GeoDiscover – a specialized search engine to discover geospatial data in the Web

Fernando Renier Gibotti[1,2], Gilberto Câmara[2], Renato Almeida Nogueira[1]

[1]Information Systems Program - Mirassol College – Mirassol – SP – Brazil

[2]Image Processing Division, National Institute of Space Research - INPE, São José dos Campos, SP, Brazil

{gibotti,gilberto}@dpi.inpe.br, renato@faimi.br

Abstract

The fast development of the internet and the growth of digital contents available led to the development of search engines that facilitate the recovery of information in the Web. However, these engines have limitations mainly when recovering specialized contents. Geospatial data are created by local governments, companies and people, but these data aren't available in a systematized way. In this context, this paper presents a specialized search engine to access and recover geospatial data in the Web, focusing on its main characteristics, architecture, technologies and performance.

1 Introduction

The world is an imperfect and unpredictable place. From its origin in the late sixties, the Internet grew rapidly and metamorphosed from a research project into a vast collection of heterogeneous documents. The web is a network composed of billions of interconnected multimedia pages (images, sounds, texts, animations, files, etc) developed in an uncoordinated way by millions of people. The fast increase in the published information and lack of structuring of the content of the documents are challenges for data retrieval.

Search engines such as Google and Yahoo aim at organizing information retrieval on the Web [Glover 2002]. Nevertheless, these engines are limited. Limits range from the amount of the indexed documents to the quantity of unwanted results, making the access to information difficult. To overcome this problem, there is a large effort to improve the quality of semantic information included in web pages. Following Berners-Lee seminal paper "The Semantic Web" [Berners-Lee, Hendler and Lassila 2001] and a renewed interest in ontologies [Gruber 1995] [Guarino and Giaretta 1995] [Wiederhold 1994], the IT community is working on proposals to organize information for easier retrieval and interoperability. This includes languages such as OWL [Masolo, Gangemi, Guarino et al. 2002]. This geospatial community is also part of this effort [Egenhofer 2002] and proposals such as OGC's GML language aim at providing standard ways for spatial data interoperability.

However, using languages such as OWL or GML requires a major effort by the data producers. Producing and organizing ontologies is a demanding task, which may prove to be beyond the skills and means of many individuals and institutions. An alternative to standardized data distributions is to provide specialized search engines. These engines know that some communities produce structured documents and that by informed guesswork a lot of semantic information can be directly retrieved. The most recent examples are engines for searching and indexing scientific papers, such as Google Scholar and Citeseer [Giles, Bollacker and Lawrence 1998]. These engines know that a scientific paper has a well-defined structure: *title, author list, abstract, text and references*. These engines provide extremely useful information, and the only effort done by the paper authors is posting the original paper on a webpage.

What about geographical data? The Web has a large quantity of geospatial data, but the traditional search engines are not specialized in recognizing them, setting up a gap between this data and the users. However, geographical data is semi-structured. Most data files share common features: they provide local data (in vector or raster format) and the matching attributes. Besides, the number of formats for distribution of geospatial data is limited. Similarly to what happens in scientific papers (where the PDF format is prevalent), geospatial data producers usually distribute their data in formats they expect the user to read easily. This includes ESRI's shapefiles, GML and raster data in GeoTIFF.

Therefore, an alternative to semantic standardization proposals, such as GML, is a specialized search engine for geospatial data. These specialized engines would consider the semi-structured nature of geographical data to make guests informed. The idea is to allow sharing of geographical data without the need for extra work on semantic annotation. Some of the principles for this type of data sharing are outlined in Onsrud et al. [2004]. Based on this motivation, this paper presents a specialized search engine for retrieval of unstructured geospatial data. Our propose includes a distributed and fast crawling technology to gather the geospatial data in the web and maintain them up to date.

In this paper, we discuss the conceptual challenges involved in designing a search engine for geographical data. The main challenges are tree: (a) Developing algorithms for information retrieval from geographical data; (b) Designing a robust architecture to deal with large data volumes; (c) Including techniques for privacy enforcement and copyright restrictions. We discuss the issue of developing algorithms for geospatial information retrieval in Section 2. In Section 3, we discuss the architecture of the search engine. In Section 4, we describe the system operation and present initial results. The issue of privacy and copyrights has been discussed in Onsrud et al. [2004] and will not be further developed in this paper.

2 Retrieving geographical data

Traditional search engines don't have specialized parsers and crawlers looking for geospatial data. Their crawlers search for hypertext documents and hyperlinks while GeoDiscover crawlers retrieve geospatial data. This section discusses the main problems involved in designing geospatial web services.

2.1 Finding spatial data

The first obvious question is how to find spatial data. In an imperfect world, we have to use approximate meanings. Our choice was to consider that geospatial data is usually distributed in a set of predefined formats associated to GIS systems. As an example, the shapefile format used in ArcView software from ESRI® is a well-established way of exchanging geospatial data. Thus, GeoDiscover tries to find typical GIS files. In our current prototype, we are targeting shapefiles, but the searcher can be easily extended to crawl other GIS formats.

2.2 Making sense of unstructured data

What is special about spatial data? We consider that most geospatial data has information that enables educated guests about its semantic content. The most obvious is the geographical coordinates. For example, shapefiles contain information about the bounding box of the data, but no specific data in projection and datum. Usually, we can infer such information. Coordinates in lat/long projection have a different range than UTM coordinates. Datum information is more difficult. In this case, Geodiscover will assign a best guess.

The second type of content that allows educated guess is *place-names*. Most geospatial data sets include place-names, either directly or by indexation (such as FIPSNO in USA or IBGE reference in Brazil). The columns for place-names are usually associated with tags such as NAME, NOME, or NAMEN. Once a column with potential place-names is found, we can compare its contents to gazetteers such as USGS's Geographic Names Information System (GNIS) (http://geonames.usgs.gov) or the NGA's GEOnet Names Server (http://earth-info.nga.mil/gns/html/). The gazetteers are used effectively in Alexandria Digital Library [Frew 1998].

The two types of information above (projection and place) are enough for a general-purpose geospatial index. For more details, a second type of guesswork is needed. The crawler needs to explore the column labels of the data and to deduce its contents by using information such as the range of the variable and its values. For example, positive integers in geospatial data are likely to be count data. Real values between -1 and +1 are possible indices. We refer to this process of discovery as wayfinding in data. In its current version, GeoDiscover does not attempt to wayfinding in data, but we hope to include such skills in the future.

2.3 Overlapping data

The existence of copies of geospatial data in different web sites is a problem to the indexation because it can induce to duplications in the database. The detection of similar data is possible if we analyze their structure and some intrinsic attributes of the files, such as name, creation date, size and type. Using this intrinsic attributes, GeoDiscover identifies similar files and avoids the storage of redundant ones.

2.4 Description of ranking of producers

Conventional search engines use different ways to classify and rank the main websites. Google utilizes the PageRank method to prioritize the results of web keyword searches [Page, Brin, Motwani et al. 1999] [Gerhart 2002]. Similar to the method to classify web pages using link information, the Citeseer [Giles, Bollacker and Lawrence 1998] classifies scientific articles as hubs and authorities based on citation graphics.

The objective of search engine is to make possible that people recover geospatial data in a safe way regarding the quality and origin. GeoDiscover identifies producers based on three aspects: quantity of available geospatial data in the web site; quantity of downloads requested by users to any data in that web site; and by indication if that web site is a hub or an authority. In this approach, hubs are web sites that recommend other web sites that contain geospatial data, and authorities are web sites that are recommended by several hubs.

When GeoDiscover returns a list of geospatial data to the user, for each data it demonstrates its origin (e.g. IBGE), the quantity of geospatial data that the producer has (e.g. 1080 shape files), and the quantity of request downloads of that producer (e.g. 10112 files downloaded).

3 System characteristics

The GeoDiscover has some features that make difference in the process of search geospatial information coming up with good results. First it classifies producers by quantity of produced geospatial data and by quantity of downloads requested by users. Other important feature is that GeoDiscover uses the additional information provided by hyperlinks to describe geospatial data. Finally it identifies overlapping data available in different websites.

3.1 Anchor text and extended anchor text

Usually the file attributes (name, creation, modification, and last access date, size and type) don't expose their contents and detailed descriptions. The lack of additional information makes the indexation of these files a difficult task and sometimes the obtained results are not desired ones.

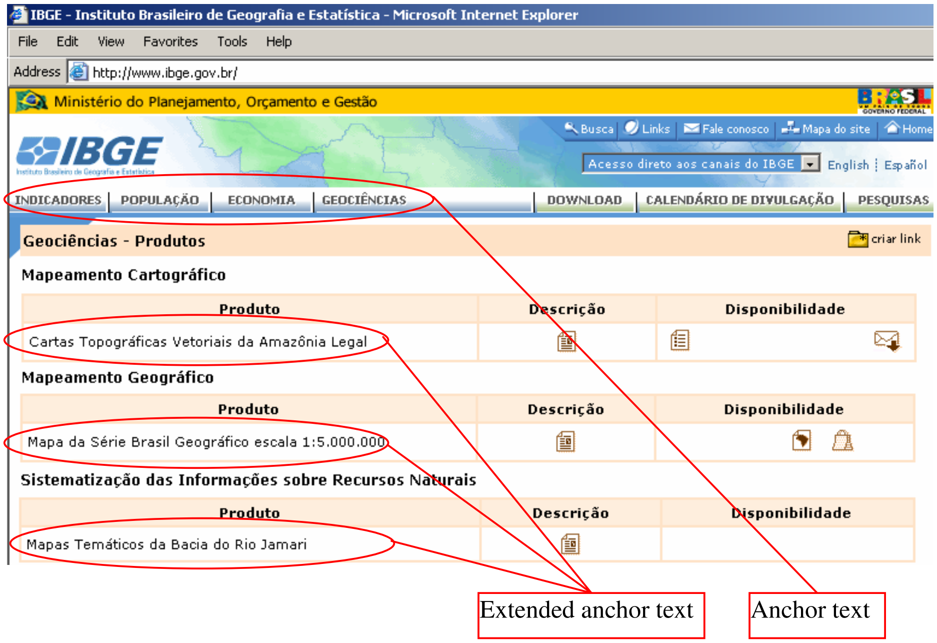
A web site can be composed by multimedia such as sounds, images, texts, files and by connections to another web sites or pages – hyperlinks. The structure created by these connections is being researched and used to improve the crawlers [Cho, Garcia-Molina and Page 1998] and the page classification process of the search engines. This has been done in order to discover web communities and organize research results in hubs and authorities. A hyperlink contains the URL of the page to which it refers and an associated anchor text that describes the link. The anchor texts can offer excellent page descriptions to which it refers. These anchor texts can be useful to describe and help in the content recovery of not indexed pages by traditional search engines, containing elements such as images, databases, and geospatial data.

The idea of utilizing anchor text was initially implemented in the World Wide Web Worm [Mcbryan 1994] especially because it helps search for non-text information.

The anchor text permits to connect words (and context) to a specific content (e.g. Click here to download the map of São José dos Campos town and neighborhood).

GeoDiscover uses the anchor text concept to help in description of context and in search results. In order to improve the search results GeoDiscover uses extended anchor text. In this case, besides the link text, the words and phrases near the links are taken into consideration to classify data with better accuracy (as illustrated by the figure 1). Thereby we can obtain additional information to describe geospatial data and, therefore increase the power of search.

Due to the size of the geospatial data files, normally these files are available in the Web in compressed formats such as zip, arj, rar files and others. These formats are not the focus of the Geodiscover crawlers. In this case, the anchor text has another important function: helping the crawlers to find compacted geospatial data files analyzing the context of web pages and recognizing the files.



Geociências - Produtos

Mapeamento Cartográfico

Produto	Descrição	Disponibilidade
Cartas Topográficas Vetoriais da Amazônia Legal		

Mapeamento Geográfico

Produto	Descrição	Disponibilidade
Mapa da Série Brasil Geográfico escala 1:5.000.000		

Sistematização das Informações sobre Recursos Naturais

Produto	Descrição	Disponibilidade
Mapas Temáticos da Bacia do Rio Jamari		

Figure 1. Anchor text and extended anchor text

3.2 Distributed processing

The question of efficiency is reflected in GeoDiscover model of distributed processing. In order to support a consistent database query, differently from all the existent search engines, GeoDiscover bases on the distributed processing with an application server centralized at web services. This server is responsible for managing the client request. Today the main search engines have approximately eight billion of indexed pages in its database, thereby the extraction, analysis and indexation process of the web sites are slower. Another problem is the time of revisiting that happens between thirty to sixty days, making the database not updated.

Using a distributed processing with collaborative users, these processes can be more efficient, faster and cheaper.

3.3 Collaborative user

The idea of open source systems organized communities (people and organizations) that have similar interests. These communities work intensively to construct new operational systems, programming languages and computer applications and make them available. The growth of this philosophy is notable and increasing daily. GeoDiscover shares this philosophy since it adopts collaborative users.

Collaborative users are clients that take part in the processing when their computers are idle. They contribute to Geodiscover crawling the web and executing the parser to find clues of geospatial data. In order to collaborate with the GeoDiscover, the user needs to download and install the software that manages tasks related to the search engine.

The main advantages of working with collaborative users are the distributed processing that makes the empowered servers unnecessary to execute functions of crawling and parsing; the reduction of investments to maintain the project; and the capability of a large growth as new users collaborate with the project.

3.4 Other characteristics

Other important feature of GeoDiscover is that its crawlers respect the crawlers' ethic code. So it avoids a specific web site to be crawled. GeoDiscover observes some meta-tags such as <*meta name* = *"robots" content* = *"noindex, nofollow"*>. This meta-tag is included in the header of the pages and the value *robots* can be changed by the name of the specific robot (e.g. *google* or *geodiscover*) that respects this tag. The value *noindex* determines to the robots that the page is not indexed. The value *nofollow* determines that the links eventually existing in the page are not followed. Any arrangement of values *index/noindex, follow/nofollow* is permitted.

4 System anatomy

The GeoDiscover is implemented in C# and the database server runs SQL Server 2000. There are two main architectures: GeoDiscover Servers and GeoDiscover Clients. The basic roles of the first are: managing the distribution of URLs, receiving, organizing and storing the captured data and files, and making the interface available so the user can perform a web query and visualize results. For the second, executed in a collaborative user's computer, the roles are: requesting a list of URLs that will be visited, crawling the web sites, downloading pages, parsing these pages, extracting the interesting data and sending the found content to the server. Figure 2 shows the system operation.

The data traffic between clients and server is performed through XML (eXtended Markup Language) and Web Services to improve the use of the application by computers protected by firewall and proxy.

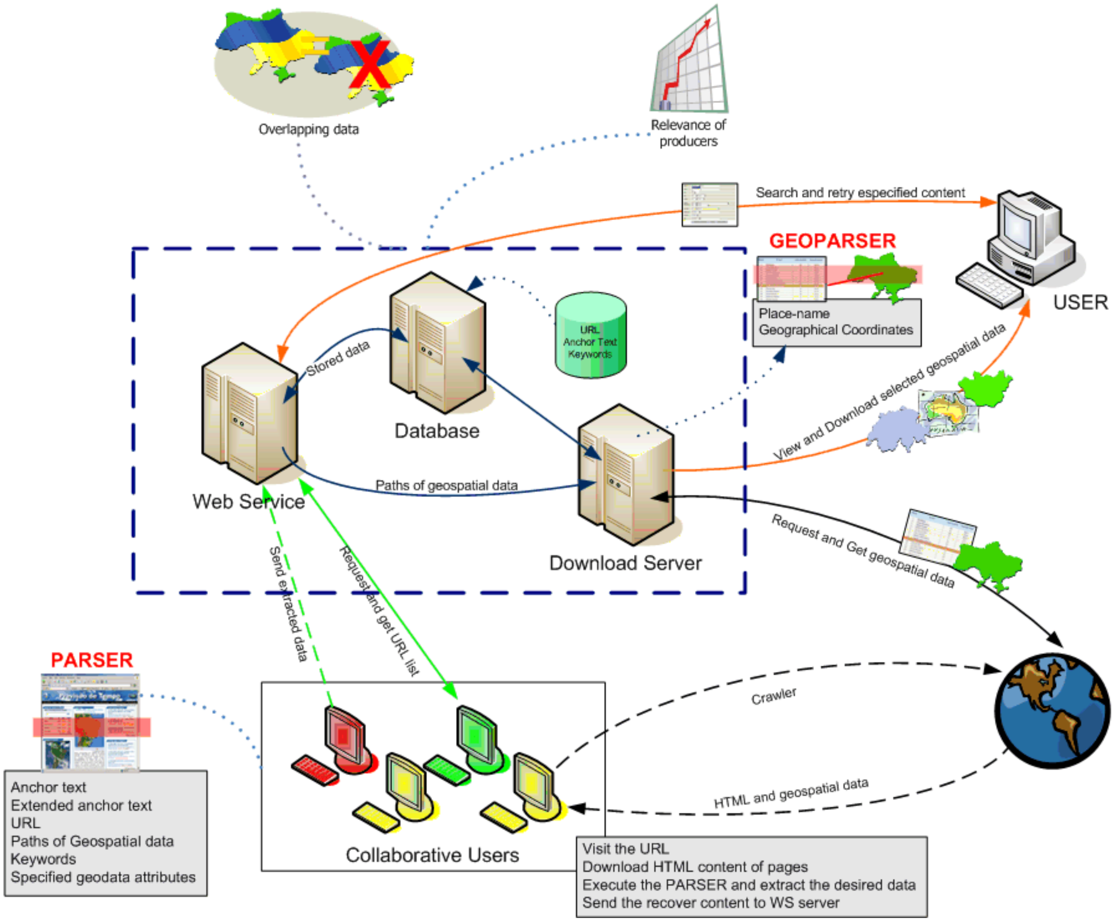


Figure 2. GeoDiscover system operation

The process of discovering geospatial data is started when a collaborative user's computer is idle. The GeoDiscover Client requests a list of URLs to be crawled. The Web Service (WS) server sends an ordered URL list to the client. The client uses the web crawler to seek the indicated web sites and recover the HTML content of the visited pages. Soon after, it parsers this content looking for geospatial data, extracting the desired data (URLs, paths of geospatial data, keywords, anchor text and extended anchor text), and sending these data to the WS Server which stores the organized data in the database (DB) server. Before being stored, the relative URLs are converted to absolute URLs.

New URLs are ordered to be visited hereafter. The indexing function is performed by the indexer. The indexer performs a number of functions. It reads and parses the repository. It parses out all the links in every web page and stores important information about them in the anchors files. This file contains enough information to determine where each link points from and to, and the text of the link.

The BD server is linked to the download server by an internal network. The download server monitors the included paths in the BD server, and it starts the download for each inclusion. The download server seeks the file in its origin local and stores it in a directory. Then, the file is compressed to optimize the storage space.

GeoDiscover uses the class *GZipStream* available in .NET Framework 2.0 to compress the downloaded geospatial files through its constructors. This class performs the compression and decompression of files to optimize the large quantity of data in the download server.

The geoparser is executed in the geospatial data to extract important information that describes the data. Today the geoparser is implemented to look for projection and place-name in the columns of the DBF tables. This information is ordered and stored in a repository of the DB server, and later it is used in the search process. Additional information present in the columns can be stored too.

4.1 Crawler

The crawler is responsible for visiting the stored addresses in Geodiscover DB server. The first task executed by the crawler is the verification of the file *robots.txt* in the server root directory to certify if there is any rule to the indexation of the web site. Detailed information about the file *robots* is available in http://www.robotstxt.org/wc/norobots.html.

After the crawler executes this task, it returns to the present state of the page and, if it is available to be visited, a copy of its content is made using a HTTP protocol to download. The crawler utilizes classes that return a stream of bytes of received data. These data are converted to ASCII characters and the page content is reconstructed so the parser can make the extraction process.

Crawling is a complex application because it demands interacting with thousands of web servers and several name servers that extrapolates the control of the system [Brin and Page 1998]. In order to visit the millions of web pages, Geodiscover has a fast distributed crawling system that runs in collaborative user computers. At the moment each crawler can visit sixty thousand web pages per day. This is an excellent number since the Geodiscover can involve hundreds of collaborative users increasing considerably the covering of available pages in the Web.

4.2 Parser

The parser is responsible for extracting the whole content of the pages. In these pages we can find the referred addresses that will be stored in a database and later they will be visited by crawlers. As soon as the crawler executes the download of a web page, all characters are converted to lower case. The parser seeks meta-tag that prohibits the indexation of the web page. After the HTML content is analyzed in order to extract some relevant information for its classification. Among this content the following meta-tags are found: *<title></titile>* that describes the title of the page, *<META NAME="Description" Content="">* that describes the content of the page, *<META*

NAME="Keywords" Content=""> that stores keywords related to page content, <*a href=""*> that indicates links to other pages and geospatial data files.

The parser removes all HTML tags, scripts, and other markups maintaining the pure text. The table 1 shows the original HTML content and the table 2 shows the result after parsing. The extracted words are stored in a table of words in the DB server. For each word one hash code is generated to improve the speed in the search process.

Table 1. Original content with all tags and scripts

```
<!DOCTYPE HTML PUBLIC “-//W3C//DTD HTML 4.0 Transitional//EM”>
<html>
   <head>
      <META NAME="KEYWORDS" CONTENT="Maps, thematic, environment, population,
        health, education, transport, GIS, SIG.">
      <META NAME="DESCRIPTION" CONTENT="Thematic maps of natural resources,
        environment, population, health, education and transport.">
      <meta name="GENERATOR" content="Microsoft FrontPage 5.0">
      <title>GIS Data - Education</title>
   </head>
   <SCRIPT>
      function link(end)
         {
          window.open(end, "nova", "height=500, width=720, left=50, top=0");
         }
      </SCRIPT>
   <P align=center><FONT face=Arial color=#000000 size=2>
      <a href="http://www.gisconsult.com.br/"></a>
   </font></p>
   <font face="Arial, Helvetica, sans-serif" color="#000080"> Click here to
      <a href="DownloadMaps/Education.shp">download </a> the Education map.
   </font>
</html>
```

Table 2. Extracted content after parser.

```
Maps, thematic, environment, population, health, education, transport, GIS, SIG.

Thematic maps of natural resources, environment, population, health, education and
transport.

GIS Data - Education

http://www.gisdata.com.br/

Clique here to http://www.gisdata.com.br/DownloadMaps/Education.shp download the
Education map.
```

4.3 Address Storage

The address storage is responsible for storing the addresses extracted by the parser in the DB server and verifying if these addresses have already been input. The addresses are used by the WS server to distribute the URLs that will be visited and analyzed by the collaborative users.

4.4 Searching

The search process is focused on quality of obtained results in GeoDiscover research. The interface is user friendly and runs in the user browser.

For the search execution, the WS server verifies the keywords related to those informed by user in the DB server. When WS server finds the results that satisfy the user query, it returns to the user a list of files that attend the content of his seek. The rank of files in the list is calculated considering the proximity of query terms with the terms found in the DB server and by the producer relevance.

For each item in the returned list of files additional information is provided such as producer relevance, quantity of geospatial data of this producer, quantity of downloads of the file. This information helps the user know better the geospatial data before starting the download.

The steps of the search are:

1. The user accesses the GeoDiscover interface and types in the desired search terms.
2. WS server parses the query.
3. The DB server is scanned through until there is a geospatial data file that matches all the searched terms.
4. The classification of that geospatial data file for the query is computed.
5. The geospatial files that have been found by classification are ordered.
6. WS server sends a list of classified geospatial data files and additional information to the user's interface.

The download of the desired geospatial data file can be executed through GeoDiscover download server or through its original site. GeoDiscover makes these two options possible to guarantee the download, because if the original site is not accessible, it is possible through the download server.

4.5 Results and performance

The quality of search results is the main measurement of a search engine. GeoDiscover produces excellent results when it manipulates geospatial data. Starting from an only URL registered in their database, GeoDiscover obtains the page, gets new URL and makes the visits in an interrupted and infinite process. In preliminary tests, GeoDiscover processed 60.000 pages by day using computers with 2.0 gigahertz processor, 512 megabytes of RAM and an internet connection of 128 megabits.

5 Conclusions

The Geodiscover is designed to make the available geospatial data in the web more accessible and usable. It was implemented in a distributed environment using collaborative users. The collaborative users are very interesting since they help in the crawling and parsing processes increasing significantly the coverage of the web. This scenario has a lot of advantages: it makes the empowered servers unnecessary to execute functions of crawling and parsing; reduction of investments to maintain the project; and the capability of a large growth as news users collaborate with the project.

In order to provide high quality in search results, Geodiscover uses techniques such as specialized crawler and parser to geospatial data, overlapping data, description of ranking of producers, anchor text and extended anchor text. It is a complete search engine for collecting, indexing, and performing search queries on geospatial data.

Some functions are being developed to complete the Geodiscover system, among them the gazetteers to compare the place-names; to extend the crawler to gather other GIS formats (Geotiff and SPR files); to include ontology to improve the results returned to the user; to implement a user-friendly visualization system to the user.

6 References

Berners-Lee, T., J. Hendler and O. Lassila (2001). "The Semantic Web." Scientific American **May**.

Brin, S. and L. Page (1998). The anatomy of a large-scale hypertextual Web search engine. Seventh International World Wide Web Conference, Brisbane, Australia.

Cho, J., H. Garcia-Molina and L. Page (1998). Efficient Crawling Through URL Ordering. Seventh International Web Conference (WWW 98). Brisbane, Australia.

Egenhofer, M. (2002). Toward the semantic geospatial web. 10th ACM international symposium on Advances in geographic information systems table of contents, McLean, Virginia, USA.

Frew, J. (1998). "The Alexandria Digital Library Architecture." International Journal on Digital Libraries **2**(4): 259-268.

Gerhart, A. (2002). "Understanding and Building Google PageRank." from http://www.searchengineguide.com/orbidex/2002/0207_orb1.html

Giles, C. L., K. Bollacker and S. Lawrence (1998). CiteSeer: An automatic citation indexing system. Digital Libraries 98 - The Third ACM Conference on Digital Libraries, Pittsburgh, PA, ACM Press.

Glover, E. J. T., K.; Lawrence, S.; Pennock, D.; Flake, G. W (2002). Using Web Structure for Classifying and Describing Web Pages. WWW2002, Honolulu, Hawaii, USA.

Gruber, T. R. (1995). "Toward Principles for the Design of Ontologies Used for Knowledge Sharing." Int. Journal of Human-Computer Studies **43**: 907-928.

Guarino, N. and P. Giaretta (1995). Ontologies and Knowledge Bases: Towards a Terminological Clarification. Towards Very Large Knowledge Bases: Knowledge Building and Knowledge Sharing 1995, Amsterdam, IOS Press.

Masolo, C., A. Gangemi, N. Guarino, et al. (2002). The WonderWeb Library of Foundational Ontologies. Padova, LADSEB-Cnr**:** 36.

Mcbryan, O. A. (1994). GENVL and WWWW: Tools for Taming the Web. First International Conference on the World Wide Web, Geneva, CERN.

Onsrud, H., G. Camara, J. Campbell, et al. (2004). Public Commons of Geographic Data: Research and Development Challenges. III International Conference on Geographical Information Science (GIScience 2004), Washington, Springer.

Page, L., S. Brin, R. Motwani, et al. (1999). "The PageRank Citation Ranking: Bringing Order to the Web." from <http://dbpubs.stanford.edu/pub/1999-66>.

Wiederhold, G. (1994). Interoperation, Mediation and Ontologies. International Symposium on Fifth Generation Computer Systems (FGCS94), Tokyo, Japan, ICOT.