Spatial data discovery using general purpose web search engines

Samy Katumba and Serena Coetzee

Centre for Geoinformation Science (CGIS), Department of Geography, Geoinformatics and Meteorology, University of Pretoria, Pretoria, South Africa

Abstract. To find spatial data, interested parties have to know where and how to access existing geospatial web services or geoportals. People in geographic information communities are aware of such online platforms, but what about others? Finding relevant spatial data that satisfy the needs of the requesters is still an issue. In this paper, we propose a method for documenting spatial data for both human and machine consumption. Through this method, a model for compiling spatial metadata based on the mapping between ISO 19115 spatial metadata standard and Dublin core is designed. The ultimate goal with this model is the documentation of spatial data in HTML for discovery by web search engines, but at the same time being understandable to users.

Keywords: spatial metadata, metadata, standards, spatial data discovery

1. Introduction

Geoportals, implemented as part of spatial data infrastructures (SDIs), allow the online dissemination of spatial data from providers. Geoportals provide web platforms to search for spatial data and associated metadata content. However, the discovery of geospatial services and spatial data remains a challenge (Lopez-Pellicer 2012). Business, legal and technological barriers result in the invisibility of geospatial web services to well-known general purpose web search engines. This hinders the visibility of geoportals and thus the discovery of spatial data. Furthermore, considerable amounts of online spatial data are of no value due to the lack of supporting information about them.

Attempts to solve the technological barriers related to the invisibility of geospatial web services have been channeled towards the development of fo-

cused crawlers with the sole mission of surfacing geospatial web contents (services and data) behind geoportals. Such efforts are centred on web crawlers' capabilities to search for geospatial web services, rather than on the ability of such resources making themselves visible to web search engines. In a number of studies, these focused crawlers use web pages listed by Bing, Google or Yahoo as seed for further search refinement (Lopez-Pellicer 2012).

The research reported in this paper is part of an experimental endeavour to empirically test the discoverability of HTML pages with information about geographic information resources by general purpose web search engines. It is motivated by current difficulties in finding spatial data with general purpose web search engines (Katumba et al 2012). The test results will lead to an improved understanding of how to prepare HTML documents (files) with spatial metadata about geographic information resources on the web. These documents are the indices for web search engines and serve as descriptions of geospatial data for users. This approach is used by web resource publishers for general purpose internet search. It has been most of the time anecdotally proven to be effective by professionals engaged in optimising web search engines and web resources' visibility. However, seldom scientific studies, such as the one by Zhang and Dimitroff (2005), empirically tested its effectiveness. Using this technique for spatial data discovery through associated spatial metadata published online in HTML format, makes the study reported in this paper distinct to others. The study aims to contribute towards an improved understanding of how spatial metadata should be embedded in HTML pages for better discovery of the corresponding spatial data by web search engines. In this paper, we describe how the results of a user study were used to develop a user-centered methodology for preparing the HTML pages for the first empirical tests.

The remainder of this paper is structured as follows: in Section 2, a survey of related studies is provided as context for the research. In Section 3, the methodology is described. In Section 4, initial results are presented and discussed. Section 5 concludes the paper.

2. Related studies

In this section, concepts related to web search engines, the deep web, metadata, search engine optimisation and web search behaviour are discussed. Emphasis is put on describing how these concepts are relevant in the context of this research.

2.1. Web search engines and the deep web

Web pages are added to web search engine databases either by crawling and indexing hypertext links or when information about web pages is submitted to the web search engine directories (Kumar et al. 2011). Web pages that are invisible to web search engines are said to be part of the "deep web" because their contents are hidden behind "query forms" or "web service interfaces" served by backend databases (Wu et al. 2006). Geoportals fall into this category because one has to enter query information, e.g. keywords in a text box, to retrieve relevant spatial contents from a database (Lopez-Pellicer 2012). It is impossible for web search engines to figure out all possible keywords for these text boxes, because web search engines are only designed to crawl and index existing web contents.

Ntoulas et al. (2005) proposed a web crawler that automatically queries hidden web page interfaces served by backend text databases. Madhavan et al. (2008) described the inclusion of web pages hidden behind HTML forms served by SQL databases into the Google search engine indexing process. In order to surface geospatial web services and associated contents, Lopez-Pellicer (2012) proposed the implementation of a crawler focused on the discovery of OGC web services.

The literature suggests that efforts towards solving the technical problems related to the invisibility of deep web contents in general and geospatial resources in particular, have been on the design and implementation of focused crawlers. Such attempts require considerable amounts of computational resources to match well established web search engines. Furthermore, anyone looking for spatial information has to know about the existence of such crawlers which presumably would be hosted inside geospatial web services or geoportals. Not knowing about the existence of geoportals in the first place, makes it impossible for anyone searching for spatial information to use these search engines (crawlers). Therefore, the use of general purpose web search engines is the optimal option for discovering geospatial web services and their associated contents.

2.2. Metadata

2.2.1. Dublin Core

Even though well-designed web pages, rich in content, are a requirement for high visibility, metadata content also facilitates the indexing and discovery of such web resources (Greenberg et al. 2001) (Zhang and Dimitroff 2005). Furthermore, the description of web resources is also provided through metadata content. Hence, a metadata standard for use by various different web publishers for web resources identification, description and discovery is needed. The Dublin Core metadata standard is used as refer-

ence description for online web resources (Dublin Core 2012). It has been considered in this study because of its simplicity and ease of use, its international scope, extensibility and most importantly for its generic terms for web resource description.

2.2.2. Spatial metadata (ISO 19115)

ISO 19115:2003, Geographic Information – Metadata, provides descriptions of geographic information and associated services with attributes for the identification, extent, quality, spatial and temporal schema, spatial reference and distribution of digital geographic data. ISO 19115:2003 was considered in this study due to the fact that most of the existing spatial metadata standards are profiles (specialisations) of it (Nogueras-Iso et al. 2004).

2.2.3. Mapping between ISO 19115 and Dublin Core

Since ISO 19115:2003 is a specialised metadata standard and Dublin Core is more generic and therefore suitable for describing web resources, a mapping between the two standards is imperative for the accomplishment of this research. In 2003, members of the European Committee for Standardisation Workshop Agreement (CWA) produced a consensus-based specification "CWA 14857:2003 (E)" that defines the mapping between Dublin Core and ISO 19115:2003. The principal aim was to enhance the discovery of geographic information in "cross metadata searches" (CEN 2003). Since then, its only usefulness has been the design and implementation of tools for automatic conversion from one standard to another (Nogueras-Iso et al. 2004). In this study we use this mapping to prepare web resources (HTML pages) with spatial metadata that are discoverable by general purpose web search engines. We also describe how spatial metadata is integrated into web resources (HTML pages) based on the CWA 1487:2003 (E) specification.

2.3. Search engine optimisation (SEO)

Since its inception, the World Wide Web has seen a growing number of web search engine users from a variety of disciplines, as well as the proliferation of web sites with web page contents changing constantly. As a result user satisfaction for web searches has deteriorated. This prompted information retrieval experts to research web search engine performance and users' web search behaviour. The performance of a web search engine is measured by the amount of relevant web sites it lists. It should not take users time and effort to get web sites of relevance in response to their queries. This is a concern for web publishers who would want their web sites to be visible on web search engine listings. Among many other considerations, a higher web page visibility in response to relevant user queries relies on SEO techniques

employed by web page designers and publishers. These techniques are meant to maximise the indexing of web sites (pages) by search engines. Elements that impact web page visibility are related to the metadata structure of the page, its content and the number of hyperlinks pointing to it (Zhang and Dimitroff 2005). The tuning of these elements through SEO contributes to the visibility of web pages. However, web publishers can only tune some of these. The number of hyperlinks and the way in which users refine their query terms, are completely out of the web publisher's control (Greenberg 2001). Nevertheless, users' web search behaviour in terms of keywords selection can be monitored to harvest the kind of keywords that users employ when searching information on a particular topic. With respect to this research, harvested keywords serve as guidance in the design of web page contents and associated metadata.

2.4. Web search behaviour

After analysing a large-scale log of web search behaviour, White et al. (2009) concluded that domain expertise greatly influences user behaviour and success when querying and retrieving information on web search engines. They evaluated domain expertise by comparing the vocabulary (textual queries) of participants with domain-specific lexicons. Other studies with smaller numbers of participants corroborate the findings of White et al. (2009): user domain expertise plays a major role in web search engine information retrieval (Holscher et al. 2000).

3. Research design

3.1. Approach

Our approach differs from the literature described in section 2 since it seeks to study and exploit the capability of making web resources with information about spatial data visible to well known web search engines. We propose that spatial data be documented in HTML in such a way that it can easily be indexed by general purpose web search engines. This is done by compiling spatial metadata contents as HTML documents which are crawlable by web search engines. This enables the discovery of spatial data without the user having to know about existing geoportals with spatial contents that are invisible to web search engines in the first place. We base this exercise on using search engine optimisation (SEO) techniques applicable to the contents of HTML web pages and associated tags such as the "meta" tag. A mapping from ISO 19115:2003 standard to Dublin Core is used to produce the end result as HTML web pages.

The figure below puts our research into context through a model that describes the spatial data search flow of a typical web user. First, the user uses a general purpose web search engine to locate information about spatial data, next the user enters the geoportal to view the spatial data and evaluates the metadata in order to decide whether it can be used for his/her purpose. Our research is concerned with the outer layer in Figure 1.

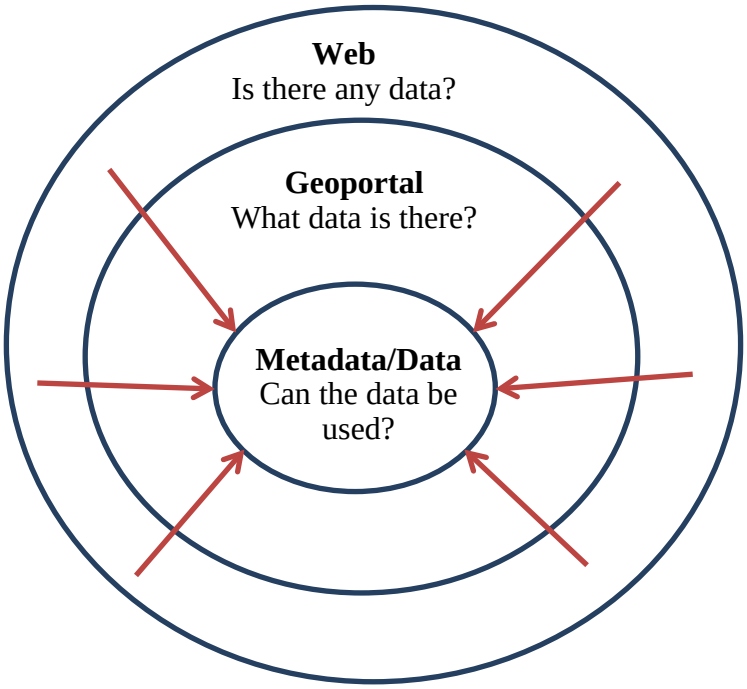

Figure 1. Description of the study context

3.2. Methodology for the research described in this paper

Since users drive online searches in searching for spatial data, it was imperative to come up with a user-centered methodology for preparing the HTML pages for the empirical tests. Furthermore, the findings of the related work as reported in section 2.4, suggest that user domain expertise plays a major role in web search engine information retrieval, irrespective of the number of users that participate in a given web search experiment. Therefore, we conducted a user study to determine which keywords users employ to search for geospatial data. Keywords collected from the user study were qualitatively analysed based on occurrence patterns in different participants' query terms. The analysis of these keywords facilitated the design of the model that defines the process of compiling spatial metadata as web resources (HTML pages).

3.2.1. Keyword experiment (user study)

A group of 17 BSc Geoinformatics students in their final year (third year) of studies participated in an experiment of two hours. They were required to search the web for spatial data to address a particular problem of their choice. They were advised to use any knowledge they had acquired in their studies. Their keyword (textual queries) selection was carefully monitored using the Mozilla Firefox web browser cache. A qualitative analysis of the collected keywords was performed to identify keywords used by all participants, excluding those specific to each individual's topic (domain) of interest. Commonality of keywords was considered because it reflects user behaviour in terms of keyword (or textual queries) selection.

4. Results

4.1. Spatial metadata compilation model (SMCM)

Selected search terms used by experiment participants are given in Table 1.

Participant ID	Search terms
P1	rwanda spatial data sets shapefile
P2	Ghana land cover shape file
P3	south africa vegetation cover spatial data
P4	spatial data of electricity distribution in rwanda
P5	raster data of south African power plan
P6	east african rift system formation in kenya
P7	land+use+data+sets+limpopo
P8	esri+shapefiles+schools
P9	Spatial data for nile river
P10	"Botswana", "water bodies",
P11	Cairo roads spatial data free
P12	mount kilimanjaro shapefile download

Table1. Sample of keywords used by participants

A qualitative analysis of participants' keyword selection suggests that spatial data search terms can be classified into three main categories as described in Figure 2.

- *Topic*: defines the subject of interest for which spatial data is being searched. Examples are: infrastructures, urban planning, water resources, and environmental management.
- *Location*: defines the spatial extent or geographic coverage of the spatial data being searched. Keywords related to the location are usually place name of geographic areas, e.g. Africa or Johannesburg.
- *Geographic feature*: defines the actual geographic object or feature of interest to the user. There are two sub-categories with respect to the two main spatial data models, namely vector and raster. Under the vector model, keywords are based on the geographic feature primitives such point, line or polygon. "shapefile" as keyword under the vector category was also used, since it is a well-known format for vector (spatial) data. The keyword "raster" is useful when looking for continuous geographic features.

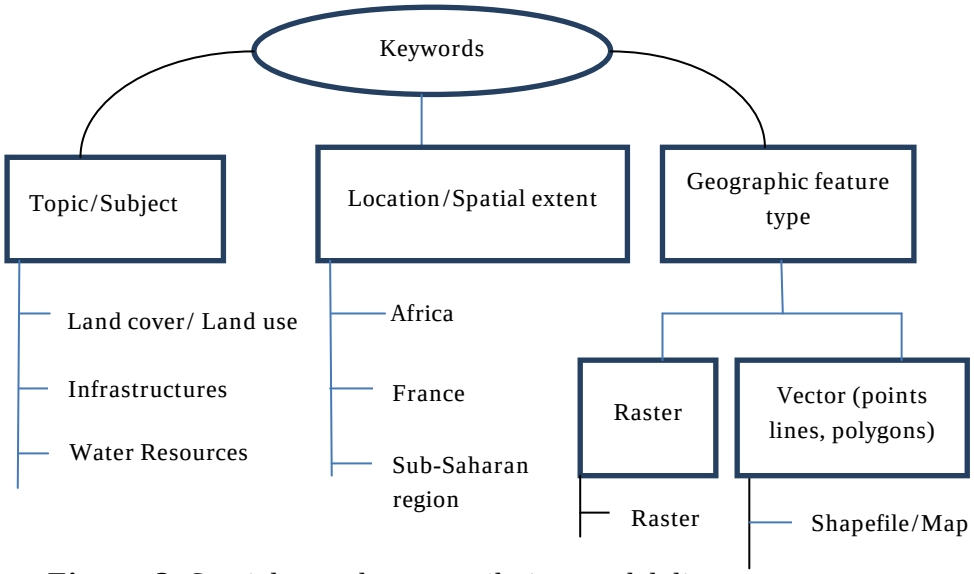

Figure 2. Spatial metadata compilation model diagram

The three main categories of keywords in the proposed model, describe how well users' keywords can be matched into the elements of the mapping between ISO 19115 and Dublin core standards. This concept (model) facili-

tates the preparation of spatial metadata for online publication not only for GIS professionals who are accustomed to ISO 19115 but for anyone who uses Dublin core.

4.2. Mapping SMCM categories to Dublin Core via ISO 19115

The SMCM categories were mapped to corresponding core elements of ISO 19115. This mapping facilitates compiling spatial metadata contents based on ISO 19115:2003 metadata standard. The core elements of ISO 19115 allow an easy understanding of geographic data by both consumers and producers (Nogueras-Iso et al. 2004). Moving from ISO 19115 to Dublin Core was the ultimate task of this exercise, since Dublin Core is the de facto standard for the description and discovery of web resources (HTML web pages). The table below describes the mapping process.

SMCM category	Corresponding ISO 19115 core elements	Corresponding Dublin Core elements
Topic/Subject	• Dataset title	TITLE
	• Dataset topic category	SUBJECT
	• Abstract describing the datasets	DESCRIPTION
Location/Spatial Extent	• Geographic location of the dataset (by four coordinates or geographic identifier)	COVERAGE: SPATIAL
	• Additional extent information for the dataset (vertical and temporal)	COVERAGE: TEMPORAL
Geographic Feature Type	• Spatial representation	TYPE
	• Distribution format	FORMAT
	• Lineage	SOURCE
	• Dataset responsible party	CREATOR
		PUBLISHER
	• Dataset reference date	DATE
	• On-line resource	IDENTIFIER
	• Dataset Language	LANGUAGE

Table 2. Mapping from SMCM to Dublin Core via ISO 19115

4.3. Application example

A practical example of the proposed model (SMCM) is provided to illustrate its application. The example describes a scenario illustrating the documentation of spatial data appropriate for discovery by participant P11 whose topic for spatial data discovery was "Cairo roads in Egypt". A dataset with the title "Egypt – Roads" on the "FAO Africover[1]" website was used, because it contains detailed metadata. Figure 3 shows an extract of the metadata web page for the spatial dataset considered.

Figure 3. FAO Africover metadata file of Egypt-Roads dataset

[1] FAO Africover project, http://www.africover.org/

The application of the proposed method for preparing spatial metadata for insertion in HTML pages is described as follows:

SMCM category and value	ISO 19115 elements and values
Topic: ***Cairo roads in Egypt***	Dataset title: ***Egypt – Roads***
	Dataset topic category: ***Roads Network of Egypt***
	Abstract describing the datasets: ***The roads of Egypt have been produced from visual interpretation of digitally enhanced LANDSAT TM images (Bands 4,3,2) acquired mainly in the year 1997.***
Location: ***Cairo, Egypt***	Geographic Location of the dataset: ***Cairo, EGYPT***
Geographic Feature types: ***Roads***	Data Format: ***ESRI ArcView Shapefile (.shp)***
	Lineage: ***The roads have been produced from visual interpretation of digitally enhanced LANDSAT TM images (Bands 4,3,2) acquired mainly in the year 1997***.
	Spatial representation: ***Vector***

Table 3. Mapping from SMCM to ISO 19115:2003

The final HTML page result in Figure 4 illustrates how the different field elements of the Dublin Core standard are filled using the "HTML meta" tag based on the mapping from ISO 19115 to Dublin core. It is possible for web search engines and their crawlers to index web resources (HTML documents prepared in this way) for optimum spatial data (metadata) discovery because web search engines are best at discovering HTML pages.

<link rel="schema.DC" href="http://purl.org/dc/elements/1.1/" />

<link rel="schema.DCTERMS" href="http://purl.org/dc/terms/" />

<meta name="DC.title" lang="English " content="***Egypt – Roads***" />

<meta name="DC.creator" content="***FAO AFRICOVER***" />

<meta name="DC.subject" lang="English " content="***Roads Network of Egypt*** " />

<meta name="DC.publisher" content="***FAO AFRICOVER***" />

<meta name="DC.description" content="***The roads of Egypt have been produced from visual interpretation of digitally enhanced LANDSAT TM images (Bands 4,3,2) acquired mainly in the year 1997***." />

<meta name="DC.date" content="2012-11-09" />

<meta name="DC.type" content="***Vecto***r" />

<meta name="DC.format" content="***ESRI ArcView Shapefile (.shp)***" />

<meta name="DC.identifier" scheme="DCTERMS.URI" content="***http://www.africover.org***" />

<meta name="DC.language" scheme="DCTERMS.URI" content="English" />

<meta name="DC.coverage" scheme="DCTERMS.URI" content="***Cairo; Egypt; Northern Africa ; Africa***" />

<meta name="DC.rights" scheme="DCTERMS.URI" content="***Copyright, FAO AFRICOVER 2012 All rights reserved***" />

Figure 4. Dublin core metadata contents of the final HTML page result

5. Conclusion

We described how the results of a user study guided the development of a user-centered methodology for adding spatial metadata to HTML pages. These pages will be used in first empirical tests of the effectiveness of search engine optimization. The proposed method is based on a model (SMCM) designed with input from an analysis of keywords obtained from an experiment in which users had to search for spatial data on the web. A mapping between ISO 19115:2003 spatial metadata and the Dublin core was used to prepare spatial metadata about geospatial resources for enhanced discovery by general purpose web search engines. The empirical tests are currently in progress and first results will be reported at the conference. First indica-

tions are that the pages are discoverable by general purpose web search engines. We plan to refine the SMCM model with results from additional studies with different user groups. This work can be extended by designing a tool to automate the process of spatial data documentation following the proposed spatial metadata data compilation model. Further empirical tests will be done by submitting HTML web pages obtained from the proposed methodology to web search engine directories. Subsequently, the HTML web pages in web search engines page rankings will be evaluated.

Acknowledgements

We would like to thank Ms Sanet Eksteen and her third year Geoinformatics students (class of 2012) for participating in the keyword experiment described in this paper.

References

CEN Workshop Agreement (2003) ftp://cenftp1.cenorm.be/PUBLIC/CWAs/e-Europe/MMI-DC/cwa14857-00-2003-Nov.pdf, (accessed: 09/11/2012)

Dublin Core Metadata Initiative (2012) http://dublincore.org/, last updated: 9/11/2012, (accessed: 15/11/2012)

Greenberg J., Maria C. P., Bijan P. and Davenport W.R. (2001) Author-generated Dublin Core Metadata for Web Resources (2001) A Baseline Study in an Organization. Journal of Digital Information, Volume 2 Issue 2 Article No. 78, November 2001.

Holscher C., Strube G., Web search behaviour of Internet experts and newbies, Computer Networks, 33, (2000), 337-346

Katumba S., Coetzee S., De la Rey A. (2012) An assessment of spatial data availability for renewable energy planning in sub-Saharan Africa, GISSA Ukubuzana 2012 Conference Proceedings, Kempton Park, South Africa, 2-4 October 2012.

Kumar R., Choundhary R. (2011) Modeling and analyse the Deep Web: Surfacing Hiden Value, International Journal of Computer and Information Security, Vol. 9, No 6, June 2011.

Lopez-Pellicer F., Béjar R., Zarazaga-Soria F. (2012) Providing Semantic Links to the Invisible Geospatial Web, Notes in Geoinformatics Research, 1st Edition, Prensas Universitarias de Zaragoza, 2012

Madhavan J., Ko D., Kot L., Ganapathy V., Rasmussen A., Halevy A.(2008) Google's Deep Web crawl, Proceedings of the VLDB Endowment, 1 (2): 1241-1252, 2008

Nogueras-Iso J., Zarazaga-Soria F.J. , Lacasta J. , Bejar R. B, Muro-Medrano P.R. Metadata standard interoperability: application in the geographic information domain, Computers, Environment and Urban Systems, 28 (2004) 611–634

Ntoulas A., Zerfos P., Cho J. (2005) Downloading textual hidden web content through keyword queries, Proceedings of the 5th ACM/IEEE-CS joint conference on Digital libraries, 2005

Sherman C., Price G., The Invisible Web: Uncovering Sources Search Engines Can't See, Thomas H. Hogan, Sr., 2001, United States of America

White R. W., Dumais S.T., Teevan J. (2009) Characterizing the influence of domain expertise on web search behaviour, Proceedings of the Second ACM International Conference on Web Search and Data Mining, 132-141, New York, USA, 2009

Wu P., Wen JR., Lui H., Ma WY.(2006) Query Selection for efficient Crawling of structured Web Sources, Proceedings of the 22nd International Conference on Data Engineering (ICDE), 2006

Zhang J., Dimitroff A.(2005) The impact of webpage content characteristics on webpage visibility in search engine results (Part I), Information Processing and Management 41 (2005) 665-690

Zhang J., Dimitroff A. (2005) The impact of webpage content characteristics on webpage visibility in search engine results (Part II), Information Processing and Management, 41, (2005), 691-715