, F. J., Zarazaga–Soria, F. J., 2011b. Generación Automática de Metadatos Geográficos de Páginas Web. In: 9th International Geomatics Week (Semana Geomática Internacional 2011). Barcelona, Spain, 15-17 March 2011.

Brickley, D., Guha, R., 2004. RDF Vocabulary Description Language 1.0: RDF Schema. W3C Recommendation.
URL http://www.w3.org/TR/rdf-schema/

Brin, S., Page, L., 1998. The anatomy of a large–scale hypertextual Web search engine. Computer Networks and ISDN Systems 30, 107–117.
URL http://dx.doi.org/10.1016/S0169-7552(98)00110-X

Brockmans, S., Celino, I., Cerizza, D., Della Valle, E., Erdmann, M., Funk, A., Lausen, H., Schoch, W., Steinmetz, N., Turati, A., 2008. Realizing Service–Finder: Web Service Discovery at Web Scale. In: The 2nd European Semantic Technology Conference (ESTC). Vienna, Austria, 2008.

Brockmans, S., Celino, I., Cerizza, D., Della Valle, E., Erdmann, M., Funk, A., Lausen, H., Schoch, W., Steinmetz, N., Turati, A., 2009. Service–Finder: First Steps toward the realization of Web Service Discovery at Web Scale. In: The 1st International Workshop on Interoperability through Semantic Data and Service Integration, Camogli, Italy, June 2009.

Broder, A., 2002. A taxonomy of web search. SIGIR Forum 36, 3–10.

Broens, T., 2004. Context–aware, ontology–based, service discovery. Master's thesis, Telematics from the University of Twente, Enschede, The Netherlands.

Broughton, V. ., 2006. The need for a faceted classification as the basis of all methods of information retrieval. In: Aslib proceedings. Vol. 58. pp. 49–72.
URL http://dx.doi.org/10.1108/00012530610648671

Bussler, C., 2002. A conceptual architecture for semantic web enabled web services. SIGMOD Record 31 (4), 24–29.
URL http://dx.doi.org/10.1145/637411.637415

Campelo, C. E., Souza Baptista, C., 2009. A Model for Geographic Knowledge Extraction on Web Documents. In: Proceedings of the ER 2009 Workshops on Advances in Conceptual Modeling – Challenging Perspectives. ER'09. Springer–Verlag, Berlin, Heidelberg, pp. 317–326.
URL http://dx.doi.org/10.1007/978-3-642-04947-7_38

Cavnar, W. B., Trenkle, J. M., 1994. N–Gram–Based Text Categorization. In: In Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval. pp. 161–175.

Cayo, M., Talbot, T., 2003. Positional error in automated geocoding of residential addresses. International Journal of Health Geographics 2 (10).

Chakrabarti, S., van den Berg, M., Dom, B., 1999. Focused crawling: a new approach to topic-specific Web resource discovery. Computer Networks: The International Journal of Computer and Telecommunications Networking 31, 1623–1640.

Chen, Wu, 2011. WSDL term tokenization methods for IR–style web services discovery. Science of Computer Programming 77 (3), 355–374.
URL http://dx.doi.org/10.1016/j.scico.2011.08.001

Chen, H., Fan, H., Chau, M., Zeng, D., 2003. Testing a cancer meta spider. International Journal of Human–Computer Studies 59 (5), 755–776.

Chin, J. P., Diehl, V. A., Norman, K. L., 1988. Development of an instrument measuring user satisfaction of the human–computer interface. In: Proceedings of the SIGCHI conference on Human factors in computing systems. CHI '88. ACM, New York, NY, USA, pp. 213–218.
URL http://dx.doi.org/10.1145/57167.57203

Chitra, S., Vidhya, K., Aghila, G., sept. 2010. Web service selection based on ranking of QoS using Naïve Bayes through ontology mapping. In: Computer and Communication Technology (ICCCT), 2010 International Conference on. pp. 782 –787.
URL http://dx.doi.org/10.1109/ICCCT.2010.5640436

Chukmol, U., 2008. A framework for web service discovery: service's reuse, quality, evolution and user's data handling. In: IDAR '08: Proceedings of the 2nd SIGMOD PhD workshop on Innovative database research. ACM, New York, NY, USA, pp. 13–18.
URL http://dx.doi.org/10.1145/1410308.1410313

Clark, J. B., 2008. Closure of OASIS UDDI Specification TC (OASIS Mailing List Archives).
URL http://lists.oasis-open.org/archives/tc-announce/200807/msg00000.html

Clement, L., Hately, A., von Riegen, C., Rogers, T., 2004. UDDI Version 3.0.2. UDDI Spec Technical Committee Draft.
URL http://uddi.org/pubs/uddi_v3.htm

Couclelis, H., 1992. People Manipulate Objects (but Cultivate Fields): Beyond the Raster–Vector Debate in GIS. In: Proceedings of the International Conference GIS - From Space to Territory: Theories and Methods of Spatio-Temporal Reasoning on Theories and Methods of Spatio-Temporal Reasoning in Geographic Space. Springer-Verlag, London, UK, pp. 65–77.

Craglia, M., 2007. Volunteered Geographic Information and Spatial Data Infrastructures: when do parallel lines converge? Vol. 12. NCGIA, pp. 13–14.

Craglia, M., Goodchild, M. F., Annoni, A., Camara, G., Gould, M., Kuhn, W., Mark, D., Masser, I., Maguire, D., Liang, S., Parsons, E., 2008. Next–generation Digital Earth: A position paper from the Vespucci Initiative for the Advancement of Geographic Information Science. International Journal of Spatial Data Infrastructures Research 3, 146–167.
URL http://dx.doi.org/10.2902/1725-0463.2008.03.art9

Cyganiak, R., Jentzsch, A., 2011. The Linking Open Data cloud diagram. Home page.
URL http://richard.cyganiak.de/2007/10/lod/

Daviel, A., 2007. Geo Tags for HTML Resource Discovery. GeoTags Website.
URL http://geotags.com

Daviel, A., Kaegi, F., 2007. Geographic registration of HTML documents. ETF Draft, no longer active.

Daviel, A., Kaegi, F., Kofahl, M., 2007. Geographic extensions for HTTP transactions. ETF Draft, no longer active.

Davis, C. A., Fonseca, F. T., Borges, K. A. V., 2003. A Flexible Addressing System for Approximate Geocoding. In: GeoInfo 2003: Proceedings of the Fifth Brazilian Symposiumon GeoInformatics. Instituto Nacional de Pesquisas Espaciais (INPE).

Davis, J. C. A., Fonseca, F. T., 2007. Assessing the Certainty of Locations Produced by an Address Geocoding System. GeoInformatica 11 (1), 103–129.

DCMI, 1995–2012. Dublin Core Metadata Initiative (DCMI). DCMI Home.
URL http://dublincore.org/

de la Beaujardiere, J., 2006. OpenGIS Web Map Server Implementation Specification. Version: 1.3.0. OGC 06-042.

Díaz, L., Granell, C., Gould, M., Huerta, J., 2011. Managing user–generated information in geospatial cyberinfrastructures. Future Generation Computer Systems 27, 304–314.

Dietze, S., Gugliotta, A., Domingue, J., 2008. Towards context–aware semantic web service discovery through conceptual situation spaces. In: CSSSIA. p. 6.

DoD, 2006. National Imagery Transmission Format Version 2.1. MIL-STD-2500C.

Dolbear, C., Hart, G., 2007. Ontological bridge building – using ontologies to merge spatial datasets. In: AAAI Spring Symposium: Semantic Scientific Knowledge Integration. AAAI, pp. 15–20.

Doulkeridis, C., Loutas, N., Vazirgiannis, M., 2006. A System Architecture for Context–Aware Service Discovery. Electronic Notes in Theoretical Computer Science 146 (1), 101–116.
URL http://dx.doi.org/10.1016/j.entcs.2005.11.010

EC, 2003. Regulation 2003/1059/EC of the European Parliament and of the Council of 26 May 2003 on the establishment of a common classification of territorial units for statistics (NUTS). Official Journal of the European Union.

EC, 2007a. Directive 2007/2/EC of the European Parliament and of the Council. Directive of the European Parliament and of the Council of 14 March 2007 establishing an Infrastructure for Spatial Information in the European Community (INSPIRE).

EC, 2007b. INSPIRE Network Services Performance Guidelines. Version 1.0.

Egenhofer, M., Mark, D., 1995. Naive Geography. In: Frank, A., Kuhn, W. (Eds.), Spatial Information Theory: A Theoretical Basis for GIS, International Conference COSIT '95, Semmering, Austria, September 21–23. Vol. 988 of Lecture Notes in Computer Science. Springer–Verlag, pp. 1–15.

Egenhofer, M. J., 2002. Toward the semantic geospatial web. In: GIS '02: Proceedings of the 10th ACM international symposium on Advances in geographic information systems. ACM, New York, NY, USA, pp. 1–4.
URL http://dx.doi.org/10.1145/585147.585148

Ellisman, M. H., 2005. Cyberinfrastructure and the Future of Collaborative Work. Issues in Science and Technology 22 (1), 43–50.

Erl, T., 2005. Service–Oriented Architecture: Concepts, Technology, and Design. Prentice Hall/-PearsonPTR.

Erling, O., Mikhailov, I., 2009. RDF Support in the Virtuoso DBMS. In: Pellegrini, T., Auer, S., Tochtermann, K., Schaffert, S. (Eds.), Networked Knowledge – Networked Media. Vol. 221 of Studies in Computational Intelligence. Springer Berlin / Heidelberg, pp. 7–24.
URL http://dx.doi.org/10.1007/978-3-642-02184-8_2

Fagan, J., 2010. Usability studies of faceted browsing: A literature review. Information Technology and Libraries 29 (2), 58–66.

Fallahi, G. R., Frank, A. U., Mesgari, M. S., Rajabifard, A., 2008. An ontological structure for semantic interoperability of GIS and environmental modeling. International Journal of Applied Earth Observation and Geoinformation 10 (3), 342 – 357.
URL http://dx.doi.org/10.1016/j.jag.2008.01.001

Faruqui, M., Padó, S., 2010. Training and Evaluating a German Named Entity Recognizer with Semantic Generalization. In: Proceedings of KONVENS 2010. Saarbrücken, Germany.

FGDC, 1998a. Content Standard for Digital Geospatial Metadata. Version 2.0. FGDC Steering Committee.

FGDC, 1998b. The Spatial Data Transfer Standard.

Finkel, J. R., Grenager, T., Manning, C., 2005. Incorporating non–local information into information extraction systems by Gibbs sampling. In: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. ACL '05. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 363–370.
URL http://dx.doi.org/10.3115/1219840.1219885

Florczyk, A. J., López-Pellicer, F. J., Béjar, R., Nogueras-Iso, J., Zarazaga-Soria, F., 2010a. Applying Semantic Linkage in the Geospatial Web. In: Painho, M., Santos, M., Pundt, H. (Eds.), Geospatial Thinking, Lecture Notes in Geoinformation and Cartography. Springer-Verlag, pp. 201–220.

Florczyk, A. J., López-Pellicer, F. J., Lacasta, J., Nogueras-Iso, J., Zarazaga-Soria, F. J., 2010b. Semantic Spatial Data Infrastructures in Action. In: INSPIRE Conference 2010: INSPIRE as a framework for cooperation. Krakow, Poland, 22–25 June 2010.

Florczyk, A. J., López-Pellicer, F. J., Muro-Medrano, P. R., Nogueras-Iso, J., Zarazaga-Soria, F. J., 2010c. Semantic Selection of Georeferencing Services for Urban Management. Journal of Information Technology in Construction 15, 111–121.

Florczyk, A. J., López-Pellicer, F. J., Rodrigo, P., Rioja, R., Muro-Medrano, P. R., 2008. Geocoder compuesto–solución hıbrida en el mundo de ruido informativo. In: JIDEE 2008 - V Jornadas Técnicas de la Infraestructura de Datos Espaciales de España.

Florczyk, A. J., López-Pellicer, F. J., Valiño, J., Béjar, R., Muro-Medrano, P. R., 2009a. INSPIRE–able Services. In: 12th AGILE International Conference on Geographic Information Science.

Florczyk, A. J., López–Pellicer, F. J., Gayán-Asensio, D., Rodrigo–Cardiel, P., Latre, M., Nogueras–Iso, J., 2009b. Compound Geocoder: get the right position. In: GSDI 11 World Conference and the 3rd INSPIRE Conference 2009, Rotterdam 15-19 June 2009.

Florczyk, A. J., López–Pellicer, F. J., Rioja, R., Nogueras-Iso, J., Zarazaga-Soria, F. J., 2009c. Enabling Geolocating via Ontologies. In: Urban Ontologies for an improved communication in urban development projects. Les Editions de l'université de Liége, pp. 85–94.

Florczyk, A. J., Maué, P., López-Pellicer, F. J., Nogueras-Iso, J., 2010d. Finding OGC Web Services in the Digital Earth. In: DE-2010 - Workshop Towards Digital Earth: Search, Discover and Share Geospatial Data 2010.

Florczyk, A. J., Nogueras-Iso, J., Zarazaga-Soria, F. J., Béjar, R., Oct. 2011. Identifying Orthoimages in Web Map Services. Computers & Geosciences.
URL http://dx.doi.org/10.1016/j.cageo.2011.10.017

Foulonneau, M., Riley, J., 2008. Metadata for digital resources: implementation, systems design and interoperability. Chandos information professional series. Chandos, Oxford, xvi, 203 p.

Gamma, E., Helm, R., Johnson, R., Vilssides, J. M., 1994. Dessign Patterns: Elements of Reusable Object-Oriented Software. Addison–Wesley.

Garofalakis, J., Panagis, Y., Sakkopoulos, E., Tsakalidis, A., 2006. Web Service Discovery Mechanisms. Journal of Web Engineering 5 (3), 265–290, rinton Press.

Garvin, D., 1988. Managing Quality: The Strategic Competitive Edge. New York: The Free Press.

Giuliani, G., Ray, N., Lehmann, A., 2011. Grid–enabled Spatial Data Infrastructure for environmental sciences: Challenges and opportunities. Future Generation Computer Systems 27, 292–303.

Goldberg, D., Wilson, J., Cockburn, M., 2010. Toward Quantitative Geocode Accuracy Metrics. In: Proceedings of the Ninth International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences, Leicester, UK. pp. 329–332.

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

Goldberg, D. W., 2008. A Geocoding Best Practices Guide. North American Association of Central Cancer Registries (NAACCR).

Golliher, S. A., 2008. Search Engine Ranking Variables and Algorithms. SEMJ.org 1, 15–19.

Gonzalez-Garcia, A. C., Sossa-Azuela, J. H., Felipe-Riveron, E. M., 2007. Image Retrieval based on Wavelet Computation and Neural Network Classification. In: Eighth International Workshop on Image Analysis for Multimedia Interactive Services 2007, Santorini, Greece. IEEE Computer Society, Los Alamitos, CA, USA, p. 44.

Goodchild, M. F., 2007. Citizens as Voluntary Sensors: Spatial data infrastructure in the World of Web 2.0. International Journal of Spatial Data Infrastructures Research 2, 24–32.

Goodchild, M. F., Yuan, M., Cova, T. J., 2007. Towards a general theory of geographic representation in GIS. International Journal of Geographical Information Science 21 (3), 239–260.
URL http://dx.doi.org/10.1080/13658810600965271

Goodchild, M. F., Zhou, J., 2003. Finding Geographic Information: Collection–Level Metadata. GeoInformatica 7 (2), 95–112.

Goodwin, J., Dolbear, C., Hart, G., 2008. Geographical Linked Data: The Administrative Geography of Great Britain on the Semantic Web. Transactions in GIS 12 (Suppl. 1), 19–30.
URL http://dx.doi.org/10.1111/j.1467-9671.2008.01133.x

Gooneratne, N., Tari, Z., 2008. Matching Independent Global Constraints for Composite Web Services. In: WWW'08.

Gould, M., 2007. Vertically interoperable geo–infrastructures and scalability. In: Specialist Meeting on Volunteered Geographic Information, Santa Barbara, December 13–14. Positional Paper.

Green, D., 2000. The evolution of Web searching. Online Information Review 24 (2), 124–137.
URL http://dx.doi.org/10.1108/14684520010330283

Greenberg, J., Pattuelli, M. C., Parsia, B., Robertson, W. D., 2001. Author–generated Dublin Core Metadata for Web Resources: A Baseline Study in an Organization. In: Proceedings of the International Conference on Dublin Core and Metadata Applications 2001. National Institute of Informatics, Tokyo, Japan, pp. 38–46.

Griesbaum, J., 2004. Evaluation of three German search engines: Altavista.de, Google.de and Lycos.de. Information Research: an international electronic journal 9 (4).

Gruber, T. R., 1993. A translation approach to portable ontology specifications. Knowledge Acquisition – Special issue: Current issues in knowledge modeling 5 (2), 199–220.
URL http://dx.doi.org/10.1006/knac.1993.1008

Hagemann, S., Letz, C., Vossen, G., 2007. Web Service Discovery – Reality Check 2.0. In: Proceedings of the Third International Conference on Next Generation Web Services Practices. NWESP '07. IEEE Computer Society, Washington, DC, USA, pp. 113–118.
URL http://dx.doi.org/10.1109/NWESP.2007.31

Hansen, B., 2008. The GeoURL ICBM Address Server. Homepage.
URL http://geourl.org/

Hao, Y., Zhang, Y., Cao, J., 2010. Web services discovery and rank: An information retrieval approach. Future Generation Computer Systems 26 (8), 1053 – 1062.
URL http://dx.doi.org/10.1016/j.future.2010.04.012

Hariharan, R., Hore, B., Mehrotra, S., 2008. Discovering gis sources on the web using summaries. In: Proceedings of the 8th ACM/IEEE–CS joint conference on Digital libraries. JCDL '08. ACM, New York, NY, USA, pp. 94–103.
URL http://dx.doi.org/10.1145/1378889.1378907

Harris, S., Seaborne, A., 2012. SPARQL 1.1 Query Language. W3C Working Draft.
URL http://www.w3.org/TR/sparql11-query/

Hausmann, J. H., Heckel, R., Lohmann, M., 2004. Model-based Discovery of Web Services. In: Proceedings of the IEEE International Conference on Web Services. pp. 6–9.

Hearst, M., Elliott, A., English, J., Sinha, R., Swearingen, K., Yee, K.-P., 2002. Finding the flow in web site search. Communications of the ACM 45 (9), 42–49.
URL http://dx.doi.org/10.1145/567498.567525

Hearst, M. A., 2000. Next Generation Web Search: Setting Our Sites. IEEE Data Engineering Bulletin 23, 38–48.

Hearst, M. A., 2006. Clustering versus faceted categories for information exploration. Communications of the ACM – Supporting exploratory search 49 (4), 59–61.
URL http://dx.doi.org/10.1145/1121949.1121983

Hearst, M. A., 2009. Search User Interfaces. Cambridge University Press.
URL http://searchuserinterfaces.com/

Heath, T., Bizer, C., 2011. Linked Data: Evolving the Web into a Global Data Space (1st edition). Vol. 1 of Synthesis Lectures on the Semantic Web: Theory and Technology. Morgan & Claypool.
URL http://dx.doi.org/10.2200/S00334ED1V01Y201102WBE001

Hick, I., 2011. HTML5. A vocabulary and associated APIs for HTML and XHTML. W3C Working Draft.
URL http://www.w3.org/TR/html5/

Hickson, I., 2011. HTML Living Standard. WHATWG Web Applications 1.0 specification.

Hill, L. L., 2006. Georeferencing: The Geographic Associations of Information (Digital Libraries and Electronic Publishing). The MIT Press.

Hill, L. L., Frew, J., Zheng, Q., 1999. Geographic Names: The Implementation of a Gazetteer in a Georeferenced Digital Library. D–Lib Magazine 5 (1).
URL http://dx.doi.org/10.1045/january99-hill

Hohpe, G., Woolf, B., 2003. Enterprise Integration Patterns. Addison–Wesley.

Humphreys, J., 2002. PhraseRate: An HTML Keyphrase Extractor. Tech. rep., University of California.

Hutchinson, M., Veenendall, B., 2005. Towards Using Intelligence to Move from Geocoding to Geolocating. In: Proceedings of 7th Annual GIS in Addressing Conference, Austin, Texas, USA.

INSPIRE DTM and EC/JRC, 2010. INSPIRE Metadata Implementing Rules: Technical Guidelines based on EN ISO 19115 and EN ISO 19119. Version 1.2.

INSPIRE NS DT, 2008. INSPIRE Network Service Architecture Version 3.0.

Inthiran, A., Alhashmi, S. M., Ahmed, P. K., 2010. A Reflection of Search Engine Strategies. Communications of the IBIMA.
URL http://dx.doi.org/10.5171/2010.126850

Ioannidis, Y., 2003. The history of histograms (abridged). In: Proceedings of the 29th International Conference on Very Large Data Bases. Vol. 29 of VLDB '2003. VLDB Endowment, pp. 19–30.

Iosifescu-Enescu, I., Hugentobler, M., Hurni, L., 2010. Web cartography with open standards – A solution to cartographic challenges of environmental management. Environmental Modelling & Software 25 (9), 988–999.

Isaac, A., Summers, E., 2009. SKOS Simple Knowledge Organization System Primer, W3C Working Group Note 18 August 2009.

ISO, 2003. ISO19115:2003, Geographic information - Metadata. Tech. Rep. 19115:2003, ISO/TC 211.

ISO, 2007a. ISO 3166–1:2006/Cor 1:2007 Codes for the representation of names of countries and their subdivisions – Part 1: Country codes.
URL http://www.iso.org/iso/country_codes/iso-3166-1_decoding_table.htm

ISO, 2007b. ISO 3166–2:2007 Codes for the representation of names of countries and their subdivisions – Part 2: Country subdivision code.
URL http://www.iso.org/iso/country_codes/iso-3166-1_decoding_table.htm

ISO, 2009. ISO 15836:2009 Information and documentation – The Dublin Core metadata element set.

ISO/TC 211, 2002. ISO 19113:2002. Geographic information – Quality principles.

ISO/TC 211, 2003. ISO 19114:2003 Geographic information – Quality evaluation procedures.

ISO/TC 211, 2006. ISO/TS 19138:2006 Geographic information – Data quality measures.

Jakobsson, A., 2006. On the Future of Topographic Base Information Management in Finland and Europe. Ph.D. thesis, Helsinki University of Technology.

Jakobsson, A., Zaccheddu, P., 2009. EuroGeoNames (EGN) – A Prototype Implementation for an INSPIRE Service. In: GSDI 11 World Conference and the 3rd INSPIRE Conference, Rotterdam 15–19 June 2009.

Janowicz, K., Schade, S., Bröring, A., Keßler, C., Maué, P., Stasch, C., Apr. 2010. Semantic Enablement for Spatial Data Infrastructures. Transactions in GIS 14 (2), 111–129.
URL http://dx.doi.org/10.1111/j.1467-9671.2010.01186.x

Janowicz, K., Schade, S., Bröring, A., Keßler, C., Stasch, C., 2009. A Transparent Semantic Enablement Layer for the Geospatial Web. In: Terra Cognita 2009 Workshop In conjunction with the 8th International Semantic Web Conference (ISWC 2009), October 26, 2009.

Järvelin, K., 2011. IR research: systems, interaction, evaluation and theories. In: Proceedings of the 33rd European conference on Advances in information retrieval. ECIR'11. Springer-Verlag, Berlin, Heidelberg, pp. 1–3.

Jefatura de estado, 2010. Law 14/2010, de 5 de julio, sobre las infraestructuras y los servicios de información geográfica en España. BOE no. 163.

Johnston, P., Powell, A., 2008. Expressing Dublin Core metadata using HTML/XHTML meta and link elements. DCMI Recommendation.
URL http://dublincore.org/documents/dc-html/

Jones, C. B., Alani, H., Tudhope, D., 2001. Geographical Information Retrieval with Ontologies of Place. In: COSIT 2001: Proceedings of the International Conference on Spatial Information Theory. Springer–Verlag, London, UK, pp. 322–335.

Jones, C. B., Purves, R. S., 2008. Geographical Information Retrieval. International Journal of Geographical Information Science 22 (3), 219–228, taylor & Francis.
URL http://dx.doi.org/10.1080/13658810701626343

Julea, A., Méger, N., Rigotti, C., Doin, M., Lasserre, C., Trouvé, E., Bolon, P., Lâzârescu, V., 2010. Extraction of frequent grouped sequential patterns from Satellite Image Time Series. In: Proceedings of IGARSS'10 – IEEE International Geoscience and Remote Sensing Symposium. pp. 3434–3437.

Kedzierski, A., 2002. Artur's Auto Annotator. Master's thesis, Department of Computer Science, University of California.

Keller, U., Lara, R., Lausen, H., Polleres, A., Predoiu, L., Toma, I., 2005. WSMX Deliverable. D10 v0.2. Semantic Web Service Discovery. Wsmx working draft, DERI.

Keßler, Carsten, J.-K., Bishr, M., 2009. An Agenda For The Next Generation Gazetteer: Geographic Information Contribution and Retrieval. In: 17th ACM SIGSPATIAL International Symposium on Advances in Geographic Information Systems, ACM-GIS 2009, November 4–6, 2009, Seattle, Washington, USA,. pp. 91–100.
URL http://dx.doi.org/10.1145/1653771.1653787

Khoo, M., Hall, C., 2010. Merging metadata: a sociotechnical study of crosswalking and interoperability. In: Proceedings of the 10th annual joint conference on Digital libraries. JCDL '10. ACM, New York, NY, USA, pp. 361–364.
URL http://dx.doi.org/10.1145/1816123.1816180

Kim, E., Lee, Y., Kim, Y., Park, H., Kim, J., Moon, B., Yun, J., Kang, G., 2011. Web Services Quality Factors Version 1.0. Committee Specification 01.

Kimler, M., 2004. Geo–Coding: Recognition of geographical references in unstructured text, and their visualisation. Ph.D. thesis, University of Applied Sciences Hof.

Klyne, G., Carroll, J. J., 2004. Resource Description Framework (RDF): Concepts and Abstract Syntax. W3C Recommendation.
URL http://www.w3.org/TR/rdf-concepts/

Kokash, N., Birukou, A., D'Andrea, V., 2007. Web service discovery based on past user experience. In: Proceedings of the 10th international conference on Business information systems. BIS'07. Springer-Verlag, Berlin, Heidelberg, pp. 95–107.

Kopecký, J., Simperl, E., 2008. Semantic web service offer discovery for e–commerce. In: ICEC '08: Proceedings of the 10th international conference on Electronic commerce. ACM, New York, NY, USA, pp. 1–6.
URL http://dx.doi.org/10.1145/1409540.1409579

Koren, J., Zhang, Y., Liu, X., 2008. Personalized Interactive Faceted Search. In: Proceedings of the 17th International Conference on World Wide Web. WWW '08. ACM, New York, NY, USA, pp. 477–486.
URL http://dx.doi.org/10.1145/1367497.1367562

Krauss, T., Reinartz, P., Lehner, M., Stilla, U., 2007. Coarse and fast modelling of urban areas from high resolution stereo satellite images. In: Urban Remote Sensing Joint Event 2007, Paris, France. pp. 1–12.

Kresse, W., Fadaie, K., 2004. ISO Standards for Geographic Information. Springer, Berlin.

Kressler, F., Steinnocher, K., Franzen, M., 2005. Object–oriented classification of orthophotos to support update of spatial databases. In: Proceedings of IGARSS'05 – IEEE International Geoscience and Remote Sensing Symposium. Vol. 1. pp. 253–256.

Kules, B., 2006. Supporting Exploratory Web Search with Meaningful and Stable Categorized Overviews. Ph.d. dissertation from the department of computer science, University of Maryland.

Kules, B., Capra, R., 2008. Creating Exploratory Tasks for a Faceted Search Interface. Vol. 19. p. 2010.

Kumar, R., Suri, P., Chauhan, R., 2005. Search Engines Evaluation. DESIDOC Bulletin of Information Technology 25 (2), 3–10.

Kwan, M.-P., Ransberger, D. M., 2010. LiDAR assisted emergency response: Detection of transport network obstructions caused by major disasters. Computers, Environment and Urban Systems 34 (3), 179–188.

Kwon, O. B., 2003. Meta web service: building web–based open decision support system based on web services. Expert Systems with Applications 24 (4), 375 – 389.
URL http://dx.doi.org/10.1016/S0957-4174(02)00187-2

La Barre, K., 2007. Faceted navigation and browsing features in new OPACs: robust support for scholarly information seeking? Knowledge Organization 34 (2), 78–90.

La Barre, K., 2010. Facet analysis. Annual Review of Information Science and Technology 44 (1), 243–284.
URL http://dx.doi.org/10.1002/aris.2010.1440440113

Lacasta, J., López-Pellicer, F. J., Muro-Medrano, P. R., Nogueras-Iso, J., Zarazaga-Soria, F. J., 2007. ThManager: an open source tool for creating and visualizing SKOS. Information Technology and Libraries 26 (3).

Lan, G., Huang, Q., 2007. Ontology-based Method for Geospatial Web Services Discovery. In: ISKE '07: Proceedings of the International Conference on Intelligent Systems and Knowledge Engineering.

Langegger, A., Wöß, W., Blöchl, M., 2008. A semantic web middleware for virtual data integration on the web. In: Proceedings of the 5th European semantic web conference on The semantic web: research and applications. ESWC'08. Springer-Verlag, Berlin, Heidelberg, pp. 493–507.

Larry Klein, A. T., 2007. HDF–EOS5 Data Model, File Format and Library. Recommended Standar.

Latre, M., Lacasta, J., Mojica, E., Nogueras-Iso, J., Zarazaga-Soria, F., 2009. An approach to facilitate the integration of hydrological data by means of ontologies and multilingual thesauri. Advances in GIScience, 155–171.

Lausen, H., Steinmetz, N., 2008. Survey of current means to discover web services. Tech. rep., Semantic Technology Institute.

Lee, K., Jeon, J., Lee, W., Jeong, S.-H., Park, S.-W., 2003. QoS for Web Services: Requirements and Possible Approaches. W3C Working Group Note 25.

Lee, Y., Kim, E., 2006. A Study for Web Service Quality Description Language (WSQDL). NCA IV-RER-04052. Not available any more.

Leidner, J. L., 2007. Toponym Resolution in Text. Ph.D. thesis, University of Edinburgh.

Leite, F. L., de Souza Baptista, C., Silva, P. D. A., Silva, E. R. D., 2006. WS–GIS: Towards a SOA-Based SDI Federation. In: Brazilian Symposium on GeoInformatics. pp. 199–214.

Leroy, G., Xu, J. J., Chung, W., Eggers, S., Chen, H., 2006. End User Evaluation of Query Formulation and Results Review Tools in Three Medical Meta–Search Engines. International Journal of Medical Informatics 11, 780–789.
URL http://dx.doi.org/10.1016/j.ijmedinf.2006.08.001

Li, W., Yang, C., 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.
URL http://dx.doi.org/10.1080/13658810903514172

Li, Z., Yang, C. P., Wu, H., Li, W., Miao, L., 2011. An optimized framework for seamlessly integrating OGC Web Services to support geospatial sciences. International Journal of Geographical Information Science 25 (4), 595–613.
URL http://dx.doi.org/10.1080/13658816.2010.484811

Lo, C., Yeung, A. K. W., 2006. Concepts and Techniques of Geographic Information Systems, 2/E. Prentice Hall.

Longueville, B. D., 2010. Community–based geoportals: The next generation? Concepts and methods for the geospatial Web 2.0. Computers, Environment and Urban Systems 34 (4), 299–308.
URL http://dx.doi.org/10.1016/j.compenvurbsys.2010.04.004

López-Pellicer, F. J., 2011. Semantic Linkage of the Invisible Geospatial Web. Ph.D. thesis, Universidad de Zaragoza.

López-Pellicer, F. J., Béjar, R., Florczyk, A. J., Muro-Medrano, P. R., Zarazaga-Soria, F., 2010a. State of Play of OGC Web Services across the Web. In: Proceedings of INSPIRE Conference 2010: INSPIRE as a framework for cooperation, Krakow, Poland, 22–25 June 2010.

López-Pellicer, F. J., Béjar, R., Florczyk, A. J., Muro-Medrano, P. R., Zarazaga-Soria, F., 2011a. A review of the implementation of OGC Web Services across Europe. International Journal of Spatial Data Infrastructures Research 6 (1), 168–186.

López-Pellicer, F. J., Béjar, R., Renteria-Agualimpia, W., Florczyk, A., Muro-Medrano, P., Zarazaga-Soria, F., 2011b. Status of INSPIRE inspired OGC Web Services. In: INSPIRE Conference 2011. European Commission Joint Research Centre.

López-Pellicer, F. J., Florczyk, A., Béjar, R., Nogueras-Iso, J., Zarazaga-Soria, F., Muro-Medrano, P., 2010b. State of Play: Spain and Portugal. In: JIIDE'2010 - I Iberian Conference on Spatial Data Infrastructures.

López-Pellicer, F. J., Florczyk, A., Renteria-Agualimpıa, W., Nogueras-Iso, J., Muro-Medrano, P., 2011c. Publishing standard geospatial catalogues in the Web of Data. In: 14th Conference of the Spanish Association for Artificial Intelligence (CAEPIA 2011). AEPIA.

López-Pellicer, F. J., Florczyk, A., Rentería-Aguaviva, W., Nogueras-Iso, J., Muro-Medrano, P., 2011d. CSW2LD: a Linked Data frontend for CSW. In: II Iberian Conference on Spatial Data Infrastructures (JIIDE 2011). Institut Cartogràfic de Catalunya.

López-Pellicer, F. J., Florczyk, A. J., Béjar, R., Muro-Medrano, P. R., Zarazaga-Soria, F. J., 2011e. Discovering geographic web services in search engines. Online Information Review 35 (6), 909–927, in press.
URL http://dx.doi.org/10.1108/14684521111193193

López-Pellicer, F. J., Florczyk, A. J., Lacasta, J., Zarazaga-Soria, F. J., Muro-Medrano, P. R., 2008. Administrative Units, an Ontological Perspective. In: Song, I.-Y. e. a. (Ed.), ER Workshops. Vol. 5232 of Lecture Notes in Computer Science. Springer, pp. 354–363.
URL http://dx.doi.org/10.1007/978-3-540-87991-6_42

López-Pellicer, F. J., Florczyk, A. J., Nogueras-Iso, J., Muro-Medrano, P. R., Zarazaga-Soria, F., 2010c. Exposing CSW catalogues as Linked Data. In: Painho, M., Santos, M., Pundt, H. (Eds.), Geospatial Thinking, Lecture Notes in Geoinformation and Cartography. Springer–Verlag, pp. 183–200.

López-Pellicer, F. J., Lacasta, J., Florczyk, A. J., Nogueras-Iso, J., Zarazaga-Soria, F., 2011f. An Ontology for the representation of Spatio–Temporal Jurisdictional Domains in Information Retrieval Systems. International Journal of Geographical Information Science In press.
URL http://dx.doi.org/10.1080/13658816.2011.599811

Luca, Padovani, 2010. Contract–based discovery of Web services modulo simple orchestrators. Theoretical Computer Science 411 (37), 3328–3347.
URL http://dx.doi.org/10.1016/j.tcs.2010.05.002

Mabrouk, M., Bychowski, T., Williams, J., Niedzwiadek, H., Bishr, Y., Gaillet, J.-F., Crisp, N., Wilbrink, W., Horhammer, M., Roy, G., Margoulies, S., Fuchs, G., Hendrey, G., 2008. OpenGIS Location Services (OpenLS): Core Services. Version:1.2. OpenGIS Interface Standard. OGC 07-074.

Mabrouk, N., Georgantas, N., Issarny, V., 2009. A semantic end–to–end QoS model for dynamic service oriented environments. In: Principles of Engineering Service Oriented Systems, 2009. PESOS 2009. ICSE Workshop on. pp. 34–41.
URL http://dx.doi.org/10.1109/PESOS.2009.5068817

MacFarlane, A., 2007. Evaluation of web search for the information practitioner. Aslib Proceedings 59 (4–5), 352–366.
URL http://dx.doi.org/10.1108/00012530710817573

Maguire, D. J., Longley, P. A., 2005. The emergence of geoportals and their role in spatial data infrastructures. Computers, Environment and Urban Systems 29 (1), 3–14.
URL http://dx.doi.org/10.1016/j.compenvurbsys.2004.05.012

Mandell, D. J., McIlraith, S. A., 2003. A Bottom–Up Approach to Automating Web Service Discovery, Customization, and Semantic Translation. In: Proceedings of the Twelfth International World Wide Web Conference Workshop on E-Services and the Semantic Web (ESSW'03).

Manikrao, U. S., Prabhakar, T., 2005. Dynamic Selection of Web Services with Recommendation System. In: Proceedings of the International Conference on Next Generation Web Services Practices. NWESP '05. IEEE Computer Society, Washington, DC, USA, pp. 117–.

Manning, C., Klein, D., 2003. Optimization, maxent models, and conditional estimation without magic. In: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: Tutorials – Volume 5. NAACL–Tutorials '03. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 8–8.
URL http://dx.doi.org/10.3115/1075168.1075176

Manolopoulos, Y., Nanopoulos, A., Papadopoulos, A. N., Theodoridis, Y., 2006. R–Trees: Theory and Applications. Advanced Information and Knowledge Processing. Springer London.
URL http://dx.doi.org/10.1007/978-1-84628-293-5

Mansourian, A., Zoje, M. V., Mohammadzadeh, A., Farnaghi, M., 2008. Design and implementation of an on–demand feature extraction web service to facilitate development of spatial data infrastructures. Computers, Environment and Urban Systems 32 (5), 377–385.

Marchionini, G., 2006. Exploratory search: from finding to understanding. Communications of the ACM 49, 41–46.
URL http://dx.doi.org/10.1145/1121949.1121979

Margoulies, S., 2001. Geocoder Service Draft Candidate Implementation Specification. Version 0.7.6. Draft Candidate. OpenGIS Consortium Discussion Paper. OGC 01-026r1.

Martin, D., Burstein, M., Mcdermott, D., Mcilraith, S., Paolucci, M., Sycara, K., Mcguinness, D. L., Sirin, E., Srinivasan, N., 2007a. Bringing Semantics to Web Services with OWL–S. World Wide Web 10 (3), 243–277.
URL http://dx.doi.org/10.1007/s11280-007-0033-x

Martin, D., Paolucci, M., Wagner, M., 2007b. Bringing Semantic Annotations to Web Services: OWL–S from the SAWSDL Perspective. In: 6th International and 2nd Asian Semantic Web Conference (ISWC2007+ASWC2007).

Mattmann, C., Zitting, J., 2011. Tika in Action. Manning Publications.

Maué, P., 2008. An extensible semantic catalogue for geospatial web services. International Journal of Spatial Data Infrastructures Research 3, 168–191.
URL http://ijsdir.jrc.ec.europa.eu/index.php/ijsdir/article/view/76

Maué, P., Schade, S., 2009. Data Integration in the Geospatial Semantic Web. Journal of Cases on Information Technology 11 (4), 100–122.
URL http://dx.doi.org/10.4018/jcit.2009072105

Maué, P., Schade, S., Duchesne, P., 2009. Semantic Annotations in OGC Standards. Version: 0.3.0. OpenGIS Discussion Paper. OGC 08-167r1.

McElroy, J. A., Remington, P. L., Trentham-Dietz, A., Robert, S. A., Newcomb, P. A., 2003. Geocoding addresses from a large population–based study: lessons learned. Epidemiology 14 (4), 399–407.
URL http://dx.doi.org/10.1097/01.EDE.0000073160.79633.c1

Medeiros, N., 2001. A craftsman and his tool: Andy Powell and the DC–dot metadata editor. OCLC Systems & Services 17 (2), 60–64.
URL http://dx.doi.org/10.1108/10650750110391939

Meisner, R., Lang, S., Jungert, E., Almer, A., Tiede, D., Sparwasser, N., Mertens, K., Gobel, R., Blaschke, T., de la Cruz, A., Stelzl, H., Silvervarg, K., 2009. Data Integration and Visualization for Crisis Applications. In: Jasani, B. e. a. (Ed.), Remote Sensing from Space. Vol. III. Springer Netherlands, pp. 141–160.

Menczer, F., 2003. Complementing search engines with online web mining agents. Decision Support Systems – Special issue: Web data mining 35, 195–212.
URL http://dx.doi.org/10.1016/S0167-9236(02)00106-9

Metatags Company Inc., 2012. Meta tags Website.
URL http://www.metatags.org/all_metatags

Miller, G., 1956. The magical number seven, plus or minus two: some limits on our capacity for processing information. Psychological Review 63 (2), 81–97.
URL http://dx.doi.org/10.1037/h0043158

Minsky, M., 1986. The society of mind. Simon & Schuster, Inc., New York, NY, USA.

Mitchell, S., 2006. Data Fountains. A National, Cooperative Information Utility for Shared Internet Resource Discovery, Metadata Application and Rich, Full-text Harvest of Value to Internet Portals, Virtual Libraries and Library Catalogs with Portal-like Capabilities. IMLS WebWise 2006 Presentation.

Mitchell, S., Mooney, M., Mason, J., Paynter, G., Ruscheinski, J., Kedzierski, A., Humphreys, K., 2003. iVia open source virtual library system. D–Lib Magazine 9 (1).

Mitsos, A., Shirakawa, T., Adam, R., Lautenbacher, C. C., 2005. Global Earth Observation System of Systems GEOSS, 10-Year Implementation Plan Reference Document. ESA Publications Division.

Mohammed, S., Fiaidhi, J., Hahn, M., 2006. A UDDI Search Engine for SVG Federated Medical Imaging Web Services. Journal of Computer Science 2 (4), 303–313.
URL http://dx.doi.org/10.3844/jcssp.2006.303.313

Nack, F., Ossenbruggen, J. v., Hardman, L., 2005. That Obscure Object of Desire: Multimedia Metadata on the Web, Part 2. IEEE MultiMedia 12 (1), 54–63.
URL http://dx.doi.org/10.1109/MMUL.2005.12

National Institute of Standards and Technology, 1995. FIPS PUB 10–4: Standard for Countries, Dependencies, Areas of Special Sovereignty and Their Principal Administrative Divisions. National Institute of Standards and Technology, Gaithersburg, MD, USA.

Nebert, D., Whiteside, A., Vretanos, P., 2007. OpenGIS Catalogue Services Specification. OGC 07-006r1.

Nebert, D. D., 2004. Developing Spatial Data Infrastructures: The SDI Cookbook. Global Spatial Data Infrastructure.

NGA, 1996. Performance Specification Digital Terrain Elevation Data (DTED). Specification.

Nielsen, J., 1989. Coordinating User Interfaces for Consistency. Academic Press, Boston, reissued 2002 by Morgan Kaufmann Publishers, San Francisco, ISBN 0-12-518400-X.

Nielsen, J., 2003. Usability 101: Definition and Fundamentals – What, Why, How. Jakob Nielsen's Alertbox.
URL http://www.useit.com/alertbox/20030825.html

Nielsen, J., 2008. Top–10 Application–Design Mistakes. Jakob Nielsen's Alertbox.
URL http://www.useit.com/alertbox/application-mistakes.html

NIMA, 2004. Department of Defense World Geodetic System 1984, Its Definition and Relationships With Local Geodetic Systems (NIMA TR8350.2), Third Edition. Tech. rep., National Imagery and Mapping Agency.

Nogueras-Iso, J., Barrera, J., Rodríguez, A., Recio, R., Laborda, C., Zarazaga-Soria, F., 2009. Development and deployment of a services catalog in compliance with the inspire metadata implementing rules. In: van Loenen, B. e. a. (Ed.), Spatial Data Infrastructure Convergence: Research, Emerging Trends, and Critical Assessment. Vol. 48 of Groene. The Netherlands Geodetic Commission (NGC), The Netherlands, pp. 21–34.

Nogueras-Iso, J., Zarazaga-Soria, F., Lacasta, J., Béjar, R., Muro-Medrano, P., 2004. Metadata standard interoperability: application in the geographic information domain. Computers, Environment and Urban Systems 28 (6), 611 – 634.
URL http://dx.doi.org/10.1016/j.compenvurbsys.2003.12.004

NSF, 2003. Revolutionizing science and engineering through cyberinfrastructure: Report of the national science foundation blue–ribbon advisory panel on cyberinfrastructure. NSF Panel reports. Tech. rep., NFS.

NSF, 2007. Cyberinfrastructure vision for 21st century discovery. Tech. rep.

OASIS, 2012. OASIS homepage.
URL http://www.oasis-open.org/org

Oscarsson, D., 2001. Simple ASCII Compatible Encoding (SACE). Internet Draft.
URL http://tools.ietf.org/id/draft-ietf-idn-sace-00.txt

Ossenbruggen, J. v., Nack, F., Hardman, L., 2004. That Obscure Object of Desire: Multimedia Metadata on the Web, Part 1. IEEE MultiMedia 11 (4), 38–48.
URL http://dx.doi.org/10.1109/MMUL.2004.36

Overell, S., Rüger, S., 2008. Using co–occurrence models for placename disambiguation. International Journal of Geographical Information Science 22 (3), 265–287.
URL http://dx.doi.org/10.1109/MMUL.2004.36

Palmonari, M., Comerio, M., De Paoli, F., 2009. Effective and Flexible NFP–Based Ranking of Web Services. In: Baresi, L., Chi, C.-H., Suzuki, J. (Eds.), Service–Oriented Computing. Vol. 5900 of Lecture Notes in Computer Science. Springer Berlin / Heidelberg, pp. 546–560.
URL http://dx.doi.org/10.1007/978-3-642-10383-4_40

Paynter, G. W., 2005. Developing practical automatic metadata assignment and evaluation tools for internet resources. In: JCDL '05: Proceedings of the 5th ACM/IEEE–CS joint conference on Digital libraries. ACM, New York, NY, USA, pp. 291–300.
URL http://dx.doi.org/10.1145/1065385.1065454

Percivall, G., 2002. OpenGIS Service Architecture (Version 4.3). The OpenGIS Abstract Specification and ISO/DIS 19119. Geographic information — Services.

Pérez-Pérez, M., Rodrigo, P., Usón, M., Fernández Ruiz, M., Morlán, V., Laiglesia, S., Florczyk, A., Lopez-Pellicer, F., 2009. IDEZar 2.0 para la Administración y Gestión de Incidencias de Policía Local. In: JIDEE 2009 - VI Jornadas de la Infraestructura de Datos Espaciales de España.

Pierce, M. E., Fox, G. C., Choi, J. Y., Guo, Z., Gao, X., Ma, Y., 2009. Using Web 2.0 for scientific applications and scientific communities. Concurrency and Computation: Practice & Experience – Web 2.0, Semantics, Knowledge and Grid 21, 583–603.

Polfreman, M., Rajbhandari, S., 2008. MetaTools – Investigating Metadata Generation Tools. Final Report. Tech. rep., Joint Information Systems Committee (JISC).

Portele, C., 2012. OGC Geography Markup Language (GML) — Extended schemas and encoding rules. Version: 3.3.0. OGC 10-129r1.

Powell, A., Nilsson, M., Naeve, A., Johnston, P., Baker, T., 2007. DCMI Abstract Model.
URL http://dublincore.org/documents/2007/04/02/abstract-model/

Price, G., Sherman, C., 2001. The Invisible Web: Uncovering Information Sources Search Engines Can't See. Information Today, Inc.

Putkey, T., 2011. Using SKOS to Express Faceted Classification on the Semantic Web. Library Philosophy and Practice.
URL http://unllib.unl.edu/LPP/putkey.htm

Raggett, D., Hors, A. L., (eds), I. J., 1999. HTML 4.01 Specification. W3C Recommendation.
URL http://www.w3.org/TR/html401/

Raghavan, S., Garcia-Molina, H., 2001. Crawling the Hidden Web. In: Proceedings of the 27th International Conference on Very Large Data Bases. VLDB '01. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp. 129–138.

Ranganathan, S., 1967. Prolegomena to library classification, 3rd edition. New York: Asia Publishing House.

Ranganathan, S., 2006. Colon classification: basic classification. Ess Ess Pub, reprint.

Ratcliffe, J. H., 2001. On the accuracy of TIGER-type geocoded address data in relation to cadastral and census areal units. International Journal of Geographical Information Science 15 (5), 473–485.
URL http://dx.doi.org/10.1080/13658810110047221

Ratcliffe, J. H., 2004. Geocoding crime and a first estimate of a minimum acceptable hit rate. International Journal of Geographical Information Science 18, 61–72.

Reed, C., 2006. An Introduction to GeoRSS: A Standards Based Approach for Geo-enabling RSS feeds. OGC White Paper. OGC 06-050r3.

Ritter, N., Ruth, M., 2000. GeoTIFF Format Specification (version: 1.8.2).
URL http://www.remotesensing.org/geotiff/spec/geotiffhome.html

Rodriguez, J. M., Crasso, M., Zunino, A., Campo, M., 2010. Improving Web Service descriptions for effective service discovery. Science of Computer Programming 75 (11), 1001–1021.
URL http://dx.doi.org/10.1016/j.scico.2010.01.002

Rose, D. E., 2006. Reconciling information–seeking behavior with search user interfaces for the Web. Journal of the American Society for Information Science and Technology 57 (6), 797–799.
URL http://dx.doi.org/10.1002/asi.20295

Rose, D. E., Levinson, D., 2004. Understanding user goals in web search. In: WWW '04: Proceedings of the 13th international conference on World Wide Web. ACM, New York, NY, USA, pp. 13–19. URL http://dx.doi.org/10.1145/988672.988675

Rose, L. C., 2004. Geospatial portal reference architecture. Discussion Paper. OGC 04-039.

Ru, Y., Horowitz, E., 2005. Indexing the invisible web: a survey. Online Information Review 29 (3), 249 – 265, emerald Group Publishing Limited.
URL http://dx.doi.org/10.1108/14684520510607579

Sample, J. T., Ladner, R., Shulman, L., Ioup, E., Petry, F., Warner, E., Shaw, K., McCreedy, F. P., 2006. Enhancing the US Navy's GIDB Portal with Web Services. IEEE Internet Computing 10, 53–60.

Saracevic, T., 1997. The Stratified Model of Information Retrieval Interaction: Extension and Applications. In: Proceedings of the ASIS Annual Meeting. Vol. 34. pp. 313–27.

Schmitt, M., Stilla, U., 2010. Utilization of airborne multi-aspect InSAR data for the generation of urban ortho-images. In: Proceedings of IGARSS'10 – IEEE International Geoscience and Remote Sensing Symposium. pp. 3937–3940.

Schoenhardt, N., Ioup, E., McCreedy, F., 2010. A system to distribute Navy Coastal Ocean Model data using the open Geospatial consortium's Web Map Service protocol. In: Innerspace: A Global Responsibility, OCEANS 2010 MTS/IEEE Seattle, Seattle, Washington, USA. pp. 1–7.

Schut, P., 2004a. Geolinked Data Access Service (GDAS). Version: 0.9.1. OpenGIS Discussion Paper. OGC 04-010r1.

Schut, P., 2004b. Geolinking Service (GLS). Version 0.9.1. OpenGIS Implementation. Abstract Specification. OGC 04-011.

Schut, P., 2007. OpenGIS Web Processing Service. Version 1.0.0. OpenGIS Standard. OGC 05-007r7.

Schut, P., 2009. Geographic Linkage Service (GLS) Specification. Version: 0.12. OGC Draft Implementation Specification. OGC 08-006r2.

Schut, P., 2010. OpenGIS Georeferenced Table Joining Service (TJS) Implementation Standard. Version: 1.0.0. OGC 10-070r2.

Seaman, C. B., G. Mendonça, M., Basili, V. R., Kim, Y.-M., 2003. User Interface Evaluation and Empirically–Based Evolution of a Prototype Experience Management Tool. IEEE Transactions on Software Engineering 29 (9), 838–850.
URL http://dx.doi.org/10.1109/TSE.2003.1232288

ShaikhAli, A., Rana, O. F., Al-Ali, R., Walker, D. W., 2003. UDDIe: An Extended Registry for Web Services. In: SAINT-W '03: Proceedings of the 2003 Symposium on Applications and the Internet Workshops (SAINT'03 Workshops). IEEE Computer Society, Washington, DC, USA, p. 85.

Silva, M. J., Martins, B., Chaves, M., Afonso, A. P., Cardoso, N., 2006. Adding Geographic Scopes to Web Resources. Computers, Environment and Urban Systems 30, 378–399.
URL http://dx.doi.org/10.1016/j.compenvurbsys.2005.08.003

Sinha, P., Jain, R., 2008. Classification and annotation of digital photos using optical context data. In: Proceedings of the 2008 International Conference on Content-based Image and Video Retrieval, Niagara Falls, Canada. CIVR'08. Association for Computing Machinery, New York, NY, USA, pp. 309–318.

Sioutas, S., Sakkopoulos, E., Makris, C., Vassiliadis, B., Tsakalidis, A., Triantafillou, P., 2009. Dynamic Web Service discovery architecture based on a novel peer based overlay network. Journal of Systems and Software 82 (5), 809–824.
URL http://dx.doi.org/10.1016/j.jss.2008.11.845

Skirvin, S. M., Kepner, W. G., Marsh, S. E., Drake, S. E., Maingi, J. K., Edmonds, C. M., Watts, C. J., Williams, D. R., 2004. Assessing the accuracy of satellite–derived land–cover classification using historical aerial photography, digital orthophoto quadrangles, and airborne video data. In: Remote Sensing and GIS Accuracy Assessment. CRC Press, London, pp. 115–131.

Sreenath, R. M., Singh, M. P., 2004. Agent–based service selection. Web Semantics: Science, Services and Agents on the World Wide Web 1 (3), 261–279.
URL http://dx.doi.org/10.1016/j.websem.2003.11.006

Sriharee, N., 2006. Semantic Web Services Discovery Using Ontology-Based Rating Model. In: WI '06: Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence. IEEE Computer Society, Washington, DC, USA, pp. 608–616.
URL http://dx.doi.org/10.1109/WI.2006.155

Stadler, C., Lehmann, J., Höffner, K., Auer, S., 2011. LinkedGeoData: A Core for a Web of Spatial Open Data. Semantic Web Journal.

Steinmetz, N., Lausen, H., 2009. Ontology–based feature aggregation for multi–valued ranking. In: Proceedings of the 2009 international conference on Service-oriented computing. ICSOC/ServiceWave'09. Springer-Verlag, Berlin, Heidelberg, pp. 258–268.

Steinmetz, N., Lausen, H., Brunner, M., 2009. Web Service Search on Large Scale. In: Proceedings of the 7th International Joint Conference on Service-Oriented Computing. ICSOC-ServiceWave '09. Springer-Verlag, Berlin, Heidelberg, pp. 437–444.
URL http://dx.doi.org/10.1007/978-3-642-10383-4_32

Stollberg, M., Norton, B., 2007. A Refined Goal Model for Semantic Web Services. In: ICIW '07 Proceedings of the Second International Conference on Internet and Web Applications and Services. ICIW '07. IEEE Computer Society, Los Alamitos, CA, USA, pp. 17–.
URL http://dx.doi.org/10.1109/ICIW.2007.7

Sycara, K., Paolucci, M., Ankolekar, A., Srinivasan, N., 2003. Automated discovery, interaction and composition of Semantic Web Services. Web Semantics: Science, Services and Agents on the World Wide Web 1, 27–46.
URL http://dx.doi.org/10.1016/j.websem.2003.07.002

Tait, M. G., 2005. Implementing geoportals: applications of distributed GIS. Computers, Environment and Urban Systems 29 (1), 33–47.
URL http://dx.doi.org/10.1016/j.compenvurbsys.2004.05.011

Talantikite, H. N., Aissani, D., Boudjlida, N., 2009. Semantic annotations for web services discovery and composition. Computer Standards & Interfaces 31 (6), 1108–1117.
URL http://dx.doi.org/10.1016/j.csi.2008.09.041

Tawileh, W., Mandl, T., Griesbaum, J., 2010. Evaluation of five web search engines in Arabic language. In: Atzmüller, M., Benz, D., Hotho, A., Stumme, G. (Eds.), Proceedings of LWA2010 - Workshop-Woche: Lernen, Wissen & Adaptivitaet. Kassel, Germany.

The White House, 1994. Executive Order 12906 of April 11, 1994, Coordinating Geographic Data Acquisition and Access: The National Spatial Data Infrastructure. Federal Register 59 (71), 17671–17674.

Tsesmetzis, D., Roussaki, I., Sykas, E., 2008. QoS–aware service evaluation and selection. European Journal of Operational Research 191 (3), 1101 – 1112.
URL http://dx.doi.org/10.1016/j.ejor.2007.07.015

Turner, A. J., 2006. Introduction to Neogeography. O'Reilly Media, Inc.

UNGIWG, 2007. Strategy for Developing and Implementing a United Nations Spatial Data Infrastructure in support of Humanitarian Response Economic Development Environmental Protection Peace and Safety.

van Oort, P., 2005. Spatial data quality: from description to application. Publications on Geodesy 60. Netherlands Geodetic Commission, Delf.

Verma, K., Sivashanmugam, K., Sheth, A., Patil, A., Oundhakar, S., Miller, J., 2004. METEOR–S WSDI: A Scalable Infrastructure of Registries for Semantic Publication and Discovery of Web Services. International Journal of Information Technology and Management 6 (1), 17–39.
URL http://dx.doi.org/10.1007/s10799-004-7773-4

Vickery, B. C., 1966. Faceted classification schemes. Vol. 5 of Rutgers series on systems for the intellectual organization of information. Graduat School of Library Service, Rutgers, The State University, New Brunswick, N.J.

Vitvar, T., Kopecky, J., Viskova, J., Fensel, D., 2008. WSMO–Lite Annotations for Web Services. In: Hauswirth, M., Koubarakis, M., Bechhofer, S. (Eds.), ESWC'08 Proceedings of the 5th European semantic web conference on The semantic web: research and applications. LNCS. Springer Verlag, Berlin, Heidelberg.

Vretanos, P., 2010a. OpenGIS Filter Encoding 2.0 Encoding Standard. Version: 2.0.0. OGC 09-026r1 and ISO/DIS 19143.

Vretanos, P. A., 2010b. OpenGIS Web Feature Service 2.0 Interface Standard. OGC 09-025r1 and ISO/DIS 19142.

W3C OWL Working Group, 2009. OWL 2 Web Ontology Language. W3C Recommendation. URL http://www.w3.org/TR/owl2-overview/

W3C SWEO, 2012. Linking Open Data project. Community Web site. URL http://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData

Walker, R., 2008. A general approach to addressing. In: ISO Workshop on address standards: Considering the issues related to an international address standard.

Wallace, J., 2007. Spatially Enabling Mortgage Markets in Australia. In: Rajabifard, A. (Ed.), Towards a Spatially Enabled Society. Melbourne: Melbourne University, pp. 119–138.

Wang, H., Shi, Y., Zhou, X., Zhou, Q., Shao, S., Bouguettaya, A., oct. 2010a. Web Service Classification Using Support Vector Machine. In: Tools with Artificial Intelligence (ICTAI), 2010 22nd IEEE International Conference on. Vol. 1. pp. 3 –6. URL http://dx.doi.org/10.1109/ICTAI.2010.9

Wang, P., Chao, K.-M., Lo, C.-C., 2010b. On optimal decision for QoS–aware composite service selection. Expert Systems with Applications 37 (1), 440–449. URL http://dx.doi.org/10.1016/j.eswa.2009.05.070

Wang, S., Zhu, X.-G., 2008. Coupling cyberinfrastructure and geographic information systems to empower ecological and environmental research. BioScience 58 (2), 94–95.

Wang, X., Vitvar, T., Kerrigan, M., Toma, I., 2006. A QoS-aware Selection Model for Semantic Web Services. In: Dan, A., Lamersdorf, W. (Eds.), ICSOC'06: 4th International Conference on Service Oriented Computing, Chicago, IL, USA, December 4–7. Vol. 4294 of Lecture Notes in Computer Science. Springer, pp. 390–401.

Wegener, M., 2000. Spatial models and GIS. In: Fotheringham, A., Wegener, M. (Eds.), Spatial Models and GIS: New Potential and New Models. GISDATA 7. London: Taylor & Francis, pp. 3–20.

White, R. W., Kules, B., Drucker, S. M., Schraefel, M., 2006. Supporting Exploratory Search, Introduction. Communications of the ACM 49 (4), 36–39.

White, R. W., Richardson, M., Bilenko, M., Heath, A. P., 2008. Enhancing web search by promoting multiple search engine use. In: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval. SIGIR '08. ACM, New York, NY, USA, pp. 43–50.
URL http://dx.doi.org/10.1145/1390334.1390344

Whiteside, A., Greenwood, J., 2010. OGC Web Services Common Standard, Version: 2.0.0. OpenGIS Implementation Standard, oGC 06-121r9.

Whitsel, E. A., Rose, K. M., Wood, J. L., Henley, A. C., Liao, D., Heiss, G., 2004. Accuracy and Repeatability of Commercial Geocoding. American Journal of Epidemiology 160 (10), 1023—1029.

Wick, M., 2012. GeoNames homepage. online.
URL http://www.geonames.org/

Williamson, I., Rajabifard, A., Holland, P., 2010. Spatially Enabled Society. In: FIG Congress 2010, Facing the Challenges - Building the Capacity, Sydney, Australia, 11-16 April 2010.

Wilson, T., Burggraf, D., Lake, R., Patch, S., Martell, R., McClendon, B., Hagemark, M. J. M. A. B., Wernecke, J., Reed, C., 2008. OGC KML. OGC Standard. Version: 2.2.0. OGC 07-147r2.

Wolf-Tilo Balke, M. W., 2003. Towards Personalized Selection of Web Services. In: Proceedings of the Twelfth International World Wide Web Conference, WWW2003, Budapest, Hungary, 20–24 May 2003.

Wu, B., Wu, X., 2010. A QoS–aware Method for Web Services Discovery. Journal of Geographic Information System 2 (1), 40–44.

Wu, H., Li, Z., Zhang, H., Yang, C., Shen, S., 2010. Monitoring and evaluating the quality of Web Map Service resources for optimizing map composition over the internet to support decision making. Computers & Geosciences 37 (4), 485–494.

Wukovitz, L. D., Jan 2001. Using internet search engines and library catalogs to locate toxicology information. Toxicology 157 (1–2), 121–39.
URL http://dx.doi.org/10.1016/S0300-483X(00)00343-7

Yang, C. P., Raskin, R., Goodchild, M. F., Gahegan, M., 2010. Geospatial Cyberinfrastructure: Past, present and future. Computers, Environment and Urban Systems 34 (4), 264–277.

Yang, D.-H., Bilaver, L. M., Hayes, O., Goerge, R., 2004. Improving Geocoding Practices: Evaluation of Geocoding Tools. Journal of Medical Systems 28, 361–370.
URL http://dx.doi.org/10.1023/B:JOMS.0000032851.76239.e3

YDN, 2012. Yahoo! GeoPlanet homepage. REST Web Service, the API is accessed via HTTP GET.
URL http://developer.yahoo.com/geo/geoplanet/

Yu, T., jay Lin, K., 2005. Service Selection Algorithms for Composing Complex Services with Multiple QoS Constraints. In: Benatallah, B., Casati, F., Traverso, P. (Eds.), ICSOC'05: 3rd International Conference on Service Oriented Computing, Amsterdam, The Netherlands, December 12–15. Vol. 3826 of Lecture Notes in Computer Science. Springer, pp. 130–143.

Yue, P., Gong, J., Di, L., He, L., Wei, Y., 2011. Integrating semantic web technologies and geospatial catalog services for geospatial information discovery and processing in cyberinfrastructure. GeoInformatica 15 (2), 273–303.
URL http://dx.doi.org/10.1007/s10707-009-0096-1

Zandbergen, P. A., 2009. Geocoding Quality and Implications for Spatial Analysis. Geography Compass 3 (2), 647–680.
URL http://dx.doi.org/10.1111/j.1749-8198.2008.00205.x

Zhang, C., Zhao, T., Li, W., 2010. Automatic search of geospatial features for disaster and emergency management. International Journal of Applied Earth Observation and Geoinformation 12 (6), 409–418.

Zhang, X., Chen, X., Zhan, Q., 2009. Study on Geographic Information Services Taxonomy. In: Computer Sciences and Convergence Information Technology, 2009. ICCIT '09. Fourth International Conference on. pp. 204–209.
URL http://dx.doi.org/10.1109/ICCIT.2009.136

Zong, W., Wu, D., Sun, A., Lim, E.-P., Goh, D. H.-L., 2005. On Assigning Place Names to Geography Related Web Pages. In: JCDL '05: Proceedings of the 5th ACM/IEEE–CS joint conference on Digital libraries. ACM, New York, NY, USA, pp. 354–362.
URL http://dx.doi.org/10.1145/1065385.1065464