MASTERS PROGRAM IN

GEOSPATIAL TECHNOLOGIES

Discovery and retrieval of Geographic data using Google

Carlos Abargues Casanova

Dissertation submitted in partial fulfilment of the requirements
for the Degree of *Master of Science in Geospatial Technologies*

UNIVERSITAT JAUME·I

WESTFÄLISCHE WILHELMS-UNIVERSITÄT MÜNSTER

Discovery and retrieval of geographic data using Google

Dissertation supervised by

PhD Rafael Berlanga Llavori

February 2008

ACKNOWLEDGMENTS

My utmost gratitude goes to Professor Rafael Berlanga, Sven Schade and Professor Marco Painho for their supervision, advises and support. Really special thanks to Eva and my family for their unconditional and continuous support, motivation and love. Also thanks to my master colleagues for their friendship despite the distance, the language or the culture. Finally I would like to thank professors Joaquin Huerta and Michael Gould for their guidance to join the M.Sc. in Geospatial Technologies.

Discovery and retrieval of geographic data using Google

ABSTRACT

The growth of content in the Internet makes the existence of effective ways to retrieve the desired information fundamental. Search engines represent applications that fulfil this need. In these last years it has been clearly increased the number of services and tools to add and use the geographic component of the content published on the World Wide Web, what represents a clear trend towards the so called GeoWeb. This web paradigm promotes the search of content based also in their geographical component. Here is presented a study about the possibilities of using the different services and tools that Google offers to discover and retrieve geographic information. The study is based in the use of Keyhole Markup Language files (KML) to express geographic data and the analysis of their discovery and indexing. This discovery process is done by crawlers and the study tried to obtain objective measures about the time and effectiveness of the process simulating a real case scenario. In the other side the different KML elements that could allocate information and metadata were analyzed. In order to better understand which of these elements are effectively used in the indexing process a test data set composed by KML files containing information in these elements were launched and the obtained results analyzed and commented. With the experiment's results the use of these services and tools are analyzed as a general solution for Geographic Information Retrieval. Finally some considerations about future studies that could improve these tools usage are exposed.

KEYWORDS

Geographic Information Systems

Google

Geographic Information Retrieval

Keyhole Markup Language

Metadata

ACRONYMS

KML – Keyhole Markup Language

WWW – World Wide Web

OGC – Open Geospatial Consortium

WMS – Web Map Service

WFS – Web Feature Service

WPS – Web Processing Service

Table of Contents

INDEX OF TABLES

INDEX OF FIGURES

1. INTRODUCTION

1.1 Motivation

In the last years some of the biggest companies in the Internet have released services and tools to visualize geographic content entering in this way into the Geographic Information (GI) market [1, 2, 3]. Among these companies it can be found Google, Yahoo or Microsoft, Some of the companies owning the most used and well known search engines. This could be the reason why their mapping applications are not limited to visualize geographic content but they can also be used to search it. In some cases, like in the case of Google, the users can also create and share with others their own geographic content [4, 5].

The users have several publication options in order to have their content appearing in the search results for other users that are looking for some specific geographic content. These options are not much different of the ways the users already have to get their websites appearing in the search results. These goes from using directly Google's services like Google Maps [1] or Google Earth [6] through their related websites to simply publish the appropriate files in a publicly accessible server and wait till the search engines systems reach the content. This last publication method represents probably the easiest way to share any content. It also implies a clear advantage concerning time and simplicity in comparison with other methods of publishing geographic content such as the actual catalogues [7] like GeoNetwork (http://www.geonetwork.net), where the users usually need to upload to a given server the geographic content.

Moreover not only the publication model is easy and accessible but also the way the creators can describe their content. Currently, when somebody wants to publish and share geographic content seems imperative the use of metadata [8] to give information about the content itself. In most of the cases, this metadata is the information used in the searches or in other words, the place where looking for the information that matches the search parameters. Now, with the new services offered by Google and the other companies, seems that the user does not need to create metadata anymore and it is the information in the same content's file the one used in searches. Sometimes this information seems reduced to a simple textual description or free text. This also could be observed as an advantage in front of the actual formal methods for metadata creation that implies the creation of a long list of attributes, where some of them usually finish being irrelevant for the final user or for their searching.

Another important aspect is that most of the companies that offer these kind of services act globally, analyzing a vast number of sites and resources on the whole World Wide Web (WWW or simply the Web). These companies are supported by huge technical infrastructures highly worth and hard to imitate.

All these factors, the publishing easiness, the simplified use of metadata and the technical resources owned by these companies make them interesting to study as

complement or alternative to the actual methods for sharing geographic information. Among all these companies Google seems the one that offers more services, with web based, desktop and mobile applications and for sure is one of the most used. For these reasons, Google has been used as use case for the following study. These services allow the use of different file types, however the most promising and probably used is the Keyhole Markup Language (KML) [9], recently declared as an Open Geospatial Consortium (OGC) [10] standard. This format has been analyzed and used along this study as standard for representing geographic content.

1.2 Problem Statement

Although these services could represent a simple way of publishing and retrieve geographic information these could not necessarily represent an effective solution. An analysis on different parameters such as the process performance and effectiveness must be performed.

Moreover the information that these systems use as basis for the searches should be discovered and analyzed. At the same time and because a great effort has been invested in creating successful ways of explaining the geographic content, most of times using some metadata standards, its integration and exploitation using these services should be explored.

1.3 Research Questions and Goals

In order to address these issues the following questions will try to be answered in this study:

(1) How much time does it take to get some geographic content indexed for a standard use case?
In order to be considered an effective solution, the geographic content not only should be correctly indexed but also it must be indexed in reasonable time.

(2) Which information is relevant within a file for its correct indexing?
The KML format allows structuring the information in different ways. The knowledge about which parts of those files are meaningful for the indexing process could provide useful information about how to design them to increase their possibilities to become successfully indexed.

(3) Do the elements within a file where the information is placed affect the file's indexation?
Continuing with the previous question, a good file design could improve its indexing but it is fundamental also to know if placing the information in other places could derive in a failure.

(4) Could existing metadata be reused using Google's services and tools for geographic discovery and retrieval? In the case this is affirmative, how could be done?

A lot of geographic information with its correspondent metadata already exists. If it is finally demonstrated that these services offer an effective method for discover and retrieve geographic content then it would be also interesting how the existing geographic content and more precisely their metadata could be adapted to use these services.

(5) Could this way of publishing be a general solution for discover and retrieve geographic information? Could it replace other techniques already in use?
Based on the answers to the above questions, the use of the services provided by Google as a general solution could be analyzed and also a comparison with actual methods of content publication such as catalogues performed.

1.4 Methodology

For this study KML files have been used to express geographic data. The decision of using KML for the study has been taken considering its recent standardization by the OGC and also because Google specially indexes this type of files. It is also important to make note of all the possibilities this flexible format represents to express and visualize geographic information.

To answer the research questions the study has been focus on three main aspects. The first one is the time the system takes to crawl and index the content since its publication in the server. This measure could demonstrate if Google offers and effective solution with acceptable performance. Secondly, the number or percentage of files that are finally indexed from the whole set of test files released in the experiment. It seems that not all the files become crawled by Google and even less finish in the index for some reason. If this quantity demonstrates to be too small probably the system cannot be considered effective as a general solution. Finally which parts within a KML file are analyzed by Google's search engine. If these elements were discovered then it would be easier to insert the correct information in the right place for an effective indexing. The study of the KML elements includes also the use of existing metadata inside a KML file and all the different elements that seem suitable to store significant information.

Google seems to divide into two different indexes the information used by the traditional web search service from the other specifically geographic services. The web search service allows the user to search geographic content expressed in KML as well. At the same time the geographic content in the other index is most of the times also represented in KML. Although this is none of the study's objectives, the results obtained using one or another index could be contrasted.

An experiment to analyze the above explained aspects was conducted. A real case scenario was simulated and the Google's advices about publishing geographic content followed [11].

To reproduce a real case scenario the different files that conformed the test data sets were uploaded and made publicly accessible in a server. These test data sets were composed by a significant number of KML files containing data in different elements

within the files. The use of the different KML elements has been based on their functionality considering those of them suitable to store representative information. These test data sets could be derived into two main groups. The first one, make use of the same information stored in different places within a KML file, all in the same position. The indexing of these files could give an idea of which KML elements are used in the process. The second set was composed by different files with different information in different elements within the file and also in different locations. The reason why a second set was used is to avoid some situations that could affect the correct indexing of the files such as the content duplication.

To obtain the different measures about performance, effectiveness and indexing of the files different queries were used in the different searching services offered by Google. Then the time spend, the number of files appearing in the results for these queries and the files that were successfully indexed were recorded and analyzed.

1.5 Structure of the Thesis

In the following section, Background, the actual context, the concepts and the ideas that have been the basis for this study are explained. Also the different services or tools and technologies used are described starting with the set of free tools and services offered by Google and used along the study. Since there exist different ways to get geographic content indexed by Google, all the different ways available for the users and the path the information follows are analyzed. Finally an overview of the OGC KML standard is presented, explaining the most important elements considered for this study and presenting some of the great possibilities this format offers to visualize and annotate geographic information. In section 3, Methodology, all the details about the publication of the test data sets and the different configurations of the files that composed them are explained. The details to obtain the tests' results including the different searching services and queries as well as the results themselves are explained in the section 4, Results. These test results include measures about the crawling time, the number of files appearing in the search results and the files that apparently give information about the elements analyzed within a KML file. These Results are explained and analyzed in more detail in the section Discussion of results. Finally, in the Conclusion section the overall study is discussed and the research questions try to be answered. As a future work, other related experiments are proposed to explore deeper the Google's services for geographic content.

2. BACKGROUND

2.1. Study Context

Since the birth of the World Wide Web in 1989 the quantity of users has been radically increased [12]. With the number of users reaching extraordinary high quantities also the quantity of information, services and other resources available to them through the Internet have experimented a huge growth. This high number of resources transforms the task of finding any of them into a difficult one, unless the users know exactly where these resources can be found. In this context it is not difficult to guess that a tool capable of helping users in finding what they are looking for would be extremely useful. In the first stages of the WWW list of categorized directories exist in order to address somehow the searches and help the users to find the resources based on their theme. However it is with the appearance of the well-known search engines, when that tool intended to help the user appears. This search engines simplify the complicated task of finding any resource in an ocean of information by, most of times, just typing the terms that better match or describe the content sought. A tool that simplifies such a tedious process is destined to succeed. This is one of the reasons why nowadays names such as Google, Yahoo or Microsoft are not unknown for the vast majority of the Web's users. There are more companies offering search engines as well, however these are some of the companies that own the most used search engines in the whole WWW recording millions of visits per day [13].

In the last years the way the people use the WWW has evolved. At the beginning the Web could be considered as a unidirectional way of content service. In this scene the producer generates content directly consumed by users. Nowadays, this Web users are not merely content consumers but also producers deriving in a bidirectional schema of content production and consumption. The Web 2.0 [14] could be described as a change in the technology and design on the actual Web focused mainly to improve its functionality, communications, information sharing, creativity and collaboration along it. Everyday more and more people write about their experiences or knowledge in their blogs, share their last trip photographs using their online photo album or create social or professional networks through different web sites. All these facts derive in an impressive growth in the quantity of content publicly accessible through a web browser. It seems logical to think that at the same time the quantity of content increases, the act of searching and discovering specific information among that content becomes more difficult. Then it also seems clear the importance of effective searching tools.

The Internet population's anxiety of content creation that appeared with the Web 2.0 popularization is not limited to share photos or thoughts in blogs. As it happens in the real world, a huge percentage of decision-making processes have a geographic factor. In this Web 2.0 era new terms that combine its principles with geographic information are appearing. Hereby terms such as GeoWeb [15] or Neogeography [16] are becoming more familiar. The GeoWeb represents the idea of merging geographic information with other types of information that are actually found in the Internet as

for example HTML web pages. Among other things, this would allow the search of content based on the content's location. The Neogeography it is a term used to describe the use of geographic information related techniques by non-professional users or for personal or community purposes. In any case this usage is done usually through web browser and give the opportunity for creating geographic content to a broader public. Both cases suggest an increase of geographic information publicly available on the web. As it already happened with the textual content expressed in simple HTML, effective techniques to find this content will be required.

The Internet and the Web's evolution inevitably have affected and still affects other fields and not only those directly related with computing. In general, apart of the appearance of the Neogeography and the GeoWeb concepts, the Geographic Information (GI) has evolved also using for its own profit the advances that the Web and Internet's evolution has brought. Nowadays it is common to hear terms like Spatial Data Infrastructure (SDI) [12], read about interoperability studies like the one represented by GEOSS [17], or study new standards and services like the ones defined by OGC. All of these topics are related with the use of the Web to transmit, create or share geographic content. As already mentioned this evolution brought also new standards like the OGC specifications, most of them based also on the Web as the main transmission medium. Several of these standards like the Web Map Service (WMS) [18], Web Feature Service (WFS) [19] or Web Processing Service (WPS) [20] are becoming or already are widely spread used. Currently it is possible to convert the web browser into a GIS application using remote applications and services such as WMS, WFS or WPS. This way of working represents the also known thin clients, and directly eliminates the need of using desktop GIS applications or heavy clients by the end user. It seems that the improvement carried by the Web and the Internet evolution has directed the GI field's evolution towards an intensification in the use of distributed resources like remote services and data against the traditional use of desktop applications with locally stored data.

This movement from the desktop to the network can be also seen on the constant release of new web services related with the GI field and also on the continuous emergence of new SDI projects. One example of these projects can be found in the European INSPIRE Directive [21]. Defining an SDI is a task that varies with the context. Some people could consider an SDI as an infrastructure that interconnects data and software in an organization. Somebody else could add that an SDI is much more than the technical part and includes also the rules that compose a work framework that defines how to work in that organization. What it is probably true in any context is that one of any SDI major goals is to allow the sharing and collaboration between their users. This is based usually on accepted standards including services for sharing geographical information between different partners using a key element that keeps all the parts in a SDI connected, the metadata.

Metadata can be defined as data about data or a service, or simply the documentation of data. The metadata is used in a broad range of applications. Probably the most well known of the metadata initiatives is the Dublin Core, initially created for the description of electronic resources. There also exist standards concerning the description of geographic information datasets. Probably the most used geographic information metadata standard is the one defined by the International Standard Organization (ISO), the ISO19115 standard [22]. This international standard defines

the model to express geographic information. It defines a set of mandatory and optional fields and allows extensions to adapt it to specialized and specific situations. These metadata standards have primarily been used to help in understanding, comparing and interchanging the content of the dataset described. Among other functions metadata standards are primarily concerned with the discovery of data [23].

The study on metadata is an active field where new studies are constantly presented. In the last years some critical studies about the actual metadata standards and their use have appeared as well. Goodchild [24] argues that the actual metadata for expressing data quality is producer-centric and requires a change towards an user-centric model. In his paper Goodchild enumerates a set of problems in the actual use of metadata for data quality and talks about the need of a second generation of standards. Bulterman [25] goes further and makes an analysis about the current utility of metadata from a multi purpose point of view. In his study he concludes that maybe metadata is not as needed as it could be considered: *"The point is this: People do not need to add metadata to text documents if documents are processed electronically. Experience has shown that the contents of text documents can be mined directly using a host of existing information retrieval technologies and that metadata descriptions are often superfluous."*

Actually some search engines already perform this file content processing. For instance, Google searches and indexes files expressed with different types. This number of types is continuously being increased adding new ones. Currently users can search content into file types such as Adobe Acrobat PDF [26], Shockwave Flash [27] or Microsoft Word [28]. Maybe saying that metadata is not needed anymore is saying too much, at least for all type of files and users. However it is true that not all users need all the information provided by conventional metadata and just base their search on the content itself and not in the information about it. An example could be found in the case where a user is looking for a given PDF document and that user is only interested in the document's content or text. No other information about the document such as author, creation date or even license that could be found in some of the metadata standards are necessary for all users. However there could exist also the case of that specialized user looking for a PDF document within specific properties such as author, license and more. In this case probably the information included in the document content or text is not enough to satisfy the search requirements and again the use of metadata that follows the document is required. These points of view could also be applied to the GI field. Most of the studies about searching in this field are based on the use of metadata standards such as the ISO19115.

Another import aspect about metadata is its creation process that probably because the complexity or extension of the used standard could derive into a tedious process. Although the existence of tools for metadata creation and promotion and also being clearly defined its importance in any SDI creation and for geographic data sharing in general, its creation seems still avoided for some content creators. This could support in some sense Goodchild's and even Bulterman's ideas. The concept of the GeoWeb implies also that people could act as sensors, creating geographic content [29]. In this context, the idea of non-specialized people creating geographic content and simultaneously its metadata using for instance the ISO19115 standard seems difficult. The major problem is that the process could become tedious and usually people try to avoid it if they have the chance. For a governmental agency the creation of metadata

in all their datasets could be a standard practice strictly followed. However for a sporadic creator of geographic content adding the right information for all the fields required by the ISO19115 seems an easy to ignore task. Probably simplifying the metadata creation process, limiting this to a simply but complete description and improving how the information retrieval system works could represent a solution. In the already coming GeoWeb most of the content will be produced by non-experts and simple ways to add metadata to the geographic content will be required.

Some of these services are the ones offered by Google that we will call Google Geo Services in the rest of the document. These free services include three-dimensional viewers like Google Earth (http://earth.google.com), web based and two-dimensional viewers like Google Maps, publishing tools like My Maps and more including APIs that allow the programmatically access to these services. All these tools facilitate the use and creation of geographic content contributing to effectively create that GeoWeb. Also with the release of its services Google promotes the use of the KML, a XML based language created for geographic information annotation and visualization. The most import aspect, from this study's point of view is that Google also crawls and indexes that geographic content in KML format allowing to search it using its services. The crawling process transforms the publication process in a task as easy as publishing the content in a publicly accessible server. At the same time the searching for the geographic content is not restricted to the use of structured metadata but textual descriptions within the files. The way it works seems to represent the previously mentioned idea of metadata creation process simplification. Certainly Google's search engine could be considered as an acceptable information retrieval system supported by a huge infrastructure. This fact makes Google the ideal candidate system that could demonstrate that such an Information Retrieval system could be still a solution for discover and retrieve geographic content without the use of structured metadata.

2.2. Google Geo Services

Google offers a set of services and tools for visualizing, creating and sharing geographic information using KML as main file format. These tools include a two-dimensional viewer that can be web-based or executed in mobile devices, a three-dimensional earth browser and finally the resources needed to use them in other applications.

Google Maps is a map service executed on the web browser. Depending on the user's location, this can show different information such as basic or customized maps, local business information or driving directions. By default this location also determines the default view for the user. Google Maps offers several types of views including traditional map, satellite imagery, terrain model, street-level imagery and traffic view, what it is identical to the traditional map view but adding information about traffic in a given area where this information is available.

The web interface allows the user to navigate in the map using either the mouse or the keyboard. Additionally the user can zoom in or out on any specified location. At the same time there also exist navigation controls that offer other options such changing the view for facing other directions or the street-level imagery activation. All this

controls are showed over the map as it appears on Figure 1. This web interface offers a left panel where information like the search results is displayed. Figure 1 shows some results for a simply query where these results are listed in the left hand side panel. Usually when clicking on a marker used to indicate a location an info window or balloon appears. This info window could display additional information about that location in text but also using images, links or even videos. Depending on the user's location, Google Maps allows searching businesses, addresses, roads and intersections, places, coordinates, geographic features, real state listings, driving directions and the most interesting of all of them, user created content. This user created content includes KML-based content. The users can restrict their searches based on any of these categories. One important service offered in Google Maps is the content creation service also called My Maps. Depending again on the user's location, they can create customized maps. However these maps are restricted to the use of Placemarks, Lines and Shapes. Once the customized map is finished, the users can share it, collaborate with other users or directly open it in Google Earth. The users can also import KML or GeoRSS [30] to their map. The GeoRSS is a set of standards to represent geographic data and it is built inside the RSS [31] family that are commonly used to publish frequently updated works like news or blog entries. Applying this, the GeoRSS could be used inside Google Geo Services to represent geographic content that is frequently updated (i.e. traffic conditions). The easiest way to share the map with other users is using the specific URL that any map owns. However there exist also the option to make the map public or unlisted. A public map is included in the search results on Google Maps and Google Earth. However the unlisted maps are more restricted being only accessible for a specified group of users. The collaboration option allows other users to edit the map. It is also possible to export that map into KML format for visualizing it directly in Google Earth or other earth viewers able to work with this format.

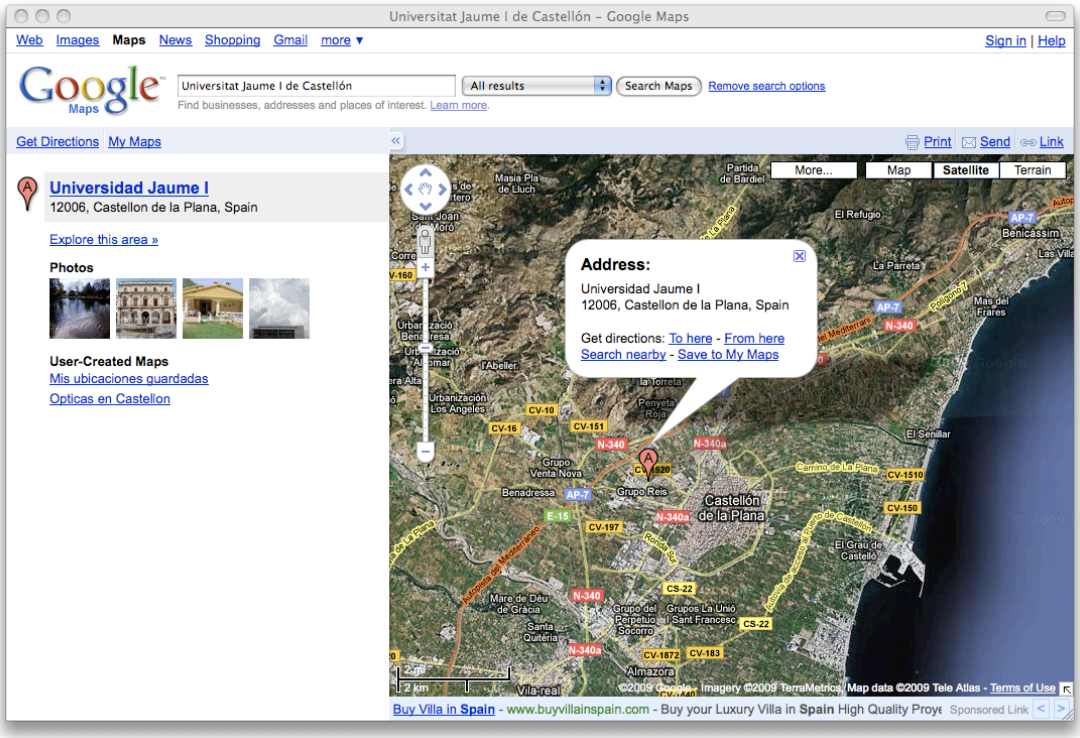

Figure 1: Google Maps is the web-based map service offered by Google.

Google Earth is the Google's virtual globe, map and geographic information desktop application. Google Earth displays satellite imagery of varying resolution of the Earth's surface allowing its visualization from different angles and perspectives as it shows Figure 2. It also allows to visualize all kind of images overlaid on the Earth's surface and can work as a Web Map Service client. The use of Google Earth is based on the use of KML and KMZ files using its visualization possibilities. The KMZ files can be defined as compressed (zipped) KML files. Basically they store internally a KML file, usually called *doc.kml,* and the other resources used by the KML features (images, photos, etc) described within this file. The KML and KMZ files allow three-dimensional visualization of data like buildings or terrain and also animations on time. The users can also search for addresses, locations and other user-created content. These searches provide the same results as the searches performed using Google Maps however in this case the application has no limitation to represent KML data that contain three-dimensional information. Finally one important point is the addition of different layers that can be loaded in Google Earth. These layers include information from a broad range of sources. For instance the user can find layers displaying Wikipedia (http://wikipedia.org) information, weather forecast, content from the Google Earth gallery and much more.

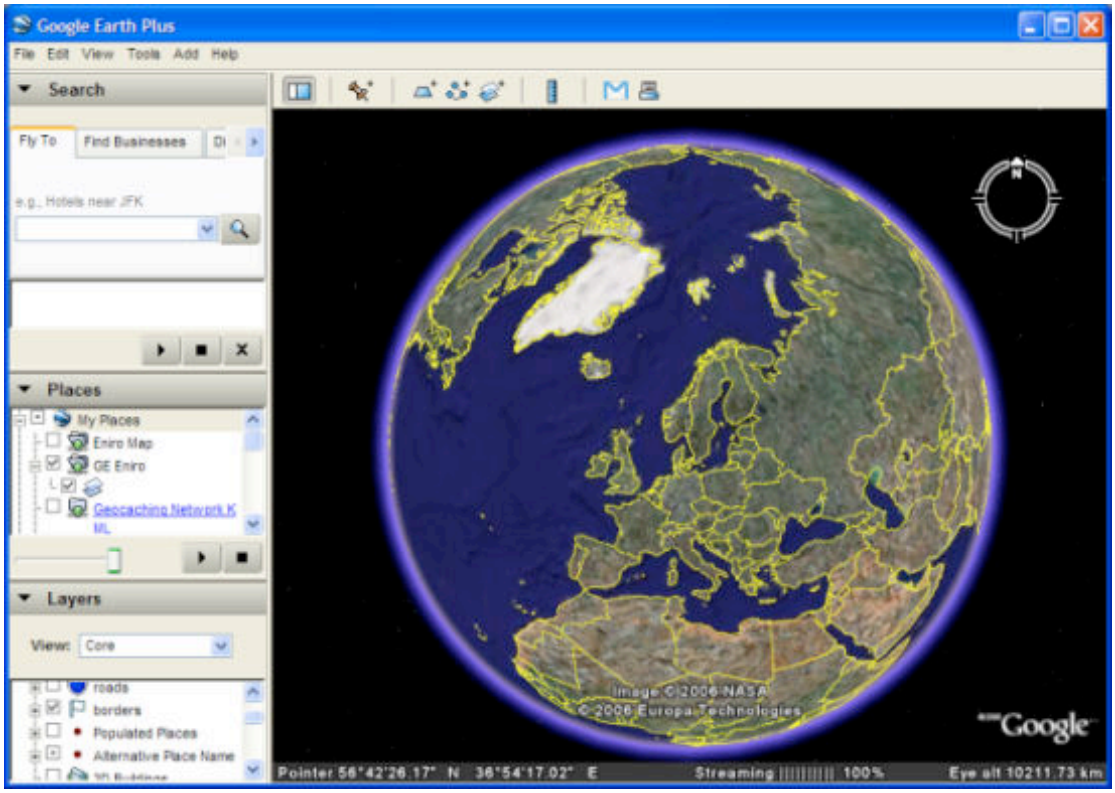

Figure 2: Google Earth allows the visualization of three-dimensional Earth surface.

Google allows the access to their services via different tools but also offering a set of Application Programming Interface (API) to create applications that make use of these services. There exist a broad range of APIs for the different services and in the case of Google Maps API (http://code.google.com/apis/maps) this lets the users to embed and use Google Maps in their own websites using JavaScript. Although there

exist some limitations like the number of uses per day, this API provides the same set of utilities that can be found on the Google Maps website. These utilities also include the search over user-created content.

2.3. Content Publication

One of the main motivations for analyzing the use of a search engine like Google for retrieving geographic content is the simplicity of the content's publication process. The basic idea of this process is to make publicly available the users' content and wait for Google to discover, index and rank that content. Unfortunately behind this simple idea there are much more parameters and aspects that finally determine if that content will ever be found in the search results.

Basically the Google tasks could be divided into three: Crawling, Indexing and Serving. The crawling is the process by which the Google's bot also known as Googlebot discovers new content to be added to the Google index and updates the old one. This process is supported by the huge technological infrastructure owned by Google and by the Googlebot programming. This programming determines parameters like which sites to crawl, how often or how many pages to crawl from each site. In this crawling process all the URLs in a site are analyzed and used to define new sites to crawl, changes to existing ones or dead links that finally update the Google index. The next task is the indexing, based on the compilation of an index with the words extracted from the sites analyzed by Googlebot and their location on the site. Actually Google supports the indexing of different data types, including KML files that express geographic information, but not other interesting formats such as ESRI's shapefiles [32]. In the case of the HTML pages not only the text is extracted to build up the index but also different tags and attributes. The last task is the process by which Google tries to serve as result the information it considers more relevant with the searching parameters. This process of sorting or ranking is based in a huge number of factors. One of the most famous of these parameters is the PageRank, which measures the importance of a given site based on the incoming links from other sites.

In the last years, Google and its search engine has become so popular that the search engine's website has become the first place to look for information for a high number of people. It is not casual that its website is one of the most visited on the Internet. It is true that thanks to its infrastructure Google is able to explore and index a big part of the Internet but of course not all. It is also true that Google does not guarantee that the published content will ever be indexed. This is true and it happens even if the information is made publicly available and specifically reachable for its web crawlers. There are several reasons that make this circumstance understandable. The first is the size of the Internet in number of resources and its growth speed. Everyday people publish content on their blog, companies close their website, professors publish in PDF format some notes on their public websites, somebody creates a public photo gallery and much more. Movements like these happen thousands or millions of times per day. It is true that Google's infrastructure is huge but discovering, organizing and maintaining updated records for such a quantity of information is still hard to imagine. It is not strange in such panorama that the web crawlers take some time to discover and add to Google's index some specific content or even avoid its discovery

and analysis in some cases. Another reason why some content does not appear on the Google's search results could be simply because the content is not “Google-friendly”. This term is usually applied to those websites that do not follow the Google's Webmaster Guidelines [33]. These guidelines are sets of recommendations about design and content, technical and quality aspects to help Google find, index and rank the websites. The guidelines about design and content give some recommendations about the use of links and descriptive content. The technical part explains the importance of creating well formed websites, some web server functions and the use of robot.txt files to avoid the incorrect crawling of the website. The robot.txt files are simply lists of directories that the crawler must or must not visit. These files help to restrict which content will become part of Google index in those cases where all the content resides in public directories. Maybe the most important of these recommendations are the ones related with the quality. In these guidelines some of the illicit practices that can make a site become penalized or deleted from the Google index are explained. These two actions mean, in most of the cases, the removal of the site from the Google search results. The guidelines about design, content or technical aspects could influence in the indexing process of some pages or the time taken to crawl the site. However the quality guidelines are extremely important since by them it can be decided if the user's content is completely deleted from any search results provided by Google. The quality guidelines are divided into basic and specific principles. In the basic principles the user can find recommendations about the importance of designing the sites for people instead of search engines, the avoidance of the participation in link schemes to increase the number of incoming links or the error of using applications that violate the Google's Terms of Service. In the other side the specific guidelines describe more precisely aspects to avoid when releasing a website. These include the avoidance of hidden text or links, cloaking or sneaky redirects, the sending of automated queries to Google, the loading of pages with irrelevant keywords and also the avoidance of duplicate content sites creation. If any site is catalogued as if it would not accomplish the guidelines it is also possible to ask Google for reconsideration after a previous modification in order to accomplish them.

The above tries to explain briefly the general case to process all kind of supported files and include them into the Google Search Index or simply Google index. However apparently this Google Search Index is not the only one. Since the release of the different Geo services offered through Google Maps and Google Earth, Google is using a new index also known as Google Geoindex. From a simplistic point of view, the Google Search Index is queried based on keywords introduced by text. It seems that when querying for geographic content using either Google Maps or Google Earth the results obtained corresponds to the terms introduced as text but also the geographic region visualized when performing the query. In other words, it seems like the bounding box that represents the region in the viewer acts as a filter for the query results. Apparently the new Geoindex can be queried not just by words but also by region. At the same time all that geographic content is catalogued depending on its possible use and source. As previously explained, in the main Google Maps website, the users can currently filter the search by *Locations*, *Businesses*, *Real State* properties, *Mapped web pages* and *User-created content*. All categories are interesting however is this last one, the user-created content, the objective of this analysis. There are several ways for trying to make the user-created geographic content appear in the search results using either Google Maps or Google Earth. In

Figure 3 a diagram created by Barry Hunter [34] shows all the current ways to publish and retrieve geographic content using Google Geo Services.

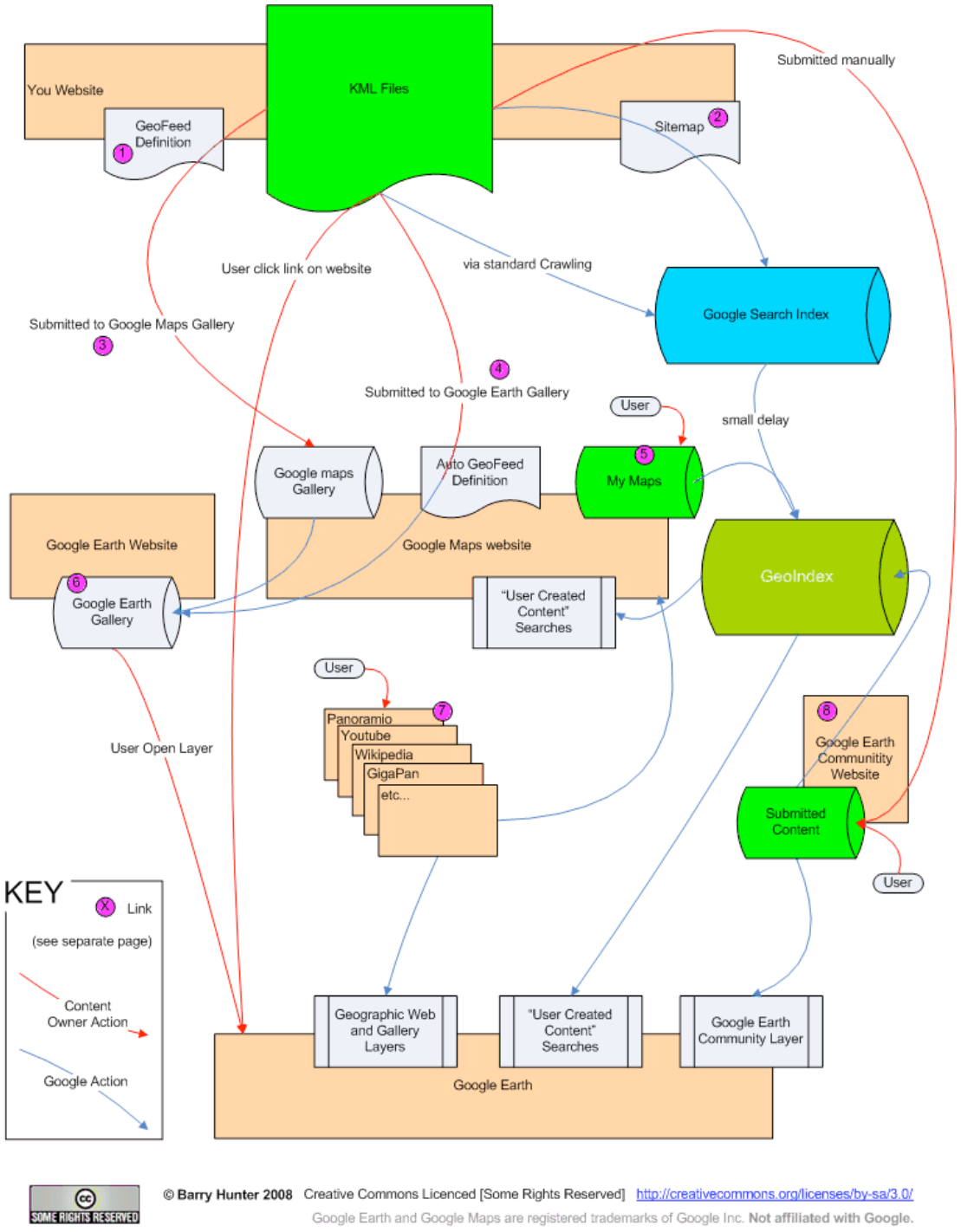

Figure 3: Barry Hunter's diagram representing all the possible options to make available geographic content using Google's services and tools.

In the diagram, in blue and green and using a cylinder, the previously explained Google Search Index and the Google Geoindex are represented. It is important to note that the different actions taken in order to retrieve the content are clearly divided into Google's actions, in blue, and content owner's or publisher actions, in red. We can

also differentiate, using rectangles in light orange, the different interfaces, either websites or applications that are part of the process. These interfaces are the Google Maps website, the Google Earth and Google Earth Community websites, the Google Earth desktop application, the set of different websites that actually georeference some of their content using Google's geo services and finally the own user's website where the content is published. In the diagram the box representing the User-Created Content Searches appears in both Google Maps website and Google Earth application. These boxes representing the searching are fed in both cases by the Google Geoindex without using in any case the Google Search Index.

The different content sources are represented in light green. In the first place, there are the KML files allocated in any server in the Internet. Secondly it is represented the option of creating personalized maps using Google Maps' *My Maps* service. And finally it is also represented the direct submission of geographic content through the Google Earth Community Website.

Considering the case of those KML files published in a server with public access, these files have several options to appear in Google's Geoindex. The options can be divided into two main groups. First those options requiring some kind of action done by the content publisher. Second the actions based on an automatic processing done by the Google's crawlers. The first group of options basically consist on submitting the KML files to Google maps or Google Earth galleries [35]. In the case of Google Earth gallery the user just need to specify a set of URLs for the KML or KMZ file and for its screenshots via a form in the gallery website. In the case of Google Maps gallery the content published is not composed by KML files but by mapplets. The mapplets could be defined as tiny applications executed inside Google maps. These are build in a similar way than the gadgets used by Google in some of their places like iGoogle, a customized Google website. There exists a complete and freely available API for these mapplets creation. Once the KML files are included in any of these galleries users can visualize the content using any of these applications. The content that compose the Google Maps gallery is automatically included in the Google Earth gallery and then becomes accessible using any of these tools. These galleries are accessible using My Maps service on Google Maps, in different layers like *Gallery* in Google Earth or simply opening the different content directly from the Google Earth gallery website.

Concerning these galleries another important aspect is the role played by the different websites that georeference part of their content using Google Geo Services. As shown in Figure 3 some of these are photo and video galleries, wikis and others. All these sites seem to share some common point, the content used is created by users, following the web 2.0 tendency. Here we have another way how users can publish geographic content on the web, however this time, without the direct use of KML or KMZ files but publishing georeferenced photos, videos or articles. Internally these websites expose all these content to Google Maps and Google Earth for its indexing probably creating KML files automatically. In the first case it is easy to find these content just performing a search or adding it through the My Maps service as mapplets. Using this service users are not limited to add and visualize content of these websites but also other content published by other users. In the case of Google Earth the users can visualize these websites content using the appropriate layers like the *Geographic Web*.

Finally it is possible the automatic processing of the publicly available geographic content via the standard crawling. This is, by this study's point of view, the most interesting of the whole set of options since it seems the simplest way to be followed by any kind of user. Like Figure 3 shows there are two ways to get a KML file crawled but both are supposed to finish with the same result, the KML file included in the Google Search Index. The first way is as simple as waiting for Googlebot to index the file. This option seems to be quite passive however the second one represents a more active behavior for the publisher.. In this case the content publisher can make use of Sitemap files [36] and Google's Webmaster Tools (http://www.google.com/webmasters/tools). The purpose now is to help Googlebot to find the content to index and then speeding up the overall process. Both, the sitemap files and the Google Webmaster Tools, are deeper explained following. Once the content becomes indexed in the Google Search Index means that this is already suitable to appear in the search results using the Google Web Search. However this content will not appear in queries made using Google Maps or Google Earth yet. To appear in there the content needs to be part of the Google Geoindex. It is supposed that, after some delay, the geographic content is also indexed and becomes part of the Geoindex and then suitable to be found using the different products. It is important to consider the time taken for the whole process, since the content is published till it appears in the Geoindex. Even helping Googlebot, the time spend can be measured in weeks. This can seem a really long time, however as we will see in the following sections this time could be decreasing, probably because the improvement of the system's resources. Again, the user can face the problems previously described concerning the possibilities that the crawling or indexing process fails.

There still remain two ways used to add content: using the My Maps service and submitting the content directly to the Google Earth Community website. One important point that represents a big difference between publishing the KML files in a public server and these last options is that these two are thought to make the content becoming part of the Geoindex faster (if not immediately) than following the standard crawling process. This is not a strange idea since using any of these steps to publish the geographic content means to store that content directly in Google's servers. Despite their effectiveness this two options have been discarded in order to test the discovery of geographic data in this study. The main reason is that these two options are too specific and dependent on Google publishing services and moreover remove completely the main point of publishing the content just making it public. As an example of the inflexibility of these methods, users need a Google account in order to use the My Maps service. This is not a big problem however the interface used to create this content is Google Maps, and this could represent complications in some cases. The KML files are not just limited to three-dimensional earth browsers, however it is with this visualization tools when all the KML potential can be exploited. Using KML in two-dimensional viewers like Google Maps implies some limitations. When creating simple geographic content Google Maps could be a solution however it is impossible to create complex or three-dimensional content with it. The next options require to take any of the KML files an submitting them through a web site. When creating a low number of files this is not a bad option since the users get a fast indexing thanks to the file’s submission. However this is not the case for those users that for instance could generate a big number of KML files in an automatic process within a day. Probably and depending on the number of files,

submitting all of them could mean a tedious and long process hard to achieve in acceptable times.

These are, at the time of writing, the different possibilities to make geographic content in KML files public and accessible using the different services for searching offered by Google. The number of options has grown since the origin of the different Google Geo Services and it is probable that the company will improve them and add new ways to facilitate the process.

Google offers to the users a set of resources to improve the discovery and indexing of their content and get at the same time information about the process status. These resources include a set of communication channels like a blog, a forum, an assistance centre and a set of different tools to help users to get their content indexed. These tools include an informative assistant to get quick information about the indexing of a given site, a set of methods to send content to the different Google services, including Google Books and Google Video, and the Google Webmaster Tools. This last one is the most interesting and useful since it offers a complete set of tools to get statistics, diagnostics and also to allow a little administration capability over the crawling and indexing process. The Google Webmaster Tools can be used to follow the process done by Googlebot, analyze the possible problems that can be found on it and fix them in order to increase the possibilities of a given site to become indexed. These tools also offer information about incoming or outgoing links in the website and the different queries that could drive traffic to it.

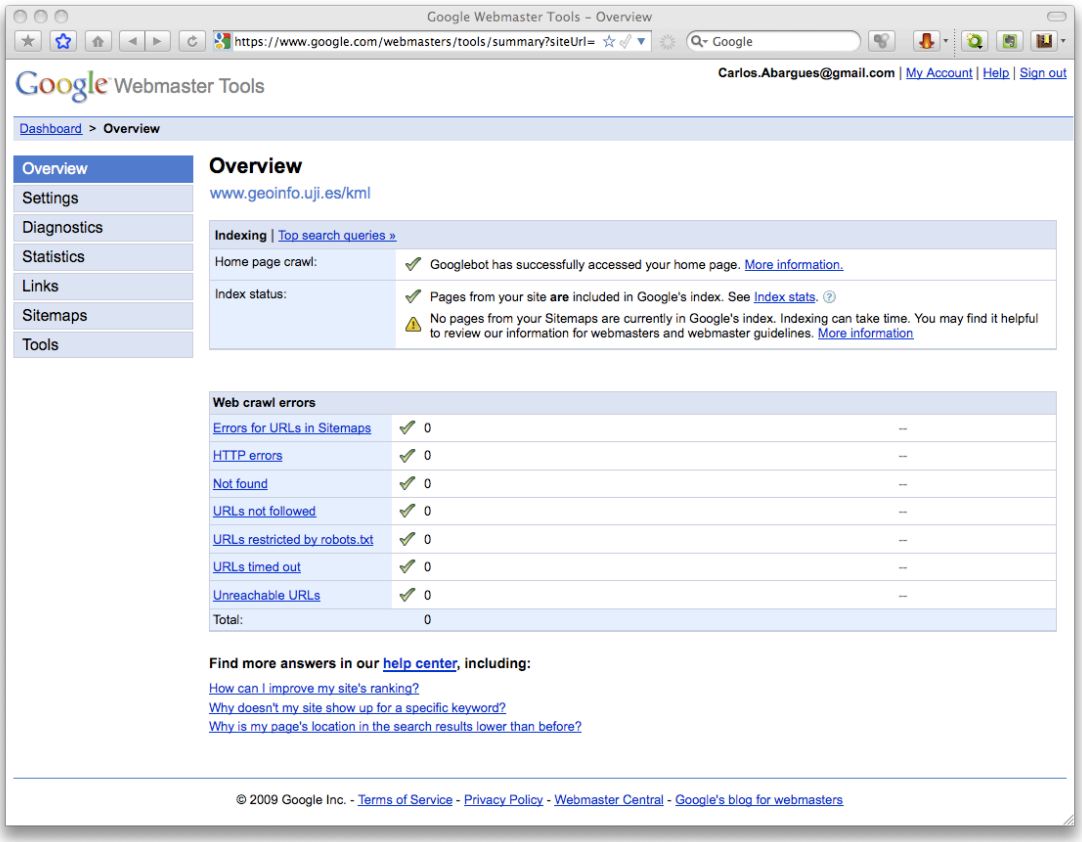

Figure 4: Overview information available on Google Webmaster Tools.

All the functions present in the Google Webmaster Tools are divided into the following sections:

Overview: As shown in Figure 4, in this section the users can find basic information about the indexing process status and the errors found during the crawling process. All this information is linked with other sections that extend this one offering more specific information about each specific issue.

Configuration: It allows the setting of some parameters like the geographic position, preferred domain, the inclusion of the site for the Google Image Labeler service and the crawling frequency. The geographic position is useful for the analysis of the website in a given country and it can only be changed in those cases where the domain is not country specific (i.e. .com, .net, .org). The preferred domain option allows the user to specify the domain used by default in a website when it has more than one. The Google Image Labeler is a service intended to improve the indexing of images relating this ones with tags. Finally the crawling frequency option allows the user to recommend a specific frequency to Google, however this has not a great impact on Googlebot actions since it crawls the website based on the number of webpages present on it.

Diagnostics: This section gives information to the user about the web and mobile crawling and the possible problems found in this process. In the case of the web errors these include HTTP errors, inaccessible URL, resources not found and more. The mobile crawling subsection shows information about problems found in the CHTML or WML/XHTML crawling, specific for mobile devices. Finally it also shows information about the website's content giving with it information about the indexing process.

Statistics: It shows to the user information about the most common queries where the website appears as a Google search result and also Googlebot, crawling and indexing statistics.

Links: As its name indicates, this section offers information about linked websites. These include websites that link to and from the user's site. It also shows information about the sites links automatically generated by Google in some cases based on the site's content.

Sitemaps: This section allows the user to have some kind of control over the crawling process thanks to the submission of sitemap files. For each sitemap sent the user can check its status, the number of URLs on it, the number of URLs indexed, the last time the file was checked by Googlebot or even the format. This format already allows the use of the special type Geo to specify geographic content. Some of these details can be found on Figure 5.

From a simple point of view the Sitemaps are XML files describing a list of URL in a website. This list helps to crawl the entire website and also helps in the discovery of resources that will not be reached by the standard crawling process. This files are recommended in those cases where the site contains dynamic content or pages difficult to crawl due the use of AJAX [37] or Flash, where the site is new an has a

reduced number of links pointing to it and also where the content distributed in different pages is not well linked between them. Sitemap files use is not restricted to Google and are used by other search engines that adopt the standards defined by sitemaps.org.

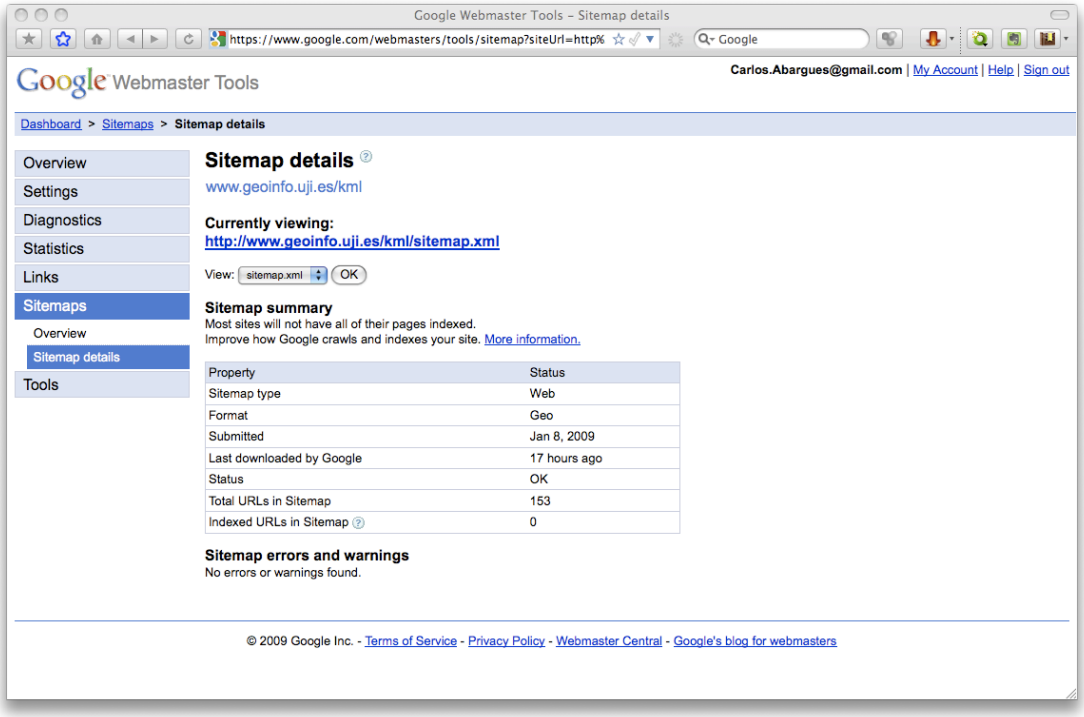

Figure 5: Sitemap.xml file's details displayed using Google Webmaster Tools

Currently Google accepts Sitemaps described using the Sitemap Protocol 0.9. This protocol is a XML dialect to describe a site's structure information including URLs available for crawling within a website. This protocol is not limited to indicate a list of URL but it also allows the addition of more information about each URL to improve the crawling process. This information includes the domain's URL, the date of the last file's modification, the frequency of changes in the file and a priority relative to other URLs on the site. This priority can be specified using a decimal number ranging from 0.0 to 1.0. The following example shows the use of all these elements to describe an URL:

```
<?xml version="1.0" encoding="UTF-8"?>
<urlset xmlns="http://www.sitemaps.org/schemas/sitemap/0.9">
<url>
<loc>http://www.example.com/</loc>
<lastmod>2005-01-01</lastmod>
<changefreq>monthly</changefreq>
<priority>0.8</priority>
</url>
</urlset>
```

The Sitemap files can specify a maximum of 50.000 URLs. In those cases where this number is not enough several Sitemap files can be created and referenced using a

Sitemap index file. The following example shows the use of these Sitemap index files:

```
<?xml version="1.0" encoding="UTF-8"?>
<sitemapindex xmlns="http://www.sitemaps.org/schemas/sitemap/0.9">
<sitemap>
<loc>http://www.example.com/sitemap1.xml.gz</loc>
<lastmod>2004-10-01T18:23:17+00:00</lastmod>
</sitemap>
<sitemap>
<loc>http://www.example.com/sitemap2.xml.gz</loc>
<lastmod>2005-01-01</lastmod>
</sitemap>
</sitemapindex>
```

Although the Sitemaps are also used for other search engines, Google allows the use of specialized Sitemaps for concrete types of content, not supported for all the rest. Currently Google supports specialized elements for specifying video, mobile sites, news, code and geographic content. This last one represents an extension of the protocol and includes a geo-specific tag. The <geo:format> specifies the format of the geo content. These formats are limited to KML, KMZ and GeoRSS format only. Following there is an example of Geo Sitemap file:

```
<urlset xmlns="http://www.sitemaps.org/schemas/sitemap/0.9"
xmlns:geo="http://www.google.com/geo/schemas/sitemap/1.0">
<url>
<loc>http://www.example.com/download?format=kml</loc>
<geo:geo>
<geo:format>kml</geo:format>
</geo:geo>
</url>

<url>
<loc>http://www.example.com/download?format=georss</loc>
<geo:geo>
<geo:format>georss</geo:format>
</geo:geo>
</url>
</urlset>
```

Google recommends the use of these files since they provide additional information for the crawling process probably resulting in more pages indexed and in less time. However Google never guarantee that a URL will be added to the Google index even if this appears in a submitted Sitemap file.

Tools: In the last section of the Google Webmaster Tools panel the user can find a set of different tools for the site administration. Using this section the users can administrate all the verified site owners, delete URL that do not exist in the website anymore or improve the 404 (HTTP code to specify a webpage is not found) pages to give information to the website users about how to find the webpage. Probably the most important tools found here are the robot.txt analysis and the robot.txt generator tools. The robot.txt or robot exclusion standard is a convention to prevent web robots like web spiders or crawlers (like Googlebot) to access parts of a website that are publicly available. These files can be easily checked to found errors or generated in

case it does not exist. These files are not required but are still useful to keep some folders invisible for Googlebot.

One important aspect about the Google Webmaster Tool is the restrictions imposed based on the use of root domains. In that case where the user is the owner of a domain like www.example.com, this user can access all the information provided by the set of statistics, tools and others. However in the case of a URL that corresponds to a directory under that root level, for instance www.example.com/subdirectory, some of the options are not available. For example this is the case of the information concerning the content analysis, crawling and Googlebot statistics among others.

2.4. Keyhole Markup Language

2.4.1. Past, Present and Future

The KML is a language designed to express geographic annotation and visualization and it is based on the XML standard. The geographic visualization includes not only the representation of the graphical data but also establishes orders or control on the navigation. KML is used for geographic content visualization in a broad range of interfaces such as web-base and two-dimensional maps (including those in mobile devices) and three-dimensional Earth browsers. Usually these applications also use KMZ.

KML was originally created by Keyhole Inc. This company was founded in 2001 and was specialized in software development for geospatial data visualization. Its main application suite was called Earth Viewer that was transformed into Google Earth in 2005 thanks to the acquisition of Keyhole Inc. by Google in 2004.

Currently a broad range of applications dedicated to visualize geographic data is using KML format. In this group of applications we can find tools such as ESRI's ArcGIS Explorer [38], OpenLayers [39], NASA's World Wind [40] or Google Earth and Google Maps among many others. KML is also used for other services that are not directly related with geographic data visualization. In this category we can find popular services like Yahoo's Flickr, a photo and video hosting and sharing service. This service allows to view geotagged photos and videos in Earth browsers such as Google Earth or even in two-dimensional viewers like Google Maps thanks to the use of KML files. This one is a clear example of how KML can be used to interconnect different services.

Another example of the increasing use of KML is the new output option in one of the most used servers in the geographic information field, GeoServer (http://www.geoserver.org). This supports both KML and KMZ output for WMS requests. This allows the end user to visualize the output of a WMS request to a GeoServer installation directly in an Earth browser thanks to the interface for KML files output offered.

Besides there exist other tools that improve the broad range of possibilities that KML offers. For example Zonum Solutions (http://www.zonums.com) offers free tools,

some of them online, to create, process and import or export KML code. Two of the most useful tools are related with the translation of KML to other formats and the other way around. KML2SHP allows users to transform KML files into some of the most used formats in the GIS field, the ESRI's shapefiles, AutoCAD (DXF) [41] and GPS (GPX) [42] files . SHP2KML allows the inverse process of translating from ESRI's shapefile to KML. These tools do not support yet all the features for these formats and need to take care about some aspects related with the data's reference system for instance. However both tools are free and continuously improved. These are just some examples but the simplicity and structure of KML allows also the transformation from other more common file formats used also by non-professional users. One example could be the transformation of a Comma Separated Value (CSV) [43] file with some kind of geographic information or coordinates to KML. The CSV express a really simple structure where values are separated by commas. This format is really common for exporting information from a broad range of databases and spreadsheet applications.

These are just a few clear examples of the possibilities and projection of this file format. However the most important fact that assures its continuity and improvement is its recent recognition as an OGC standard.

On April 14th 2008 the Open Geospatial Consortium adopted KML version 2.2. as an OGC standard what means the first step in the attempt for its harmonization with other relevant OGC standards that compose the OGC standards baseline. The most related existing OGC standards, also complementary with the new KML standard are the Geography Markup Language (GML) [44], Web Feature Service (WFS) and Web Map Service (WMS). In fact, there are some common points between GML and KML, using this last one, some geometry elements such as point or line string derived from GML version 2.1.2. This current harmonization between the two standards is planed to be increased arriving even to use exactly the same geometry representation in a future. The interoperability of KML with standard services such as WMS could be found in actual tools such as Google Earth. Using it the users can visualize and link WMS with KML files. The OGC have four major objectives regarding KML:

- *That there be one international standard language for expressing geographic annotation and visualization on existing or future web-based online and mobile maps (two dimensions) and earth browsers (three dimensions).*
- *That KML be aligned with international best practices and standards, thereby enabling greater uptake and interoperability of earth browser implementations.*
- *That the OGC and Google will work collaboratively to ensure that KML community is kept informed of progress and issues.*
- *That the OGC process will be used to ensure proper life-cycle management of the KML Standard, including such issues as backwards compatibility.*

The extensive use of KML, in part thanks to its adoption as default format by Google for representing geographic content, would suggest that KML format has a productive and probably long future in front. At the same time with the release of new products it seems that there is a trend in the use of three-dimensional earth viewers to work with geographic data using most of them KML as default file format. Its use in applications

and services is a good indicator however its standardization as an OGC standard is what seems to guarantee its continuity and improvement.

2.4.2. Structure and Elements

The KML is based on the XML standard [45]. As any other XML-based language, KML is composed by different tags or elements with a name and attributes that at the same time can contain others creating nested structures. XML files are structured and easy to process by automated agents and at the same time easy to read by people. Probably this is one of the best characteristics of XML and it is also present in all its derived languages like KML or GML.

KML is currently in its version 2.2 and its schema is publicly available under the Open Geospatial Consortium schema repository as any of their standards (http://schemas.opengis.net/kml/2.2.0). This is not only useful for learning about the structure of the language but also to validate this type of files when using some editors. In the OGC repository the KML schema is accompanied by another one, the *atom_author_link* schema. This represents a subset of the Atom Syndication Format and Publishing Protocol that is designed to support publishing and syndication of text content and media resources. This subset is used in the KML schema to represent some information about the file's author.

Figure 6 shows an object-oriented class tree diagram representing the principal elements in the actual KML schema. In this diagram, elements represented inside a dotted rectangle represent abstract elements. These elements are not directly used in the implementation of any KML file but are useful for information design purpose. Also elements to the right on a branch represent and extension or specialization of the elements they have on left. For instance the element <*Placemark*> derives or is an extension of <*Feature*> and at the same time <*Feature*> derives from <*Object*>. This hierarchy establishes that elements derived from another inherit its properties and specific child elements. For instance when defining a <*Placemark*> or <*NetworkLink*> the same properties that define a <*Feature*> are available for both. At the same time KML is a XML grammar and because of this it inherits XML properties and restrictions. Two of the most important of these restrictions are the use of case-sensitive tag names and the order in which tags appear in a KML file. In other words, in a KML file such a tag with name <*placemark*> will be wrong since this tag is defined in the schema as <*Placemark*> with capital p. As we will see later the order and structure of a KML file is also important. For instance we cannot define a <*Document*> element inside a <*Placemark*> but it is possible to define the opposite.

As we can see in the diagram represented in Figure 6, each element in a KML file derives from the basic element <*Object*> and then any element in a KML file inherits its attributes and child elements. The most important of these element's attributes is its identifier or commonly known as id. This id is used in KML to identify uniquely any element within a file and also to apply some other actions like a shared visual style or assign the updating preferences for linked resources. This id corresponds with the XML definition for identifier and its value could be represented by a string.

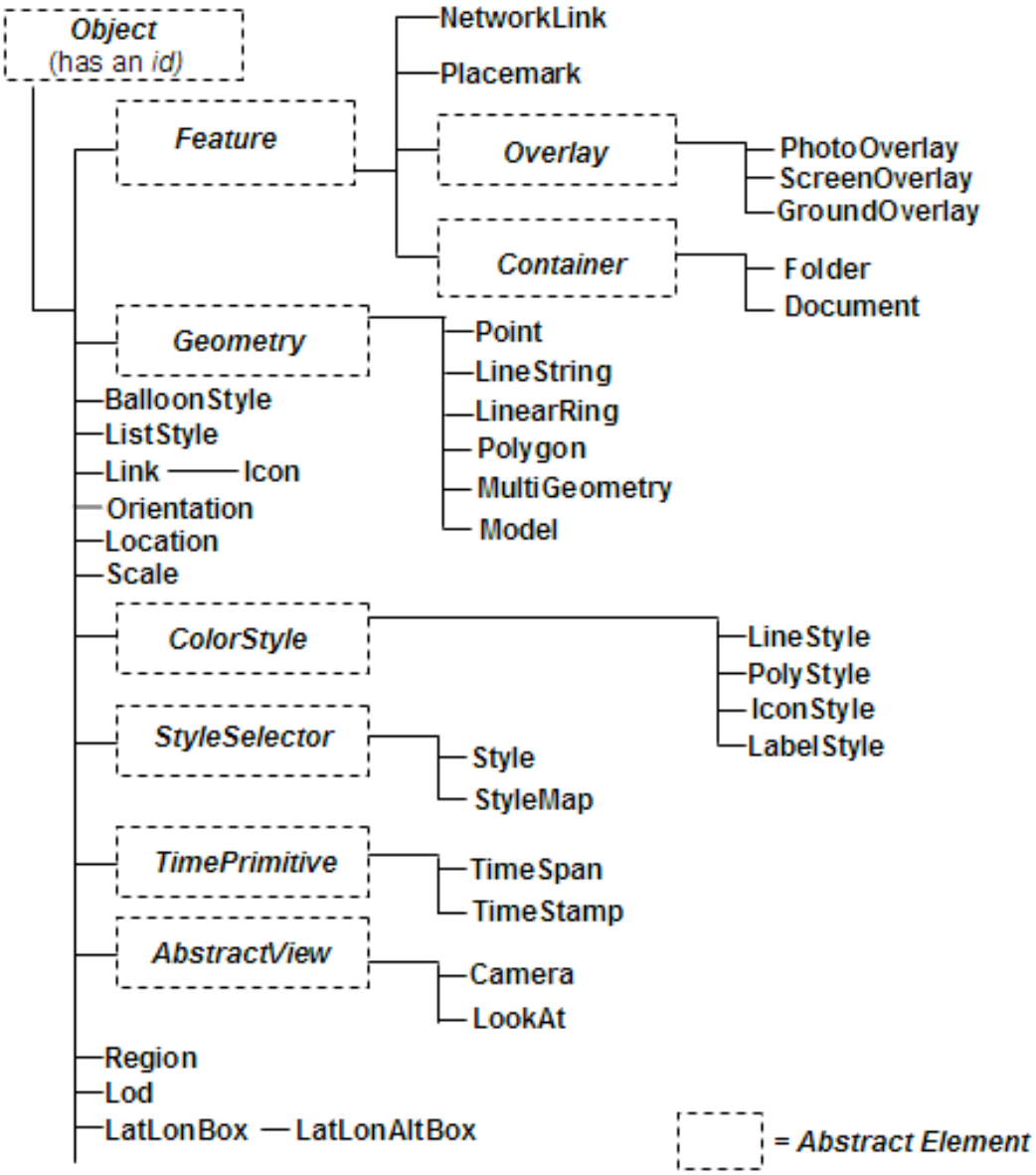

Figure 6: KML elements hierarchy

Below the <*Object*> level abstract and non-abstract elements are found and a subdivision based on the purpose of each element could be done. The two main usages of KML seem geographic content annotation and visualization. For instance we have elements such as <*Placemark*> to annotate geographic content and <*LookAt*> or <*Camera*> that define some characteristics about the visualization of this content in three-dimensional earth browsers. This categorization based on the purpose gives also a first reference about the elements suitable to store and transport descriptive information or metadata about the content. Reviewing the KML specification seems that all those elements used to specify visualization parameters like <*ColorStyle*>, <*StyleSelector*>, <*TimePrimitive*> or <*AbstractView*> and all its derived elements do not offer any way to carry information that could be used in searches, unless the user would need to search features on time where the *TimePrimitive* derived elements could be used. For this reason all those elements,

strictly related with the specification of characteristics for content visualization are not explored in this study.

In the following tables (Table 1, Table 2, Table 3, Table 4, Table 5, Table 6 and Table 7) the principal or most interesting KML elements for annotating geographic information that were used to store descriptive information in the experiment performed are presented. However some other elements, although they can be used to annotate geographic content and carry descriptive information are not described. Some of these are for instance the *Overlay*-derived elements. Along with their name and description some of the most important derived or specific elements and attributes are also explained.

Feature (inherits from Object)	
Description	Abstract element that sets some general structure for its inherited elements.
Major elements	***name***
	Specifies the Feature's name. It is used as the object's label in the three-dimensional viewers.
	visibility
	Specifies if the Feature is open in the viewer when it is loaded.
	open
	Similar to visibility but specifying if it appears open in the list of features (i.e. Places panel in Google Earth). It is only applicable to Container-derived elements (Folder and Document).
	atom:author
	Element extracted from the Atom Syndication Format specification (http://atompub.org) used to specify the content's author.
	atom:name
	Specific element of atom:author used to indicate the author's name.
	atom:link
	Through its attribute *href* it specifies the URL to the website containing the KML or KMZ file.
	description
	Describes the Feature. It can contain plain text and a subset of HTML formatting elements such as tables, images or even videos. It is also possible to insert links to other elements configuring some options. Among these elements the user can specify scripts that return KML or KMZ files.
	snippet
	Represents a short description for the Feature. It does not allow the use of HTML tags and contains the attribute *maxLines* to specify the maximum number of lines to visualize.

	address
	Represents an unstructured address. Its use is restricted based on the functionality given by Google in the country where the Feature is situated.
	xal:AddressDetails
	Represents a structured address, formatted using the international standard for address formatting extensible Address Language (http://www.oasis-open.org/committees/ciq/ciq.html#6). It can be used for geocoding in Google Maps.

Table 1: KML Feature element details.

NetworkLink (inherits from Feature)	
Description	References a KML or KMZ file on a local or remote network.
Major elements	***refreshVisibility***
	Indicates if the visibility of the Feature referenced should be reset automatically in the earth browser every time the link is refreshed.
	flyToView
	Indicates if the viewer should move the virtual camera as specified by other KML elements specifically designed for visualization purposes.
	Link
	Specifies the location of remote resources (KML or KMZ files). It has specific elements to set the URL and other parameters related for example with the refreshment rate. With this element it is also possible to visualize WMS services output in earth browsers allowing the specification of some WMS specific parameters like BBOX (Bounding Box).

Table 2: KML NetworkLink element details.

Placemark (inherits from Feature)	
Description	Describes a Feature with an associate geometry.
Major elements	***Geometry***
	Specifies the geometry used by the Placemark. This can be any of the Geometry-derived elements.

Table 3: KML Placemark element details.

Container (inherits from Feature)	
Description	Abstract class that is used as base for other derived elements that hold one or more Features allowing the creation of nested hierarchies.

Table 4: KML Container element details.

Folder (inherits from Container)	
Description	Contains other features and allows their arrangement creating hierarchical structures.

Table 5: KML Folder element details.

Document (inherits from Container)	
Description	Contains other features allowing their hierarchical arrangement and the organization of the different KML elements. It is required when using shared styles between different elements. It is also required to define the *Schema* element that allows the definition of custom XML schemas within a file.

Table 6: KML Document element details.

Geometry (inherits from Object)	
Description	Abstract element used as basis for all the elements defining geometry in KML: *Point*, *LineString*, *LinearRing*, *Polygon*, *LinearRing*, *MultiGeometry* and *Model*.

Table 7: KML Geometry element details.

KML is a really complete standard and define much more elements. Probably one of the most interesting elements in the KML specification is the <*ExtendedData*> element. This new element in version 2.2 allows the user to add custom XML data to any *Feature*-derived element. This can be done using three techniques:

1. Adding data/value pairs using the <*Data*> element.
2. Defining custom KML schemas with the <*Schema*> element and using them in any <*Feature*> via the <*SchemaData*> element.
3. Using XML elements or schemas defined in other namespaces (i.e. other files) by referencing the external namespace within the KML file.

Also, these three can be combined together to express different data in different parts within a KML file.

The addition of data/value pairs is the simplest of the three techniques. This pairs are expressed in KML using the element <*Data*>. This element allows the creation of this kind of pairs but does not allow specifying any type for them. Its basic structure is the following:

```
<Placemark>
        <name>Club house</name>
        <ExtendedData>
                <Data name="holeNumber">
                        <value>1</value>
                </Data>
        </ExtendedData>
</Placemark>
```

Basically the <*Data*> element specifies the pair name using an attribute *name* and the value using its child element <*value*>. It is also possible to specify the pair's name using the element <*displayName*>. This allows the user to create a formatted version of the name that includes the use of HTML tags. Usually <*displayName*> is used for visualization purposes on earth browsers since this value is displayed within the <*Feature*> element's info window when this is available.

The second technique requires the use of the elements <*Schema*> and <*SchemaData*>. The <*Schema*> element is used to define custom KML schemas that can be used to add custom data to KML <*Features*> elements by the element <*SchemaData*> element. These <*Schema*> elements use to be declared as child elements of <*Document*> and have two attributes, *name* and *id.* This last one must be unique and it is used to reference the <*Schema*>. Inside the element the user can declare different custom fields using the <*SimpleField*> element. This element allows the specification of the field's type and name as attributes. Again the <*displayName*> element can be used for the same purpose. Attributes *type* and *name* must be specified or the custom field will be ignored. The types the user can specify are some of the most common ones: *string*, *int*, *uint*, *short*, *ushort*, *float*, *double* and *bool*. The next code shows an example:

```
<Schema name="string" id="ID">
        <SimpleField type="string" name="string">
                <displayName>...</displayName>
        </SimpleField>
</Schema>
```

The <*Schema*> elements declared are used to add custom data to a <*Feature*> using the element <*SchemaData*>. From an object-oriented point of view, the <*Schema*> element could be understood as the object declaration, being the <*SchemaData*> element the instance of that object. The <*SchemaData*> element refers to a specific <*Schema*> using its attribute *schemaUrl*. This attribute can contain a string representing a full URL for referencing other files or a <*Schema*> *id* defined also in other or in the same KML file. An example of its use can be found following:

```
<SchemaData schemaUrl="http://host.com/zclv.kml#my-schema-id">
<SchemaData schemaUrl="zclv.kml#my-schema-id">
<SchemaData schemaUrl="#schema-id">
```

To use the custom fields specified in the <*Schema*> element declaration for adding information to any <*Feature*>, the <*SchemaData*> element defines its child element <*SimpleData*>. The use of this element is limited to reference the custom fields declared in the <*Schema*> using the attribute *name*, and add its value like in the following example:

```
<ExtendedData>
        <SchemaData schemaUrl="#language-schema-id">
                <SimpleData name="LanguageCode">EN</SimpleData>
```

The last technique implies the use of already existing schemas. To make use of them the user just need to reference these schemas and assign a namespace with a prefix. This namespace prefix will be used along the KML file when adding elements defined in these external schemas. One important difference with respect the other techniques is that the information represented using this one, is not visualized in Google Earth. We can see how this method is intended to provide a way to transport information without any aim of visualization. The following lines represent a possible use:

```
<ExtendedData xmlns:prefix="my_own_metadata">
        <my_own_metadata:lang>EN</my_own_metadata:lang>
        <my_own_metadata:author>John Doe</my_own_metadata:author>
        <my_own_metadata:points>12</my_own_metadata:points>
</ExtendedData>
```

The <*ExtendedData*> element replaces the deprecated element <*Metadata*>. This new element opens new ways to insert any type of structured information within a KML file. In the following sections we will analyze its value when inserting structured metadata already declared in other standards such as ISO19115.

This is just a brief introduction to the different elements that compose the KML specification and some elements have not been explained. The OGC publish and maintains the official OGC KML schema. This schema is publicly available and can be checked in order to have a better understanding of the complete schema and the whole set of elements.

3. METHODOLOGY

3.1. Introduction

In order to achieve all the goals of the study, the best and probably unique solution seems to release a set of files and analyze the process since their publication to their hypothetical appearance in search results. The experiment could be divided into two main tasks: the content publication and the content indexing.

It has been explained in section 2.3 that there exist several ways to publish geographic content. How the files are published influence in the time spent by the crawling process. Discover which information is really used by Googlebot to create the Geoindex would allow to improve the design of KML files in order to obtain a successful indexing of the content. Not only the content and where it is allocated influence in the indexing process but also other factors could affect it. One these problems could be the content duplication that, in some cases could be avoided however in other seems inevitable. To obtain the results specific queries over the different search services that operate with the different indexes were performed. In the following sections a detailed description of each process is exposed.

3.2. Content Publication

There are several options to try to make the user's geographic content become part of the Google's Geoindex. Some of these options seem more effective than others in some aspects but they fail in others. For instance, the use of My Maps service assures the indexing of the content faster than any other way since the content is directly stored in Google's servers. Unfortunately the KML files created have some limitations because the use a two-dimensional viewers like the web browser. Since one of the goals of the study is the analysis of the KML files and their indexing a technique that implies a reduction in the number of available KML elements should be discarded. The submission of KML files to any Google's gallery allow the user to create KML files as complex as they want. Unfortunately this process does not seem to be effective in those cases where the number of files to publish is high. It seems that the best solution is the one that coincides with the recommendations given by Google in order to make user's geographic content public:

1. *Create the KML or GeoRSS content. Be sure to add attribution tags, which will appear in the Google Search results for your content.*
2. *Post your files on a public web server.*
3. *Create the Sitemap file. Copy this file to the directory of your website.*
4. *Submit your Sitemap to Google.*

For sure this is the slowest possibility however it seems the most practical, simple and probably the most used for the Google Geo Services users.

The server www.geoinfo.uji.es was used for the experiment. Googlebot already indexed this server some time before the experiment was performed and had good number of links in both directions. Inside this server a new folder was created for storing the different files and subfolders used as test data set for the study accessible in www.geoinfo.uji.es/kml. Once all the files were correctly uploaded to the server the address www.geoinfo.uji.es/kml was registered using the Google Webmaster Tools and the corresponding Sitemap element uploaded. This Sitemap file described the list of all elements that composed the test data sets using the special element *Geo* to indicate that each URL was referencing geographic content. These URL included KML files and PHP scripts that also returned KML files as output for their execution. These last ones were also included in the Sitemap using the *Geo* tag.

3.3. Content Indexing

The first of the Google's recommendation about the geographic content within a KML file concerns its authoring. As explained in section 2.4.2 it is possible to use some elements to support this purpose in KML. These elements are imported from the Atom specification: *atom:author*, *atom:name*, *atom:link* and *href*. This authoring information does not offer any security measure and could be easily a fake but it can be still useful for some users. However this aspect should be taken in consideration for those companies or organization publishing geographic content that need a more secure authoring method. This is not a big deal if the user knows the URLs used by the organization since these KML files are probably allocated in any of them. It is clear the great hole there still exists concerning the authoring of content. This problem does not affect exclusively the geographic content on the Internet and it is out of scope for this study. In this experiment all the files used in the test battery contained fake information about their authoring. The name *John Doe*, a typical English name to design people with unknown name, was used to indicate the author of the content. Although the name was not real the URL where the content was published was correctly provided.

In KML it is possible to organize the different elements like <*Placemark*> and <*NetworkLink*> into hierarchical structures using other *Container*-derived elements (<*Document*> and <*Folder*>). All these elements are derived from the same <*Feature*> element and then all of them share common elements. Among these elements there are suitable ones to store information or metadata. Then the KML specification presents the possibility of specifying metadata at different levels within a file. An example could be represented by those KML files containing a <*Document*> element that at the same time contain a set of several <*Placemark*> elements. In this case, the content creator could specify metadata or just descriptive information about the entire document at <*Document*> level. Also from a more specific point of view the user can do the same with each one of the <*Placemark*> contained in that <*Document*> element. From another perspective this could be used also to mix in a unique file content from different fields or with different information. A good example could be such a KML file specifying a <*Folder*> element containing two <*Document*> elements. All of them could describe <*Placemark*> elements where their descriptive information has nothing in common between them. In this scenario each <*Document*> contains its descriptive information or metadata separately from the information or metadata concerning the other <*Document*> element. These

different possibilities carry with them the need of analyzing at which level Googlebot extracts the descriptive information to be added to the index. In other words, discover if the information at <*Container*> and at <*Feature*> levels are both used. Since both share almost the same elements, files with the same information in the same elements but at different levels were used in the experiment.

Three different sets of information were created for the test, each one in one different language. The reason has nothing to do with the content's indexing but it is meant to facilitate the results recollection. These sets include information in English, Spanish and in Catalan, going from a more broadly used language to a less one. The results hypothetically obtained for the indexing of the Catalan information set would appear in a higher position in the search results and so these would be easier to find and the opposite could happen for the English set. This is a normal behaviour if it is considered the quantity of content in any of these languages. For each information set, a point represented by a <*Placemark*> in the KML files was created, all of them over Europe (two over Spain and one over Germany).

Another interesting aspect to test was the indexing of KML content dynamically generated. This is the case of scripts in languages like Python or PHP that generate a KML output or the case of the KML output option in GeoServer. In the experiment a PHP script that generated a KML for any of the different information sets was created. This script was executed given its URL and a special parameter indicating what information set should be used. The KML generated file contained information in both elements, <*Document*> and <*Placemark*> and using the elements <*name*> and <*description*>. It is possible that Googlebot recognizes geographic content based on the file's extension however it seems difficult that it executes and recognizes that a script's output contains geographic information unless the user indicates this fact specifically. For this reason the Sitemap and its Geo extension were used to help in the crawling process. The following code represents an example of how the script output was specified in the Sitemap file for its crawling and indexing:

```
<url>
        <loc>
                http://www.geoinfo.uji.es/kml/getKml.php?op=forest
        </loc>
        <geo:geo>
                <geo:format>kml</geo:format>
        </geo:geo>
</url>
```

Coming back to Google's advices about how to place the meaningful information within a KML file, there are four more recommendations:

- *Give your <Document> a meaningful <name>.*
- *Provide a relevant <description> for each <Placemark> so that the user can see the context of the search results.*
- *If you have a big quantity of data, divide it into topic-specific layers.*
- *Give each <Feature> an "id" so that the search result can link directly to it.*

These recommendations just talk about a really small number of elements in comparison with the whole set described in the OGC KML 2.2 specification. This

specification talks about a higher number of fields that can be suitable to store information or metadata. By suitable is meant that can contain information about the content itself either referring to a *Feature*-derived, *Container*-derived or any other KML element. At the same time this information can be represented as free or structured text in an XML-alike structure. Based on these assumptions the different elements chosen for the study were divided into four categories: Standard, Snippet, NetworkLink and ExtendedData KML elements.

3.3.1. Standard KML Elements

By standard elements are meant those KML elements that are specifically recommended by Google for storing content's information. These elements are <*name*> and <*description*>. They have a purely descriptive purpose in KML and effectively seem the best choice to store the most descriptive information about the content. Their use is recommended at different levels and it seems that for different purposes. The element <*name*> is recommended to be used within a <*Document*> however this element can be used for other nested elements derived from <*Feature*>. Also the use of the <*description*> for <*Placemark*> is recommended by Google in order to offer to the end user more information about the context of the search results. It is not clear then if the information stored in this element is used for the file's indexing. As it happens in the previous case this element can be also used at different levels.

Descriptive names and descriptions were inserted into the elements <*name*> and <*description*> for the experiment's test data sets. At the same time these elements were used separately and all together and also at different levels. With these combinations the test data sets contained 9 different files using the standard KML elements for each information set.

It is evident that existed some kind of redundancy in the content's distribution. It seems right to think that if the file containing information in <*name*> at <*Document*> level is indexed, the element containing information in the <*name*> and <*description*> elements at the same level will also be indexed. If this were not the case this would demonstrate that elements with the same information at same level were exposed to be excluded of the index. In other words, even a file with information in the right elements could be also rejected.

3.3.2. Snippet KML Element

The <*Snippet*> element is used to give a short description of the KML or KMZ file that is displayed in the Places panel when using Google Earth. In fact, when this element is not specified the first lines of the <*Description*> element are used as replacement. This field means a place for meaningful data that could be used to index the file. The only negative point of this element is its length, usually limited to a couple of lines. The <*Snippet*> field can contain meaningful and indexable data although this limitation. This field can be used for a short description of the content but also to store for example a set of key words related with it.

The <*Snippet*> element is specific of *Feature*-derived elements, so this could be used in <*Document*> as well as in <*Placemark*> or <*NetworkLink*> elements. Based on the idea of different levels where specifying the content's information this element can also be used in any of them. For the experiment's test files the information inserted in this element was the same used for the <*Description*> elements.

3.3.3. NetworkLink KML Element

The <*NetworkLink*> element within a KML file allows loading and visualizing information specified in another file. Thanks to this element the user can for instance reuse information in other files composing new ones. This KML element inherits from <*Feature*> and for this reason it has some already explained elements suitable to store information (<*name*>, <*description*> *and* <*Snippet*>). Although <*NetworkLink*> can also use those elements this is not the way it was used in the experiment.

As explained before, when using this element other file's data is loaded and visualized. If these loaded files have content stored in any tag suitable for being indexed maybe by a chain effect the file with the <*NetworkLink*> would become indexed. So the question would be which of these files would become part of the index. Would become the first, the second or both files? This is an interesting point and will provide information about how Googlebot uses the links within a KML file, if it explores them and how index the linked files.

For each information set a KML file containing a <*NetworkLink*> was created. These files linked to the PHP scripts previously explained. If effectively the content produced by the script is indexed through the <*NetworkLink*> file, this element can be applied to structure and publish automatically generated information. In the case that some geographic content is dynamically generated using a reasonable big quantity of scripts. An easy solution for its publication and indexing would be a single KML file containing the different <*NetworkLink*> elements linking to the different scripts.

3.3.4. ExtendedData KML Element

Probably the most interesting element studied in the experiment is the <*ExtendedData*> that allows the use of custom data inside a KML file.

This element allows the user to add content's information in a structured way, as the actual metadata standards for geographic information do. This KML element represents the possibility of bringing to KML files all the descriptive capabilities that the metadata standards like the ISO19115 represent.

In this study the ISO19115 was chosen as reference for metadata format in order to insert existing metadata in a KML file. This ISO standard defines the schema required for describing geographic information. It provides information about the identification, the extent, the quality, the spatial and temporal schema, spatial reference, and distribution of digital geographic data. This metadata schema is applicable to a broad range of activities including the cataloguing of datasets,

clearinghouse activities, the full description of datasets, geographic datasets, dataset series and more.

This standard probably exceeds the needs of the average Google Geo Services users. These users probably do not require specifying such a quantity of information about the content they are publishing. However from a more professional point of view the metadata is a key element in the actual perspective. The metadata results a basic element in the development of any SDI or simply to share and to publish geographic information in any other way.

Currently there exist tools capable to perform conversions from shapefiles to KML. The ESRI's shapefiles are one of the most used file types for geographic information. For all that users that take care about creating the correspondent metadata, these shapefiles have associated metadata content. If the <*ExtendedData*> element would represent a suitable place within a KML file where store this metadata, this could be included in the translation process. If this could be achieved and Googlebot effectively would access to the <*ExtendedData*> element and index its content this would probably be one of the easiest ways to publish a high quantity of the already existing geographic information. This would not only be an easy publishing technique but also would offer a huge infrastructure to search on those files. It is clear how important and useful could be the fact that Googlebot would extract information from this KML element for the indexing process.

To perform the experiment an ISO19115 schema stored in a public server and available on the URL http://www.geoinfo.uji.es/kml/schemas/iso19115/schema.xsd was used. In this study the three techniques available to insert custom XML data were used. However the ISO standard defines a nested structure impossible to recreate using the <*Data*> or <*Schema*> / <*SchemaData*> elements. For this reason, in these cases a simplified structure was created trying to represent a subset of the information contained in those metadata files.

The <*Data*> element can be used at any level either in <*Container*> or <*Placemark*>. Then for each information set three different files, one per each level at which the element <*Data*> could be applied, were created. As it happened with the standard fields the third file corresponds to that one that had the information at both levels. Some of the elements that were included using <*Data*> and trying to emulate the ISO19115 representing information about the file's title, description, topic category, contact information, creation date, language and spatial extent. It is also important to note that the name for the different elements were arbitrary chosen to express its purpose without following any convention. The code representing the use of the <*ExtendedData*> and <*Data*> elements for the files used in the experiment can be found in Annex I.

In the case of <*Schema*> and <*SchemaData*>, two different files were created for each information set. In the first one, the <*Schema*> element was defined inside the <*Document*> element. Using now the element <*SimpleData*>, at <*Placemark*> level the same information used for the files with the <*Data*> element was represented using the schema defined in <*Schema*>. This schema and how it was applied using the <*SimpleData*> element is showed in Annex I.

In the second file the same <*Schema*> was used however this time this schema was not declared directly in the file but referenced using the *schemaUrl* attribute in the following way:

```
<SchemaData
schemaUrl="http://www.geoinfo.uji.es/kml/files/forest_soil_chemistry_schema.kml#my
MetadataSchema"
>
```

The last technique used to insert custom XML data with the <*ExtendedData*> element makes use of an external and custom XML schema. This just requires declaring a namespace for the imported schema in the file's *kml* declaration (at the file's beginning). This technique allows the user to directly use the ISO19115 schema within a KML file. Now all the structure of the ISO standard can be directly introduced including nested hierarchies. In the same way a hypothetical translator from other file types with associated metadata could import the schema and encapsulate that metadata within the KML file. Here the only task is applying the specified namespace to each element. This information can be used also at any level. For this reason three different files containing the same information again at different levels were created. An example of the code used in the experiment can be found in Annex I.

3.3.5. Test Data Sets

For each one of the above elements and combinations one single element was created. Therefore the test data set for the experiment was composed by the following combinations of KML elements at different levels and in different files:

- <*name*> at <*Document*> level.
- <*name*> at <*Placemark*> level.
- <*name*> at both levels.
- <*description*> at <*Document*> level.
- <*description*> at <*Feature*> level.
- <*description*> at both levels.
- <*name*> and <*description*> at <*Document*> level
- <*name*> and <*description*> at <*Feature*> level.
- <*name*> and <*description*> at both levels.
- <snipped> at <*Document*> level.
- <*snipped*> at <*Placemark*> level.
- <*snipped*> at <*Document*> and <*Feature*> levels.
- <*NetworkLink*> linking a element with <*name*> and <*description*> at both levels.
- <*Data*> at <*Document*> level.
- <*Data*> at <*Placemark*> level.
- <*Data*> at both levels.
- <*Schema*> at <*Document*> level and <*SchemaData*> at <*Placemark*> level.
- <*SchemaData*> at <*Placemark*> level importing the <*Schema*>
- Custom XML data in ISO19115 format at <*Document*> level.
- Custom XML data in ISO19115 format at <*Placemark*> level.
- Custom XML data in ISO19115 format at both levels.

Finally when talking about the use of a meaningful <*name*> for the <*Document*> in the Google's recommendations this makes reference to those elements in a KML file. However when using Google Web Search if the keywords introduced as query parameters coincide with a file or domain name this could appear in the results for that search. It seems that the files' name is used by Googlebot when crawling and indexing websites for the Google Search Index. Maybe this could be the case also for the Geoindex. In the test batteries, each file was duplicated. One of them had assigned a descriptive name indicating its content and where this content was located within the KML file. The other had a name composed with the first letters of the words composing the others name. For instance a couple of files for one of the information sets used in the experiment were: *fsc_std_name_feature.kml* and *forest_soil_chemistry_std_name_feature.kml*. This would also help in the results recollection making easier the process of identifying which files would be part of the Geoindex. At the same time this would give information about the importance of a descriptive file name for KML files.

With all these variations and if the crawling process would work correctly with the appropriate files, the results of this test battery could give a clearer idea about what elements are useful and at what level. In total each information set had 42 KML files plus one script, what means that the experiment was composed by **129** KML elements to be indexed. All the files created for the study were validated previously to their publication. The validator used was the free online KML validator provided by the company Galdos (http://kmlvalidator.com) that supports the current OGC KML 2.2 specification. The experiment, deeply explained in the following section, was based in the search of the above explained files using the services offered by Google. By these searches' results measures about time, number of indexed files and KML elements present in the indexed files were obtained.

4. EXPERIMENT

4.1. Introduction

Basically the only way to test the effective indexing of the files was to search for them using the different Google's tools and services. All the different Google Geo Services and tools seem to access the same index, the Geoindex. In the other hand all files found through web searches are indexed in the Google Search Index. For a KML file to become part of the Geoindex seems to mean to become also part of the Search Index previously. This situation brings the possibility of comparing both indexes and so compare which index can contain more geographic content.

Using either Google Maps or Google Earth the users access the Geoindex for any search. This search at the same time is restricted by three factors:

- **Purpose or type of the geographic content**. This option is just available using Google Maps and allows the users to chose from directions to user-created content. It is also possible to choose all results without specifying a category. In the different search performed in the experiment the options used were *all content* and *user-created content*. In fact all the information published belongs to this last category since it has been crawled by Googlebot without using any of the services Google offers for companies, which information could be part of the *businesses* category.
- **Region.** It seems that when using Google Maps or Google Earth for searching any type of geographic content the search also uses the information about the region that is being visualized. If there exist results within the specified region these are shown. If there are no results in that region but in another one, Google Maps or Google Earth change the visualization to that region. Finally if there exist no results for the query the corresponding message is shown. Then it seems that the active region or the actual bounding box could be used by Google as a filtering or sorting parameter based on the number of results in the given zone. This option in Google Maps and Google Earth has been used when performing the different searches for the experiment in order to be more precise and avoid unnecessary search results.
- **Free text**. This is the basic element for the search. In the searches performed different words and sentences used in the KML elements to test were used. Also, searches including part of the KML file's name were also performed in order to discover the relevance of this value in the indexing.

Google Web Search allows the filtering of a search by file type. Among other file types Google allows the user to search for KML files. This option allows checking the files from the test data set that were also indexed and became part of the Google Search Index. In order to get a fair comparison between both indexes, the same information was used when performing the search queries.

4.2. Results

The first test data set was launched on October 30th 2008. Meaning by launching the file's publication in a publicly accessible server and the submission of the corresponding Sitemap file to Google using the Google Webmaster Tools. Search queries were performed daily to check the appearance of any of the test data set's files in the results provided by Google. The first results were found searching on Google Web Search on November 21st 2008 and within the same day some files were successfully found using the search capabilities in Google Maps. More results started to appear within the following days. This indicates a crawling and indexing time of approximately three weeks. In the following three months after these results appearance, weekly queries were performed in order to discover any change in the number of results annotating no changes.

In a first moment the total number of files created for the experiment was of **129**, including KML files and PHP scripts for the three information sets. One effective way to find all the files that were part of any of the indexes was the use of the search query restricted to those files in the experiment's specific domain. The search query *site:www.geoinfo.uji.es/kml* was used to retrieve all the indexed files coming from the used domain. Using Google Web Search the search was restricted to KML files only. In a first moment the number of results returned by Google were two. This result was caused by the omission of similar results that Google performs by default. Repeating the search, this time including those similar files, the number of results in the search was **56**. This represents the **43%** of the files in the test battery. Performing the query with the same keywords on Google Maps, indicating a search over the user-created content the number of results where **7,** what represents the **5,42%** of the overall number of files created for the experiment. Figure 7 shows a graphical representation of these results.

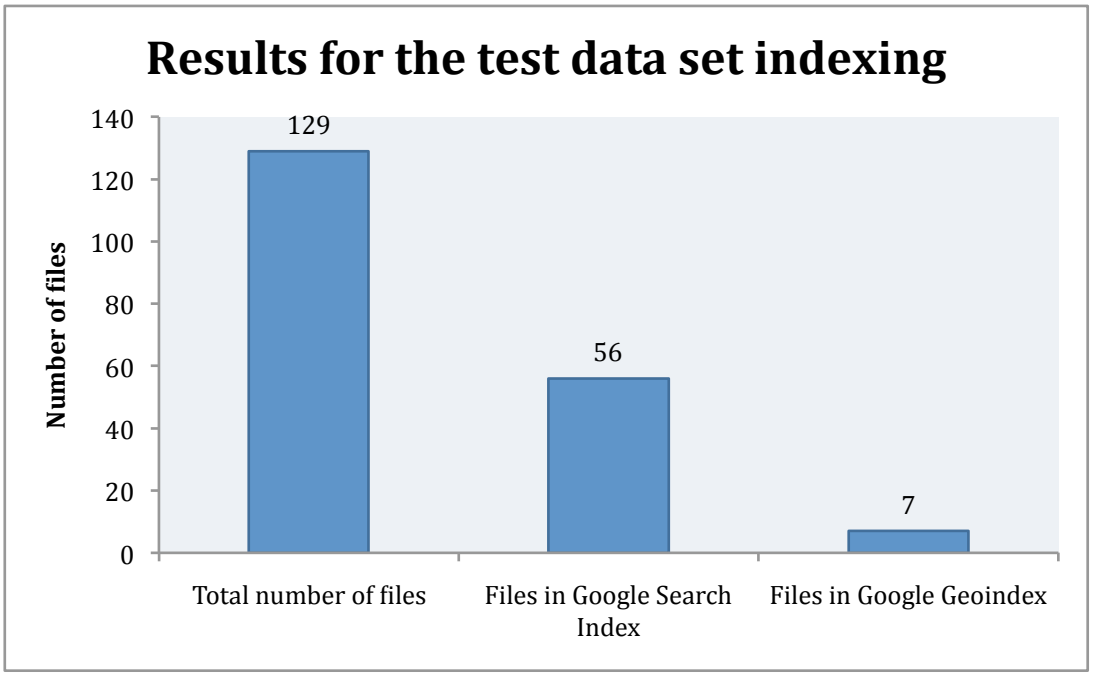

Figure 7: Results for the test data set indexing.

Other more specific searches based on textual information contained in the KML files confirmed these first values. For instance, using words and sentences for any of the information sets returned the same files obtained in the first search.

The seven KML files successfully indexed presented their information in the following KML elements and levels:

Information set 1:
- File 1: Information in elements <*name*> and <*description*> at <*Placemark*> level.
- File 2: Information in element <*name*> at <*Placemark*> level.

Information set 2:
- File 1: Information in elements <*name*> and <*description*> at both levels.
- File 2: Information in element <*name*> at <*Placemark*> level.
- File 3: Information in element <*name*> at <*Placemark*> level.

Information set 3:
- File 1: Information in elements <*name*> and <*description*> at both levels.
- File 2: Information in element <*name*> at <*Placemark*> level.

After obtaining these results and because the low number of files indexed a second test data set was released. This new test data set was created to avoid some possible issues that could affect the indexing of the first files. This issues include the filter done by Googlebot when detecting duplicated content. The files composing it had information in the same elements and with the same combinations as the files in the first data set. However now, the information contained in each file was unique for each one of them. Also, the coordinates or situation specified in each one was not shared between them. In other words, each file differs from the rest in content, KML elements used and geographic situation. Finally in this case, all files get an arbitrary name since the utility of a specific file name can be observed with the first test data set. The publication process was almost the same. The files were uploaded and their references added to the existing Sitemap file that was sent again to Google using its Google Webmaster Tools. Obviously, all of the URL specified for the new files made use of the special tag Geo to indicate the type of content.

The crawling and indexing times were similar to the previous experiment with the first test data set however other big and important differences were found. In the first place, the most significant difference seems the fact that the files indexed appear using any of the Google Geo Services but not in the search results when using Google Web Search. Concerning the number of results obtained using the Geo Services, the number of files indexed were of **four** from an original number of 23 KML files plus one PHP script, what represents more than a **16%** of the files.

In this case the files indexed presented the following configurations concerning KML elements and levels:

- File 1: Information in elements <*name*> and <*description*> at <*Placemark*> level.
- File 2: Information in elements <*name*> and <*description*> at both levels.
- File 3: Information in element <*name*> at <*Placemark*> level.
- File 4: Information in element <*name*> at both levels.

5. DISCUSSION OF RESULTS

5.1. Time

Checking in Google's forums related with Google geographic products and services, some of their users reported their experience concerning the time taken some months previously to this study. These users employed the different services offered by Google either for personal or professional uses but not for any scientific study. Although their opinion could be understood as personal these are still taken in consideration as references or indicators. Most of them indicated that the average time for crawling and indexing the geographic content was of six weeks. Apparently this process time is decreasing and hopefully it will decrease even more in the future. Google Geo Services are relatively new. Google owns a massive infrastructure however not all the resources are applied to its Geo Services. We could expect that as the time goes on and the number of users and possibilities of these services increase, not just from a technical perspective but also economical, the company will probably assign more resources. The apparent trend that is being observed in the Web concerning the creation of georeferenced content could support this idea. This aspect combined with the free cost of the most of the services Google offers seems to indicate that their users and contents will keep also being increased.

However although the trend shows that the crawling and indexing time is decreasing the truth is that three weeks could still be too much time for a high number of uses. For all those users that for instance just want to publish a KML file with their last holidays' pictures, a period of time between three and six weeks is maybe not too much time to wait for being able to search and find their content using Google Maps. However these services are not limited to those uses. The world is a dynamic place, continuously changing and these changes affect any kind of data including the geographic information. For instance it could exist the case where somebody, some organization or company needs to publish some kind of geographic content that represents one of these dynamic processes. Depending on the process itself it is possible that the information could become obsolete before its hypothetical discovery by its potential users. It is also true that these are free services and then they offer no guarantee either in time or crawling effectiveness. Probably in the cases where a fast access is required other solutions should be studied.

5.2. Effectiveness

For the first test data set the final number of files returned in the searches were not high using Google Web Search and much lower when using Google Maps. The reason why such a small number of elements were found in the index could be caused by problems in the crawling or indexing processes or even in both. The Google Webmaster Tool did not indicate any problem in the Sitemap submitted or in the crawling process. It is also true that because of the use of a subdirectory and not a root directory, Google Webmaster Tool did not show all the possible information. However this situation could be usual for some users storing data in subdomains or

subdirectories within a domain. Also, there were no robots.txt files that could block the crawling process for any directory and the structure was easily accessible without any rich media format file that could cause problems in the indexing process. Although none of these negative aspects were found in the site not all the files were crawled. Since the KML files are text based the problems in the indexing process could be related with the content's relevancy and usefulness. It is known that one of the factors that Google analyzes in order to assign a relevancy is the number of links to and from a site. In the experiment this number was especially low, for all the files inside the subdirectory used. Among other reason the one that makes more sense in the case of this experiment was the existence of duplicated content. One of the quality guidelines recommended by Google is the avoidance of sites with substantially duplicated content. Google refers to duplicate content as *substantive blocks of content within or across domains that either completely match other content or are appreciably similar*. Some examples could be found in the discussion forums, store items shown or linked via multiple URLs or printer-only versions of web pages. The reason why Google recommend to do not duplicate content is because, in some cases, this duplication of content is done to drive traffic to websites and manipulate search engines rankings. These are some reasons why a site could be penalized or even removed from Google's index. Usually when Google detects duplicate content it chooses one of the sources to list. However Google "can perceive" that content that has been duplicated for the above explained reasons and apply the corresponding measures to the site. Google gives some more specific advices about how to avoid the problems with the duplicated content [46].

In the experiment, for each information set, all the different KML files, although in different places within the file, contained the same or really similar textual information. Reading the duplicated content description it seems possible that Googlebot evaluated as duplicated content many of the different files discarding some of them for their indexing. This situation would explain why just the 43% of the KML files appeared in the Google Web Search Index. However the number of KML files appearing in the Geoindex was still much lower. Then the idea of an additional prune process between the Geoindex and the Google Search Index was taken in consideration.

The first hypothesis that maybe could explain this prune or filtering process had relation with the KML elements used by Googlebot to create the Geoindex. It would be possible that some of the KML files that were successfully indexed and included in the Search Index did not contain valuable information in the fields that Googlebot is supposed to analyze for the geographic content. In this case it is possible that some of these files were discarded for their insertion in the Geoindex because the useful KML elements were not filled with information.

The second hypothesis about the low number of indexed files came from an error present in some of the files that composed the test battery and that was discovered after the files' indexing. Three of the KML files containing information about one of the information sets had an error in the coordinates that represented their position. All the files but these three ones were situated in a point over Spain. The error in the coordinates situated these three over Kenya. For the other two information sets, just two KML files were successfully indexed. In the case of this third set (the one with the wrongly situated files) three files, including one of the KML files with wrong

coordinates were found when searching the appropriate words and visualizing the entire globe. Probably this result is just a casualness however brings the idea of other type of filters or prune, this time based on the region. Maybe when creating the Geoindex, Google also filters and discard the duplicated content in the same or similar points or regions. This is just a hypothesis and further experiments had to be done to confirm it.

The fact that maybe a percentage of the files in the first test data set were not indexed because their duplicated content and the hypothesis of a filter of those files with similar content and similar geographic position brought the need to extend the study. This extension meant the confection of a new test data set with KML files that this time did not share either content or position. If the rate of indexed files would be the same this could mean that none of the previous hypothesis (duplicated content and duplicated position) would explain the low rate. Discarding those ones the main reason for a successful indexing would become the appropriate or inappropriate use of the KML elements. The results of the second test data set showed just four files representing more than the 16% of the total files in this test data set. In this case the content was unique in each file and also the geographic position. Again, the number of files indexed is not high. However checking which files were finally indexed a clear pattern that will be analyzed following could be extracted. This pattern demonstrate that all the files indexed have information in the element <*Name*> at <*Feature*> level. This demonstrates that effectively there is a prune of KML files based on where the information is placed inside a file. However, these indexed files do not represent all the files in both test data sets that contain information in that location within the KML file. For instance, in the first test data set, there were five files that include information in the right place within a file:

- <*name*> at both <*Document*> and <*Feature*> levels.
- <*name*> at <*Feature*> level.
- <*name*> and <*description*> at <*Feature*> level.
- <*name*> and <*description*> at <*Document*> and <*Feature*> levels.
- PHP script.

Considering that the files are duplicated because the analysis of the file's name that finally have demonstrated useless the number of files suitable to be indexed per each information set in the first test data set is nine. The first information set got two files indexed (22,22%), the second again two files and the third (the one with wrong coordinates in some files) three files indexed (33,33%). In the second test data set just five files were suitable for indexing due to the files were not duplicated to test the effectiveness of the file's name. In this case over five files four were indexed (80%). Figure 7 represents a comparison between the percentages of suitable files indexed in each data set:

Observing the graph and considering that the content in the files composing the second test data set was not duplicated it is clear that the duplicated content was probably a cause also for the low number of files indexed in the first test data set. By these results it is impossible to assure that a hypothetical pruning process based on location was performed. Although more analysis with a bigger number of files should be performed in order to confirm these values, seems that the KML elements used and

the existence or not of duplicated content affects the results in the indexing process for the Google Geo Services.

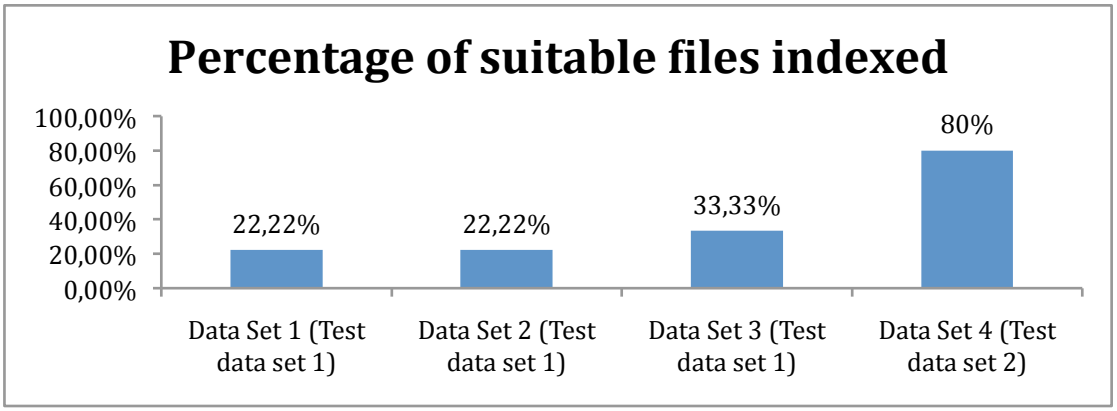

Figure 8: Comparison of percentage of suitable files indexed.

Another important result obtained with the second test data set is the fact that none of its files appears as search result using the Google Web Search. This means that probably, these files do not appear in the Google's Search Index. This fact needs further investigation however could demonstrate that the Google Geoindex is updated separately from the Google Search Index. Maybe the Geoindex is just feed for the first time by the Search Index and the updating of the sites that already appear in the Geoindex are directly performed without using the other index. As already mentioned these are just hypothesis and would need of further investigation and experiments. These new ones could reproduce the experiment here performed with different configurations regarding the number of files, Sitemap file options or updating period among other factors.

5.3. Elements for the Indexing.

The following table summarizes the results obtained concerning the elements used in the KML files:

Test Data Set 1			**Test Data Set 2**
Data Set 1	*Data Set 2*	*Data Set 3*	*Data Set 4*
<name> & *<description>* at *<Feature>* level.			*<name>* & *<description>* at *<Feature>* level
<name> at *<Feature>* level.	*<name>* at *<Feature>* level. (2 files)	*<name>* at *<Feature>* level.	*<name>* at *<Feature>* level.
	<name> & *<description>* at both levels.	*<name>* & *<description>* at both levels.	*<name>* & *<description>* at both levels.
			<name> at both levels.

Table 8: KML elements used in the indexed files.

Looking at the results one clear pattern can be found. Those files with the information in the element <*name*> at <*Placemark*> level become part of the index. This seems right since the KML files with the information allocated in that element at that level have been indexed for all the information sets. The rest of results could indicate that using the element <*description*> and placing it at different levels could also derive in the file's indexing. However it is not possible to assure this fact since in those files there is also information in the element <*name*> at <*Placemark*> level what could really had influenced in the indexing of these files. What it is truth is that when using the Google Maps search option the text contained in the <*description*> element is used. This could mean that in this case just the elements with descriptive information in the element <*name*> at <*Placemark*> level have been indexed however all the content in the file or at least the content appearing on the element <*description*> have been added to the Geoindex. The name given to the file did not seem to be used for either the file indexing or as element to analyze when performing searches. Finally some of the PHP scripts used in test data sets were successfully indexed. This means that when these scripts are referenced in the Sitemap file as geographic content, their output is treated as a KML file and then suitable to be indexed if it contains the information in the right element.

At this stage seems that the use of the KML element <*name*>, at <*Feature*> level and the submission of a correct Sitemap file using the Google Webmaster Tool offers a high effectiveness concerning the indexing of a KML file. Also the avoidance of duplicated content has been demonstrated to be a requisite for a good indexing ratio. This name can or cannot contain descriptive information that would be used in posterior queries. The reason is that once the file is indexed, its content or at least the information stored in the element <*description*> is added to the index and analyzed when performing a query. This is strange since those files that contain information in that element but do not make use of the element <*name*> do not become indexed. This could be applied also to other elements. Then further studies analyzing the combination of the element <*name*> with the rest of KML elements would have to be done.

With these results maybe a reconsideration of the Google's advices should be done. These recommendations advice the use of the element <*name*> at <*Document*> level however it has been demonstrated that this element is useful at <*Feature*> level for a correct indexation. Also the recommendations talk about the use of the element <*description*> to inform the user about the file's context. However, since the information contained in this element is also searchable this seems more important than a mere context indicator.

5.4. Non-technical Aspects

Even if all the technical aspects of Google Maps and Google Earth and the technology behind them satisfies the requirements for a feasible search engine for geographic data there is still, at least one fundamental point to take care about. This is nothing related with the technology but how people can use the service and what is supposed to agree with when using it, in other words, the Terms of Use. The Google Maps/Google Earth APIs Terms of Use [47] were updated for the last time on

November 2008. These last updates (several in the same month) carried some discussions and uncertainty in a great number of users [48, 49, 50, 51]. One of the main controversial points is found in section 11, Licenses from You to Google:

> *Content License. Google claims no ownership over Your Content, and You retain copyright and any other rights you already hold in Your Content. By submitting, posting or displaying Your Content in the Service, you give Google a perpetual, irrevocable, worldwide, royalty-free, and non-exclusive license to reproduce, adapt, modify, translate, publicly perform, publicly display and distribute Your Content through the Service and as search results through Google Services. This license is solely for the purpose of enabling Google to operate the Service, to promote the Service (including through public presentations), and to index and serve such content as search results through Google Services. If you are unable or unwilling to provide such a license to Your Content, please see the FAQ for information on configuring your Maps API Implementation to opt out.*
>
> *Brand Features License. You grant to Google a nontransferable, nonexclusive license during the Term to use Your Brand Features to advertise that you are using the Service.*
>
> *Authority to Grant Licenses. You confirm and warrant to Google that you have all the rights, power and authority necessary to grant the above licenses.*

From this section the main and most problematic points are clearly the first and the last one. In the first point it is meant that Google does not want any right over the content you publish. However it seems that Google wants to own the right to use that content always (perpetually), in any place (worldwide) and finally for free. It is also true that they also indicate their purpose, basically for marketing and to serve this contents as search results. This is a reasonable measure taken by Google that wants to do their job without any legal issue opened and at the same time increase and improve the service with more and more data. However this is problematic for all that people that use non-free or already licensed (and incompatible with Google's Terms of Use) data through the service.

The third point brings a problem for all of those who do not have legal rights over the data they work with. This is not a strange scenario. In the Web 2.0 the merge of different sources of information into what is called mashups to create new applications is a spread used technique. A clear example is that of all those who create mashups combining different sources of information (i.e. properties to sell, schools in a given zone, shops, etc) that provide data which license is not own by the mashup developer. Somebody could also think that some of the data used by the mashups is not even stored in Google's server so then probably is not affected by these terms. Actually even if the data is not stored in Google's servers is showed or processed through the Google Maps service and then suitable to be affected by the terms.

Maybe the best example of problems derived from such legal issues is the one presented by the British Ordnance Survey. With the release of the new terms of use in November, the Ordnance Survey informed the Local Governments in United Kingdom [52] about its position against the use of Ordnance Survey’s derived data

through Google Maps service. Basically the problem by the Ordnance Survey was the incompatibility of the Google's Term of Use with the actual copyright license for Ordnance Survey data. In other words, the Ordnance Survey did not agree with the idea of Google Maps users granting Google a license of Ordnance Survey based or derived data. This represents a huge problem for all Google Maps API in the United Kingdom since the vast majority of geographic data is produced or derived from Ordnance Survey data. Polemics apart, this is a good example of how problematic this legal issues can become.

After all, even if one of the most famous Google's advices is “do not be evil”, Google is finally a private company with its own interests. This does not mean that Google “is evil” with the users, however as any other company has some interests and those interests change over the time and this changes could be translated into changes in their products or services licenses. Usually this changes benefit the user experience however with small changes such the ones done in November 2008 could bring big problems. Finally Google have been offering for some time a free service without any return, so in some sense it is normal that it expects some return to its invest.

6. CONCLUSIONS AND FUTURE WORK

6.1. Conclusions

During the last years Google has become a point of reference in the World Wide Web and its power and importance is no doubted. Its image as company but mostly the set of free services and tools that usually satisfy user's expectations have made Google to gain a huge number of users. With the advent of the company into the geographic information market with the release of Google Maps and Google Earth, the company opened more the door for the geographic content creation and sharing. As in the case of its other products, the simplicity, the fact that are free and also the quality have made that Google Geo Services and tools gain more and more users over the last years. This increase in users has been directly translated into an increase of geographic content with more or less interest but all publicly available on the Internet. It is popular that sentence that says that Google tries to organize the world's information and it seems it tries do it discovering and indexing all that content. In fact, in the first days of the company, its unique product was a search engine that actually it is a quite effective and satisfactory one. It seems logical to the evolution of the product allowing the processing of new types of information including geographic content. Google Web Search is for most of users an effective tool for searching on the Internet all that information they are seeking. The question is then if Google would provide and effective solution for discovering and retrieving geographic content in the Internet.

The OGC KML specification defines a language with a great number of possibilities for visualizing geographic data using either two-dimensional or three-dimensional viewers. However the KML language is not limited for visualization and it has been demonstrated that offers to the user great possibilities to transport information. In this study different KML elements have been reviewed and in the experiment some of them used for carrying information. The actual specification allows the use of several elements where inserting descriptive information. One interesting aspect is the possibility of introducing information or metadata at different levels within a KML file. These levels could be seen as the different levels in a hierarchical relationship that is possible to build thanks to the KML structure. For example, KML would allow the user to specify general information at <*Document*> or <*Folder*> level and more specifically about each <*Feature*> represented at its own level. It has been also seen that KML could be integrated easily with the actual metadata actions and standards like the ISO19115 standard. Information in this format could be easily encapsulated within a KML file using the new element <*ExtendedData*>.

The experiment tried to reproduce a real case scenario following most of the Google's recommendations to facilitate the crawling and indexing of the published geographic content. Though maybe some of the parameters, configurations or other elements could influenced the results, however it is probable that in the real world not all the recommendations can be followed either, what reaffirms the validity of the experiment.

The crawling process has presented an average of 3 weeks time. This crawling times could be acceptable for some users and applications however in some other cases this could mean that the content is already obsolete when it becomes available using Google's search services. Based on this aspect the system could not be effective for some uses.

It has result that the only relevant information within a KML file for its correct indexing is the presence of the element <*name*> for the <*Feature*> elements described in the file. This element is fundamental and seems required in order to get the file correctly indexed. It has been also demonstrated that once the file is indexed other information like the one contained in the element <*description*> is also used and checked when performing queries, however this is not relevant for the file's indexing. This has been proved so far however further studies are required for checking the combination of the <*name*> with other KML elements and then analyze if these elements are also used when performing searches.

Answering one of the research questions, the place where the information is stored within a KML file effectively affects its indexing. In fact this has been demonstrated an important factor for its indexing but not the only one. In a first stage the low number of indexed files was though to be caused by the content duplication problem that also affects the indexing of normal web sites. However the second part of the experiment has demonstrated that apparently this behaviour is still present in the Google Geo Services but is not the most important. It seems that what really makes a KML file become part of the Geoindex is not the information itself but where it is present within the file. Obviously other aspects such as the submission of a correct Sitemap file affect the process.

The OGC KML file format has demonstrated its flexibility and utility. With the new <*ExtendedData*> element this format is capable of carrying information described by custom KML or other arbitrary XML schemas. This can be the case of the metadata described using the ISO19115 format. It has been proved with their validation the correctness of these files containing such information, however it hasn't been proved that this information is used and inserted in the Geoindex. Then we can conclude that the file format allows the reuse of existing metadata however that the Google system could make use of that information still needs to be proved.

There are some reasons that could make think that Google could not be always the best solution against other options such as the actual catalogues like Geonetwork. Unfortunately one of the main reasons has nothing to do with technical aspects. While Google seems a more than promising solution for discover and retrieve geographic content, as seen in section 5.4 there exist some problematic concerning the rights of the content itself. Both parts, the service users and the service provider, in this case Google, have their own reasons to agree or disagree with the terms of use concerning the rights. It is understandable that Google needs rights over the content published using their services but at the same time it is also understandable that this could bring a conflict with previous copyrighted content. Another problem that is not present in the use of catalogues is the authoring of the content. This has not been treat in this study because its complexity and length and deserve big studies about it. However the fake information in the geographic content published using Google Geo Services is present and probably increasing in number. Then effective methods to control the

authoring of the crawled geographic content must be created. Another negative point is the time required to index the geographic content although there is an apparent trend for its reduction. This cannot be fairly compared with the time required to publish geographic content using a catalogue service since the publishing method is totally different. While the use of catalogues usually implies the submission of files by the user the Googlebot allows the user an easier and automatic publication method. However this simplicity is paid in time, what it does not necessary mean a handicap in all the cases. What could really be observed as an impediment is the fact that Google never guarantee the successful indexing of the content. This has been experienced in the experiments performed. Another important problem related to this one is the appearance of duplicated content that significantly reduces the number of files indexed. This like the issue with the time should be taking in consideration when planning to use these services. Finally although the use of Google Geo Services represent a deep change in the metadata confection, this does not necessary means a lose of information. Instead of distributing the metadata in a long list of different elements within a file, all the descriptive and important information could be placed in one single element (or more if the other KML elements become part of the index). Then Google Geo Services could be considered as an alternative for all those cases where the content does not need to be immediately available, the content among different files is substantially different, its authoring does not need to be confirmed and don't exist problems with the copyright.

6.2. Future Work

Although some of the basic aspects in the use of KML and Google services to discover and retrieve geographic information have been presented some further investigation needs to be done in order to understand and analyze all the possibilities of these elements.

It has been demonstrated those KML files with information in the element <*name*> at feature level become part of the Google's Geoindex when they have no duplicated content. The use of Sitemap files specifying the geographic content, could improve the number of files indexed and for sure it should improve the crawling process. In a future experiment the combined use of the <*name*> element with other KML elements such the ones related with <*ExtendedData*> should be performed. If with this future experiment it is demonstrated that just the information contained in <*description*> becomes part of the index, other ways to add structured or semi-structured metadata could be studied. A possible experiment could be the use of microformat with the <*description*> element. This element allows the use of HTML tags and if it would allow the use of microformats as well a way to add structured and indexable metadata would be possible. This technique would allow the correct indexing of the content within a KML file and would keep the information structured for its use by other applications.

Another aspect that needs more investigation is the use of the Google Search Index as base for the Geoindex. In a first moment it seemed that both were related and the fist fed the second that filter the files to index based on where the information was contained within the KML file. However with the second test data set, the Geoindex recover all the suitable files without using the search index since none of those new

files appear when searching for them using Google Web Search. This really needs to be further analyzed and demonstrate that this has not been an isolated case. In the case this is actually an isolated case, the use of one or another index could be compared in other studies since the number of results using the general index was much higher than with the use of the Geoindex in the first part of the experiment. In this hypothetical study the benefits of one or another could be studied and even the combination of both. This combination could actually be created using the available APIs that can perform queries on both indexes. The main point of study here would be how combine the results obtained in each index to offer a single list of results.

At the same time, the experiment has been performed in a really reduced and concrete environment. Therefore it should have to be reproduced in more machines with different configurations in order to assure that the results could be considered global.

BIBLIOGRAPHIC REFERENCES

[1] Google Maps Website (URL: http://maps.google.com, accessed September, 15th 2008)

[2] Yahoo Local Maps Website (URL: http://maps.yahoo.com, accessed September 15th 2008).

[3] Live Maps Website (URL: http://maps.live.com, accessed September 15th 2008)

[4] Google Maps - My Maps (only for registered users in Google Maps, URL: http://maps.google.com, accessed September 15th 2008)

[5] Google Map Maker (URL: http://www.google.com/mapmaker, accessed September 15th 2008)

[6] Google Earth Website (URL: http://earth.google.com, accessed September, 15th 2008)

[7] OGC Catalogue Service Specification (URL: http://www.opengeospatial.org/standards/cat, accessed September, 13th 2008)

[8] The SDI Cookbook, version 2.0 (URL: http://www.gsdi.org/docs2004/Cookbook/cookbookV2.0.pdf, accessed September, 14th)

[9] OGC KML Specification. Open Geospatial Consortium Portal(URL: http://www.opengeospatial.org/standards/kml/, accessed September, 13th 2008)

[10] Open Geospatial Consortium Portal. (URL: http://www.opengeospatial.org Accessed September, 13th 2008)

[11] Submit Your Geo Content to Google. Google Code website (URL: http://code.google.com/intl/en-EN/apis/kml/documentation/kmlSearch.html, accessed October, 1st 2008)

[12] Douglas E. Comer. Computer Networks and Internets: With Internet Applications (page 20). Edition 5. Prentice Hall, 2008. ISBN 0136061273, 9780136061274

[13] Alexa the Web Information Company (URL: http://www.alexa.com/site/ds/top_sites Accessed October, 22nd 2008)

[14] T. O'Reilly. What is Web 2.0 – design patterns and Business models for the next generation of software, 2005. URL: http://oreilly.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html Accessed November, 3rd 2008

[15] Scharl, A. (2007). "Towards the GeospatialWeb:Media Platforms for Managing Geotagged Knowledge Repositories", Eds. A. Scharl, K. Tochtermann. London: Springer. 3-14. The GeospatialWeb - How Geo-Browsers, Social Software and the Web 2.0 are Shaping the Network Society. ISBN 1-84628-826-6

[16] Andrew Turner. Introduction to Neogeography. O'Reilly, 2006. ISBN 0596529953, 9780596529956

[17] Global Earth Observation System of Systems (GEOSS) Website. (URL: http://www.epa.gov/geoss/, accessed September, 21st 2008)

[18] OGC Web Map Service Specification. Open Geospatial Consortium Portal. (URL: http://www.opengeospatial.org/standards/wms Accessed September, 13th 2008)

[19] OGC Web Feature Service Specification, Open Geospatial Consortium Portal. (URL: http://www.opengeospatial.org/standards/wfs Accessed September, 13th 2008)

[20] OGC Web Processing Service Specification. Open Geospatial Consortium Portal. (URL: http://www.opengeospatial.org/standards/wps Accessed September 13th 2008)

[21] INSPIRE Directive. Official web site of the European INSPIRE directive (URL: http://inspire.jrc.ec.europa.eu , accessed September, 15th 2008).

[22] ISO19115 specification web site (URL: http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=26020 , accessed September, 16th 2008)

[23] Comber, J. Alexi, Fisher, F. Peter, Wadsworth, and A. Richard (2008, June). Semantics, metadata, geographical information and users. Transactions in GIS 12 (3), 287-291.

[24] Goodchild, M. Beyond Metadata: Towards user-centric description of data quality. (URL: http://www.geog.ucsb.edu/~good/papers/435.pdf, accessed September, 15th 2008)

[25] Bulterman, D. C. A. (2004, October). Is it time for a moratorium on metadata? IEEE MultiMedia 11 (4), 10-17.

[26] Adobe Developer Connection – PDF Technology Center (URL: http://www.adobe.com/devnet/pdf/, accessed November, 3rd 2008)

[27] Adobe Developer Connection – Shockwave Flash Technology Center (URL: http://www.adobe.com/devnet/swf/, accessed November, 3rd 2008)

[28] Microsoft Office Binary (doc, xls, ppt) File Formats (URL: http://www.microsoft.com/interop/docs/OfficeBinaryFormats.mspx , accessed November, 4th 2008)

[29] Goodchild M.(2007). Citizens as sensors: the World of volunteered geography. GeoJournal Volume 69, Number 4, 211-221

[30] GeoRSS: Geographically Encoded Objects for RSS feeds Website (URL: http://georss.org, accessed October, 20th 2008)

[31] RSS Advisory Board Website (URL: http://www.rssboard.org, accessed October, 20th 2008)

[32] ESRI's Shapefile Technical Description – 1998 (URL: http://www.eseri.com/library/whitepapers/pdfs/shapefile.pdf, accessed November, 20th 2008)

[33] Google Webmaster Guidelines (URL: http://www.google.com/support/webmasters/bin/answer.py?hl=en&answer=35769 , accessed October, 5th 2008)

[34] Nearby.org.uk Blog (URL: http://www.nearby.org.uk/blog/, accessed September, 29th 2008)

[35] Google Earth Gallery Website (URL: http://www.google.com/gadgets/directory?synd=earth&cat=featured, accessed September, 29th 2008)

[36] Sitemaps.org Website (URL: http://sitemaps.org, accessed November, 2nd 2008)

[37] Anthony T. Holdener III, Ajax: The Definitive Guide. January2008. O'Reilly. ISBN 10: 0-596-52838-8 | ISBN 13: 9780596528386

[38] ArcGIS Explorer Overview Website (URL: http://www.esri.com/software/arcgis/explorer/index.html , accessed November, 16th 2008)

[39] OpenLayers Website (URL: http://openlayers.org, accessed November, 10th 2008)

[40] NASA's World Wind Website /URL: http://worldwind.arc.nasa.gov, accessed November, 10th 2008)

[41] Autodesk - AutoCAD Services & Support : DXF Reference (URL: http://usa.autodesk.com/adsk/servlet/item?siteID=123112&id=12272454&linkID=10809853 , accessed November, 8th 2008)

[42] GPS; The GPS Exchange Format Website (URL: http://www.topografix.com/gpx.asp, accessed November, 10th 2008)

[43] RFC4180 specification at the Internet Engineering Task Force (IETF) Website (URL: http://tools.ietf.org/html/rfc4180, accessed November, 9th 2008)

[44] OpenGIS Geographic Markup Language, Open Geospatial Consortium Portal. (URL: http://www.opengeospatial.org/standards/gml Accessed September, 13th 2008)

[45] Extensible Markup Language (XML). Word Wide Web Consortium website. (URL: http://www.w3.org/XML/, accessed November, 3rd 2008)

[46] Google Webmasters/Site owners Help (URL: http://www.google.com/support/webmasters/bin/answer.py?answer=66359, accessed October, 12nd2008)

[47] Google Maps / Google Earth Terms of Use (URL: http://code.google.com/intl/es-ES/apis/maps/terms.html. accessed December, 15th January 2009)

[48] Google Issue Tracker. Google Code website. (URL: http://code.google.com/p/gmaps-api-issues/issues/detail?id=852, accessed December, 15th 2008

[49] Google Maps API Group discussion thread (URL: http://groups.google.com/group/Google-Maps-API/browse_thread/thread/3b0bd5922c7115f0/39bff5518d9a96a4, accessed December, 15th 2008)

[50] Google Maps API Group discussion thread .http://groups.google.com/group/Google-Maps-API/browse_thread/thread/3ec81216566a3e16#, accessed December, 15th 2008)

[51] Google Maps API Group discussion thread. (URL: http://groups.google.com/group/Google-Maps-API/browse_thread/thread/d4956d6126bd3b01#, accessed December, 15th 2008)

[52] Use of Google Maps by the Ordnance Survey. (URL: http://www.freeourdata.org.uk/docs/use-of-google-maps-for-display-and-promotion.pdf, accessed January, 10th 2009)

Annex I

1. <ExtendedData> and <*Data*> elements usage example.

```
<ExtendedData>

        <!-- Name & Title -->
        <Data name="Title">
                <displayName>Title</displayName>
                <value>Title</value>
        </Data>
        <Data name="Name">
                <displayName>Name</displayName>
                <value>Name</value>
        </Data>

        <!-- Abstract & Description -->

        <Data name="Abstract">
                <displayName>Abstract</displayName>
                <value>Abstract</value>
        </Data>
        <Data name="Description">
                <displayName>Description</displayName>
                <value>Description</value>
        </Data>

        <!-- Topic category -->
        <Data name="Topic_Category">
                <displayName>Topic Category</displayName>
                <value> Environment </value>
        </Data>
        <Data name="Category">
                <displayName> Category </displayName>
                <value> Environment </value>
        </Data>
        <Data name="Topic">
                <displayName> Topic </displayName>
                <value> Environment </value>
        </Data>

        <!-- Authoring and Contact Data -->
        <Data name="Contact">
                <displayName>Contact</displayName>
                <value>John Doe, Universitat Jaume I, Castellón de la Plana,
                Castellón, Spain</value>
        </Data>
        <Data name="Author">
                <displayName>Author</displayName>
                <value>Joh Doe</value>
        </Data>
        <Data name="Author_name">
                <displayName>Author name</displayName>
                <value>Joh Doe</value>
        </Data>
        <Data name="Address">
```

```
        <displayName>Address</displayName>
        <value>Universitat Jaume I, Castellón de la Plana, Castellón, Spain</value>
</Data>
<Data name="Telephone">
        <displayName>Telephone</displayName>
        <value>555-555-555</value>
</Data>
<Data name="Link">
        <displayName>Link</displayName>
        <value>http://www.geoinfo.uji.es</value>
</Data>


<!-- Date Data -->
<Data name="Date">
        <displayName>Date</displayName>
        <value>20070514</value>
</Data>
<Data name="a-date">
        <displayName>a-date</displayName>
        <value>20070514</value>
</Data>

<!-- Language Data -->
<Data name="Language_1">
        <displayName>Language</displayName>
        <value> EN </value>
</Data>
<Data name="Language_2">
        <displayName>Language</displayName>
        <value> Inglés </value>
</Data>
<Data name="Language_3">
        <displayName>Language</displayName>
        <value> English </value>
</Data>

<!-- Spatial Extent/BBOX Data -->
<Data name="Spatial_Extent">
        <displayName>Spatial Extent</displayName>
        <value> 70.3 , 27.4 , 35 , -28.2 </value>
</Data>
<Data name="Spatial_Extent_North">
        <displayName>Spatial Extent North</displayName>
        <value> 70.3 </value>
</Data>
<Data name="Spatial_Extent_South">
        <displayName>Spatial Extent South</displayName>
        <value>27.4 </value>
</Data>
<Data name="Spatial_Extent _East">
        <displayName>Spatial Extent East</displayName>
        <value>35</value>
</Data>
<Data name="Spatial_Extent_West">
        <displayName>Spatial Extent West</displayName>
        <value>-28.2</value>
</Data>
```

```
</ExtendedData>
```

2. <ExtendedData> and <*Schema*> definition example.

```
<Schema name="myMetadataSchema" id="myMetadataSchema">
        <SimpleField type="xsd:string" name="Title">
                <displayName>Title</displayName>
        </SimpleField>
        <SimpleField type="xsd:string" name="Name">
                <displayName>Name</displayName>
        </SimpleField>
        <SimpleField type="xsd:string" name="Abstract">
                <displayName>Abstract</displayName>
        </SimpleField>
        <SimpleField type="xsd:string" name="Description">
                <displayName>Description</displayName>
        </SimpleField>
        <SimpleField type="xsd:string" name="Topic_Category">
                <displayName>Topic_Category</displayName>
        </SimpleField>
        <SimpleField type="xsd:string" name="Category">
                <displayName>Category</displayName>
        </SimpleField>
        <SimpleField type="xsd:string" name="Topic">
                <displayName>Topic</displayName>
        </SimpleField>
        <SimpleField type="xsd:string" name="Contact">
                <displayName>Contact</displayName>
        </SimpleField>
        <SimpleField type="xsd:string" name="Author">
                <displayName>Author</displayName>
        </SimpleField>
        <SimpleField type="xsd:string" name="Author_name">
                <displayName>Author_name</displayName>
        </SimpleField>
        <SimpleField type="xsd:string" name="Address">
                <displayName>Address</displayName>
        </SimpleField>
        <SimpleField type="xsd:string" name="Telephone">
                <displayName>Telephone</displayName>
        </SimpleField>
        <SimpleField type="xsd:string" name="Link">
                <displayName>Link</displayName>
        </SimpleField>
        <SimpleField type="xsd:string" name="Date">
                <displayName>Date</displayName>
        </SimpleField>
        <SimpleField type="xsd:string" name="a-date">
                <displayName>a-date</displayName>
        </SimpleField>
        <SimpleField type="xsd:string" name="Language_1">
                <displayName>Language</displayName>
        </SimpleField>
        <SimpleField type="xsd:string" name="Language_2">
                <displayName>Language</displayName>
        </SimpleField>
        <SimpleField type="xsd:string" name="Language_3">
```

```
            <displayName>Language</displayName>
        </SimpleField>
        <SimpleField type="xsd:string" name="Spatial_Extent">
            <displayName>Spatial Extent</displayName>
        </SimpleField>
        <SimpleField type="xsd:double" name="Spatial_Extent_North">
            <displayName>Spatial Extent North</displayName>
        </SimpleField>
        <SimpleField type="xsd:double" name="Spatial_Extent_South">
            <displayName>Spatial Extent South</displayName>
        </SimpleField>
        <SimpleField type="xsd:double" name="Spatial_Extent_East">
            <displayName>Spatial Extent East</displayName>
        </SimpleField>
        <SimpleField type="xsd:double" name="Spatial_Extent_West">
            <displayName>Spatial Extent West</displayName>
        </SimpleField>
</Schema>
```

3. <ExtendedData> and <SimpleData> element example.

```
<ExtendedData>
    <SchemaData schemaUrl="#myMetadataSchema">
        <SimpleData name="Title">Forest Focus Level 1 Database: Soil
        Chemistry</SimpleData>
        <SimpleData name="Name">Forest Focus Level 1 Database: Soil
        Chemistry</SimpleData>
        <SimpleData name="Abstract">Abstract</SimpleData>
        <SimpleData name="Description">Description</SimpleData>
        <SimpleData name="Topic_Category">Environment </SimpleData>
        <SimpleData name="Category">Environment </SimpleData>
        <SimpleData name="Topic">Environment</SimpleData>
        <SimpleData name="Contact">John Doe, Universitat Jaume I,
        Castellón, Spain</SimpleData>
        <SimpleData name="Author">John Doe</SimpleData>
        <SimpleData name="Author_name">John Doe</SimpleData>
        <SimpleData name="Address">Universitat Jaume I, Castellón,
        Spain</SimpleData>
        <SimpleData name="Telephone">555-555-555</SimpleData>
        <SimpleData
        name="Link">http://www.geoinfo.uji.es/kml</SimpleData>
        <SimpleData name="Date">20070514</SimpleData>
        <SimpleData name="a-date">20070513</SimpleData>
        <SimpleData name="Language_1">EN</SimpleData>
        <SimpleData name="Language_2">Inglés</SimpleData>
        <SimpleData name="Language_3">English</SimpleData>
        <SimpleData name="Spatial_Extent"> 70.3 , 27.4 , 35 , -28.2
        </SimpleData>
        <SimpleData name="Spatial_Extent_North">70.3</SimpleData>
        <SimpleData name="Spatial_Extent_South">27.4</SimpleData>
        <SimpleData name="Spatial_Extent_East">35</SimpleData>
        <SimpleData name="Spatial_Extent_West">-28.2</SimpleData>
    </SchemaData>
</ExtendedData>
```

4. <ExtendedData> using an external schema example based on ISO19115

```
<ExtendedData
xmlns:meta="http://www.geoinfo.uji.es/kml/schemas/iso19115/schema.xsd">
        <meta:Metadata>
                <meta:mdFileID>Forest Focus Level 1 Database: Soil
                Chemistry</meta:mdFileID>
                <meta:mdLang>
                        <meta:languageCode value="es"/>
                </meta:mdLang>
                <meta:mdChar>
                        <meta:CharSetCd value="utf8"/>
                </meta:mdChar>
                <meta:mdContact>
                        <meta:rpIndName>John Doe</meta:rpIndName>
                        <meta:rpOrgName>Universitat Jaume I</meta:rpOrgName>
                        <meta:rpPosName>Becario</meta:rpPosName>
                        <meta:rpCntInfo>
                                <meta:cntPhone>
                                        <meta:voiceNum>555-555-
                                        555</meta:voiceNum>
                                        <meta:faxNum>555-555-
                                        555</meta:faxNum>
                                </meta:cntPhone>
                                <meta:cntAddress>
                                        <meta:delPoint>Universitat Jaume I,
                                        Castellón de la Plana, Castellón,
                                        Spain</meta:delPoint>
                                        <meta:city>Castellón</meta:city>
                                        <meta:adminArea>Castellón</meta:adminAr
                                        ea>
                                        <meta:postCode>12004</meta:postCode>
                                        <meta:country>es</meta:country>
                                        <meta:eMailAdd>john.doe@uji.es</meta:eM
                                        ailAdd>
                                </meta:cntAddress>
                                <meta:cntOnLineRes>
                                <meta:linkage>http://www.geoinfo.uji.es/kml/files/for
                                est_soil_chemistry_iso_all.kml</meta:linkage>
                                <meta:protocol>Protocolo</meta:protocol>
                                <meta:appProfile>Perfil de
                                aplicación</meta:appProfile>
                                        <meta:orName>http://www.geoinfo.u
                                        ji.es/kml/files/forest_soil_chemistry_i
                                        so_all.kml</meta:orName>
                                <meta:orDesc>Description </meta:orDesc>
                                <meta:orFunct>
                                        <meta:OnFunctCd value="search "/>
                                </meta:orFunct>
                                </meta:cntOnLineRes>
                                <meta:cntHours>Horario de atención
                                </meta:cntHours>
                                <meta:cntInstr>Instrucciones para contacto
                                </meta:cntInstr>
                        </meta:rpCntInfo>
                        <meta:role>
                                <meta:RoleCd value="user"/>
                        </meta:role>
```

```
                </meta:mdContact>
                <meta:mdDateSt>2008-10-09 T 01:40:25</meta:mdDateSt>
                <meta:mdStanName>ISO 19115 Core</meta:mdStanName>
                <meta:mdStanVer>FDIS</meta:mdStanVer>
                <meta:distInfo/>
                <meta:dataIdInfo>
                        <meta:idCitation>
                                <meta:resTitle>Forest Focus Level 1 Database: Soil
                                Chemistry</meta:resTitle>
                                <meta:resRefDate>
                                        <meta:refDate>2008-10-09                T
                                        01:40:25</meta:refDate>
                                        <meta:refDateType>
                                                <meta:DateTypCd
                                                value="publication"/>
                                        </meta:refDateType>
                                </meta:resRefDate>
                        </meta:idCitation>
                        <meta:idAbs>Abstract</meta:idAbs>
                        <meta:dataLang>
                                <meta:languageCode value="en"/>
                        </meta:dataLang>
                        <meta:dataChar>
                                <meta:CharSetCd value="utf8"/>
                        </meta:dataChar>
                        <meta:tpCat>
                                <meta:TopicCatCd value="environment"/>
                        </meta:tpCat>
                        <meta:geoBox>
                                <meta:westBL>-28.2</meta:westBL>
                                <meta:eastBL>35</meta:eastBL>
                                <meta:southBL>27.4</meta:southBL>
                                <meta:northBL>70.3</meta:northBL>
                        </meta:geoBox>
                        <meta:dataExt>
                                <meta:tempEle>
                                        <meta:TempExtent>
                                                <meta:exTemp>
                                                <meta:TM_GeometricPrimitive>
                                                <meta:TM_Period>
                                                <meta:begin>2008-10-09          T
                                                01:40:25</meta:begin>
                                                <meta:end/>
                                                </meta:TM_Period>
                                                </meta:TM_GeometricPrimitive>
                                                </meta:exTemp>
                                        </meta:TempExtent>
                                </meta:tempEle>
                        </meta:dataExt>
                </meta:dataIdInfo>
                <meta:dqInfo/>
                <meta:refSysInfo/>
        </meta:Metadata>
</ExtendedData>
```