UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Empirical tests using search engine optimisation techniques to compare the effectiveness of two metadata vocabularies for geospatial data discovery on the Web

by

Samy Kabangu Katumba

Submitted in partial fulfilment of the requirements for the degree

Master of Science (Geoinformatics)

In the Faculty of Natural & Agricultural Sciences

University of Pretoria

Pretoria

30th November 2015

© University of Pretoria

Declaration

I, Samy Kabangu Katumba, declare that this dissertation, which I hereby submit for the degree MSc (Geoinformatics) at the University of Pretoria, is my own work and has not previously been submitted by me for a degree at this or any other tertiary institution.

Date…….

Signature……..

© University of Pretoria

Abstract

With the increasing use of geographical information science and technology in a variety of knowledge domains and disciplines, the need to discover and access suitable geospatial data is imperative. To date, spatial data infrastructures (SDI) implemented at local, national, regional or international levels provide frameworks through which geospatial data is disseminated and shared between providers and consumers. Described as the most visible part of SDIs, geoportals are online web platforms for searching and discovering geospatial data and associated metadata. However, besides being known only to people in geoinformation communities, current geoportals as entry points to SDIs present some technological issues that make it difficult for web search engines to discover and index geospatial data catalogued in (or registered) with a geoportal. This hinders the visibility of geoportals and ultimately the discovery of geospatial data. Most geoportal implementations make use of web catalogue services for the registration and discovery of geospatial resources, making the catalogue opaque to web search engines. In this study, metadata about geospatial resources was published on the Web in the form of HTML web pages. Empirical tests with search engine optimisation techniques compared Dublin Core to Schema.org in order to evaluate web search engines' retrieval effectiveness. The process of constructing HTML pages was guided by a chain of mappings from a taxonomy of search terms to the *ISO 19115:2003, Geographic Information – Metadata* standard, to the Dublin Core metadata standard and finally to the Schema.org vocabulary. The taxonomy of search terms was constructed from a qualitative and quantitative analysis of search terms employed by users who were instructed to search for geospatial data online using web search engines of their choice. This taxonomy helped to understand the kind of terms that users employ when searching for geospatial data on the Web. The terms in the taxonomy were used to describe geospatial resources (metadata) in the HTML pages. Two sets of HTML web pages were registered with the Bing and Google web search engines. One set was marked-up with the Dublin Core vocabulary, and a second set was marked-up with the Schema.org vocabulary. The number of HTML pages retrieved and the average position of pages on the search results pages of the two web search engines were analysed using descriptive and inferential statistical techniques. Results show that metadata about geospatial resources contained in HTML web pages marked-up with Dublin Core and Schema.org vocabularies are discoverable on both search engines (Bing and Google). Overall, the results showed that Google was more effective in retrieving these HTML web pages as compared to Bing. HTML web pages marked-up with Schema.org vocabulary were effectively retrieved as compared to HTML web pages marked-up with Dublin Core. Furthermore, the test statistics results were significant in most of the tests performed. The user study resulting in a taxonomy of search terms presents an approach for understanding search terms when searching for geospatial data on the Web. The mapping from the taxonomy to *ISO 19115:2003, Geographic Information – Metadata*, and from Dublin Core to Schema.org vocabulary as discussed and proposed in this research demonstrates a process for embedding and encoding geospatial metadata in HTML web pages for optimum web search engine retrieval effectiveness of geospatial data.

© University of Pretoria

Acknowledgments

I am profoundly grateful to my supervisor Professor Serena Coetzee. Her guidance, patience, words of encouragement and invaluable advice greatly helped me complete this dissertation.

I am thankful to my friends, colleagues and fellow postgraduate students Noh Tewolde, Victoria Rautenbach and Wiafe Banahene. I very much appreciate your readiness and willingness to help and assist me. The various discussions we had in the computer laboratory contributed to helping me understand my research topic.

I would like to thank Professor Francois Steffens for his invaluable contribution to this dissertation. I would not have made it without your guidance with respect to the choice of statistical analysis tests to perform in this dissertation.

I am grateful to Dr. Antony Cooper whose advice and moral support really came at a moment when I needed it the most. Thank you.

I acknowledge Dr. Koech Cheruiyot from the Gauteng City-Region Observatory for assisting me with the interpretation of the statistical analyses performed in this dissertation.

I am grateful to my friend Namhla Tshisela for her willingness to proofread my dissertation.

To my siblings, I thank you for your prayers, love and support.

Last but not least, I am immensely grateful to my parents Louis and Julienne Kabangu for the love and support you never cease to provide me.

© University of Pretoria

Table of Contents

Declaration ii

Abstract iii

Acknowledgments iv

Table of Figures vii

List of Tables viii

Chapter 1 9

1 Introduction 9

1.1 Background 9

1.2 Statement of the problem 9

1.3 Research motivation and aim 10

1.4 Research objectives 11

1.5 Research significance 12

1.6 Research limitation and assumptions 14

1.7 Ethical considerations 14

1.8 Publication from this research 14

1.9 Overview of the remainder of the dissertation 15

Chapter 2 17

2 Literature Review 17

2.1 Introduction 17

2.2 Geographic information systems (GIS) 17

2.3 Geospatial data 18

2.3.1 Vector and raster representation of geospatial data 19

2.3.2 The Value and need of geospatial data 20

2.4 The Internet: a platform for spatial data sharing and dissemination 21

© University of Pretoria

2.4.1 Spatial data infrastructures (SDIs) 22
2.4.2 Geoportals 24
2.4.3 Architecture of a geoportal 25
2.4.4 Current challenges with geoportals in terms of geospatial data discovery 28
2.5 Metadata standards 29
2.5.1 ISO 19115:2003, Geographic Information – Metadata 30
2.5.2 Dublin Core 33
2.6 Schema.org vocabulary 36
2.7 Web search engines, Search engine optimisation (SEO) and online web resources' visibility 40
Chapter 3 43
3 Related work 43
3.1 Introduction 43
3.2 Users' web search behaviour 43
3.3 Modelling users' web search behaviour in mainstream information retrieval 44
3.4 Modelling users' query terms through classification or taxonomy 47
3.5 Modelling users' behaviour in geographic information retrieval (GIR) 48
3.6 Application of SEO techniques 50
3.7 Related literature on the evaluation of web search engines 56
3.8 Application of Dublin Core 59
3.9 Mapping and interoperability between standards 60
Chapter 4 63
4 Research methodology 63
4.1 Introduction 63
4.2 Approach 63

© University of Pretoria

4.3 Research design 65
4.4 Research instruments 66
4.4.1 First experiment: User experiment 66
4.4.2 Second experiment: Bing and Google webmaster tools experiment 67
4.5 Research analysis 68
Chapter 5 70
5 Analysis of search terms and taxonomy construction 70
5.1 Introduction 70
5.2 Analysis of search terms 70
5.3 Categories of the taxonomy 71
5.4 User experiment results 74
5.4.1 First year geoinformatics students 75
5.4.2 Third year geoinformatics students 75
5.4.3 Professionals 76
5.4.4 Summary of user experiment results 76
5.5 Results of the Bing webmaster tools experiment 77
5.6 Proposed taxonomy of search terms 78
5.7 Merits and limitations of the proposed taxonomy 80
5.8 Chapter conclusion 81
Chapter 6 82
6 Mapping among the taxonomy search terms, ISO 19115:2003, Geographic Information – metadata standard, Dublin Core and Schema.org 82
6.1 Introduction 82
6.2 Overview of the mappings (crosswalks) involved 82
6.3 Mapping the categories of the taxonomy to Dublin Core via ISO 19115:2003, Geographic Information – Metadata 83

© University of Pretoria

6.4 Understanding the elements involved in the mapping (from the taxonomy of search terms to Dublin Core via ISO 19115:2003, Geographic Information – Metadata) 84

6.4.1 Description of ISO 19115:2003 elements involved in the mapping 85

6.4.2 Description of Dublin Core elements involved in the mapping 86

6.4.3 Incorporating spatial metadata into HTML web pages based on the mappings between the taxonomy of search terms, ISO 19115:2003, Geographic Information – Metadata and Dublin Core 86

6.5 Mapping the categories of the taxonomy to Schema.org vocabulary via ISO 19115:2003, Geographic Information – Metadata and Dublin Core 91

6.6 Chapter conclusion 93

Chapter 7 95

7 Results of submitted HTML web pages: retrieval effectiveness evaluation and comparison 95

7.1 Introduction 95

7.2 Construction of HTML web pages with spatial metadata 95

7.2.1 Experiments set-up 95

7.2.2 Webmaster tools (for Bing and Google) 96

7.3 Markup validation 96

7.3.1 Markup validation in Bing webmaster tools 96

7.3.2 Markup validation in Google webmaster tools 98

7.4 Comparison of the retrieval effectiveness between the two sets of HTML web pages: Dublin Core vs. Schema.org 100

7.5 Analysis and comparison of results 102

7.6 Comparison between the two search engines based on metadata vocabularies (Dublin Core and Schema.org) 105

7.6.1 Dublin Core results: Bing versus Google 105

7.6.2 Schema.org results: Bing versus Google 108

© University of Pretoria

7.7 Comparison within each web search engine based on web search engine results (Bing and Google) 111
7.7.1 Bing Results: Dublin Core versus Schema.org 112
7.7.2 Google Results: Dublin Core versus Schema.org 114
7.8 Summary of results and chapter conclusion 118
7.8.1 Results by metadata vocabulary 118
7.8.2 Results by web search engine 119
Chapter 8 121
8 Conclusion 121
8.1 Introduction 121
8.2 Summary of main findings 121
8.2.1 Analysis of search terms and taxonomy construction 122
8.2.2 Mappings between standards (ISO 19115:2003, Geographic Information – Metadata & Dublin Core) and Schema.org 122
8.2.3 Comparison of the retrieval effectiveness of HTML web pages with spatial metadata content 123
8.3 Research contribution to science and technology 124
8.4 Possible future work 125
References 126
Appendix A – Instructions to participants in the user experiment 132
Appendix B – ISO 19115:2003, Geographic Information – Metadata (Core Elements) 133
Appendix C – Dublin Core Metadata Elements 138
Appendix D – Selected Schema.org Types and Properties 140
Appendix E – Mapping between the taxonomy of search terms and metadata standards 143
Appendix F – ISO 19115:2003 spatial metadata standard elements' description 146
Appendix G – HTML web page marked-up with Schema.org vocabulary 148

© University of Pretoria

Appendix H – HTML web pages retrieved by Bing 152

Appendix I - HTML web pages retrieved by Google 156

Appendix J – Queries obtained from participants in the user experiment 161

Appendix K – Queries obtained from Bing webmaster tools experiment (as collected from the log of Bing search engine) 175

© University of Pretoria

Table of Figures

Figure 1: Geoportal architecture (Nebert et al. 2004) 25
Figure 2: Publish-Find-Bind pattern applied to a geoportal (Nebert et al. 2004) 26
Figure 3: Different geoportal distributed configurations (Nebert et al. 2004) 27
Figure 4: Multi-tier architecture applied to a geoportal (Nebert et al. 2004) 27
Figure 5: UML diagram describing Relationship between ISO 19115:2003, Geographic Information – Metadata entities (Source: ISO 19115: 2003) 31
Figure 6: Core elements of ISO 19115:2003, Geographic Information – Metadata (ISO 19115: 2003) 32
Figure 7: Incorporating Dublin Core vocabulary into an HTML web page 36
Figure 8: Hierarchical structure of selected Schema.org types (http://schema.org/) 39
Figure 9: Microdata code example 40
Figure 10: Research approach 64
Figure 11: Hierarchy of metadata (Source: ISO 19115:2003) 71
Figure 12: User experiment results 75
Figure 13: Ratio of number of participants to number of queries 77
Figure 14: Experiment results – Bing webmaster tools application 78
Figure 15: Taxonomy of search terms 80
Figure 16: Mapping among standards and vocabularies 83
Figure 17: Example of a web page with spatial metadata contents 87
Figure 18: Example of a web page with spatial metadata contents 89
Figure 19: HTML web page with spatial metadata content marked-up with Dublin Core vocabulary 90
Figure 20: HTML web page with spatial metadata content marked-up with Schema.org vocabulary 92
Figure 21: Bing Markup Validator 97
Figure 22: HTML page marked-up with Schema.org validated by Bing 98
Figure 23: Structured data testing tool (Google) 99
Figure 24: HTML web pages marked-up with Schema.org vocabulary validated by Google 100
Figure 25: HTML web pages marked-up with Dublin Core listed on Google Webmaster Tools 103
Figure 26: HTML web pages marked-up with Dublin Core listed on Bing Webmaster Tools 103
Figure 27: HTML web pages marked-up with Schema.org listed on Google Webmaster Tools 104
Figure 28: HTML web pages marked-up with Schema.org listed on Bing Webmaster Tools 105

© University of Pretoria

List of Tables

Table 1: Attributes of the Dublin Core elements (Green and Bossomaier 2002) 34
Table 2: Description of the 'meta' tag attributes (Green and Bossomaier 2002) 35
Table 3: Properties of type 'Thing' (http://schema.org/) 37
Table 4: Properties of 'Dataset' type (http://schema.org/) 38
Table 5: Terminology 70
Table 6: Search term categorisation 74
Table 7: User experiment results – First years 74
Table 8: Experiment results from the Bing webmaster tools 77
Table 9: User experiment results – Average number of terms per category 79
Table 10: Bing webmaster tools experiment results – Average number of terms per category 79
Table 11: Mapping between taxonomy of search terms and *ISO 19115:2003, Geographic Information – Metadata* 88
Table 12: Spatial metadata insertion into an HTML web page based on Schema.org vocabulary 93
Table 13: Description of measures for web search engines performance evaluation 101
Table 14: Comparison within each web search engine 102
Table 15: Comparison between the two search engines 102
Table 16: Dublin Core results (Bing versus Google) 106
Table 17: Descriptive Statistics – Average positions of web pages marked up with Dublin Core (Bing versus Google) 107
Table 18: Description of ranks – Average positions of web pages marked-up with Dublin Core (Bing versus Google) 107
Table 19: Test Statistics – Average positions of web pages marked up with Dublin Core (Bing versus Google) 108
Table 20: Schema.org results (Bing versus Google) 109
Table 21: Descriptive Statistics – Average positions of web pages marked up with Schema.org (Bing versus Google) 110
Table 22: Description of ranks – Average positions of web pages marked up with Schema.org (Bing versus Google) 110
Table 23: Test Statistics – Average positions of web pages marked up with Schema.org (Bing versus Google) 111
Table 24: Bing Results(Dublin Core versus Schema.org) 112
Table 25: Descriptive Statistics results - Bing Results (Dublin Core versus Schema.org) 113
Table 26: Description of ranks - Bing Results (Dublin Core versus Schema.org) 114
Table 27: Test Statistics - Bing Results (Dublin Core versus Schema.org) 114
Table 28: Google Results (Dublin Core versus Schema.org) 115
Table 29: Descriptive Statistics results - Bing Results (Dublin Core versus Schema.org) 116
Table 30: Description of ranks - Bing Results (Dublin Core versus Schema.org) 117
Table 31: Test Statistics - Bing Results (Dublin Core versus Schema.org) 117
Table 32: Retrieval effectiveness: summary of results 118
Table 33: Retrieval effectiveness: detailed summary of results by vocabularies 118
Table 34: Retrieval effectiveness: detailed summary of results by search engines 120

© University of Pretoria

Chapter 1

1 Introduction

1.1 Background

The use of geospatial data and geographic information systems (GIS) has become ubiquitous. A variety of knowledge domains and specialised disciplines are making use of GIS whether on desktop computers, mobile devices or over the internet (web) for analysing and sharing geospatial data to answer questions closely related to location: 'Where?'. Such domains of knowledge include transportation planning, logistics, business, environmental monitoring, natural resource management and mining to name a few. However, a GIS can only assist provided it has been supplied with suitable geospatial data ready to be processed. This means that geospatial data need to be discoverable and available. Initiatives geared towards making geospatial data available have resulted in the implementation of frameworks such as spatial data infrastructures (SDIs) which facilitate the sharing and dissemination of geospatial data.

Implemented as part of SDIs, geoportals allow the online dissemination of geospatial data from providers. Considered as being the "most visible part of SDIs" (De Longueville 2010), geoportals provide web platforms to search for geospatial data and associated metadata content. However, the discovery of geospatial data accessible through geospatial web services hosted in geoportals remains a challenge (Lopez-Pellicer *et al.* 2012). Business, legal and technological barriers result in the invisibility of geospatial web services and geospatial data to well-known general purpose web search engines. This hinders the visibility of geoportals and thus the discovery of geospatial data. Furthermore, considerable amounts of online geospatial data are of no value due to the lack of supporting information describing their contents or details on how they can be used.

1.2 Statement of the problem

To find geospatial data, interested parties have to know where and how to access existing geospatial web services hosted in geoportals. People in geographic information communities may be aware of such online platforms, but what about those outside these communities who

© University of Pretoria

want to use geospatial data in their domains of expertise? Not being aware of the existence of geoportals in the first place makes the use of such online platforms impossible. Furthermore, current geoportals which are mainly built on Open Geospatial Consortium (OGC) web services and standards present challenges in terms of the online discovery of geospatial data when using web search engines. These challenges are related to economical, legal and technological considerations (Lopez-Pellicer *et al*. 2012) that affect the online sharing of geospatial data (some of these technological challenges are discussed in more detail in section 2.4.4). Finding relevant geospatial data that satisfy the needs of the requesters is still an issue.

Attempts to solve the technological challenges related to the invisibility of web services and geospatial web services, in particular, have resulted in the development of focused web search engines (crawlers) with the sole mission of retrieving geospatial web contents (services and data) hidden behind geoportals. A number of these works from the literature have been discussed in the work done by Lopez-Pellicer *et al*. (2012). Such efforts are focused on web crawlers' capabilities to search for geospatial web services (or geoportals), rather than on the ability of such resources making themselves visible to web search engines. In many cases, these focused crawlers use web pages retrieved by Bing, Google or Yahoo as seeds for further search refinement (Lopez-Pellicer *et al.* 2012).

Furthermore, when looking at the available literature, research that focuses on exploiting the capabilities and properties of geospatial data to make such resources visible to web search engines is limited. Most research as mentioned in the above paragraph focuses on customising web search engines for finding geospatial data on the web. Possible reasons are that SDIs have typically published geospatial data in geoportals and therefore these researchers focus on increasing the visibility of geospatial data already published in geoportals; also, if the crawler can detect already published geospatial resources, the resources do not have to be republished. Techniques such as web search engine optimisation (SEO) which is applied in mainstream online information retrieval, have not yet been explored for the discovery of online geospatial data.

1.3 Research motivation and aim

The research reported in this dissertation presents the results of empirical tests on the discoverability of HTML web pages with spatial metadata content by general purpose web search engines. It is motivated by current difficulties in finding geospatial data with general

© University of Pretoria

purpose web search engines (Katumba and Coetzee 2012). The study describes two identical sets of HTML web pages: one marked-up with the 'Dublin Core metadata standard' vocabulary and the other with selected types of 'Schema.org' vocabulary. Over 15 years, the Dublin Core metadata standard has been established as the standard for interoperability suitable for the discovery of online web resources (Harper 2010). Schema.org is a collection of schemas (or types) for describing the content of web resources. Schema.org has been endorsed by major search engines such as Google, Bing, Yahoo and Yandex (Krutil *et al*. 2012). Search engine optimisation (SEO) techniques are taken into consideration during the design process of HTML web pages for the empirical tests. The main aim of the study is to compare the effectiveness of the two vocabularies (Dublin and Schema.org) for facilitating the discovery of geospatial data on the web. The term 'vocabulary' refers to both the Dublin Core metadata standard elements, and properties of selected types (schemas) contained in Schema.org. With reference to the World Wide Web Consortium (W3C) website, vocabularies "define the concepts and relationships (also referred to as 'terms') used to describe an area of concern" (W3C 2013).

1.4 Research objectives

This research pursues three main objectives listed as follows:

- Objective 1 -To design and perform experiments that result in the construction of a taxonomy of search terms. The constructed taxonomy provides an understanding of how geospatial data are searched using web search engines. The steps involved to achieve this objective are given as follows:
    - To identify search terms that are used to search for geospatial data on the web using general purpose web search engines.
    - To subjectively analyse and assign the different collected search terms to appropriate categories of the taxonomy.
    - To compute simple descriptive statistics for the purpose of validating or determining the consistency the constructed taxonomy.
- Objective 2 – To establish mappings (crosswalks) and also use existing mappings (or crosswalks) among the taxonomy (constructed in Objective 1), *ISO 19115:2003, Geographic Information – Metadata* standard, Dublin Core vocabulary and Schema.org vocabulary. The steps involved to achieve this objective are given as follows:

© University of Pretoria

- To establish a mapping between the categories of the taxonomy and selected elements of *ISO 19115:2003, Geographic Information – Metadata* standard.
- To use the existing mapping (crosswalk) between selected elements of Dublin Core vocabulary and selected elements of *ISO 19115:2003, Geographic Information – Metadata* standard.
- To establish a mapping (crosswalk) between selected elements of Dublin Core vocabulary and selected types (schemas) of Schema.org vocabulary.
- To design (construct) two sets of HTML web pages (with similar spatial metadata information as content) by marking-up one set of HTML web pages with Dublin Core vocabulary and the other set with Schema.org vocabulary. The process of marking-up these HTML web pages is guided by the established and existing mappings as per the previous steps mentioned in the attainment of this objective.
- To upload the constructed HTML web pages (with spatial metadata content) onto a web server (website), and register the website with Bing and Google web search engines via their respective 'webmaster tools' online applications.

- Objective 3: To quantify and evaluate the retrieval effectiveness of the submitted HTML web pages based on search results obtained from general purpose web search engines (Bing and Google webmaster tools), and compare these results based on the two vocabularies (Dublin Core and Schema.org) and the two web search engines (Bing and Google). In essence, the last objective involves performing a web search engine performance evaluation.

1.5 Research significance

Current spatial data infrastructures (SDIs) and geoportals for sharing geospatial data suffer from issues that impede the efficient dissemination and discovery of geospatial data on the web. Besides the technological challenges hindering the effective success of SDIs and related web services, is the fact that many people outside geoinformation communities are unaware of such infrastructures and frameworks. The research described in this dissertation presents an improved approach for finding geospatial data on the web. It suggests the use of general purpose web search engines as means of searching online web resources (or web pages) describing geospatial

© University of Pretoria

data and the use of search engine optimisation (SEO) techniques in the process of creating web resources (in this case web pages) that describe such (geospatial) data.

SEO techniques are used as a strategy for enhancing the visibility of web pages in the search results of web search engines (Ochoa 2012, Swati *et al*. 2013, Onaifo and Rasmussen 2013, and Mustafa *et al*. 2015). In the context of this research, this approach has been found worthy given the fact that web search engines are extensively used to find information about almost any subject (topic) of interest on the web (internet). According to the Pew Internet Project (2011), close to 92% of all current activities happening on the internet involve the use of search engines. According to McGee (2010), over a billion searches are conducted on Google on a daily basis. According to Cahill and Chalut (2009), for most people, Google has become not only the first port of entry for information seeking, but the only port of call. This is a proof of how many people rely on web search engines to find information on the web. Improving the visibility of web pages with information about geospatial data can thus have a significant positive impact on the visibility and discovery of geospatial data. In future, the expensive development and maintenance of customised geoportals could be replaced by widely used HTML pages with SEO-focused information about geospatial data.

Furthermore, it should be noted that this research comes not as a way of replacing geoportals, but rather as a means of complementing geoportals to achieve their intended goal. This will become evident as HTML web pages with geospatial data descriptions retrieved by web search engines, will be leading web traffic to appropriate and corresponding geoportals where the relevant geospatial data can be downloaded.

A taxonomy for the purpose of understanding search terms that users employ when searching for geospatial data using web search engines is proposed. This research also proposes a mapping between Dublin Core metadata standard elements and selected properties of types (schemas) contained in Schema.org. Such a mapping does not currently exist and it has been specifically devised for creating HTML web pages with contents that serve as a description of geospatial data (spatial metadata). Web pages constructed in this manner serve as proxies for accessing geospatial data given that they (such web pages) are retrievable by web search engines. Finally, the study compares two well-known metadata vocabularies used in mainstream online information retrieval and presents the results on performance evaluation of two well-known web

© University of Pretoria

search engines (Bing and Google). Unlike many other studies, this research makes use of existing standards and online web search tools (web search engines) to propose a solution to the problem of geospatial data discovery on the web.

1.6 Research limitation and assumptions

The user study experiment was done with a small number of participants which is justified on the assumptions that the number of participants has no bearing on the experiment outcome. What matters is the participants' expertise in the domain in which the search is performed. This is reflected in the quality of search terms or key phrases they employed.

The proposed mapping (crosswalk) between Dublin Core metadata standard elements and the properties of selected types (schemas) contained in Schema.org is purely motivated by the need of evaluating Schema.org effectiveness in terms of enhancing the visibility of HTML web pages with spatial metadata contents by web search engines. This was necessary as there is currently no available official mapping between the two vocabularies. However, a mapping initiative between the two vocabularies is underway (Dublin Core 2012).

The statistical analyses performed as a means of comparing results of the retrieval effectiveness of the considered HTML web pages in this research are not to be compared to some of the advanced and detailed analysis exercises described in the literature where the performance of web search engines is evaluated. However, the statistical analyses from this study are to be considered in the context and set up of the experiments performed.

1.7 Ethical considerations

Experiments performed in the study described in this dissertation involved a voluntary participation of human subjects referred to as participants. The procedures were fully explained to participants and a clearance to conduct the user experiment was obtained from the Faculty of Natural and Agricultural Sciences Ethics Committee at the University of Pretoria. The document containing the instructions given to participants in the experiment can be found in Appendix A.

1.8 Publication from this research

Two published papers from this research are:

© University of Pretoria

1. Katumba S and Coetzee S (2013). Spatial data discovery using general purpose web search engines, *26th International Cartographic Conference, Dresden, Germany*, 25 - 30 August 2013, which addresses a part of objective 1 (user experiment).
2. Katumba S and Coetzee S (2015). Enhancing the online discovery of geospatial data through taxonomy, folksonomy and semantic annotations, South African Journal of Geomatics, vol. 4(3), pp. 339-350, which addresses both objective 1 and objective 2 as set in this dissertation.

The papers make up a significant part of chapter 5.

1.9 Overview of the remainder of the dissertation

The remainder of this dissertation is structured as follows:

Chapter 2 – Literature review. In this chapter, the background concepts pertaining to the understanding of the content of this dissertation are discussed. These include geographic information systems, geospatial data, the internet as a platform for disseminating geospatial data, spatial data infrastructures, geoportals and metadata standards and vocabularies. Web search engines and search engines optimisation techniques and web resources' visibility are also discussed in chapter 2.

Chapter 3 – Related work. This chapter discusses work from the literature related to modelling users' web search behaviour in general and in terms of query terms' selection. Work related to the creation of taxonomy with respect to web searches, modelling users' behaviour in geographic information retrieval, application of search engine techniques, web search engine performance evaluation and use of metadata standards and mapping (crosswalks) between metadata standards are also discussed in chapter 3.

Chapter 4 – Research methodology. The methodology and approach adopted in this dissertation are discussed in chapter 4.The instruments (software) used to collect and analyse data, experiments setups and analyses performed in this dissertation are also discussed.

Chapter 5 – Analysis of search terms and taxonomy construction. This chapter provides details involved from the analysis of search terms through to the construction of the proposed taxonomy of search terms.

© University of Pretoria

Chapter 6 – Mapping among the taxonomy search terms, ISO 19115:2003, Geographic Information – Metadata standard, Dublin Core and Schema.org. This chapter elaborates on a chain of mappings from the taxonomy of search terms to the *ISO 19115:2003, Geographic Information – Metadata* standard to the Dublin Core metadata standard and finally to the Schema.org vocabulary. This chapter also discusses how these mappings guided the construction of HTML web pages with spatial metadata content.

Chapter 7 – Results of submitted HTML web pages: retrieval effectiveness evaluation and comparison. This chapter provides feedback on the evaluation and comparison of the retrieval effectiveness results between web pages marked-up with Dublin Core vocabulary and Schema.org vocabulary as obtained from Bing and Google web search engines.

Chapter 8 – Conclusion. This chapter provides a closure to this dissertation.

© University of Pretoria

Chapter 2

2 Literature Review

2.1 Introduction

This chapter discusses the background concepts pertaining to the research described in this dissertation. A narrative on the history of GIS followed by a description of geospatial data, their digital representation, their importance and role within a GIS are discussed in this chapter. The online sharing and dissemination of geospatial data over the internet facilitated by platforms such as spatial data infrastructures and geoportals are reviewed. The two metadata standards namely *ISO 19115: 2003, Geographic Information-Metadata* and Dublin Core, as well as Schema.org vocabulary, which is a collection of schemas (types) for marking-up online web resources, are discussed. Finally, concepts related to web search engines, search engines optimisation (SEO) and online web resources' visibility are reviewed.

2.2 Geographic information systems (GIS)

To begin with, the history behind the inception of GIS is as uncertain as its definition. This is evident as many of the advances in GIS came about with the development of computer-based tools to solve specific problems in a variety of disciplines. It is believed that the first ever GIS was the Canada Geographic Information System (CGIS) which was operational in the mid-1960s (Longley *et al*. 2005). This was a computer-based system used for measuring areas from maps. It was followed by the Dual Independent Map Encoding (DIME), a programme that oversaw the creation of digital records of all United States' streets for the purpose of automatically assigning and aggregating the 1970's census records. One of the notable milestones in the history of GIS was the first 'computer-made map' published by the United Kingdom (UK) Experimental Cartography Unit (ECU) in 1973. Such an achievement proved critical as it brought an end to the tedious and error-prone process of manual map making.

GIS are well known for "the intrinsic multidisciplinarity with which they have evolved" (Giacomeli 2006). Digital cartography, environmental management, ecology, urban planning and geological survey are just a few of the many disciplines where GIS find their applications. As defined by Green and Bossomaier (2002), GIS are "computer programs for acquiring, storing,

© University of Pretoria

interpreting, and displaying spatially organised information". A more encompassing definition is the one provided by Dueker and Kjerne (1989) and Chrisman (1999), who define a GIS not just as a mere software tool but as "a system of hardware, software, data, people, organisation and institutional arrangements for collecting, storing, analysing and disseminating information about areas of the Earth"; that is, collecting, storing, analysing and disseminating geospatial data. Such a definition appears to have stood the test of time particularly in an era where GIS are no longer just desktop applications that can only be operated by an individual expert. The advent of online map applications such as Google Earth, Google Maps, Bing Maps and location-based services have provided simple means by which a common user can easily interact with a GIS that supports the 'behind the scenes' functioning of such web-based platforms. However, for a GIS to be of any use it needs to be supplied with geospatial data.

2.3 Geospatial data

The notion behind geospatial data is central to the study described in this dissertation. This is because this research seeks to find ways of easing the online discovery of geospatial data. Hence, the technicalities related to the concepts of geospatial data are to be explained. This will be followed by a discussion on the importance of geospatial data and the factors impeding its effective discovery.

In the context of this study, the term 'geospatial data' is referred to as a digital or computerised representation of real world features or phenomena. It requires some level of abstraction for converting physical features or phenomena into graphic features so they can be visualised on a flat computer screen. To do so, the field of GIS (or geoinformatics) has borrowed related concepts from the field of mathematics (geometry) and computer graphics (Longley *et al*. 2005). In a GIS, a given real world feature or a phenomenon constitutes a topic of interest which can be referred to as a 'theme'. Examples of themes could be buildings of a given city, land parcels within a neighbourhood, land cover or land use to name a few. A geographic object is a model representation of a given real-world entity (theme) comprised of two parts: a spatial component and a descriptive component. The descriptive component mainly contains attributes that describe the real world features being represented by the spatial component (Longley *et al*.2005). For example, if the real world features being represented is about a road network, corresponding

© University of Pretoria

attributes may be information about the name of the roads, types, length, or the current conditions of the road network.

The physical nature of geographic features or phenomena greatly influences their model representation as spatial objects in a GIS. Geographic features are represented either as 'discrete objects' or 'continuous fields'. Man-made features such as buildings, roads, magisterial district boundaries to name a few, are usually represented as discrete objects whereas continuous fields are natural features or phenomena such as forests, rainfall, elevation, etc.

2.3.1 Vector and raster representation of geospatial data

As already described in the previous sections, the process of representing the real world into graphics and digital computer files in a computer system such as a GIS, requires some degree of abstraction and modelling. Geospatial data can either be represented as vector data or raster data (Longley *et al*. 2005). Each representation method or data structure is chosen on whether the geographic features or phenomena being represented are modelled as discrete objects or continuous fields.

Although GIS software offers options for transforming from one representation to another, discrete objects are more likely to be represented as vector data and continuous fields tend to be represented as raster data (Longley *et al*. 2005).

In a vector data structure, spatial objects are represented as points, lines or polygons:

- Points are 'zero-dimensional objects' suitable for representing geographic features or phenomena whose dimension is relatively small and can be negligible. A geographic feature is represented as a point depending on its relative size to other features being represented. For example, on a small-scale map showing the different continents of the world, cities and towns may be represented as points.
- Lines are 'one-dimensional objects or linear objects' which are mostly used to represent geographic features that form networks such as roads, hydrology, electricity cables, and pipelines.
- Polygons are 'two-dimensional objects' used to represent geographic features that are relatively larger compared to other features being represented. For example, countries are more likely to be represented as polygons on a map showing continents.

© University of Pretoria

The raster data structure presents an array of regular shapes called pixels whose values are related to the attributes or properties of the geographic phenomenon or feature being represented. The variation of the pixel values is reflected in the colour of the feature or phenomenon being represented. Raster data structures are often used for representing geographic phenomena that are modelled as continuous fields. Remote sensing constitutes one of the main sources of raster data which are stored as images.

2.3.2 The Value and need of geospatial data

Besides the fact that the GIS software industry accounts for over a billion US dollars in annual sales worldwide, the sale of geospatial data also accounts for a significant amount. The cost of geospatial data is mainly related to the process of acquiring, creating, maintaining and disseminating such geospatial data (Longley *et al*. 2005). At each stage, involved resources in terms of human labour, software and hardware equipment justify the high cost of geospatial data. Geospatial data can be collected through ground and field surveys using instruments such as the Global Positioning System (GPS). Dedicated sensors aboard satellite platforms may also be employed to remotely collect spatial information about various features above the earth's landscape. Such features may be comprised of man-made infrastructures such as buildings, roads or natural features such as water resources (rivers, dams or mountains) and vegetation (such as forests). From such collected (geospatial) data, a wealth of information can be derived to support decision-making processes in various disciplines and knowledge domains critically important to improve people's lives. Such domains of knowledge may include agricultural productivity, water resource management, transportation, environmental management, and urban planning to name a few. This resonates with the vision of Ian Masser (2005) who was well aware of considering geospatial data as "a national asset" to be put on the same pedestal with other infrastructures such as buildings, roads or railway networks. The list below provides a number of themes of geospatial data that are considered of common use and importance in various countries in the world (Nebert *et al*. 2004):

- Cadastral information
- Geodetic control
- Geographic feature names
- Orthoimagery

© University of Pretoria

- Elevation
- Transportation
- Hydrography (surface water networks)
- Governmental units

2.4 The Internet: a platform for spatial data sharing and dissemination

Since its inception, the World Wide Web (internet) has revolutionised the way in which people communicate and how organisations perform their daily operations. Knowledge domains of various kinds are affected in an unprecedented manner. The field of geospatial information is not exempted. The internet provides a platform where a single computer can be connected to a thousand other computers irrespective of their disparate geographic locations. This facilitates the sharing of information between and within organisations. The internet has blossomed to the point where software applications that were confined on single desktop computers are now accessible through appropriate computer network systems. For example, the internet makes it possible for the result of a computational task performed by a machine at one end of the globe, to be retrieved by another computer at the other end. Under normal circumstances, such a process can happen in a fraction of seconds. The need to use the internet as a platform for sharing and disseminating geospatial data emanates from the many benefits the internet offers. The development of the internet has also influenced to some extent the dissemination of geospatial data on the Web (internet). The availability of geospatial data has greatly influenced the development of online GIS applications whether on desktop computers or across a network of mobile devices.

A variety of GIS applications across different disciplines or knowledge domains brings about a diversity in the theme (subject), type, format and provenance of geospatial data in use. To date, the amount of geospatial data available online seems to be enormous to quantify (Lopez-Pellicer *et al*. 2012). The increasing number of producers and providers of geospatial data, whether from public institutions or in the private sector (commercial), posits the need of coordinating efforts involved in the production and dissemination of geospatial data (Nebert *et al*. 2004). Furthermore, the lack of adequate or complete documentation makes it difficult to discover geospatial data on the web (Lopez-Pellicer *et al*. 2012). As a remedy, the concept and materialisation of spatial data infrastructures came into existence.

© University of Pretoria

2.4.1 Spatial data infrastructures (SDIs)

Concepts related to the establishment of SDIs started to emerge in the early 1990s in Europe and in the United States of America (Foley *et al*. 2009). The term spatial data infrastructure itself originates from the United State National Research Council. An SDI defines a framework which encompasses technologies, policies and institutional arrangements meant to enable the creation, exchange, and use of geospatial data and related information resources within a given information-sharing community. In 1994, through the Federal Geographic Data Committee (FGDC), the then president Bill Clinton (of the United States of America) formally endorsed the development of a National Spatial Infrastructure in the United States of America.

An SDI can be referred to as a blend of legislations, institutional arrangements (policies), people and technologies to assure the availability, access and sharing of geospatial data (Nebert *et al*. 2004). As it has been already elaborated (refer section 2.3.2), the value and importance of geospatial data can no longer be over emphasised. However, assuring a timely and effective distribution of such (geospatial) data is critical in maintaining a uniform and balanced approach in enabling all concerned parties to take advantage of (geospatial) data (information). An SDI enables geospatial data producers to put their data on a shared platform whereby interested consumers will be able to easily discover such data and use it. Different activities undertaken in the implementation of an SDI includes the coordination of efforts pertaining to the production (creation), storage, documentation, access, sharing and distribution of geospatial data. Even more crucial, the concept of establishing an SDI goes beyond the one of a GIS to ensure that standards are applied to produce authoritative (geospatial) datasets and policies to govern all the involved activities. This adds to the number of benefits offered through the implementation and functioning of SDIs. Among many others, such benefits include (Nebert *et al*. 2004):

- Elimination of redundancy in the production of geospatial data, which leads to saving time, efforts and resources to perform such a task. In other words, the production of geospatial data occurs only once at a specific location, but consumption happens repeatedly at a wide scale by multiple applications (or people) in disparate locations.
- Ensuring the discovery of geospatial data and related spatial information and services. This amounts to raising awareness among interested parties of the existence of geospatial

© University of Pretoria

data whereby effective mechanisms are put in place to ensure that geospatial data are well advertised and can easily be found.

All these benefits reflect the key components of an SDI as described by Foley *et al*. (2009) namely legislative frameworks, content and data frameworks, metadata and standards, quality issues, geoportals, human factors and dissemination and implementation. Some of these key components are described as follows:

- Legislative frameworks: when different institutions, people and resources (geospatial data and technology) are combined to obtain an infrastructure such as an SDI, the adoption of laws and policies becomes a requirement to ensure a smooth and effective interaction between the various institutions involved. Given that an SDI may be implemented at different levels, namely local, national and global levels, at each one of these levels, laws may need to be adopted to ensure that all the parties involved are in agreement. Some of the issues that may need regulations when implementing SDIs includes copyright, ownership and data protection. Countries around the world are designing SDIs Acts and other legislations to regulate the functioning of SDIs. In South Africa, for example, the SDI Act regulates the functioning of the South African National Spatial Data Infrastructure (SASDI).
- Content and Data Frameworks: At the heart of any given GIS lies the raw data which need to be processed to obtain geographic or spatial information crucial in decision-making processes. An adequately implemented SDI is essential in materialising such an endeavour.
- Human Factors: Human intervention in the success of an SDI is pivotal. Hence, capacity building geared towards establishing SDIs becomes vital. However, besides being the actual implementers of SDI initiatives, people and various institutions in which they belong may be involved in various conflicts of interest susceptible to hinder the functioning of SDI's activities. Such conflicts could relate to legal matters (copyrights issues, data custodianship and ownership, intellectual property rights), commercial and economic factors or even national security reasons.

The two other SDI key components relevant to this dissertation namely, 'geoportal' and 'metadata and standards' are discussed separately in subsequent sections.

© University of Pretoria

2.4.2 Geoportals

As defined by the Open Geospatial Consortium (OGC), a geoportal is ‘‘a human interface to a collection of online geospatial information resources, including datasets and services” (Nebert *et al*. 2004). Quite a number of authors have assimilated geoportals to catalogue services within an SDI. Akinci and Comert (2007) regard such a view of geoportals to be incomplete as it reduces their role to that of catalogue services; which is to provide access to geospatial data. The definition of a geoportal goes beyond the one of a catalogue service in that a geoportal enables the consolidation of geospatial data of various types and formats from a variety of sources, technological infrastructures and geographic locations. In the same perspective, De Longueville (2010) refers to geoportals as being “World Wide Web gateways” for organising geospatial content and services.

Through the use of geoportals, standards that enable the interoperability of different disparate geospatial platforms are implemented to make the discovery of geospatial data and related services possible. Furthermore, geoportals provide means by which the gathered geospatial data can be visualised, evaluated and used through various services that parties involved in an SDI offer to potential consumers (users).

Authors such as Maguire and Longley (2005) classify geoportals as ‘application geoportals’ and ‘catalogue geoportals’. Application geoportals provide services such as web mapping, geoprocessing services or customised services tailored to satisfy a given target audience. Catalogue geoportals enable the discovery of geospatial data whereby different providers register their data and services with SDIs by supplying metadata description through an existing catalogue service interface within that given geoportal. This is referred to in the SDI Cookbook (Nebert *et al*. 2004) as ‘Distributed Catalogue Concepts’. These concepts involve the collections of distributed geospatial data (information) to be searched through the user interface of catalogue portals (gateways). The searching process is achieved through the use of a common vocabulary or metadata information with which the geospatial data (information) in question have been documented and registered with the SDI through the catalogue interface (which can also be referred to as ‘catalogue geoportal’ or ‘geospatial data catalogue’). The catalogue interface also provisions functionalities for geospatial data browsing in forms of maps and other means of information visualisation and geoprocessing. This can also be referred to as an application

© University of Pretoria

geoportal. For the sake of consistency and clarity, the terms: 'geospatial data catalogue' and 'catalogue portal', are referred to as 'geoportal' in this dissertation.

2.4.3 Architecture of a geoportal

A more descriptive way of understanding the functioning of a geoportal would be to define the components (objects, actors, operations and interactions) involved in a geoportal as illustrated in Figure 1.

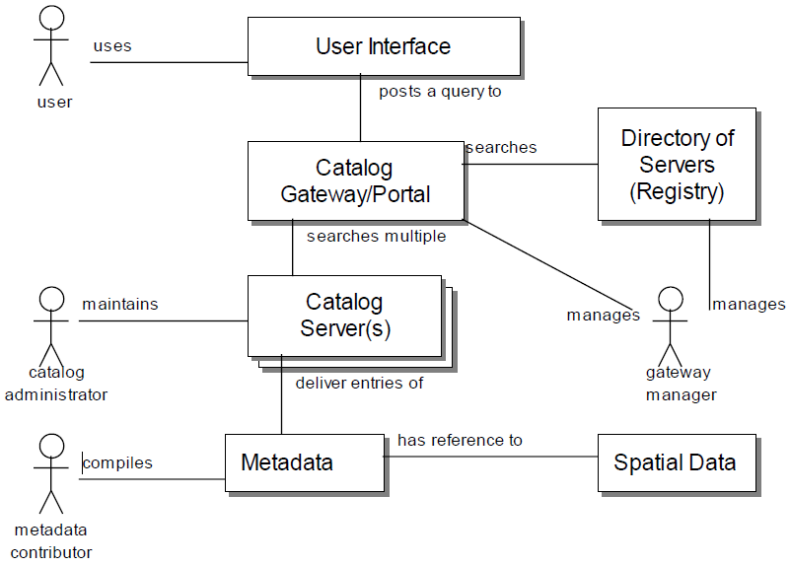

Figure 1: Geoportal architecture (Nebert et al. 2004)

The four main role players as described in the Figure 1 are:

- Metadata contributor: compiles metadata with reference to geospatial data.
- Catalogue administrator: assembles metadata as compiled by the metadata contributor, performs quality control as required by the organisation in question. Upon acceptance, the gathered metadata is pushed into the registered catalogue server as an entry ready to be searched by interested users.
- Gateway manager: has the role of hosting and maintaining searches originating from users distributed across different locations.
- Users: submit queries through the user interface to retrieve geospatial data or information that meet their search keywords or criteria. The stored metadata is searched through this process.

© University of Pretoria

Tait (2005) highlights the fact that geoportals are implemented based on distributed GIS technologies. The architecture of a geoportal as presented in Figure 1 illustrates such a fact. The concept of service-oriented architecture (SOA), which refers to a computer architecture that enables the provisions and sharing of software as services over computer networks, has been adopted as a practical way of developing and deploying geoportals. Through an SOA type of architecture, services are discovered following the Publish-Find-Bind pattern whereby a requestor (client) uses a geoportal (registry) client interface to find geospatial data or services bound to a specific (GIS) service provider. In Figure 2, the SOA concept applied in the context of geoportals is illustrated.

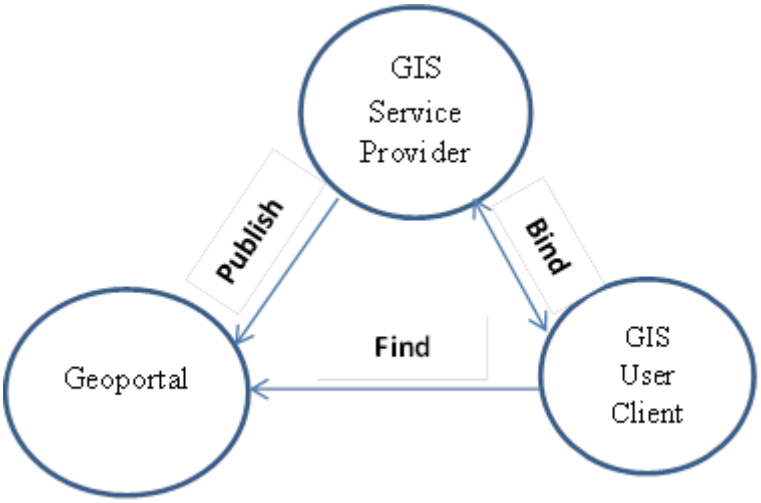

Figure 2: Publish-Find-Bind pattern applied to a geoportal (Nebert et al. 2004)

It becomes clear to understand how geoportals follow the SOA architecture in reading the definition of a geoportal by Maguire and Longley (2005). They define a geoportal as a web interface (site) linked to a database(s) of metadata information about geospatial data and services exposed and distributed using a web server(s) (or catalogue servers). Transactions between the web interface and various distributed catalogue web servers occur through messages encoded in eXtensible Markup Language (XML) carried over Hyper Text Transfer Protocol (HTTP).

The coupling between the client interface and the geoportal can be done in three different distributed configurations depending on the technological requirements of the concerned SDI(s) (refer Figure 3).

© University of Pretoria

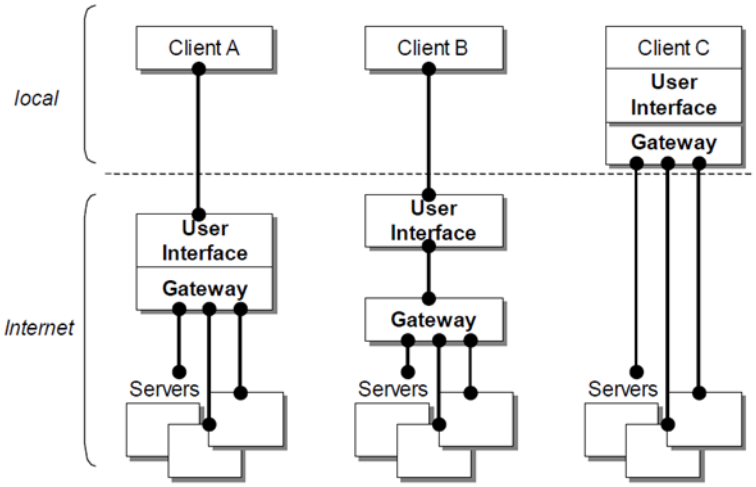

Figure 3: Different geoportal distributed configurations (Nebert *et al*. 2004)

As illustrated in Figure 3, with the first instance (client A), there is online access to the user interface which is attached to the geoportal and both the user interface and geoportal are combined in the same location. In the second instance (client B), there is online access to the geoportal by the client from the user interface sitting in a different location. In the last instance (client C), both the user interface and the geoportal are combined and are accessed locally while the server is sitting on a remote location.

The sequence 'client-user interface/gateway-servers' (refer Figure 4), describes a 'multi-tier software architecture' with a 'client tier', a middleware (or gateway tier) and a server.

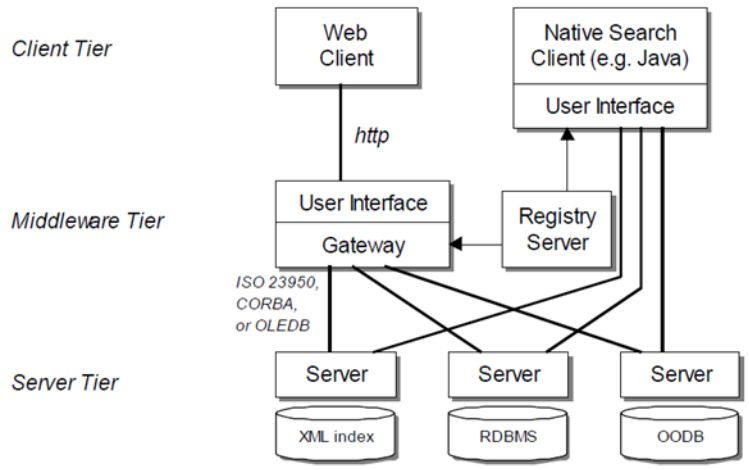

Figure 4: Multi-tier architecture applied to a geoportal (Nebert *et al*. 2004)

© University of Pretoria

The client is usually made up of a traditional web browser client application that communicates with the middleware tier (gateway tier) via (HTTP) messages. A customised (native) search client application can also be used in place of a web browser. In that case, communication can be done directly with the server using a protocol based on a specific standard such as *ISO 23950—Information retrieval (Z39.50)*. The gateway tier (middleware tier) enables a parallel search of multiple distributed (catalogue) servers based on HTTP POST OR GET requests that originate from the web client (web browsers). Hence, the need for managing these distributed catalogue servers through a directory of servers also known as registry. The Server tier is made up of one or more catalogue servers linked to repositories of metadata information structured using a specific spatial metadata standard such as the 'GEO' profile of *ISO 23950*. Such repositories could be XML documents files or databases, relational database systems containing structured spatial metadata information.

2.4.4 Current challenges with geoportals in terms of geospatial data discovery

It is relevant to mention that most of the existing geoportals implement the Catalogue Specification (CS) standard of the OGC namely the Catalogue Service for the Web (CSW) (Vockner and Mittlböck 2014). The CSW is an HTTP binding CS standard of the OGC that was designed for the purpose of enabling the discovery and retrieval of spatial data and services metadata (OCG 2007). However, the technological construct of the CSW does not make geoportals suitable for mainstream geospatial data discovery where web search engines are mostly used for data or information retrieval. Among other issues, Lopez-Pellicer *et al*. (2010) mentioned that the CSW catalogue specification is difficult to be crawled by web search engines. As a result, geospatial data get to be hidden behind geoportals that implement the CSW, and that is the case for the majority of currently existing geoportals. This makes geospatial data behind geoportals to be considered as part of the 'Deep Web' which is the term used to refer to online content that cannot be crawled by web search engines. To solve such a problem, Lopez-Pellicer *et al*. (2010) suggest the use of a toolkit (CSW2LD) that takes SDI metadata contained in CSW catalogues, to be published as Dublin Core metadata while conforming to the RDF data model and the Linked data principles. Unlike the approach taken by Lopez-Pellicer *et al*. (2010), the study described in this dissertation adopts a direct approach, which is, the use of plain HTML pages embedded with spatial metadata content and marked-up accordingly using the Dublin Core and Schema.org vocabularies. In doing so, this research also employs SEO techniques which are

© University of Pretoria

procedures known in mainstream information retrieval for enhancing the discovery of online resources (HTML pages) by web search engines (Malaga 2010).

2.5 Metadata standards

The word 'metadata' can be referred to as 'data about data'. As defined by the ISO Technical Committee 211, *Geographic information/Geomatics* (TC 211), metadata can be referred to as data about "the content, condition, quality, and other characteristics of data" (Aalders 2005). To ensure that the syntactic and semantic contexts of such information are kept and shared in an agreed upon structured way, the notion of metadata standards becomes relevant.

Once established, a geoportal needs to be known for it to act as a gateway. This is critical given that it needs to be effectively used by custom and potential geospatial data and information seekers. Uniformity in terms of the technologies used by the data providers, geoportal (SDI) implementers and the number of distributed end users is crucial. This is the reason why the implementation and adoption of suitable standards are useful in assuring interoperability between the technologies used by the different parties involved. The ultimate goal is to assure seamless sharing and dissemination of geospatial data and related services irrespective of the underlying technologies used.

Two international organisations that oversee the development of geospatial standards are the 'Open Geospatial Consortium (OGC)' and the 'International Organisation for Standardisation (ISO)' under its 'Technical Committee 211, Geographic information/Geomatics (TC 211)'. Some of the geospatial standards developed by these two organisations may be dependent on standards developed by the World Wide Web Consortium (W3C). Furthermore, computer software developed based on geospatial standards used to build SDIs (or geoportals) are constrained to also implement W3C standards to enable the sharing and dissemination of geospatial data and services over the internet to reach a wide audience.

The main geospatial standard relevant for the study described in this dissertation is:

- *ISO 19115:2003: Geographic information – Metadata*

The main non-geospatial standard relevant for this study is:

- Dublin Core

© University of Pretoria

Besides the two metadata standards mentioned, 'Schema.org', which is a collection of schemas for describing web page contents, is also considered in this study.

2.5.1 ISO 19115:2003, Geographic Information – Metadata

Spatial metadata facilitates the discovery of geospatial data and enables the assessment of such data for their fitness of use. Metadata constitutes "the primary means for locating geospatial data resources via the internet" (Govorov 2008). Such a statement adds to the value and importance of spatial metadata in facilitating and achieving the online discovery of geospatial data. The information contained in metadata standards needs to support the discovery of geospatial data, access to geospatial data, the fitness of use and the transfer of geospatial data.

ISO 19115:2003, Geographic Information – Metadata, is the standard that 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. The inception of *ISO 19115:2003, Geographic Information – Metadata*, came about to address the inconsistency and limitations in terms of interoperability that existed with national and regional standards for describing geographic datasets (Aalders 2005). Published as an international standard in 2003, *ISO 19115:2003, Geographic Information – Metadata* is considered as the best-known standard of the ISO 19100 family (Govorov 2008). *ISO 19115:2003, Geographic Information – Metadata* is also implemented in many OGC standards and ESRI products such as the Web Coverage Service (WCS), Web Feature Service (WFS), ArcIMS Image Service and ArcGIS. Furthermore, a number of existing national spatial metadata standards are profiles of *ISO 19115:2003, Geographic Information – Metadata* (Nogueras-Iso *et al*. 2004). This demonstrates how widely *ISO 19115:2003, Geographic Information – Metadata* standard is used.

Although *ISO 19115:2003, Geographic Information – Metadata* standard does not specify an encoding or exchange scheme for metadata, it provides guidance in terms of documenting digital geospatial data such as:

- Independent datasets
- Aggregations of datasets
- Individual geographic features and Attributes

© University of Pretoria

- Classes of objects that compose a feature

The Unified Modelling Language (UML) is used to describe the structure of *ISO 19115:2003 Geographic Information – Metadata* where metadata information is structured into UML packages. Each package is comprised of one or more entities (UML classes) which contain metadata elements (UML class attributes) that describe single metadata components.

The relationship between metadata entities is described in Figure 5. The *MD_Metadata* constitutes an aggregation of all the remaining entities contained in the standard. Within each of these entities, are metadata elements (attributes) which can be used to describe geospatial datasets.

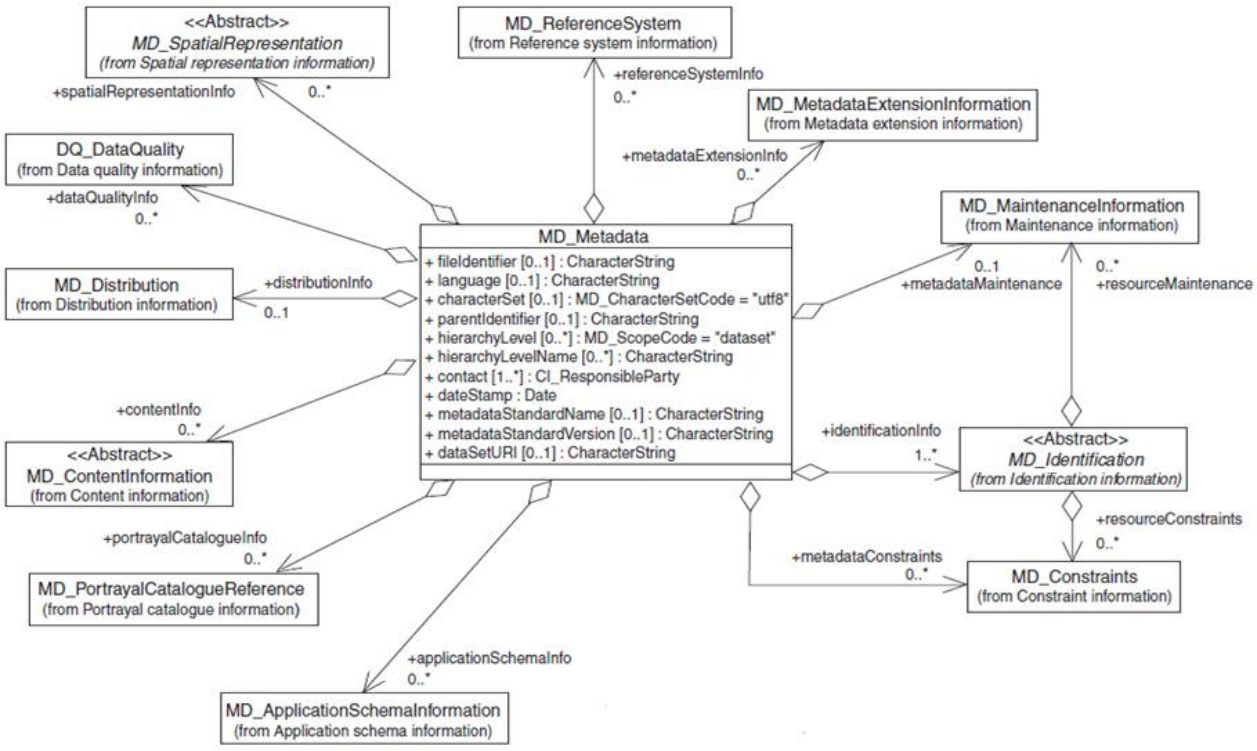

Figure 5: UML diagram describing Relationship between *ISO 19115:2003, Geographic Information – Metadata* entities (Source: ISO 19115: 2003)

For the purpose of this research, only core metadata elements are considered. This is because the core elements constitute a minimum set of elements suitable enough to identify a dataset. The core elements provide information on the topic (‘what’), the location (‘where’), the time stamp (‘when’) and the point of contact (‘who’) to enquire about the dataset (or data). The use of core

© University of Pretoria

elements simplifies the standard, eliminates ambiguity and enhances interoperability. Furthermore, it is recommended (by the standard *ISO 19115:2003, Geographic Information – Metadata*) that dataset metadata profiles include all the core elements.

Metadata elements of the standard (*ISO 19115:2003, Geographic Information – Metadata*) can be mandatory (M), mandatory under certain conditions (C), or optional (O). In Figure 6, a list of all the *ISO 19115:2003, Geographic Information – Metadata,* core elements together with the packages in which they respectively belong is provided.

Dataset title (M) (MD_Metadata > MD_DataIdentification.citation > CI_Citation.title)	**Spatial representation type** (O) (MD_Metadata > MD_DataIdentification.spatialRepresentationType)
Dataset reference date (M) (MD_Metadata > MD_DataIdentification.citation > CI_Citation.date)	**Reference system** (O) (MD_Metadata > MD_ReferenceSystem)
Dataset responsible party (O) (MD_Metadata > MD_DataIdentification.pointOfContact > CI_ResponsibleParty)	**Lineage** (O) (MD_Metadata > DQ_DataQuality.lineage > LI_Lineage)
Geographic location of the dataset (by four coordinates or by geographic identifier) (C) (MD_Metadata > MD_DataIdentification.extent > EX_Extent > EX_GeographicExtent > EX_GeographicBoundingBox or EX_GeographicDescription)	**On-line resource** (O) (MD_Metadata > MD_Distribution > MD_DigitalTransferOption.onLine > CI_OnlineResource)
Dataset language (M) (MD_Metadata > MD_DataIdentification.language)	**Metadata file identifier** (O) (MD_Metadata.fileIdentifier)
Dataset character set (C) (MD_Metadata > MD_DataIdentification.characterSet)	**Metadata standard name** (O) (MD_Metadata.metadataStandardName)
Dataset topic category (M) (MD_Metadata > MD_DataIdentification.topicCategory)	**Metadata standard version** (O) (MD_Metadata.metadataStandardVersion)
Spatial resolution of the dataset (O) (MD_Metadata > MD_DataIdentification.spatialResolution > MD_Resolution.equivalentScale or MD_Resolution.distance)	**Metadata language** (C) (MD_Metadata.language)
Abstract describing the dataset (M) (MD_Metadata > MD_DataIdentification.abstract)	**Metadata character set** (C) (MD_Metadata.characterSet)
Distribution format (O) (MD_Metadata > MD_Distribution > MD_Format.name and MD_Format.version)	**Metadata point of contact** (M) (MD_Metadata.contact > CI_ResponsibleParty)
Additional extent information for the dataset (vertical and temporal) (O) (MD_Metadata > MD_DataIdentification.extent > EX_Extent > EX_TemporalExtent or EX_VerticalExtent)	**Metadata date stamp** (M) (MD_Metadata.dateStamp)

Figure 6: Core elements of *ISO 19115:2003, Geographic Information – Metadata* (ISO 19115: 2003)

In appendix B, a table detailing the name of each core element, its definition, repeatability number (maximum occurrence), data type and domain (of possible supported values) is given. Such a table is essential in the mapping (crosswalk) process.

© University of Pretoria

2.5.2 Dublin Core

The Dublin Core provides a standard set of elements for describing and discovering digital resources on the internet (DCMI 2012). It has gained wide popularity due to its international and multidisciplinary scope. The Dublin Core is also an ISO standard described as a standard "for cross-domain resource description" with *ISO 15836:2009—The Dublin Core metadata element set* being the latest revision (ISO 15836:2009). Dublin Core does not provide any restriction in terms of what a resource might be. The Dublin Core's intended use is to facilitate 'cross-discipline discovery' and assuring interoperability across various disciplines of interest (Weibel 2005). Dublin Core metadata elements can be encoded using RDF/XML, XML and HTML 'meta' tag. The simplicity and ease of use combined with the generic nature of terms (elements) employed in the Dublin Core standard make it appropriate for the purpose of the study described in this dissertation. The existing mapping between Dublin Core and *ISO 19115: 2003, Geographic Information – Metadata* provides an evidence of the flexibility of the Dublin Core as a metadata standard for interoperability. Hence, in this study, the mapping between *ISO 19115:2003, Geographic Information – Metadata* and Dublin Core is used to embed spatial metadata into HTML web pages based on the elements of the Dublin Core (using the HTML 'meta' tag).

Dublin Core provides 15 elements that are part of a set of vocabularies managed by the Dublin Core Metadata Initiative (DCMI). Each of the Dublin Core metadata element has 10 attributes (or properties) of which six have fixed values and four have values that can change with respect to the resource being described (Green and Bossomaier 2002). Table 1 lists attributes of a Dublin Core element.

© University of Pretoria

Attributes with fixed values	**Fixed Value**
Version	1.1
Registration Authority	Dublin Core Metadata Initiative
Language	en (English, according to ISO639)
Obligation	Optional
Datatype	Character string
Maximum Occurrence	Unlimited
Attributes with variable values	**Value**
Name	e.g. Creator
Identifier	e.g. Creator
Definition	The agent(s) responsible for the content of the resource
Comment	Examples of a creator insert

Table 1: Attributes of the Dublin Core elements (Green and Bossomaier 2002)

A detailed description of the element set (15 elements) of the Dublin Core metadata standard is given in the table in appendix C where the name of each element, the URI that uniquely identifies each element in the namespace, the definition of each element and some additional comments are provided.

Parallel to the Dublin Core properties defined in the Dublin Core Metadata Element Set (DCMES) version 1.1, the DCMI released the 'DCMI Metadata Terms' document. This document provides a reference version of the fifteen DCMI elements for which formal domains and ranges have been defined. Domains and ranges determine the kind of values that can be assigned to a given element or property thereby enabling inferences to be made about the resource(s) being described (DCMI 2012).

Dublin Core metadata elements and contents need to be inserted into HTML web page to serve their purpose of describing online web resources. To do so previous and current versions of HTML support the 'Meta' tag as a means for incorporating metadata into HTML web pages. Multiple occurrences of the 'Meta' tag <meta> is allowed in order to contain all the necessary elements of the Dublin Core that are required to describe the HTML web page at hand. Attributes of the <meta> tag are described in Table 2 (Green and Bossomaier 2002).

© University of Pretoria

Attributes (of the <meta> tag	**Description (of the attribute)**
name	refers to the name of the metadata property, e.g. the author of the document
content	is the information specified in the name attribute, e.g. the author's name which is the value of the property or attribute
http-equiv	is related to the to the protocol for retrieving the HTML web page
scheme	refers to some external information that defines the value of the content of a given property (or attribute).
Lang	refers to the language of the content of the property (or attribute)

Table 2: Description of the 'meta' tag attributes (Green and Bossomaier 2002)

The Dublin Core metadata vocabulary is referenced in the in the HTML web page through the use of the '<link>' HTML tag. The example in Figure 7 illustrates how a few selected Dublin Core metadata elements are incorporated in an HTML web page.

© University of Pretoria

```
<!DOCTYPE html>
<html>
<head>
<meta charset="UTF-8">
<link rel = "schema.DC" href = "http://purl.org/DC/elements/1.1">

<meta name = "DC.Creator" content = "Samy Katumba">
<meta name = "DC. Title" content = "My HTML web page">
<meta name = "DC.Language" content = "en" scheme = "ISO639">
<meta name = "DC.Type content = "Web page">

<title>Title of the document</title>
</head>
<body>
Content of the document.....
</body>
</html>
```

Figure 7: Incorporating Dublin Core vocabulary into an HTML web page

As illustrated in Figure 7, Dublin Core elements are contained in the head section of the HTML web page being described.

2.6 Schema.org vocabulary

The need to enhance the discoverability of online web resources and web pages in particular by general purpose web search engines has led to the process of embedding structured metadata into HTML documents. Such a process involves structuring web pages contents meant for human consumption, to also be understood by machines (which in this case refer to web search engines). Shared vocabularies are used to mark-up contents of HTML web pages with metadata to make them understandable and easy to discover by web search engines. One such vocabulary is 'Schema.org' which is an initiative endorsed by Google, Bing and Yahoo search engines (Krutil *et al*. 2012). Schema.org provides a shared vocabulary that is generic and broad enough to accommodate popular concepts with the aim of enhancing the discoverability of web resources (or web pages) by web search engines. Microdata is the recommended mark-up language for embedding structured metadata into HTML documents using Schema.org vocabulary (Krutil *et al*. 2012).

© University of Pretoria

Schema.org defines a collection of schemas (or set of types) maintained in a hierarchical structure, starting with the type 'Thing'. The four proprieties of interest for type 'Thing' are given in Table 3.

Type: Thing		
Property	Data type	Description
Description	Text	a short description of the item being marked-up
Name	Text	a uniform resource locator (URL) of an image of the item being marked-up
Image	URL or ImageObject (type)	a uniform resource locator (URL) of an image of the item being marked-up
URL (uniform resource locator)	URL	the uniform resource locator (URL) address of the item being marked-up

Table 3: Properties of type 'Thing' (http://schema.org/)

All the other remaining types in Schema.org, including the ones considered in this study, descend from the type 'Thing'. The type (class) that is of particular interest in his study is the 'Dataset' type. This is because it contains generic properties that are suitable to describe a given dataset. Hence, properties of the 'Dataset' type are used to describe selected geospatial data considered for this research.

Details about selected properties and subtypes (subclasses) of type 'Dataset' relevant to this research are given in Table 4.

© University of Pretoria

Type: **Dataset**		
Property	**Expected Type/Data Type**	**Description**
Catalog	DataCatalog	a data catalog which contains a dataset
Distribution	DataDownload	a downloadable form of this dataset, at a specific location, in a specific format, e.g. mp4, shp (for shapefile)
Spatial	Place	the range of spatial applicability of a dataset, e.g. for a dataset of New York weather, the state of New York
Temporal	DateTime	the range of temporal applicability of a dataset, e.g. for a 2011 census dataset, the year 2011 (in ISO 8601 time interval format)

Table 4: Properties of 'Dataset' type (http://schema.org/)

The listed properties in Table 4 directly describe the Dataset type. However, the Dataset type itself is a subtype of type 'CreativeWork' which contains relevant properties that are necessary for describing a given dataset. The 'CreativeWork' type in turn inherits properties from super type 'Thing'.

The hierarchical structure of selected Schema.org types considered in this study is presented in Figure 8.

© University of Pretoria

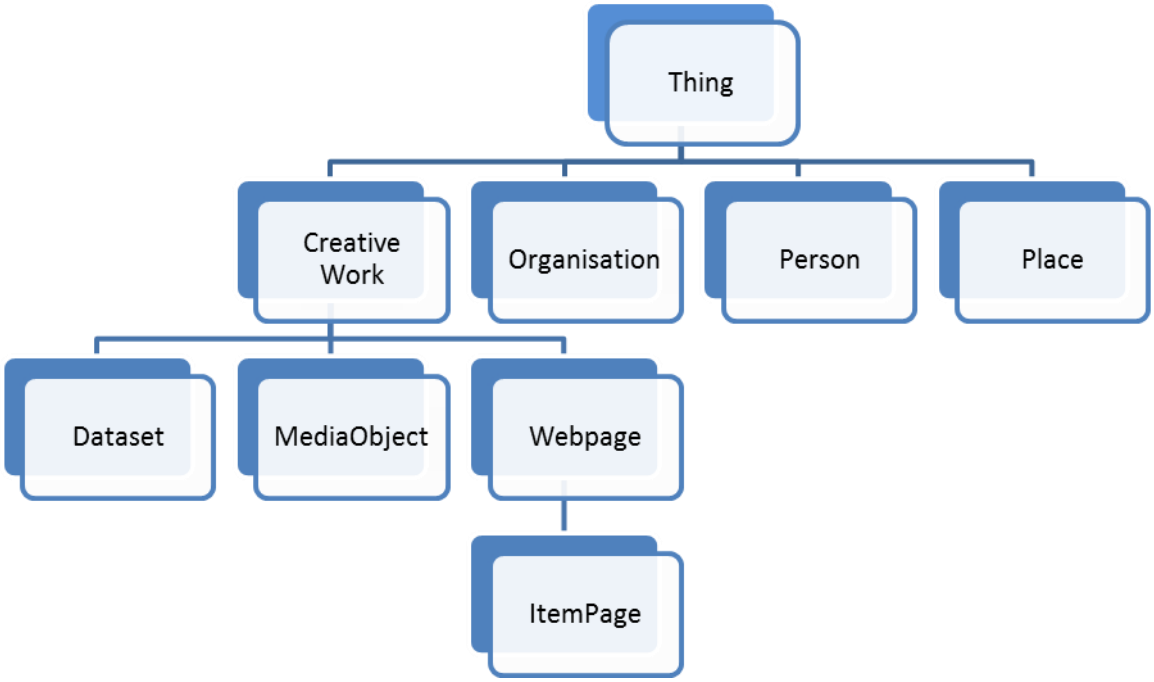

Figure 8: Hierarchical structure of selected Schema.org types (http://schema.org/)

In Appendix D, a table containing a detailed description of selected Schema.org types and properties relevant for the study described in this dissertation is provided.

To mark-up HTML documents with Schema.org vocabulary terms, major web search engines such as Google recommend microdata as the mark-up language of choice. HTML version 5 (latest version) which is used as the hypertext mark-up language for constructing web pages in this research, provides support for microdata.

Microdata defines groups of name-value pairs (properties) called items. To instantiate a new item within a given HTML page, the 'itemscope' attribute is used. To define a property for a given item type, the 'item prop' attribute is used. For each item (itemscope), the 'itemtype' attribute specifies the class or type in the vocabulary that provides the properties of the item being described. The code example in Figure 9 illustrates how microdata is used in conjunction with HTML 5 mark-up tags for encoding a web page document with Schema.org vocabulary.

© University of Pretoria

```
1<!DOCTYPE html>
2<html>
 3<body>
  4 <div itemscope itemtype="http://schema.org/ItemPage">
      5 <h2> <span itemprop="name"> Administrative Boundaries </span></h2>
      6 <div itemprop="about" itemscope itemtype="http://schema.org/Dataset">
          7 <p>Title:<span itemprop="name">Administrative Boundaries of South Africa </span></p>
      8 </div>
    9</div>
  10</body>
11</html>
```

Figure 9: Microdata code example

Line 4: The use of 'ItemPage' type specifies to web search engines that the resource (HTML web page) in question is a page specifically designed to describe a single item. In this instance, the item being described is a dataset.

Line 5: The property 'name' is used to refer to the name or title of the page which is: 'Administrative Boundaries'.

Line 6: The property 'about' is used to specify that the item page (page) for this resource (web page) is about a dataset. The 'itemscope' is used to define the type 'Dataset'.

Line 7: The 'itemprop' is used to specify the name or title of the dataset. In this instance, the name of the defined dataset is 'Administrative Boundaries of South Africa'.

The 'Dataset' type is nested inside the 'ItemPage' type. Both the 'ItemPage' and 'Dataset' types can inherit properties from 'CreativeWork' and 'Thing' types. A web page marked-up in such a manner contains semantic annotations that enable machines such as web search engines to understand its contents.

2.7 Web search engines, Search engine optimisation (SEO) and online web resources' visibility

A web search engine can be defined as a repository of indexed web pages that provides mechanisms of finding such pages through searching techniques. In order to ensure redundancy and fault tolerance, large web search engines rely on a geographically distributed farm of data centres made by clustering well-chosen computers. The choice of criteria is based on the price of

© University of Pretoria

the computers and the hardware capabilities such as the CPU speed, the size of the memory and the disc. Each computer or server within the data centre may perform a specific task such as indexing, crawling, query processing, result caching or inserting content to be advertised. Large search engines such as Google, Yahoo and Microsoft (Bing) index multiple thousands of web pages and respond reliably to billions of queries a day within less than a second (Hawking 2006). This is just an evidence of how popular such web search engines are and how many people rely on them to search for resources on the Web (Internet).

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 websites with web page contents changing constantly. Web resource visibility has always been at the forefront in the design and implementation of web search engines. When tuning web resources to improve visibility on the web, various techniques are adopted based on the type and nature of the information being shared and the business operations of the organisations authoring such resources and contents.

Often commercial organisations would want more visibility of their web resources by web search engines so that they can reach a wide audience in order to make their businesses successful. They are prepared to pay for web search engines to favour their resources by displaying them at the top of search engines' result pages. Organisations without the means of paying for such services are somehow disadvantaged (Zhang and Dimitroff 2005a). However, major web search engines such as Google provide other means of enhancing the visibility of web resources (websites) without their owners having to pay (Google 2014). Search results obtained in such a manner are categorised as 'organic'.

To improve the visibility of websites (web pages) in the organic search category, web search engine owners recommend the use of Search Engine Optimisation (SEO) techniques. The whole concept of SEO came about to improve the visibility of web resources. 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 (Malaga 2010). These techniques are meant to maximise the indexing of websites (pages) by search engines (Ochoa 2012, Swati *et al*. 2013, Onaifo and Rasmussen 2013, and Mustafa *et al.* 2015). Elements that impact on the visibility of web pages are related to the metadata structure of the page, its content and the number of hyperlinks pointing to the web page in question (Zhang and Dimitroff 2005a, Zhang

© University of Pretoria

and Dimitroff 2005b, Swati *et al*. 2013, Onaifo and Rasmussen 2013, and Mustafa *et al*. 2015). The tuning of these elements through SEO contributes to the visibility of web pages. However, web publishers can only tune some of these elements. The number of hyperlinks and the way in which users refine their search query terms are out of the web publisher's control (Malaga 2010). Nevertheless, users' web search behaviour in terms of search terms selection can be monitored with the aim of collecting (recording) the kind of keywords that they employ when searching information on a particular topic.

© University of Pretoria

Chapter 3

3 Related work

3.1 Introduction

Selected works from the literature that are related to this dissertation are discussed in this chapter. These include works related to:

- Users' web search behaviour with respect to search queries/search terms
- Taxonomy construction from search terms/keywords/keyphrases
- Search Engine Optimisation techniques
- Web search engine performance evaluation
- Mapping between metadata standards and vocabularies

3.2 Users' web search behaviour

One way of understanding users' web searching behaviour is to know their intent or the task they need to perform while using web search engines. Taxonomies have been described in the literature to classify users' web search intent (Broder 2002). Carmel *et al*. (1992), mentioned three types of web browsing:

- Search-oriented browsing: which occurs when the focus of the user is on searching information to accomplish a specific task
- Review browsing: which involves scanning in order to find information of a particular interest
- Scan browsing: which occurs when the user scans the search results with no intention of reviewing or integrating such information.

According to Sellen *et al*. (2002) the web information search process can be categorised as: - 'finding', 'information gathering', 'browsing' and 'transacting'. A taxonomy of users' web search intents that has been largely used in the literature is the one proposed by Broder (2002). Broder, (2002) categorised web search intent based on the types of queries users employ while using web search engines. The three classes that encapsulate users' web search intents according to Broder (2002) are:

© University of Pretoria

- Navigational: when the user's intent is to access a particular website of interest.
- Informational: when the user's intent is to acquire information of a particular interest while using the web search engine.
- Transactional: when the user's intent is to "perform some web-mediated activity".

Equally important, Rose and Levinson (2004) broadened Broder's taxonomy by including sub-categories and replacing the 'Transactional' category with the 'Resource' category. This latter category covers users' intent of acquiring something after the search, such as downloading a file. Such a task might not be of an informational or navigational intent in nature.

3.3 Modelling users' web search behaviour in mainstream information retrieval

Having established users' intent as a driver of web search behaviour, literature related to modelling such behaviours is explored in this section. Modelling users' web search behaviour explores how users interact with web search engines. Holscher and Strube (2000), provide an explicit model of how users interact with web search engines while seeking for information. They conducted two 'user experiments' and came up with the conclusion that web experience combined with domain-specific background, are essential in achieving a successful web search experience. Similar to the observations made by Holscher and Strube (2000), White *et al*. (2009) also affirmed that domain expertise influences user behaviour and success when querying and retrieving information using web search engines. This was concluded after analysing a large-scale log of web search behaviour by comparing textual queries of participants with domain-specific lexicons. However, Thatcher (2008) argues that besides web experience and domain-specific knowledge, cognitive search strategies also have an influence on how users interact with web search engines. Cognitive search strategies are related to different methods that web search engines' users employ while seeking for information. Such strategies include actions such as:

- Restricting oneself within familiar web environment using familiar web search engines,
- Using search terms as defined in the search task also known as 'Safe player' strategy,
- Opening more than one browser window to conduct different search at the same time (also known as parallel strategy),
- Giving exclusive reliance on hyperlinks to move between pages (also known as 'Link-dependent strategy'),

© University of Pretoria

- Using specific search terms with the intention of landing to the desired result web page (also known as 'To-the-point' strategy),
- Bypassing a web search engines altogether to going directly to the known web page where the information being sought is expected (also known as 'Known address' strategy)
- Using search terms that may not be related to the search task at hand with the intention of reasoning through one's understanding of the problem domain (also known as 'Deductive reasoning' strategy).

From the work of Thatcher (2008), it can be inferred that domain-specific knowledge influences some of the cognitive search strategies such as 'safe player' strategy or 'Deductive reasoning' strategy where search terms (queries) reformulation may be required.

While most of the studies described above have been focused on modelling users' web search behaviour before and during the search experience, Agichtein *et al*. (2006), developed a method for modelling the 'post-search' user behaviour with the aim of forecasting users' preferred search results. The parameters (or features) that feed into the model apply for each considered 'query-URL' pair. These parameters are comprised of:

- Query-text features: features related to the query text are employed to model occurrences of similarities between, for example, the user's query terms with the web page title or snippet text (short text highlighting the content of the web page displayed in the web search engine result lists).
- Browsing features: features related to how the user interacts with the web search engine interface. This includes, for example, the amount of time (dwell time) user spends on a web page for a given query for example.
- Click-through features: features related to click-through provide the number (frequency) of clicks made by the user for a given query-URL pair.

Given that search engines perform keywords matching, a closer look at the terms making up the query formulation is relevant for modelling users' web searching behaviours. One way of capturing users' web search behaviour is through understanding the types of keywords (search terms) they combine in the formulation of their queries. A number of studies present results on

© University of Pretoria

the monitoring of users' behaviour in terms of their query terms' selection. As a response to the lack of personalised web search engines' results from similar query terms submitted by different users, Sugiyama *et al*. (2004), propose methods (models) for aligning search results according to respective single user's need without their active participation in the process. In their proposed model based on collaborative filtering and monitoring individual user's browsing history, Sugiyama *et al*. (2004) explicitly include a component that measures the frequency of terms employed by users. This provides an indication of web pages' content that are relevant to a specific user. Such information gets to be automatically incorporated into the user's profile at each time the user initiates a web search session.

After revising different methods such as the categorisation of user's queries (topic-based representation) and the suggestion of both queries and Uniform Resource Locators (URLs) for personalising users' web search results, Sontag *et al*. (2012) propose a method for enhancing personalised web search results based on user profiles for modelling long term users' web search behaviours based on keywords (search terms) extracted from the Bing web search engine server log. This came after observing that in most studies, user's profiles for personalisation are constructed based on words and terms identified in browsers' search history or servers' logs. In most instances, a method such as TF-IDF (term frequency–inverse document frequency) has been used in these approaches. As such, Sontag *et al*. (2012) devised a "probabilistic framework for personalisation" that enables the estimation of the topic in which a user performs the web search based on the corresponding query, the user profile as well as the calculation of a personalised ranking from previous knowledge gathered about the user and other suitable information.

Shen *et al*. (2005) state that capturing users' behaviour by digging in browsers' cached history or server logs to identify frequent terms for inferring search interest of individual users (profiles), may not be an adequate model to respond to immediate short term information need. As such, they propose a method for modelling the immediate context in which individual users perform their search thereby enhancing web search engines results in terms of ad hoc queries. In the model proposed by Shen *et al*. (2005), while the user performs the search, search terms of the current query are adapted based on similarities with immediate previous query terms and

© University of Pretoria

corresponding search results. Search result pages that have been already known or viewed are re-ranked to accommodate the user's immediate information need.

3.4 Modelling users' query terms through classification or taxonomy

Sanchez and Moreno (2004) developed an automatic method for constructing a taxonomy of web resources that are representative of a domain of interest based on keywords (or key phrases). Sanchez and Moreno (2004) classified websites based on keywords that frequently appear in the pages of such websites. They also employed 'polysemy detection' and 'semantic clustering' to deal with polysemous words and synonyms. In another study, Liu *et al*. (2012) proposed an automatic approach for constructing taxonomies from a set of keywords instead of a text corpus. Liu *et al*. (2012) note that not only keyphrases (or keywords) are easy to obtain, they also allow to easily characterise a domain. Similarly, Saracevic and Kantor (1997), developed a taxonomy for understanding the value of using library and information services. Their taxonomy was derived based on the analysis and classification of vocabulary and terminology employed by users of library and information services who participated in the interviews. Saracevic and Kantor (1997) argued that their taxonomy satisfied the 'grounded theory approach' which suggests that models should be constructed from the ground up using raw data.

A study that is closely related to the research described in this dissertation is the one by Chuang and Chien (2002) who proposed an automatic way of constructing a taxonomy of search query made of a hierarchy of clusters. A given cluster within the hierarchy constitutes a categorisation (classification or grouping) of similar users' queries and represents a domain (or concept) under which a given search query was made. Queries that triggered the retrieval of similar web contents (web pages) are grouped in clusters. Chuang and Chien (2002) argue that such a taxonomy enables in-depth analyses of domain specific terminology (vocabulary) as well as an understanding of users' search interests. In a similar study, Limam *et al*. (2010) construct a taxonomy of search terms by making use of the semantic (meaning) of terms from a search query log in order to assign such terms into clusters.

Concepts involved in the construction of a taxonomy of search terms are essential for the purpose of the study described in this research as it entails the classification of search terms (keywords or keyphrases) obtained from the experiment phase. Since participants are specifically requested to search for geospatial data on the web, it is assumed that characteristics of geospatial

© University of Pretoria

data are reflected in their search terms. Hence, clusters or categories of the taxonomy are defined accordingly to contain these search terms. Some novelty brought in this research is that the taxonomy is mainly constructed with the purpose of understanding the kind of search terms users employ when they search for geospatial data using general purpose web search engines.

3.5 Modelling users' behaviour in geographic information retrieval (GIR)

Geographic information retrieval (GIR) systems are specifically designed to index, retrieve and visualise geo-referenced information, that is, information related to specific geographic locations (Bordogna *et al*. 2012). As in mainstream information retrieval, queries used against the information retrieval system are crucial for retrieving sought after contents and resources. Zhang *et al*. (2014), referred to spatial keyword queries as queries containing spatial information for finding spatial objects that meet location conditions.

In the field of geographic information retrieval, efforts towards easing the discovery of geographic information can be noticed as well. As described by Bordogna *et al*. (2012), information retrieval within GIR systems can be achieved by employing techniques used in general purpose web search engines with no consideration in terms of spatial indexing and reasoning. Such Techniques include the application of Natural Language Processing (NLP) part-of-speech tagging for recognising geographic names. Other techniques used in GIR include the use of spatial relationships from the texts and queries, data mining and machine learning, and the use of geographic ontologies and gazetteers.

To address issues related to the ambiguity, synonymy and homonymy of the geographic names (terms) and the vagueness of terms (queries) related to spatial conditions (relationships), Bordogna *et al*. (2012), developed a GIR system named 'Geo-Finder' in which a model for identifying, disambiguating 'geo' or 'non-geo' terms and the coding of geographic reasoning through contextual heuristic rules, was implemented.

Similarly, Rauch *et al*. (2003) employed confidence-based disambiguation coupled with methods for estimating measures of relevancy to a user's query in order to tackle the problem of ambiguity when searching for geoinformation. Rauch *et al*. (2003) identified two kinds of relevance: 'traditional textual query relevance' which is common in information retrieval, in general, and 'geo-relevance' which has a spatial dimension (e.g. the name of a place). The term

© University of Pretoria

'geo-relevance' can be textual depending on the emphasis it carries in the document (e.g. it can be in the title or header, or it can be in bold). Rauch *et al*. (2003), note that a balance between the two relevancies greatly influences the ranking of outcome search results. The methods used by Rauch *et al*. (2003) for disambiguation use techniques related to NLP, such as 'local document context' to draw the context of geographical terms based on the other terms used in the same user's queries. Data mining and the use of domain knowledge repositories were also employed. These techniques and methods were implemented within a GIR system called 'MetaCarta' which provides a geographic text search interface.

Clough (2005) describes 'geo-parsing' and 'geo-coding' as methods for extracting geospatial information from web pages using a GIR system called SPIRIT. Geo-parsing refers to the process of finding geographic references in documents. Geo-coding consists in associating geographic references with corresponding spatial coordinates. The indexing of web pages is achieved through the extraction of spatial references ("spatial footprint") and textual keywords contained in query terms. Clough (2005) argues that geographic information such as geographic references contained in web documents can be used to enable spatial awareness in information systems.

In an attempt to improve geographical search also known as 'local search', Gan *et al*. (2008), analysed a set of 36 million queries and classified them (queries) into two categories (geo and non-geo queries) and suggested a taxonomy for categorising geographic search queries. The approach taken by Gan *et al* (2008) is similar to the one adopted in this study as it seeks to understand how users search for spatial information (in this case geospatial data) by analysing the search query terms they employed.

After investigating users' behaviour in terms of how they reformulate their geographic queries (spatial keyword queries), Aloteibi (2014) suggests that studies related to query reformulation be tailored to specific types of queries. This is because such queries (e.g. geographic queries) exhibit their own properties as compared to general query terms. The assertion made by Aloteibi (2014) relates to the context of the study described in this dissertation as it focuses on analysing search query terms employed by users with the intent of searching and downloading geospatial data.

© University of Pretoria

The approach adopted by Bordogna *et al*. (2012) of developing a complete information retrieval system from scratch as mentioned in this section, differs from the one adopted in this study which suggests the use of general purpose web search engines for searching spatial information (in this case geospatial data). With the endorsement of 'Schema.org' by well-known search engines such as Bing and Google, the approach followed in this research is justified in terms of dealing with the issues related to the ambiguity, synonymy and homonymy of not only geographic terms but also terms related to the thematic context of the search. Schema.org allows the marking-up of HTML web pages with shared concepts in the form of types (schemas) thereby enabling the embedding of semantics in the content of such web pages for machines' (i.e. web search engines) consumption.

3.6 Application of SEO techniques

There is currently little work in academic literature that discusses SEO. The majority of work on SEO techniques resides in non-academic platforms such as blogs, online discussion forums, websites (belonging to web search engines owners and others) and anecdotes (Malaga 2010). The four main SEO processes or techniques as mention by Malaga (2010) include:

- Keyword research: consists of establishing a list of keywords (terms or key phrases) in order to determine their popularity, level of competition for each term and the commercial viability of such terms. In terms of keywords popularity, search engines like Google provides a tool that generates a list of keywords with their frequency of use (or average search volume). A search engine optimiser needs to insert keywords from his or her predefined list of keywords in the tool in order to obtain a list of popular keywords (key phrases or terms) similar to the ones in his or her list. To determine the level of competition for a particular keyword, term or key phrase, a search engine optimiser may need to calculate the ratio of the number of results to the number of searches over a given period of time for that particular keyword. The ideal situation is to determine keywords that have a high number search volume and lower number of corresponding results.
- Indexing: consists of enhancing the process of getting a website (web pages) to be indexed by web search engines. Contrary to search engines owners' recommendation of submitting or registering websites (web pages) to their search engine directories, Malaga (2010) advises against such a practice and state that "it appears that major search engines

© University of Pretoria

prefer discovering a new site". Instead, Malaga (2010) suggests that search engine optimisers make use of user-generated content platforms such as blogs and wikis, to have links directly pointing to their websites. Furthermore, search engine optimisers should also make use of online bookmarking applications (e.g. Delicious, Digg, and Stumble upon) to bookmark their web pages with tags (terms or keywords) relevant to the content of their respective web pages. In the context of the experiments performed in the study described in this dissertation, the constructed web pages are individually submitted to (registered with) web search engines. This is because web search engines such as Google and Bing provide mechanisms for tracking the registered web pages so as to provide information in terms of how HTML web pages are retrieved and ranked on search engine results pages.

- On-site optimisation: consists of procedures and techniques that the search engine optimiser can employ on his or her website (web pages) to enhance its visibility to web search engine crawlers. These include customisation of content and the use of appropriate HTML tags such as title tag, meta description tag, H1 tag and bold text tag. Keywords that appear in the title and in the contents enhance the visibility of web pages by search engines' crawlers. With respect to how the scanning of web page contents happen, search engines such as Google employs 'Latent Semantic Indexing (LSI)' which is a natural language processing technique that analyses the pattern in the combination of words of a document (in this case web page), to derive concepts (subjects or themes) that may be associated with the contents of that particular web page being scanned. This shows that search engines do not limit themselves to keywords matching. The semantics (meanings) of words or phrases around keywords also have an impact in terms of determining the relevance of a given web page against users' queries. It can be noted that on-site optimisation as described by Malaga (2010) mainly relates to the design of website (web page). In that perspective, Onaifo and Rasmussen (2013) suggests to libraries to consider creating their webpages in formats that can be crawled, indexed and ultimately be searchable by web search engines. Onaifo and Rasmussen (2013) went further to suggest that existing invisible online library resources be replicated (duplicated) in formats that make them suitable for being discovered by general purpose web search engines. This approach is similar to the approach suggested in this dissertation to consider creating

© University of Pretoria

HTML web pages with content information about geospatial data so as to enhance the findability of geospatial data on the web.

- Off-site optimisation: consists of procedures beyond the search engine optimiser's control. These includes backlinks which are links from other websites (web pages) pointing to the target web pages. Google, in particular, assigns a 'Page Rank (PR)' value to the target web pages based on backlinks. As such, SEO techniques suggest that incoming links from sites (web pages) with higher PR would improve the ranking of the target website (web pages). Internal links also account in the overall PR of a website (web pages).

Among a few works in scientific literature that explicitly tested the concept of search engine optimisation (SEO), feature two studies (part I and part II) done by Zhang and Dimitroff (2005a); Zhang and Dimitroff (2005a). They evaluated the impact of both web page contents and metadata implementation on the visibility of web pages by web search engines. In part I of their study, Zhang and Dimitroff (2005a) designed a set of web pages by removing or adding keywords in the title and/or content (full-text) of individual web pages depending on the characteristics being explored. Web page ranking positions of individual web pages on search engine results pages over a period of 21 weeks was used as a factor in the statistical analysis in order to establish a comparison. They concluded their study by stating that the frequency of keywords appearing in the title, the content (full-text) of a web page and in both the title and the content, enhances web pages' visibility by web search engines. In part II of their study, Zhang and Dimitroff (2005b) evaluate the impact that the frequency of keywords appearing in the 'meta' tags (metadata) has on the visibility of web pages by web search engines. Individual web pages were designed with varying frequency of keywords in the content (full-text) of the web page, metadata title field, and the metadata subject field. Sets of web pages with and without metadata elements were submitted on a public web server to be crawled by web search engines. The ranking positions of individual web pages in web search engines results pages over a period of 21 weeks served as a factor on which statistical analyses were performed. The study found that the use of metadata elements improves the visibility of web pages by web search engines. The 'subject' metadata element has a more positive impact in the visibility of web pages. Such results can be enhanced if the keywords appearing in the subject 'meta' tag also appears in the title and content of web pages. The study by Zhang and Dimitroff (2005a & 2005b) is relevant to

© University of Pretoria

the study described in this dissertation since it (this dissertation) seeks to evaluate the impact that two metadata vocabularies namely Dublin Core and Schema.org, have on the visibility of web pages with spatial metadata contents by web search engines.

In a more recent study, in an effort to enhance the findability of online library resources on the web, Onaifo and Rasmussen (2013) performed statistical tests to examine the SEO assumption which suggests that the number of links pointing to a target website has an influence on the ranking of that particular website on search engines results pages. They also statistically examined the relationship existing between the total number of indexed web pages by web search engine within a website and the total number of daily page views that such website received. Ranking positions and reputation (backlinks) results were obtained from Alexa.com website, and the number of web pages indexed and the number of daily pages viewed were obtained from Google. Their findings showed an existing correlation in the mentioned relationships. That is, the higher the number of backlinks, the lower (better) the ranking of that particular website (web page) on Alexa.com. Furthermore, the higher the number of indexed web pages, the higher number of daily page views. These findings led Onaifo and Rasmussen (2013) to recommend that SEO techniques be considered in the design and creation of online library resources to improve their findability on the internet by web search engines.

Based on an empirical study, Ochoa (2013) tested the effectiveness of selected SEO techniques on the rankings of web pages on the search engine results pages of Google web search engine. SEO techniques considered by Ochoa (2013) include:

- Topic website research
- Domain name selection and website setup
- Keyword research (in the research reported in this dissertation, a taxonomy of search terms was constructed in order to study the kind of keywords users employed to search for geospatial data on the web).
- Data collection (using Google Analytics tool). In the study conducted in this research, the Webmaster tools application of both Bing and Google were used to collect the data (average raking positions)

© University of Pretoria

- On-page SEO strategies implementation: This included the use of 'title tag', 'heading tags', 'image alt attribute' and 'description metadata tag', optimisation of URLs, the use of 'robots.txt' and sitemaps. Some of these SEO strategies are also employed in this research.
- Off-page SEO strategies implementation: This strategy involved the process of getting backlinks (external links pointing to the website or webpage in question).

According to Ochoa (2013), the application of these SEO techniques improved the visibility of his tested web pages on the search results pages of Google by noticing an increase in the number of visitors and page views. The experimented website increased in its ranking position to No. 6 in the search results for the keyword 'beneficios del alpiste' (the target keyword).

Wang *et al.* (2011), grouped SEO techniques into four categories to test the effectiveness of these techniques on the number of index pages, page views, reach and page view per user collected from 'Alexa.com' website and computed page interest as measures for quantifying the retrieval of websites (web pages) on the search result pages of web search engines. The four categories of SEO techniques as considered by Wang *et al.* (2011) are:

- Structure optimisation: which includes a customisation of the '404 Error pages', layer number for web directory and page size of the web pages.
- Keywords optimisation: which includes keyword density.
- Content optimisation: which considers the length of the web page's title.
- Link optimisation: which considers the overall links (internal and external) of the website under test.

Wang *et al*. (2011), performed a regression analysis to examine the correlation existing between these selected SEO techniques and the considered measures. They noticed that 'overall links', 'page size' and 'customised versions of 404 Error pages' had a significant positive effect on page views, reach, number of independent IP (internet protocol) addresses and page interest. They also noticed that the size of web pages and overall links had a significant positive effect on indexed pages. Lastly, Wang *et al*. (2011) observed that the length of the title, the number of layers for a web directory and the size of web pages had a significant positive effect on page views per user.

© University of Pretoria

One of the major events that influenced SEO is the adoption of Schema.org which provides a generic set of vocabularies to annotate text content of HTML web pages thereby instilling some semantics into web pages. Major search engines like Google, Bing, Yahoo and Yandex have all endorsed and supported Schema.org vocabulary (Krutil *et al*. 2012). The ultimate goal is for web page contents marked-up with Schema.org to be easy to understand by search engine crawlers. Schema.org provides different themes (types) under which web page contents can be categorised. Such categories include products, people, organisations, places, events, cooking recipes, etc., in which web page contents can be classified. The embedding of vocabularies in the contents of these pages is done through encoding standards (mark-up languages) such as 'microformat', 'RDFa' or 'microdata'.

Muhleisen and Bizer (2012), provide statistics on the use of encoding standards and some details on the types of published data encoded using encoding standards. Muhleisen and Bizer (2012) analysed two corpora of extracted web data encoded in microdata, microformat and RDFa from the Common Crawl web corpus which is the largest corpus of web data made available through the Web Data Commons project. The bulk of web data were HTML pages of which 148 million were extracted from September 2009 to September 2010, and 189 million extracted in February 2012 were structured data (HTML pages encoded with the three encoding standards). Although the results provided by Muhleisen and Bizer (2012) revealed an increased popularity in the use of RDFa and Microformat, the frequency of use of microdata greatly increased from 56, 964 (in the period 2009/2010) to 26, 929 865 in 2012. The two authors attributed such an increase to the fact that microdata is the recommended encoding standard for Schema.org vocabulary. In terms of the types of data (content of the web pages being annotated), both microdata and RDFa were used to annotate embedded data about people, organisations, media files, commercial products. Data about geographic information (i.e. addresses and geographic coordinates) were the least frequently annotated type of data. The results also show that 8% of the 2012 corpus were 'Geodata' annotated with microdata.

A practical example of the use of microdata is given in the study by Ambiah and Lukose (2012) who implemented a tool for automatically annotating web page contents with Schema.org vocabulary. Ambiah and Lukose based their experiment on annotating web pages about patent information. They modelled patent information by providing a formal RDF representation of

© University of Pretoria

patent metadata. Given the fact that Schema.org terms are also represented using RDFS, a basis for matching patent metadata terms with their corresponding Schema.org terms was possible. The study by Ambiah and Lukose (2012) is similar to the study described in this dissertation in that, instead of patent information, HTML web pages are annotated with information about geospatial data (spatial metadata) with Schema.org vocabulary using microdata. A formalisation of geospatial information (metadata) is provided through the use of both *ISO 19115:2003, Geographic Information – Metadata* and Dublin Core elements (or terms) which are matched to corresponding Schema.org terms. Unlike in Ambiah and Lukose (2012) and similar studies in the literature that are limited in applying Schema.org vocabulary and microdata on HTML web pages, the study described in this dissertation goes further in evaluating the retrieval performance (or effectiveness) of the marked-up HTML web pages.

3.7 Related literature on the evaluation of web search engines

A study done by Spiteri and Richard (1999) ascertains that the approach taken in the evaluation of the effectiveness of internet search engines is similar to the one used in Information Retrieval Systems (IRSs). This was concluded after reviewing 31 studies that focused on evaluating web search engines. Spiteri and Richard (1999) noticed that the relevance of the items (web pages) retrieved was among the criteria for evaluating the effectiveness of internet search engines as it is the case with IRSs. Measures for determining the retrieval effectiveness include ‘precision’, ‘recall’, ‘relevance’ and ‘searcher satisfaction’. Two methods used in evaluating web search engines inherited from IRSs include ‘experimental methodology’ when experts submit queries to the search engines and ‘operational methodology’ when common users (novices) submit queries to search engines. Spiteri and Richard (1999) argue that the experimental methodology may work well for IRSs such as library systems given that they are controlled systems, have a target audience and also possess a defined vocabulary for easing the search. As such, the experimental methodology may not be appropriate for internet search engines since they are uncontrolled, have a variety of audiences and have no defined vocabulary. However, Spiteri and Richard (1999) noticed that a number of studies appear to adopt the experimental methodology mostly suited for evaluating the retrieval effectiveness of IRSs.

Different approaches of measuring the retrieval effectiveness of web information retrieval systems are also reviewed. These include a study by Vaughan (2004) who proposes a continuous

© University of Pretoria

ranking measure in the evaluation of the relevance of web pages reflected in the precision and recall measures that are employed in most web search engine performance evaluations. Vaughan (2004) argues that instead of determining relevance solely based on a binary decision indicating whether a web page is relevant or not, a continuum of measures estimating the level of relevance of web page results from the most relevant to the least relevant should be adopted.

In a more recent study that provides an overview of different methods used in web search engines evaluation, Ali and Sufyan Beg (2011) classified relevant studies into eight groups given as follows:

1. Relevance based evaluation
2. Ranking based evaluation
3. User satisfaction based evaluation
4. Size/coverage of web-based evaluation
5. Dynamics of search results based evaluation
6. Few relevant/known item based evaluation
7. Specific topic/domain based evaluation
8. Automatic evaluation

The three methods that are of interest to the study described in this dissertation are ‘relevance based evaluation’, ’specific topic/domain’ and ‘ranking based evaluation’. This is because sets of web pages with spatial metadata information content are submitted to two search engine (Bing and Google) databases. These web pages are considered as relevant and their positions on the web search engine results pages are evaluated.

Given the predominance of precision and recall as measures for evaluating the relevance of web search engine results, an understanding of the two measures is appropriate in the context of the study described in this research. The two measures were first introduced in the Cranfield studies as measures of comparing search results from information retrieval systems (Croft *et al.* 2009). Recall refers to the ratio of relevant retrieved documents to all relevant documents present in the

© University of Pretoria

information system. Precision is the ratio of relevant retrieved documents to all retrieved documents (Croft *et al*. 2009). One of the ways of using precision and recall is to rank the retrieved results in order to compute the two measures at different specified ranking positions. This enables search results from two or more different search engines to be compared. However, such a comparison can only be possible provided that the settings of the experiments and the data used are fixed. In this manner, performance evaluation experiments are fair and repeatable. Recall and precision measures at different defined ranking positions for a given query can be summarised by computing an average to obtain a single number that summarises the search engine performance for that particular query. It usually happens that a set of queries is used during an experiment. In such cases, if for example precision is the measure of interest, the mean average precision (MAP) which is the average of all the individual average precision for each query in the set, is computed. The application of such a measure for the purpose of comparing the retrieval effectiveness of different search engines may require the involvement of some relevance judgement procedures (Croft *et al*. 2009). Kumar and Prakash (2009) for example evaluate the precision and recall of Google and Yahoo web search engines against a set of 15 queries for which the mean precision and the mean relative recall measures were used to compare the retrieval effectiveness of the two web search engines. Human judgement was used to determine the relevance of the search results.

Another method for computing recall or precision measures is through the use of statistical significance tests with sets of comparable search results from the concerned web search engines. A number of studies from the literature used this approach in terms of comparing retrieval performance results of web search engines. A number of such studies are mentioned by Ali and Sufyan Beg (2011). In a study that compared two methods, an automatic and manual one (done by humans) for determining the relevance of web search results, Can *et al*. (2004) performed an analysis of variance (ANOVA) based on precision and recall measures obtained from eight different web search engines. Precision and recall results at different predefined document cut-off values (ranking positions) from a set of 25 queries against each of the eight considered web search engines were compared using different statistical significant tests (e.g. ANOVA and Turkey's highest significant differences (HSD)). Similarly, among the studies that are of particular interest in this dissertation in terms of employing statistical significance tests, are the two studies by Zhang and Dimitroff (2005). In part II of Zhang and Dimitroff (2005b)'s work, a

© University of Pretoria

set of 46 web pages were posted and submitted to 19 search engines. These submitted pages differed from each other based on keyword duplication in the Title, Full-text, Metadata title field, Metadata subject field and Metadata description field. The number of these web pages appeared varied across the different web search engines. In other instances, none of the web pages appeared in some of the considered search engines. Zhang and Dimitroff (2005b) employed three statistical techniques to test their proposed hypotheses. The three statistical tests were: one-way ANOVA, two-way ANOVA, and independent-sample T-test. Their assumptions were that the involved dependent variables (which are the positions of retrieved web pages in search engines) are normally distributed, the population variances of the dependent variables are the same and the values of the dependent variable are independent of each other. However, Zhang and Dimitroff (2005) did not explicitly conduct statistical tests to validate their assumptions of normality in the distribution of the positions of web pages in the search engine results lists.

3.8 Application of Dublin Core

The Dublin Core Metadata Initiative (DCMI) has received criticisms over the years due to its inadequacy for describing complex resources and lack of detailed criteria in terms of how elements should be used in the context of local (custom) specific applications. Many of such criticisms hailed from the library community (Harper 2010). A criticism that is of concern to the study described in this dissertation is the fact that major web search engines do not consider Dublin Core metadata embedded in web pages (Beall 2008). Such an assertion has been neither explicitly confirmed nor denied by web search engine owners. For example, Google supports certain HTML meta tags elements such as title, description and charset (Google Webmaster tools 2014). Seldom scientific studies have been done to test whether web search engines consider Dublin Core metadata elements. However, a study by Zhang and Dimitroff (2004) statistically compared the visibility by web search engines of two sets of web pages: a first set containing web pages without metadata, and a second set containing web pages with Dublin Core metadata elements. The hypothesis under test was “there is no difference for search engine with respect to search engine visibility performance between the two groups; that is web pages where metadata elements are not implemented and web pages where the metadata elements are implemented”. Seven web search engines were considered: AllTheWeb, EntireWeb, Google, Lycos, AltaVista, Infospace and Yahoo which is now powered by Bing (Sullivan 2009). The statistical results showed that there was a difference in web search visibility between the two sets of web pages.

© University of Pretoria

Web pages with Dublin Core metadata outperformed those without metadata except for Altavista where the difference was not significant. However for AllTheWeb, the reverse was true; web pages without metadata outperformed those that implemented Dublin Core metadata. The study by Zhang and Dimitroff (2004) shows how different web search engines' retrieval algorithms approach the concept of metadata which in this particular instance refers to Dublin Core.

3.9 Mapping and interoperability between standards

Nogueras-Iso *et al*. (2004) highlight the fact that the concept of standard mapping makes it possible for communities of users of one standard to gain access to contents or resources described in another standard. To keep the semantic interoperability between metadata standards, two main approaches are adopted: (1) the use of ontologies which consists of creating relationships between the vocabularies of the different metadata standards involved, or (2) the creation of specific crosswalks for one-to-one mapping. Nogueras-Iso *et al*. (2004) argue that the latter approach (crosswalks) is more appropriate for metadata standards with syntaxes and semantics that are restricted and fixed such as Dublin Core or a geographic information system metadata standard namely *ISO 19115:2003, Geographic Information – Metadata*. As defined in the Dublin Core Metadata Glossary, a crosswalk is a table that maps the relationships and equivalencies between two or more metadata formats (DCMI 2001; Nogueras-Iso *et al*. 2004). The construction of crosswalks which in this context also refer to the mapping of metadata standards (vocabularies) is said to facilitate search engines to effectively search online web resources kept in different repositories. As such, crosswalks enable interoperability (Nogueras-Iso et al. 2004).

It is of great importance to study the processes involved in the construction of crosswalks as it enables: (1) the understanding of the existing mapping between the elements of Dublin Core metadata standard and *ISO 19115:2003, Geographic Information – Metadata* standard, (2) the construction of crosswalks (mapping) of vocabularies (elements or terms) between:

- Dublin Core metadata standard and Schema.org, and
- Categories of the taxonomy search terms and corresponding elements of *ISO 19115: 2003, Geographic Information – Metadata* standard.

© University of Pretoria

The processes involved in the creation of crosswalks are (Nogueras-Iso *et al*. 2004):

1. Harmonisation: Although the majority of metadata standards employ relatively identical properties in the way that the content of their elements is defined, such properties, however, are not expressed in a fixed way. For example, different data types or multiplicity could be employed for elements with similarly defined properties with respect to their content. Hence, some harmonisation is necessary by creating tables detailing the semantics of the elements of the concerned metadata standard or vocabulary. Such tables should include: an assigned number (or identifier) of the elements with respect to the level of each element in the hierarchy in the metadata standard, the name assigned to each element in the metadata standard, mandatory and multiplicity restrictions defined on each element by the metadata standard, specification of each element's semantics and data types.
2. Semantic Mapping: This process involves the mapping of elements from one metadata standard to their respective semantically equivalent elements in another metadata standard. It is considered as being the most important step in the establishment of crosswalks. A table containing the semantics of the corresponding elements is usually produced during this step.
3. Additional rules for metadata conversion: Besides the table providing the semantics of corresponding elements, conversion specifications need to be described as well. Such a specification should contain some of the transformations that need to occur in order for the produced crosswalks to be in conformity with both the source and the target metadata standards. Additional rules for metadata conversion are related to:
    - Data conversion: A metadata standard may define specific data types, the range of values, a controlled vocabulary or free texts for every single element in the metadata standard. As such, specific rules need to be established in order to deal with the mismatches that may exist between the expected content values of the corresponding metadata elements. When mapping between elements with different controlled vocabularies, it is advisable to approach the mapping on a 'one to one' relationship basis (between elements).
    - Element to element mapping: A metadata standard may define properties of an element to be repeatable or non-repeatable, mandatory or optional. A one-to-one

© University of Pretoria

map can occur between elements with identical properties. A one-to-many mapping occurs when an element in the source metadata standard gets to be associated with two or more elements in the target metadata standard. A many-to-one map occurs when more than one element from the source metadata standard are associated with only one element in the target metadata standard. However, rules must be specified in other to define how the grouping of the elements is established, especially if all the elements of the source metadata standard are considered. If the mapping is done with only a unique element from the source metadata, rules must also be specified in terms of dealing with the loss of information. When extra elements in the source metadata standard exist and do not have equivalent elements in the target metadata standard, rules must be established in terms of catering for extensions or additional information provided that the target metadata standard possess such capabilities (or properties). In the case of unresolved mandatory elements in the target metadata standard where mandatory elements do not have mapping elements in the source metadata standard, rules in the crosswalk must be provided with respect to having these elements filled with suitable values.

- Hierarchy: Given that the majority of metadata standards are implemented in a hierarchical structure, differences in the hierarchy of elements involved in the mapping between standards must be taken into consideration during the creation of crosswalks.

4. Mapping implementation: This last step consists of applying the mapping tables as described in the steps above. Nogueras-Iso *et al*. (2004) used eXtensible Style Sheet Language (XSL) to implement the crosswalk between Dublin Core and *ISO 19115:2003, Geographic Information – Metadata*. This was more appropriate given the fact that metadata standards use eXtensible Mark up Language (XML) as a representation format. And XSL is used for the manipulation and transformation of documents in XML format. However for the purpose of the research described in this dissertation, the proposed crosswalk (between Dublin Core and Schema.org) will be achieved only at the level of mapping tables.

© University of Pretoria

Chapter 4

4 Research methodology

4.1 Introduction

This chapter provides the methodology adopted to achieve the aim and objectives assigned in this dissertation. The adopted approach, the research design, experiment setups as well as the instruments used to collect the results are explained. Lastly, an overview of the analyses performed in this dissertation is provided. Some limitations and assumptions considered in this research are also mentioned in this chapter.

4.2 Approach

This research evaluates the effectiveness of two metadata vocabularies in terms of retrieving HTML web pages with *ISO 19115: 2003, Geographic Information – Metadata* content. Two sets of web pages are considered: one set of HTML web pages is described with the elements of the Dublin Core metadata standard and the other set of HTML web pages is described with properties of types defined in Schema.org vocabulary. In order to achieve this aim, steps illustrated in Figure 10 in accordance with the objectives as described in section 1.4 (chapter 1) are followed.

© University of Pretoria

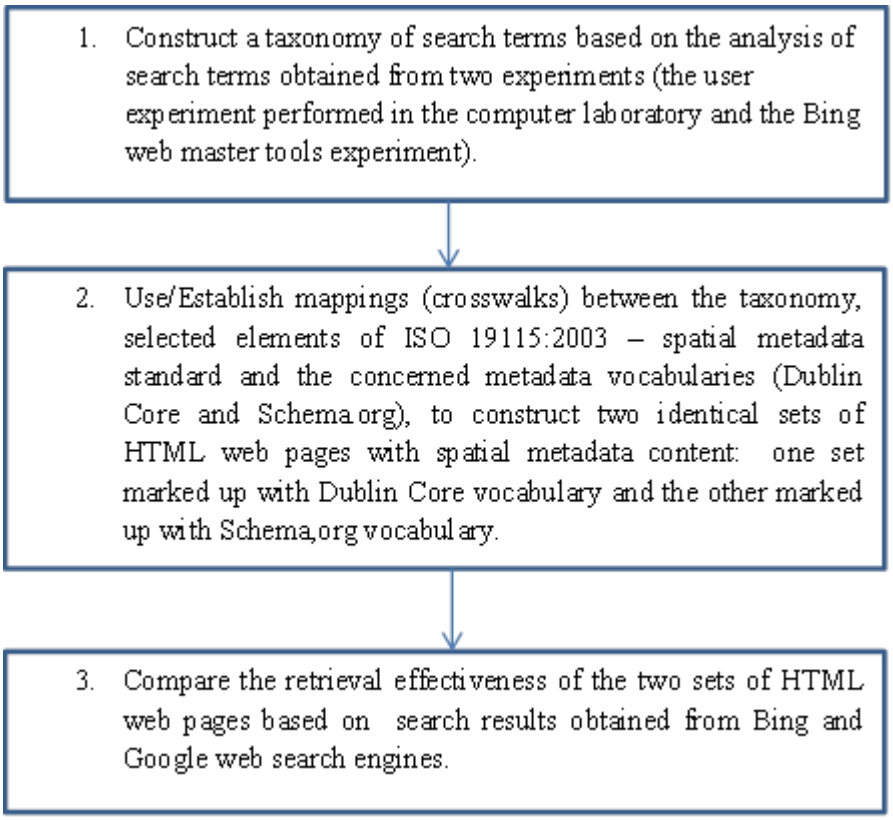

Figure 10: Research approach

Categories of the taxonomy obtained after analysing the experiments' results were mapped to equivalent elements of *ISO 19115:2003, Geographic Information – Metadata* standard. This was done given the fact that the search terms collected display characteristics or properties of geospatial data. Such properties can only be adequately described using an appropriate spatial metadata standard. *ISO 19115:2003, Geographic Information – Metadata* is a metadata standard specifically designed for describing geospatial data.

After the mapping of the constructed taxonomy with corresponding *ISO 19115:2003, Geographic Information – Metadata* standard elements, the existing mapping between the elements of *ISO 19115:2003, Geographic Information – Metadata* and Dublin Core was used to construct HTML web pages that were marked-up with the Dublin Core metadata vocabulary. Based on the methodology employed by Nogueras-Iso *et al*. (2004) described in section 3.9 (chapter 3), a mapping (crosswalk) between the Dublin Core metadata standard and selected types of Schema.org was created. The obtained crosswalk informed the construction of HTML web pages that were marked-up with Schema.org vocabulary. In essence, this also

© University of Pretoria

implies that search engine optimisation techniques were being employed. The constructed HTML web pages from the two sets were uploaded to a web server exposed on a public domain (internet) so that the web pages can be crawled by web search engine robots. The web server was registered with both Google and Bing via their respective webmaster tools.

4.3 Research design

The three main steps (as described in Figure 10) were achieved following these research techniques:

- Extended literature reviews
- Experiments
- Statistical analyses
- Interdisciplinary research

Given that this research is interdisciplinary in nature, extended literature reviews were necessary in order to understand the underlying relevant concepts involved. An overview of the nature and properties of geospatial data was reviewed. Such a review facilitated the process of understanding how geospatial data are documented for the purpose of enabling their online discovery. Hence, *ISO 19115:2003, Geographic Information – Metadata* standard as a widely used metadata standard that guides the documentation of geospatial data was studied. It was also necessary to review literature related to the use and applications of Dublin Core metadata and Schema.org vocabularies. This is because these two vocabularies were involved in the mappings (crosswalks) that guided the design and construction of HTML web pages with spatial metadata content. The approach that guided the mapping process itself involved the creation of crosswalks or 'one-to-one' mappings. This approach is described in details by Nogueras-ISO *et al*. (2004). It is said to be suitable for the mapping between metadata standards (or vocabularies).

Two types of experiments were conducted in this research. The first one, also referred to as user experiment, was a more controlled experiment conducted in a computer laboratory involving human subjects (referred to as participants). Participants were instructed to search for geospatial data using general purpose web search engines of their choice. The aim of this experiment was to collect search terms used by participants for the purpose of constructing a taxonomy of search terms. The constructed taxonomy of search terms provides an understanding of how users search

© University of Pretoria

for geospatial data online with respect to the keywords (search terms) they employ. The second experiment served two purposes:

- The collection of search terms that triggered the appearances of HTML web pages on the search results pages of Bing. The Bing webmaster tools online application was used to collect these search terms.
- The collection of results related to the retrieval effectiveness of HTML web pages (with spatial metadata content) registered with the webmaster tools applications of both Bing and Google web search engines.

Given that the submitted HTML web pages contained information about geospatial data (that is, spatial metadata), search terms that triggered the appearance of these web pages on the search results pages of Bing, were assumed to reflect how geospatial data are searched on the web using web search engines. Hence, the analysis of search terms obtained from the Bing (via the Bing webmaster tools), was compared to the analysis of search terms collected from the user experiment performed in the computer laboratory. The idea was to find some commonalities in terms of how search terms from the two experiments fit the different categories of the taxonomy. It made sense to do such a comparison in order to evaluate the consistency of the taxonomy.

Some interdisciplinary research techniques were applied in this study whereby concepts like SEO techniques and mappings (crosswalks) between metadata standards borrowed from the field of Information Science (specifically the area of online information retrieval) were applied in the field of Geoinformatics. This is evident given that such concepts were used in the construction of HTML web pages (web resources) with content (spatial metadata) describing geospatial data with the intent of improving the discovery of geospatial data by a wide audience. Statistical techniques were also employed in the analysis of results obtained from the experiments.

4.4 Research instruments

4.4.1 First experiment: User experiment

Three groups of users participated in the experiment: 1) a group of 21 BSc Geoinformatics first year students from the University of Pretoria (UP); 2) a group of 17 BSc Geoinformatics students in their final year (third year) of studies, also from UP; and 3) a group of nine South African professionals in the field of geoinformatics. These three groups represent three different levels of

© University of Pretoria

expertise in geoinformatics. This choice was made with respect to the study by Holscher and Strube (2000) who concluded that user expertise in the knowledge domain of the search influences success in web searches. Furthermore, a user-centred methodology for constructing the taxonomy is justified by the fact that users are the ones driving online searches when they search for geospatial data. The experiment was repeated with each group on a different day and lasted for two hours. Participants were instructed as follows: "Search for spatial data to address a particular need (or to solve a given problem) of your choice for specific countries or cities in Africa. One should be able to visualise (and/or analyse) such spatial data on a GIS desktop software (in e.g. ArcGIS or QGIS)." They were advised to use any knowledge they had acquired throughout their studies and/or profession. The keywords used in their search queries were monitored and collected using the Mozilla Firefox web browser cache. A detailed document describing instructions given to participants in the user experiment is given in Appendix A.

4.4.2 Second experiment: Bing and Google webmaster tools experiment

The second experiment involved the preparation of HTML web pages with spatial metadata contents. These HTML web pages were then registered with Bing and Google web search engines via their respective webmaster tools online applications. The second experiment served two main purposes which are:

1. The analysis of search terms collected from the Bing webmaster tools application provided a means for comparison with the analysis of search terms from the first experiment (user experiment). Such a comparison enabled the evaluation of the consistency of the taxonomy.
2. The results on impressions (appearances of web pages on the search results pages of web search engines) and average appearance position of the submitted web pages provided a means for evaluating and comparing the retrieval effectiveness of web pages submitted to both Google and Bing web search engines. It is important to note that these web pages have as contents, spatial metadata information in *ISO 19115:2003, Geographic Information – Metadata* standard format. They were marked-up with selected Dublin Core tags and microdata tags with Schema.org vocabulary respectively.

The webmaster tools applications for both Bing and Google web search engines provided reports on page traffic, search engine optimisation (SEO), crawl information and search keywords that

© University of Pretoria

triggered appearances of HTML web pages on their search results pages. As research instruments in this study, the webmaster tools applications (for both Bing and Google) were used to collect results related to web pages appearance on the search results pages of the two web search engines (Bing and Google).

It is important to highlight that unlike the user experiment which was a more controlled experiment where participants were instructed in terms of what to search. There was no such control over the experiment where the webmaster tools of the two web search engines (Bing and Google) were used. Compared to the results obtained from the user experiment performed in the computer laboratory, there was no control in terms of:

- The kind of queries that triggered the appearance of HTML web pages on the search results pages of Bing and Google,
- The frequency (number of times) Bing or Google was used,
- The number and background of people who submitted their search terms (search queries) to Bing or Google to retrieve HTML web pages on the search results pages of the two search engines.

4.5 Research analysis

The process of constructing the taxonomy involved a qualitative and a quantitative analysis of search terms collected from the experiments. Search terms were obtained from both the user experiment conducted in the computer laboratory and the Bing webmaster tools experiment. The qualitative analysis performed, entailed a subjective exercise whereby search terms dissected from search queries were assigned to appropriate categories of the taxonomy. One of the challenges in the creation of the taxonomy of search terms is the inconsistencies that may arise in the assignment of terms to appropriate categories. Some of these challenges are discussed in the work of Chuang and Chien (2002). In this study, a comparison between the analysis of search terms from the two experiments (user experiment and Bing webmaster tools experiment) was performed in order to evaluate the consistency of the constructed taxonomy.

To compare the retrieval effectiveness between the set of HTML web pages marked-up with Dublin Core and the other set of HTML web pages marked-up with Schema.org vocabulary,

© University of Pretoria

descriptive and inferential statistics analyses were employed and are reported in chapter 7. The following were determined:

- The frequencies of appearance of HTML web pages (with spatial metadata contents) in the search results pages of the concerned web search engines (Bing and Google) were analysed.
- The total number of retrieved web pages for each of the two sets was computed and compared.
- The 'recall' measure, which is the ratio of the number of web pages retrieved to the total number of all the web pages uploaded and registered with the server, was also computed and used as the basis for comparison.
- Bing and Google via their webmaster tools online applications provided the average position of individually retrieved web pages on the search results pages. The lower the position number, the more visible (on average) is that particular web page on the search results pages.

In order to determine whether there was a statistical difference in the average positions (ranks) in the two sets of results (that is, HTML web pages marked-up with Dublin Core vocabulary and those marked-up with Schema.org vocabulary), the Wilcoxon Signed Ranks test was employed. The reasons being:

- The null hypothesis of the test is that both sets of ranks have identical data distributions. That is, each page has a similar rank independent of the annotation vocabulary used. Rejecting this null hypothesis means that the use of an annotation vocabulary has an impact in the distribution of ranks. If so, additional analyses were performed in order to establish which of both vocabularies is the most appropriate.
- Given that not all the submitted HTML web pages appeared on the search engines' results pages (of Google or Bing), to allow for the 'one to one' pairing of average positions between HTML webpages, HTML web pages that did not appear, were assigned an average position value of 1000. This assumes that such page appear at the very lowest (worst) position on that particular search engine's results pages.

© University of Pretoria

Chapter 5

5 Analysis of search terms and taxonomy construction

5.1 Introduction

This chapter provides details on a taxonomy constructed from search terms that users employed when searching for geospatial data using general purpose web search engines. Results on the two types of experiments that led to the construction of this taxonomy namely the user experiment conducted in the computer laboratory and the Bing webmaster tools experiment are also discussed in this chapter. Parts of this chapter were published in Katumba and Coetzee (2013), and Katumba and Coetzee (2015).

5.2 Analysis of search terms

After completion of experiments, an analysis of the collected search terms was performed to create categories of search terms for the taxonomy. Hence, the constructed taxonomy is comprised of categories (classes) of search terms. The frequency of search terms was considered because it facilitates the creation of clusters of search terms to make up a category. The results of the analysis of the first experiment were compared to the results of the second experiment as a way of evaluating the consistency of the taxonomy. The experiments were conducted in English only. Since this experiment focussed on how users search for data described using metadata, the reliability and validity of the actual data (geospatial data) were not evaluated.

Table 5 provides definitions for terminology used in the analysis of search terms obtained from the experiments.

	Definition
Search term	Refers to a whole string of alphanumeric characters entered by a participant (web searcher), e.g. ‘namibia’ or ‘shapefile’.
Query	Is an occurrence of a single search term or a combination of more than one search term submitted to the web search engine. e.g. ‘namibia shapefile’.
Session	Refers to the duration of a web search experience of a single user.

Table 5: Terminology

© University of Pretoria

5.3 Categories of the taxonomy

The construction of the taxonomy was informed by the hierarchy of different levels of granularity of metadata described in *ISO 19115:2003, Geographic Information – Metadata* standard (Figure 11):

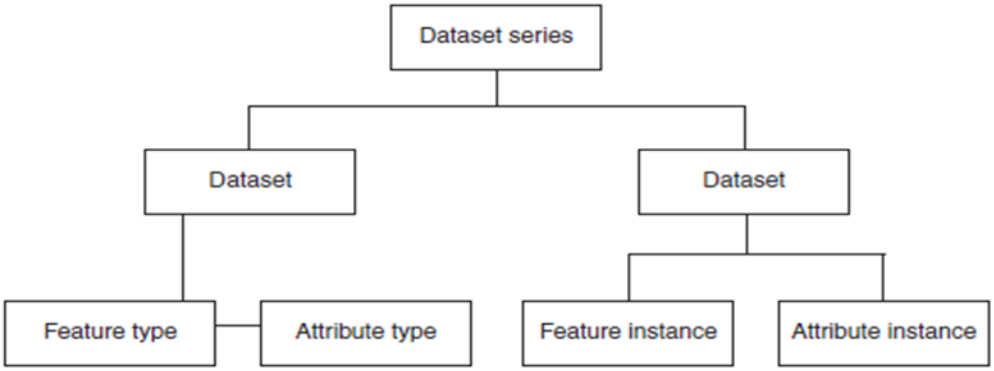

Figure 11: Hierarchy of metadata (Source: ISO 19115:2003)

- Dataset series: a collection of spatial data with similar characteristics of theme, source date, resolution, and methodology, for example, a collection of raster map data captured from a common series of paper, or a collection of vector datasets depicting surface hydrography with associated attribution for multiple administrative areas within a country.
- Dataset: a consistent instance of a spatial data product which may be composed of a set of identified feature types and instances, and attributes types and instances.
- Feature type: are feature constructs with common characteristics, for example, all bridges within a dataset.
- Feature type instance: features (spatial constructs) that have a direct correspondence with a real-world object, for example, The Nelson Mandela bridge in Johannesburg.
- Attribute type: a digital parameter that describes a common aspect of grouped spatial primitives (i.e. 0-dimensional (points), 1-dimensional (lines), 2-dimensional (polygons) and 3-dimensional geometric objects). For example, overhead clearance associated with a bridge.
- Attribute instance: a digital parameter that describes an aspect of a feature instance, for example, the overhead associated with a specific bridge across a road.

© University of Pretoria

This hierarchy (as described in Figure 11) defines different views of metadata to describe a dataset, feature type, instance of a feature or instance of an attribute at specific levels of detail. The motive behind adopting this approach was to determine how keywords employed by users relate to a feature type, attribute type, instance of a feature type or instance of an attribute, that is, how users' keywords (search terms) relates to the model representation of geospatial data in their primitive constructs: Vector (as point, line and polygon) and Raster, with both of these representations accompanied with their corresponding attributes as descriptions.

Each query was broken into individual terms and each term was placed within a suitable category. The results suggest that web search terms employed by participants when searching for geospatial data using general purpose web search engines can be classified into four main categories (or clusters), which are defined based on selected core elements of *ISO 19115: 2003, Geographic Information – Metadata* standard as follows:

- Application Domain: This category defines the subject, the theme of the general topic of interest for which the geospatial data is being searched. Query terms (keywords) that fall under this category could be related, for example, to infrastructures, urban planning, water resources topics or subjects for which a given user is searching. *ISO 19115: 2003, Geographic Information – Metadata* elements associated with this category are 'Dataset topic category', 'Dataset title', 'Abstract describing the dataset' and 'Lineage'. Search terms falling under the Application Domain category fit one of these four metadata elements.
- Location: This category defines the spatial extent or geographic coverage for which the spatial data in question is being searched. Query terms (keywords) relate to location, place names or location types or classes. Examples are 'Africa', 'Johannesburg', 'Paris', 'Sub-Sahara', 'South America', etc. as place names and 'city', 'country', etc. as location types. This definition also includes land demarcations, such as municipalities or district. *ISO 19115:2003, Geographic Information – Metadata* elements associated with this category are 'Geographic location of the dataset'. The standard provides for the location to be described by four (geographic) coordinates or by a geographic identifier. None of the search queries from the experiments contained geographic coordinates. Search terms

© University of Pretoria

falling under the Location category, fit the 'EX_GeographicDescription' metadata element.

- Feature Type: This category defines the actual geographic object, feature or phenomenon of interest to the user performing the search. Examples of search terms that fall under this category include roads, buildings, bridges and power lines. The *ISO 19115:2003, Geographic Information – Metadata* element associated with this category is 'Spatial representation type'.
- Data model: This category relates to the data representational model that spatial data employ. Spatial data can be represented as vector or raster. Under the vector model, query terms (keywords) display characteristics of primitive geographic features (i.e. point, line or polygon). A common search term that falls under this category is 'shapefile', referring to the well-known vector (spatial) data representation model. The term 'raster' is used when searching for continuous geographic features (e.g. rainfall, temperature and forests). The *ISO 19115:2003, Geographic Information – Metadata* element associated with this category is 'Distribution format'.

As an illustration, Table 6 shows how a sample of four search queries were dissected and categorised. For example, 'Ghana land cover shape file' has five search terms, two terms ('land' and 'cover') fall under the 'Application Domain' category, one term ('Ghana') falls under the 'Location' category, none of the terms falls under the 'Feature type' category and two terms ('shape' and 'file') fall under the 'Model' category. It should be noted that although the terms 'shape' and 'file' may refer to a single term that is 'shapefile', in this instance, the two terms were counted separately to maintain consistency when counting search terms (refer to Table 5). The 'Ignored' column shows the number search terms that were not classified under any of the four categories. Examples for such terms include adjectives, pronouns, conjunctions, adverbs and verbs, as well as words such as 'free', 'download', 'spatial', 'data', etc.

© University of Pretoria

Search terms	Total	Ignored	Category			
			Application Domain	Location	Feature Type	Data Model
'Ghana land cover shape file'	5		2	1		2
'uganda administrative boundaries shapefile'	4		1	1	1	1
'tanzania infrastructures shapefiles'	3		1	1		1
'south african roads and shapefiles'	5	1		2	1	1

Note: Search terms are included as entered by users, with proper nouns in lower case.

Table 6: Search term categorisation

5.4 User experiment results

The analysis results for each of the groups of participants in the user experiment, as well as the overall result, are provided in Table 7 and in Figure 12. These results are discussed in subsequent sections.

Group	Total number of search terms	Ignored	Category			
			Application Domain	Location	Feature Type	Data Model
First years	866	510 (59%)	76 (9%)	150 (17%)	59 (7%)	71 (8%)
Third years	363	107 (29%)	76 (21%)	90 (25%)	35 (10%)	55 (15%)
Professionals	201	93 (46%)	44 (22%)	37 (18%)	18 (9%)	9 (5%)
Overall	1430	710 (50%)	196 (14%)	277 (19%)	112 (8%)	135 (9%)

Table 7: User experiment results – First years

© University of Pretoria

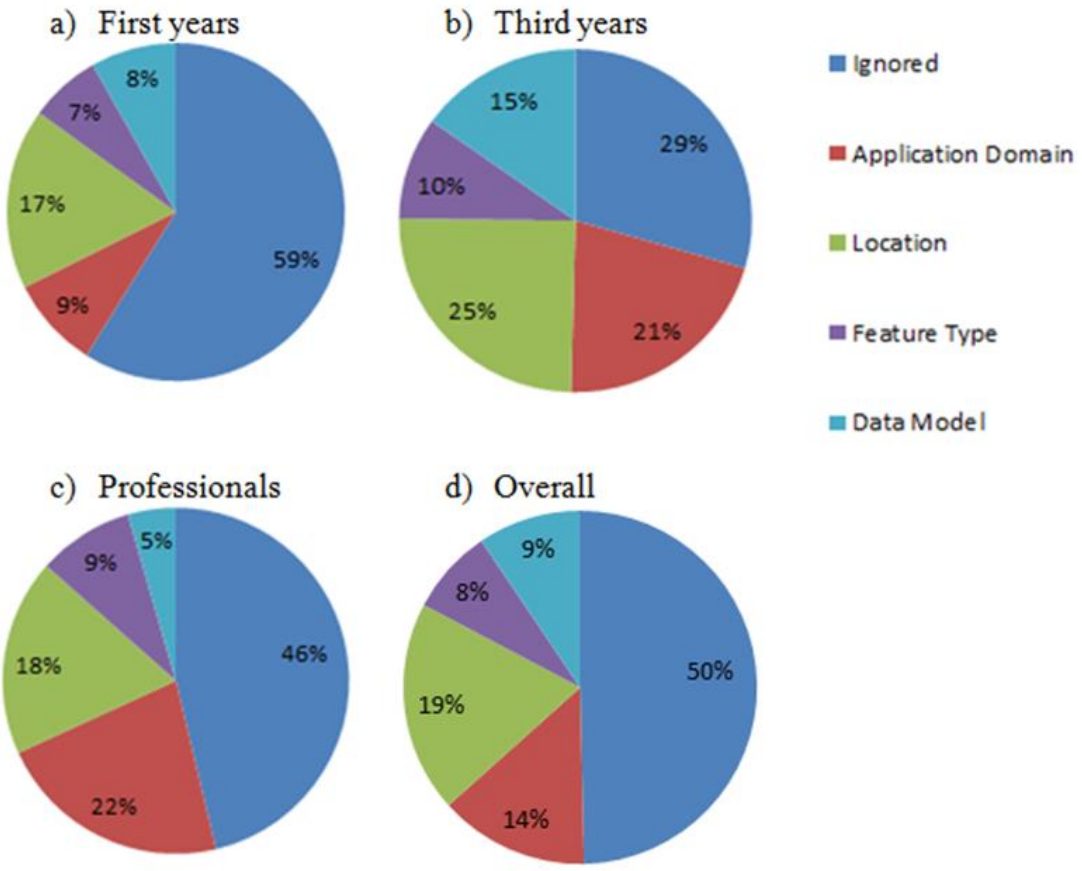

Figure 12: User experiment results

5.4.1 First year geoinformatics students

Being the largest group (21 participants) in the experiment, the first years submitted a total number of 159 unique search queries that were dissected into 866 search terms. Table 7 and Figure 12a show the results.

5.4.2 Third year geoinformatics students

A total number of 363 search terms were obtained from 88 unique search queries submitted to web search engines by 15 third year Geoinformatics students. As described in Table 7 and Figure 12b, 21% of the terms fell under the 'Application Domain' category, 25% of the terms fell under the 'Location' category, 15% of the terms fell under the 'Data Model' and 10% of the 'Feature Type'. 29% of the terms that fell under the 'Ignored' category were not related to any of the four categories of the taxonomy.

© University of Pretoria

5.4.3 Professionals

A total number of 47 unique search queries were obtained from a group of nine professionals specialising in the field of geoinformatics (GIS). A total number of 201 search terms were derived from these search queries as categorised in Table 7. As it can be seen from Figure 12c., 22% of the query terms employed by professionals were related to the application domain or subject for which the search for spatial data was being quested. 18%, 9% and 5% of the query terms were related to 'Location', 'Feature Type' and 'Data Model' respectively. 46% of the query terms that fell in the 'Ignored' were not related to any of the category described in the taxonomy.

5.4.4 Summary of user experiment results

A consolidated set of results of the search terms collected from the three different groups of participants (First year geoinformatics students, Third year geoinformatics students and geoinformatics Professionals) is summarised in Table 7. The 1430 search terms were derived from a total number of 294 unique search queries participants entered in the text search boxes of general purpose web search engines of their choice. From Figure 12d, it can be seen that 50% of the terms employed by participants fall under the 'Ignored' category leaving the other 50% shared between the four categories of the taxonomy. The category that had a high frequency of terms (19%) is the 'Location', followed by the 'Application Domain' category with 14% of terms. The 'Feature Type' category and the 'Data Model' category had 8% and 9% of the terms respectively.

The ratios of the number of participants to the number of search queries employed by the three groups are provided in Figure 13. The group of professionals has the highest ratio (0.19), followed by the third years (0.17) and the first years with the smallest ratio (0.13). These ratios can be expressed in terms of ten participants: ten First year students submitted 13 queries, ten Third year students submitted 17 queries, and ten Professionals submitted 19 queries (although only nine Professionals participated in the experiment). A possible explanation could be that the first years had to submit a higher number of queries to web search engines due to the fact that they were not obtaining desirable search results. Such a fact could be linked to the observation by Holsche and Strube (2000) that user expertise in the knowledge domain of the search influences success in web searches.

© University of Pretoria

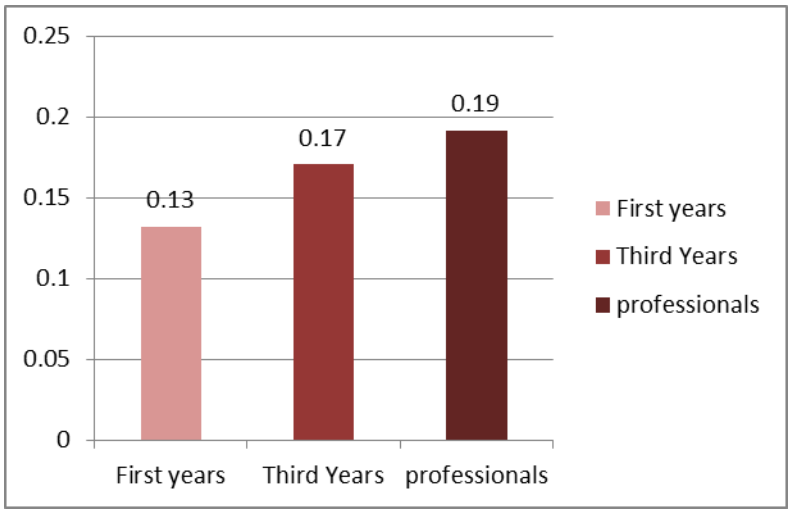

Figure 13: Ratio of number of participants to number of queries[1]

5.5 Results of the Bing webmaster tools experiment

A total number of 90 search queries were obtained from the Bing web search engine webmaster tools between 23 November 2014 and 18 February 2015 or, 87 days. HTML web pages, for which the 90 queries were recorded, were marked-up with selected types of the Schema.org vocabulary. After dissecting each of these queries as depicted in Figure 14, 23% of the terms fell under the 'Location' category, 18% of the terms fell under the 'Application Domain' category, 9% of the terms fell under the 'Feature Type' category and 5% of the terms fell under the 'Data Model' category. The 'Ignored' category was made of 45% of the total number of terms.

Total number of search terms	**Ignored**	**Category**			
		Application Domain	**Location**	**Feature Type**	**Data Model**
330	148 (45%)	58 (18%)	77 (23%)	30 (9%)	17 (5%)

Table 8: Experiment results from the Bing webmaster tools

[1] The equivalent figure in Katumba and Coetzee (2015) shows 0.14 for First years because accidentally an incorrect number of participants (22) was used in the calculation. However, the difference does not change the results.

© University of Pretoria

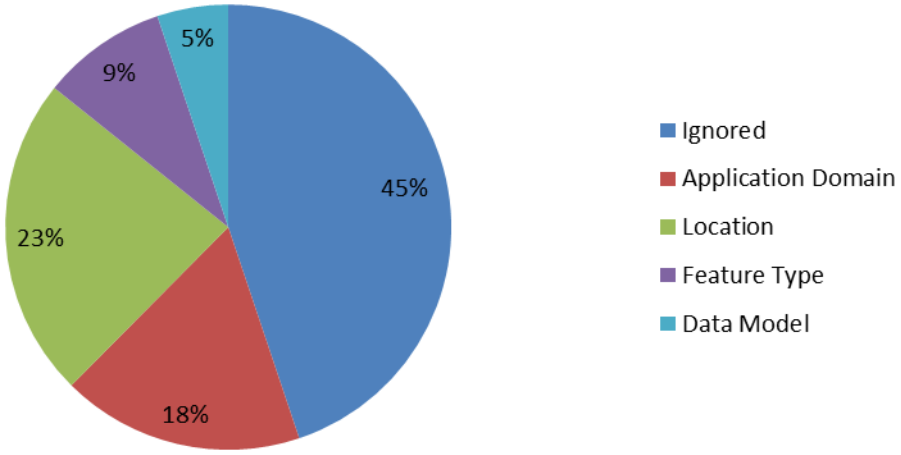

Figure 14: Experiment results – Bing webmaster tools application

5.6 Proposed taxonomy of search terms

It can be seen from the pie diagrams of both sets of results (Figure 12d and Figure 14) that the 'Location' category has the highest proportion of terms, followed by the 'Application Domain' category. This reveals the strong association with 'location' inherent in search queries with the intent of finding geospatial data. The 'Application Domain' with the second highest number of search terms, demonstrates that the quest for geospatial data is typically guided by the need to use such data in a given application domain. The 'Feature Type' and the 'Data Model' categories had the lowest proportions of terms in both sets of results as compared to the other two categories. Based on these similarities observed between the results of the two experiments, it can be deduced that the constructed taxonomy of search terms is consistent.

Summaries of both sets of results as illustrated in Table 9 reveal that the average number of terms per unique query of participants in the user experiment is five and the average number of terms per unique search query from the Bing webmaster tools is four (refer Table 10). In both sets of results, the average number of terms related to each of the four categories ('Application Domain', 'Location', 'Feature Type' and 'Data Model') is one.

© University of Pretoria

	Number of terms per query	Ignored	Category			
			Application Domain	Location	Feature Type	Data Model
Average	5	3	1	1	1	1

Table 9: User experiment results – Average number of terms per category

	Number of terms per query	Ignored	Category			
			Application Domain	Location	Feature Type	Data Model
Average	4	2	1	1	1	1

Table 10: Bing webmaster tools experiment results – Average number of terms per category

The results of the two experiments lead one to conclude that on average, a search query employed with the intent of finding geospatial data using web search engines will be made of at least one term from the four categories provided in the taxonomy. Furthermore, on average, the longest search query is between four terms (user experiment conducted in the lab) and five terms (Bing webmaster tool). Such averages of the number of terms per query could be affected by a few outliers in each search query set. For example, while on average the number of terms per search query from the user experiment was four, the set of search queries collected contained a query with 13 terms. Similarly, the longest search query from the Bing webmaster tools experiment contained 11 terms. Hence, the mode which provides the most frequently occurring value in a data set can be used as a descriptive measure for comparing the results of the two experiments. For example, the mode for both the individual experiments and the two experiments combined in terms of the number of search terms per query is three. This implies that a search query employed for finding geospatial data on the web is frequently comprised of three terms. Furthermore, given that the 'Application Domain' and 'Location' categories have the highest number or proportions of terms; search terms employed by users to find geospatial data are more likely to fall under these two categories. Based on the analysis of the search terms thus conducted, a taxonomy of search terms is proposed and illustrated in Figure 15.

© University of Pretoria

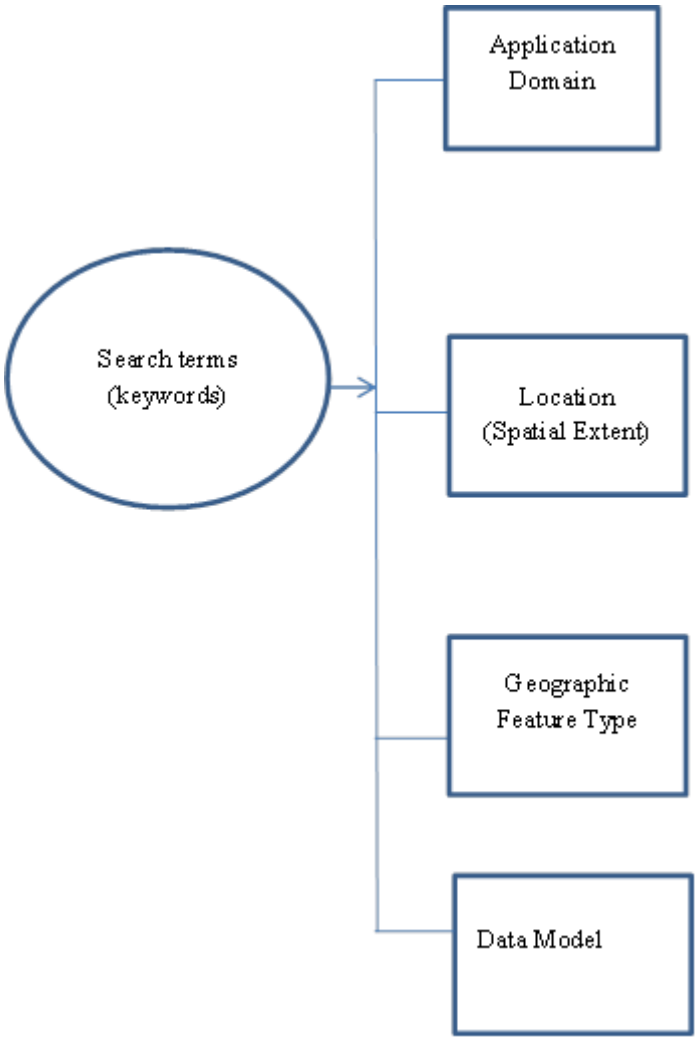

Figure 15: Taxonomy of search terms

5.7 Merits and limitations of the proposed taxonomy

The sole purpose of constructing a taxonomy in this research is to understand how users search for geospatial data online. This fits in the broader endeavour of applying SEO techniques for enhancing the findability of geospatial data online as assigned in this research. The concept of understanding keywords from users' searches by creating a taxonomy is proposed in this dissertation as an option to consider in current SEO techniques in general and in the findability of geospatial data online in particular.

It is important to note that the taxonomy proposed in this dissertation is not exempted from the inconsistencies that may arise in the attribution of terms to appropriate categories. Such challenges are common in the creation of taxonomies (refer to Chuang and Chien (2002)).

© University of Pretoria

Furthermore, the way in which users formulate their queries (keywords, key-phrases or search terms) greatly influenced search terms' assignment to corresponding categories of the taxonomy. A list of all the queries (search terms or keywords) employed by participants during the user experiment and of the ones (search terms) obtained from the Bing experiment is provided in appendix J and appendix K to shed more light in terms of why nearly 50% of terms were excluded from the taxonomy. Many of such terms were adjectives, pronouns, conjunctions, adverbs and verbs.

5.8 Chapter conclusion

The analysis of search terms obtained from the experiments conducted in this research has been discussed in this chapter. The user experiment performed in the computer laboratory involved the participation of human subjects that were instructed to search for geospatial data using general purpose web search engines. The Bing webmaster tools experiment consisted of collecting search terms from the Bing web search engine log (webmaster tools online application) that triggered the appearances of HTML web pages with spatial metadata content on the search results pages of Bing web search engine. The Collected search terms from the two experiments were analysed and assigned to appropriate categories of the taxonomy. The four categories in the taxonomy for which the collected search terms were grouped are: 'Subject (theme)', 'Location (spatial extent)', 'Geographic feature type' and 'Data Model'. The taxonomy thus constructed, serves in the understanding of search terms that users employed when searching for geospatial data on the web using general purpose web search engines.

© University of Pretoria

Chapter 6

6 Mapping among the taxonomy search terms, ISO 19115:2003, Geographic Information – metadata standard, Dublin Core and Schema.org

6.1 Introduction

The second objective attained in this dissertation as mentioned in section 1.4 (Research objectives), is discussed in this chapter. The different categories of the taxonomy, selected elements of *ISO 19115:2003 – Geographic Information Metadata* standard, elements of Dublin Core vocabulary and selected properties of types contained in Schema.org vocabulary are discussed. The existing mapping between Dublin Core and *ISO 19115:2003, Geographic Information – Metadata* is reviewed in more details and the established mappings from the taxonomy search terms to *ISO 19115:2003, Geographic Information – Metadata* and from the Dublin Core to Schema.org are explained.

6.2 Overview of the mappings (crosswalks) involved

Through the construction of HTML web pages, this research seeks to study and exploit the capability of making web resources with information about geospatial data visible to well-known web search engines. Hence, the idea to document geospatial data using HTML web pages in such a way that they can easily be indexed by general purpose web search engines. This is done by compiling geospatial metadata as contents of HTML web page documents which can be crawled by web search engines. This enables the discovery of geospatial data without the user having to know about existing geoportals registered with geospatial resources that are invisible to web search engines in the first place. Search engine optimisation (SEO) techniques are employed in the process of creating these HTML web pages with the use of appropriate mark-up tags such as the ‘meta’ tag and ‘microdata’ tags. Besides employing the already existing mapping (crosswalk) between *ISO 19115:2003, Geographic Information – Metadata* and Dublin Core standards, two different mappings (crosswalks) are created to facilitate both: the insertion of geospatial metadata into HTML web pages and the process of marking-up these web pages with Dublin Core and Schema.org vocabularies. The two proposed crosswalks (also referred to in Figure 16) are:

© University of Pretoria

- A mapping between the categories of the taxonomy of search terms and appropriate elements of *ISO 19115:2003, Geographic Information – Metadata* standard.
- A mapping between the elements of Dublin Core metadata standard and proprieties of appropriate types of Schema.org vocabulary.

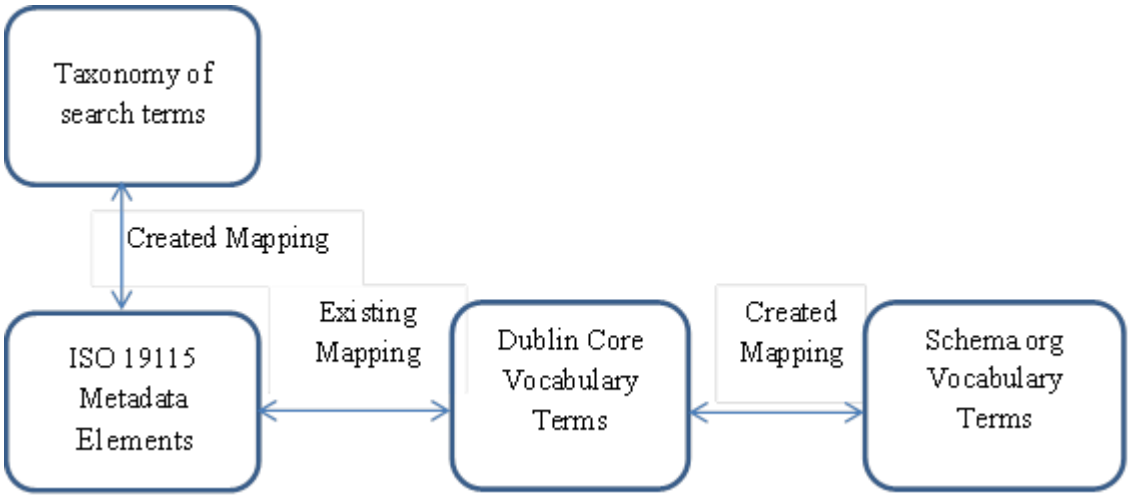

Figure 16: Mapping among standards and vocabularies

6.3 Mapping the categories of the taxonomy to Dublin Core via ISO 19115:2003, Geographic Information – Metadata

As already mentioned in chapter 5, the purpose of creating the taxonomy is to understand how users search for geospatial data online. The immediate benefit of having such an understanding is evident in terms of preparing HTML web page contents as part of the SEO strategy to have search engines efficiently index such web pages (content). The fact of knowing or having some ideas of what users type in search engines' search boxes in search for geospatial data, helps in terms of preparing the kind of corresponding content that will be retrieved by web search engines. The spatial metadata standard *ISO 19115:2003, Geographic Information – Metadata* is the metadata standard that is suitable for structuring information that describes geospatial data. Hence, it is against this background that it made sense to map the categories of the taxonomy (which provides an understanding of how users' search for geospatial data online based on their keywords) to corresponding core elements of *ISO 19115:2003, Geographic Information – Metadata*. The established mapping (categories of the taxonomy to corresponding elements of

© University of Pretoria

ISO 19115:2003, Geographic Information – Metadata) provides a means for understanding how keyword (search terms) employed by users, fit into corresponding elements of *ISO 19115: 2003, Geographic Information – Metadata* standard.

Categories of the taxonomy were mapped to corresponding selected core elements of the *ISO 19115:2003, Geographic Information – Metadata* standard. Search terms making up each of the categories of the taxonomy are assumed to exhibit characteristics or descriptions of geospatial data. This is because when typed into search boxes of web search engines, users' search terms lead to the online retrieval of web resources containing geospatial data. Furthermore, core elements of *ISO 19115:2003, Geographic Information – Metadata* allow an easy understanding of geographic data by both consumers and producers (Nogueras-Iso *et al*. 2004).

The mapping (crosswalk) employed between the categories of the taxonomy of search terms, the elements of *ISO 19115:2003, Geographic Information – Metadata* standard and Dublin Core metadata standard is illustrated in Appendix E. Elements are also defined based on the existing mapping between *ISO 19115:2003 – Geographic Information - Metadata* and Dublin Core as provided in the *CWA 14857:2003 (E)* document.

6.4 Understanding the elements involved in the mapping (from the taxonomy of search terms to Dublin Core via ISO 19115:2003, Geographic Information – Metadata)

The existing mapping between Dublin Core and *ISO 19115:2003, Geographic Information - Metadata (CWA 14857:2003 (E))*, was used to facilitate the integration of spatial metadata into HTML web pages. Although withdrawn since 2011, the creation of *CWA 14857:2003 (E)* stemmed from the need to enhance the discovery of geographic information in cross metadata searches (CEN 2003). In the research described in this dissertation, *CWA 14857:2003 (E)* serves as a reference for finding existing similarities between the elements of the Dublin Core and the elements of *ISO 19115:2003, Geographic Information – Metadata* standard. Other *ISO 19115:2003, Geographic Information – Metadata* elements such as Dataset responsible party (or CREATOR), Dataset reference date (or DATE), On-line resource (or IDENTIFIER), Dataset language (or LANGUAGE) didn't have corresponding categories in the taxonomy of search terms but their inclusion in the construction of HTML web pages was relevant.

© University of Pretoria

Using *CWA 14857:2003 (E)* as starting point, a more detailed study of selected elements of *ISO 19115:2003, Geographic Information – Metadata* that were mapped with the categories of the taxonomy of search terms, was performed. The aim was mainly to understand the definition of each particular element (or property) and the data type and domain associated with the value of such elements. Domains describe the kind of resource and values assigned to elements (or properties) of *ISO 19115:2003, Geographic Information – Metadata*, Dublin Core or Schema.org vocabulary.

6.4.1 Description of ISO 19115:2003 elements involved in the mapping

A comprehensive knowledge of these selected elements of *ISO 19115:2003, Geographic Information – Metadata* is important in order to understand the mapping that exists between *ISO 19115:2003, Geographic Information – Metadata* and Dublin Core. Definitions, maximum occurrences, data types and domains (or ranges) of values for each of the selected *ISO 19115:2003, Geographic Information – Metadata* elements involved in the mapping with Dublin Core are provided in Appendix F. Out of the seven selected elements (of *ISO 19115:2003, Geographic Information – Metadata*), four elements namely 'Dataset title', 'Abstract', 'Lineage (statement)' and 'Distribution format (name)' are of data type character string and 'free-text' domain. *ISO 19115: 2003, Geographic Information – Metadata* recommends that values for the 'Dataset topic category' element be of class *MD_TopicCategoryCode* that contains an enumerated controlled vocabulary of terms that represent predefined categories of topics. The standard (*ISO 19115:2003, Geographic Information – Metadata*) also recommends that the value for the 'geographicIdentifier' element that defines the geographic location of the dataset using a geographic identifier, be of class *MD_Identifier*. However the *MD_Identifier* contains an element (property) named 'code' which has 'free-text' as a domain. Such an element is useful in describing geographic location names that may not have corresponding identifiers (codes) contained in a controlled vocabulary of terms due to the various names of locations that exist in the world. The standard (*ISO 19115:2003, Geographic Information – Metadata*) recommends values for the 'Spatial representation type' element be drawn from *MD_SpatialRepresentationTypeCode* element code list. It is a domain that assigns codes to different data structures in which geospatial data are represented. For example, code: 001 is used for vector data, code 004 for tin, etc. (refer *ISO 19115:2003, Geographic Information – Metadata* document).

© University of Pretoria

6.4.2 Description of Dublin Core elements involved in the mapping

The seven selected Dublin Core elements involved in the mapping are: 'TITLE', 'SUBJECT', 'DESCRIPTION', 'SOURCE', 'COVERAGE', 'TYPE' and 'FORMAT'. The Dublin Core standard doesn't recommend any controlled vocabulary for values describing the 'TITLE', 'DESCRIPTION' and 'SOURCE' elements. Furthermore, best practices suggest a controlled vocabulary to be employed when assigning values to 'SUBJECT', 'COVERAGE', 'FORMAT' and 'TYPE' elements. In reference to section 2.5.2 as described in the literature review chapter, the 'DCMI Metadata Terms' caters for domains and ranges that provides controlled vocabularies for values of properties that describe resources published on the web. However, the DCMES version 1.1 was chosen in the creation of the HTML web pages with spatial metadata content instead of the recommended 'DCMI Metadata Terms'. This was because search terms employed by users during the experiment phase of this research informed the creation of HTML web pages with spatial metadata contents. As such, one cannot restrict or control the kind of search terms that users type on web search engines' search boxes.

6.4.3 Incorporating spatial metadata into HTML web pages based on the mappings between the taxonomy of search terms, ISO 19115:2003, Geographic Information – Metadata and Dublin Core

Through a use case scenario, a practical example is provided to illustrate the construction of an HTML web page with spatial metadata; i.e., this HTML web page is to match the content from the analysis of search terms employed by a participant whose topic for geospatial data search was 'Cairo roads in Egypt'. It is an adaption from the layer titled 'Egypt – Roads' published on the 'FAO Africover' geoportal (http://www.africover.org) (refer Figure 17).

© University of Pretoria

FOOD AND AGRICULTURE ORGANIZATION of The United Nations

METADATA

[Print]

National Focal Point Institution:	
Organisation:	FAO - Africover
Data Set Description Fields:	
Title:	Egypt - Roads (Africover)
Dataset Reference Date:	2002-04-04
Dataset Reference Date Type:	Publication
Dataset Edition:	First
Presentation Format:	mapDigital
Abstract:	The roads have been produced from visual interpretation of digitally enhanced LANDSAT TM images (Bands 4,3,2) acquired mainly in the year 1997.
Purpose:	The roads have been included for orientation purposes and should not be seen as comprehensive.
Completedness/Progress Code:	Complete
Theme Keywords:	orientation, roads
Place Keyword:	Egypt
ISO Topic Category:	Earth Cover
Supplemental Infomation:	
Direct Spatial Reference Method:	Vector
Data Format:	ESRI ArcView Shapefile (.shp)
Scale of the Dataset:	1:100 000
Dataset Language:	English
Dataset Character Set:	usAscii
Resource Provider:	Mr. Antonio Di Gregorio - FAO Africover
Point of Contact:	Mr. Antonio Di Gregorio - FAO Africover
Custodian:	Dr. Nabil El Mowelhi - Soil and Water Research Institute - Ministry of Agriculture
Owner:	Dr. Nabil El Mowelhi - Soil and Water Research Institute - Ministry of Agriculture
Originator:	Mr. Antonio Di Gregorio - FAO Africover
Processor:	Mr. Antonio Di Gregorio - FAO Africover
Publisher:	Mr. Antonio Di Gregorio - FAO Africover

Figure 17: Example of a web page with spatial metadata contents

© University of Pretoria

An illustration of how contents of the four categories of the taxonomy of search terms translate into contents of selected elements of *ISO 19115:2003, Geographic Information – Metadata* standard is provided in Table 11.

Categories of taxonomy of search terms and values	**ISO 19115 elements and values**
Subject (Theme)	*Dataset title*: Egypt - Roads
	Dataset topic category: Roads Network of Egypt
	Abstract: 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.
	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.
Location (Spatial Extent)	*Geographic location of the dataset by geographic identifier*: Egypt
Geographic Feature Type	*Spatial representation type*: Dataset, Vector
Data Model	*Distribution format*: Shapefile (.shp)

Table 11: Mapping between taxonomy of search terms and *ISO 19115:2003, Geographic Information – Metadata*

The HTML web page result in Figure 18 illustrates how the different field elements of the Dublin Core standard are filled with spatial metadata contents using the HTML 'meta' tag based on the mapping from *ISO 19115:2003, Geographic Information – Metadata* to Dublin Core. It is possible for web search engines and their crawlers to index HTML web pages prepared in this way for optimum geospatial data (metadata) discovery. This is because web search engines are best at discovering HTML pages.

© University of Pretoria

```
<head>
        <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" content="Roads of Egypt" lang="en">
        <meta name="DC.Subject" content="Town and Cities Infrastructures" lang="en">
        <meta name="DC.Description" content="Egypt's ESRI Shapefiles of Roads" lang="en">
        <meta name="DC.Type" scheme="DCTERMS.DCMIType" content="Dataset">
        <meta name="DC.Type" content="Vector">
        <meta name="DC.Format" scheme="DCTERMS.IMT" content="text/xml">
        <meta name="DC.Format"content="Shapefile">
        <meta name="DC.Coverage" scheme="DCTERMS.TGN" content="Egypt">
        <meta name="DC.Creator" content="Antonio Di Gregorio">
        <meta name="DC.Creator" content="FAO-UN">
        <meta name="DC.publisher" content="FAO Africover">
        <meta name="DC.Date" scheme="DCTERMS.W3CTDF" content="2002-04-04">
        <meta name="DC.Language" scheme="DCTERMS.ISO639-2" content="en">
        <title> Roads of Egypt</title>
</head>
```

Figure 18: Example of a web page with spatial metadata contents

Given that the DCMI recommends the use of 'DCMI Metadata Terms' because they support domains and ranges that define values for selected appropriate elements, both the Dublin Core namespace (http://purl.org/dc/elements/1.1/) in version 1.1 and the Dublin Core terms described in the 'DCMI Metadata Terms' namespace were considered. Dublin Core elements involved in the mapping for which the DCMI recommends controlled vocabularies are 'Subject', 'Type', 'Format' and 'Coverage'. Because not all the subjects for which HTML web pages were published have domains that are supported by the DCMI, the 'Subject' element only contained free text values. The 'Type' element appeared twice in the head section of each of the published HTML web pages. In the first instance, the 'Type' element used 'Dataset' as a value from the DCMI Type controlled vocabulary. In the second instance, the 'Type' element used 'Vector' as a free text value for each of the published HTML web pages. The 'Format' element had two occurrences in the head section of each of the published HTML web pages. In the first instance, the 'text/xml' from the Media Type (IMT) controlled vocabulary was used as a value for the 'Format' element while, in the second instance, the word 'Shapefile' was used as free text. Lastly, the 'Coverage' element was set to use values from the 'Getty Thesaurus of Geographic Names (TGN)'. It can be noticed that the use of both the Dublin Core namespace (http://purl.org/dc/elements/1.1/) in version 1.1 and the Dublin Core terms described in the 'DCMI Metadata Terms' namespace, makes it possible to encode specific values (e,g, code list)

© University of Pretoria

while also catering for free-text values. Such a flexibility provides for alternative elements in the annotation of web pages to be used (refer Figure 18 where the value 'Dataset' from the 'DCMI Type vocabulary' and the free-text value 'Vector' are both used for the 'type' element respectively). Consequently, the possibility of information loss is minimized. More descriptions about this mapping is provided in Appendix E and also in the mapping specification itself *(CWA 14857:2003 (E))*. In Figure 19, an illustration of how the body content of an HTML web page with spatial metadata content looks like when displayed on a browser is provided.

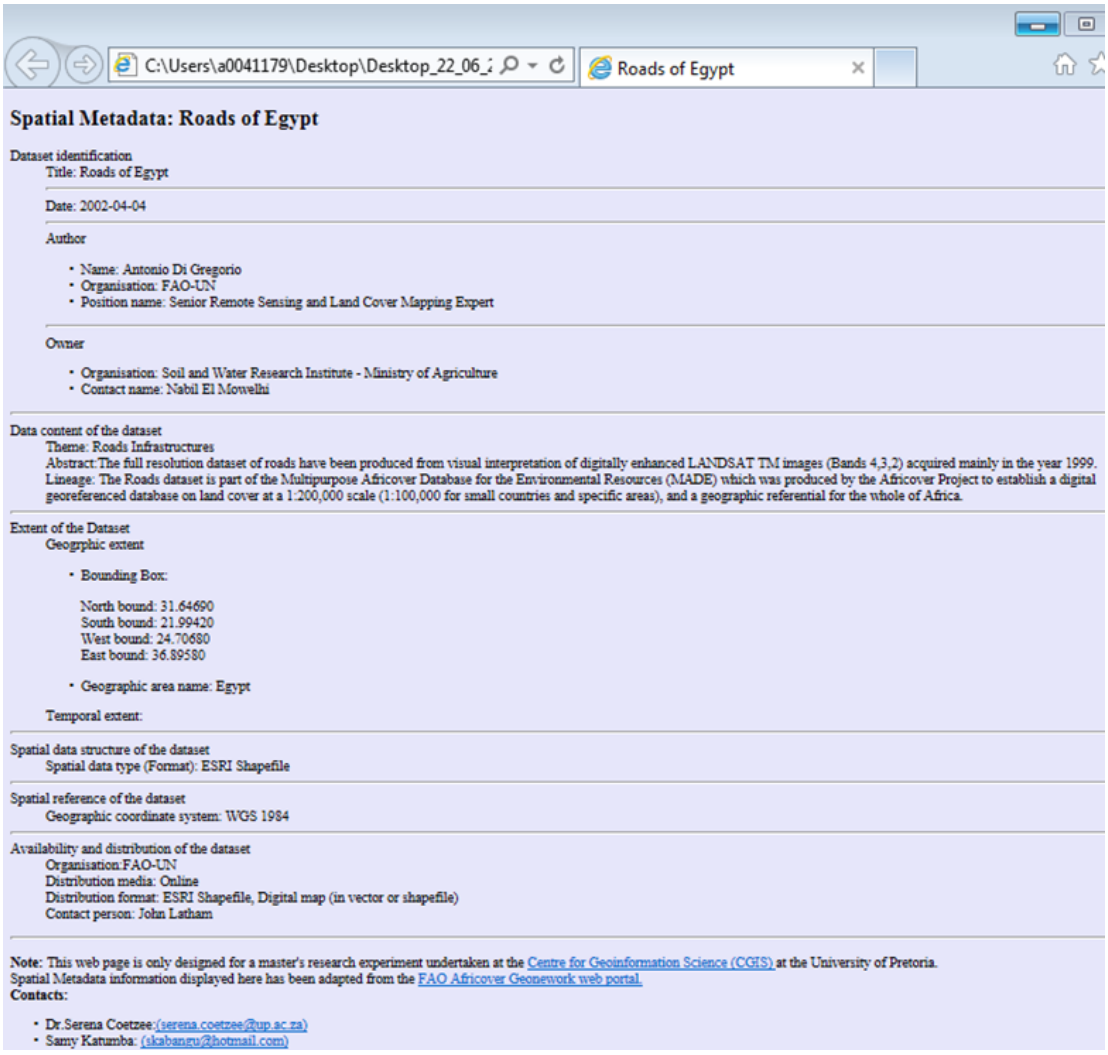

Figure 19: HTML web page with spatial metadata content marked-up with Dublin Core vocabulary

© University of Pretoria

6.5 Mapping the categories of the taxonomy to Schema.org vocabulary via ISO 19115:2003, Geographic Information – Metadata and Dublin Core

Given that the mapping from the taxonomy of search terms to Dublin Core via *ISO 19115:2003, Geographic Information – Metadata* standard has already been elaborated, properties of types from Schema.org vocabulary are directly mapped to corresponding elements of Dublin Core. This mapping of the terms between Schema.org and Dublin Core vocabularies is solely based on matching corresponding definitions or descriptions (or semantics) between the elements (properties) of the two vocabularies. Such a process is also known as 'crosswalk' (refer Nogueras-Iso *et al*. 2004). It should be noted that there is currently no official standard mapping between the Dublin Core and Schema.org vocabularies. However, initiatives about mapping the two vocabularies are underway and are led by the Dublin Core Metadata Initiative (Dublin Core Metadata Initiative 2012).

The table that illustrates the mapping of selected properties (elements) of Schema.org's types to corresponding elements in Dublin Core based on the description of these elements is provided in Appendix E. Schema.org properties: 'name', 'description' and 'about' matched the Dublin Core elements 'title', 'description' and 'subject' respectively. Four elements from Schema.org namely, 'spatial', 'geo', 'latitude' and 'longitude' corresponded to the 'coverage' element in Dublin Core. Two elements of Schema.org vocabulary namely 'distribution', and 'encodingFormat' correspond to the 'format' element in Dublin Core. With reference to the mapping between the elements of *ISO 19115:2003, Geographic Information – Metadata* and Dublin Core where the 'spatial representation type' element of ISO *19115:2003, Geographic Information – Metadata* is mapped to the 'type' element of Dublin Core, the closest match of 'type' in Schema.org vocabulary as described in this research is the 'encoding' property. An extract of an HTML web page containing spatial metadata marked-up with Schema.org vocabulary is illustrated in Figure 20.

© University of Pretoria

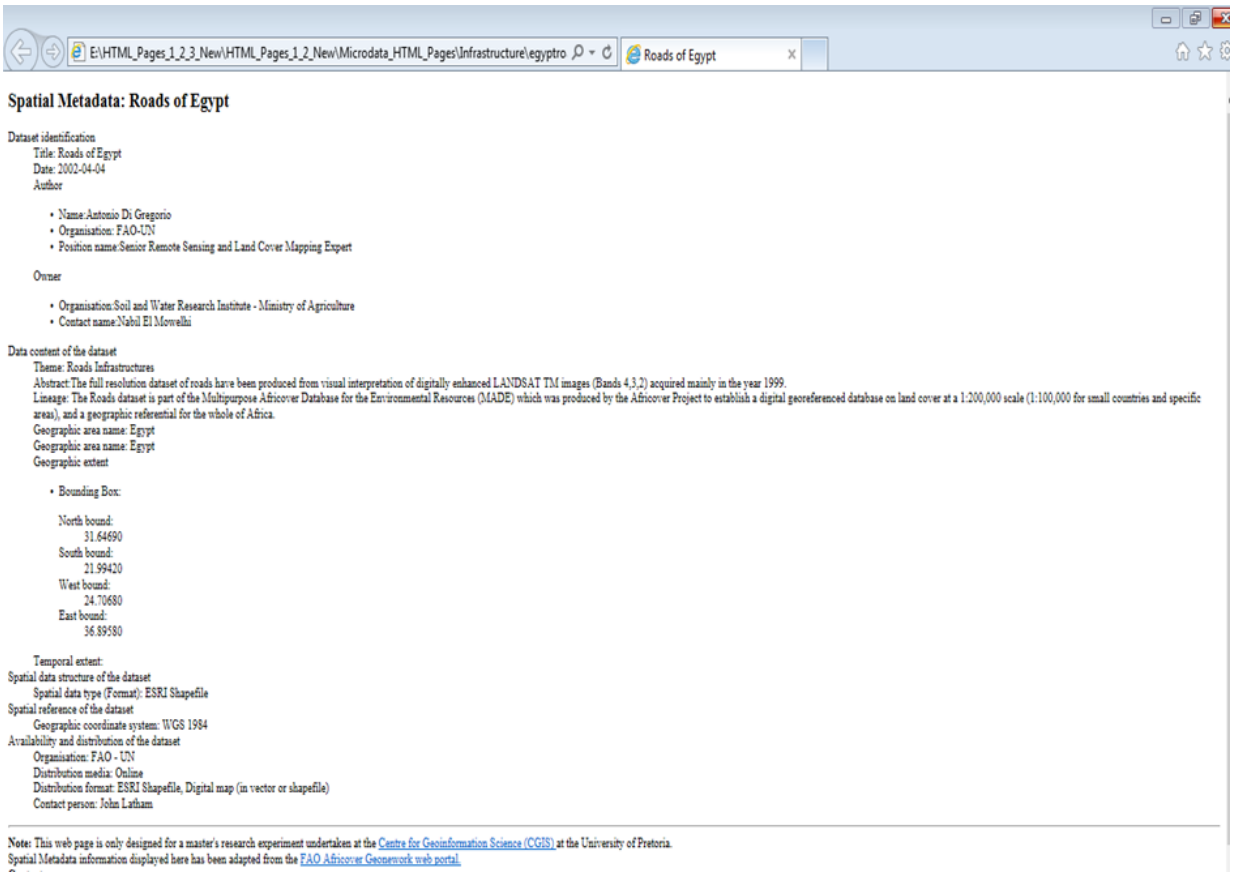

Figure 20: HTML web page with spatial metadata content marked-up with Schema.org vocabulary

© University of Pretoria

In Table 12, an illustration of how spatial metadata contents of selected elements of Dublin Core translate into contents of corresponding selected properties (elements) of Schema.org is provided.

Dublin Core elements with spatial metadata contents	**Corresponding Schema.org properties (elements) with spatial metadata contents**
TITLE: Roads of Egypt	NAME: Roads of Egypt
SUBJECT: Town and Cities Infrastructures	ABOUT: Town and Cities Infrastructures
DESCRIPTION: Egypt's ESRI Shapefiles of Roads	DESCRIPTION: Egypt's ESRI Shapefiles of Roads
COVERAGE: Egypt	SPATIAL: Egypt
	geo, latitude: 31.64690 (Northbound)
	geo, latitude: 21.99420 (Southbound)
	geo, longitude: 24.70680 (Westbound)
	geo, longitude: 36.89580 (Eastbound)
FORMAT: Shapefile	ENCODING: Shapefile
	encodingFormat: shp

Table 12: Spatial metadata insertion into an HTML web page based on Schema.org vocabulary

The HTML web page result in Appendix G illustrates how the different sections of the HTML web page with spatial metadata content are marked-up with selected properties of Schema.org vocabulary. It is possible for web search engines and their crawlers to understand the content of HTML web pages prepared in this way and index them thereafter for optimum geospatial data (metadata) discovery.

6.6 Chapter conclusion

Different mappings (crosswalks) of standards and vocabulary employed in this research have been discussed in this chapter. The existing mapping between Dublin Core and *ISO 19115:2003, Geographic Information – Metadata* standard has been discussed. Furthermore, based on the creation of specific crosswalks for one-to-one mapping, suggested by Nogueras-Iso *et al*. (2004),

© University of Pretoria

two mappings (crosswalks) have been created. The first mapping consists of matching the definitions of the categories of the taxonomy of search terms with selected corresponding elements of *ISO 19115:2003, Geographic Information – Metadata* standard. The purpose of this crosswalk was to understand how search terms used for searching geospatial data on the web can serve in the documentation of geospatial data namely the creation of spatial metadata based on *ISO 19115:2003, Geographic Information – Metadata* standard. The existing mapping between *ISO 19115:2003, Geographic Information – Metadata* and Dublin Core enabled the translation of spatial metadata content into Dublin Core metadata content suitable for online web resources (web pages)' discovery. The second crosswalk created, consisted of creating a 'one to one mapping' between the Dublin Core and Schema.org vocabularies based on the definitions of elements (of Dublin Core) and types (of Schema.org).

Ultimately, the mappings (crosswalks) described in this chapter led to the construction of two sets of HTML web pages with spatial metadata content: one set of HTML web pages marked-up with selected elements of Dublin Core vocabulary using the HTML 'meta' tag, and the other set of HTML web pages marked-up with selected properties of Schema.org's types using microdata. A comparison of the retrieval effectiveness results of the two sets of HTML web pages with spatial metadata is discussed in the next chapter (chapter 7).

The logic behind establishing the chain of mapping (from the taxonomy of search terms, *ISO 19115:2003, Geographic Information – Metadata* standard, Dublin Core through to Schema.org) was to understand how keywords (search terms) captured in the taxonomy relate to the different respective corresponding elements provided by these standards (*ISO 19115:2003, Geographic Information – Metadata* and Dublin Core) and vocabulary (Schema.org). This logic is in support of one of the main objectives assigned in this research which is to exploit the proprieties and capabilities of geospatial data to make them findable by web search engines by employing SEO techniques. The proposed chain of mapping adopted this research served in the process of creating HTML web pages that were used in the experiment reported in chapter 7.

© University of Pretoria

Chapter 7

7 Results of submitted HTML web pages: retrieval effectiveness evaluation and comparison

7.1 Introduction

This chapter provides results of empirical tests on evaluating and comparing the retrieval effectiveness between HTML web pages with spatial metadata content marked-up with Dublin Core and Schema.org. It starts with a detailed description of the experiments set-up and collection of results. The chapter ends with a discussion on the analysis and comparison of results.

7.2 Construction of HTML web pages with spatial metadata

Before commencing with the analysis of results illustrating the appearance or visibility of HTML web pages with spatial metadata submitted to both web search engines (Bing and Google), a report on the experiment set-up and a brief overview on the webmaster tools are provided.

7.2.1 Experiments set-up

The content of the constructed HTML web pages is an adaptation of selected spatial metadata of datasets made available on the 'FAO Africover' geoportal website. A set of 58 constructed HTML web pages with information about spatial data (spatial metadata) was uploaded onto a web server (website). An index page with direct links to the 58 pages was also uploaded. Spatial metadata information contained in HTML web pages was structured according to *ISO 19115:2003, Geographic Information – Metadata* standard. The process of uploading HTML web pages happened in two phases. The first phase consisted of uploading the 58 HTML webpages marked-up with Dublin Core and the index HTML web page containing direct links to the 58 HTML web pages onto the web server. In the second phase, the 58 HTML web pages previously marked-up with Dublin Core were replaced with 58 HTML webpages marked-up with 'microdata' tags using selected types of Schema.org vocabulary. The index HTML web page containing direct links to the 58 HTML web pages (marked-up with Schema.org vocabulary) was also uploaded onto the web server. The content of the uploaded HTML web pages at each of the two phases remained the same. The only difference is the mark-up syntax

© University of Pretoria

used to annotate the HTML web pages. The HTML 'meta' tag was used to annotate HTML web pages with Dublin Core vocabulary, and 'microdata' tags were used to annotate the same HTML web pages with Schema.org vocabulary.

7.2.2 Webmaster tools (for Bing and Google)

At each of the two phases of the experiment, the 58 HTML web pages were registered with the webmaster tools of Bing and Google. By registering the 58 HTML webpages with each of the web search engines via their respective webmaster tools applications and marking those up with Dublin Core and Schema.org vocabularies, two of the four main SEO processes are applied in this research namely 'Indexing' and 'On-site optimisation'. These two main SEO processes are described in more details in section 3.6 (Application of SEO techniques). Furthermore, a sitemap in XML format that contains a list of URL links to all the 58 HTML web pages was uploaded onto the server and registered with each of the web search engines via their respective webmaster tools applications. The submission of sitemaps to web search engines is recommended as an SEO practice given that it enhances the crawling process (refer https://support.google.com/webmasters/answer/156184?hl=en or http://www.bing.com/webmaster/help/how-to-submit-sitemaps-82a15bd4).

7.3 Markup validation

A description of how the two web search engines (Bing and Google) interpreted HTML web pages marked-up with both the Dublin Core and Schema.org vocabularies is provided in this section. The 'Markup validator' found in Bing webmaster tools and the 'Structured Data Testing Tool' of Google webmaster tools were employed. It is assumed that the retrieval effectiveness of the marked-up HTML web pages could possibly be influenced by how the two web search engines (Bing and Google) recognise the two vocabularies (Dublin Core and Schema.org).

7.3.1 Markup validation in Bing webmaster tools

At the time of testing, the Bing web search engine did not identify any kind of mark-ups or tags in all HTML web pages that were marked-up with Dublin Core vocabulary using the 'meta' tag. An excerpt of the source code (HML code) of the web page named 'ugandaboundaries.html', tested in Bing Markup validator is illustrated in Figure 21.

© University of Pretoria

.

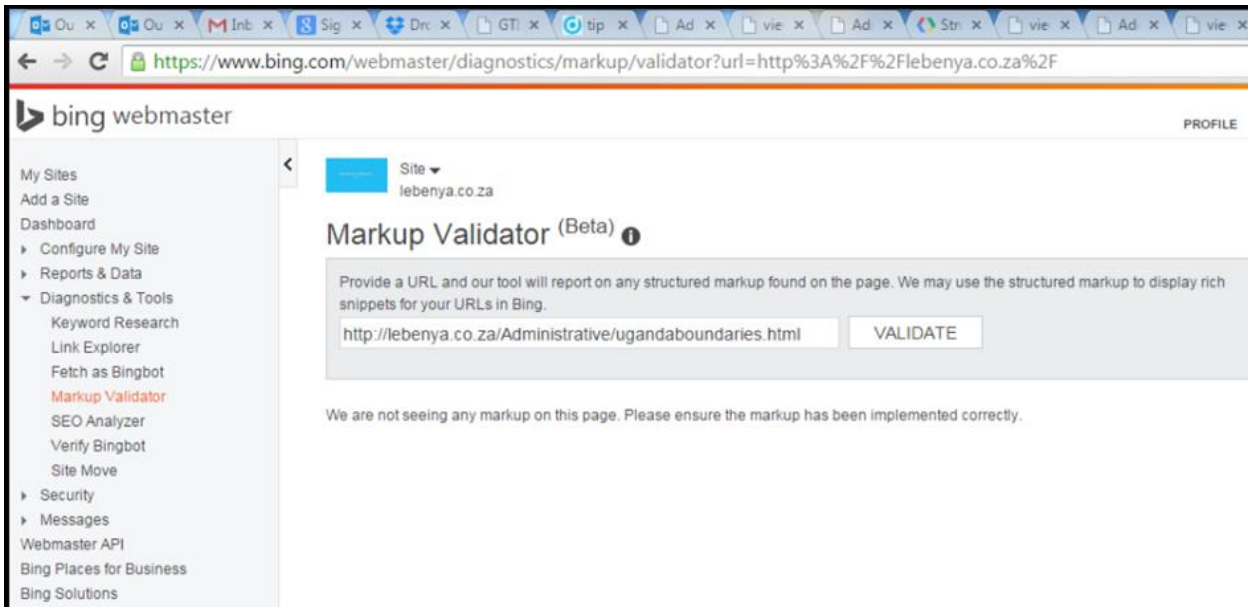

Figure 21: Bing Markup Validator

The Bing web search engine did not recognise any form of markups with this web page and in all the HTML web pages marked-up with Dublin Core vocabulary (refer Figure 21). However, HTML web pages marked up with microdata tags using selected types of Schema.org vocabulary were recognised by Bing web search engine as illustrated in Figure 22.

© University of Pretoria

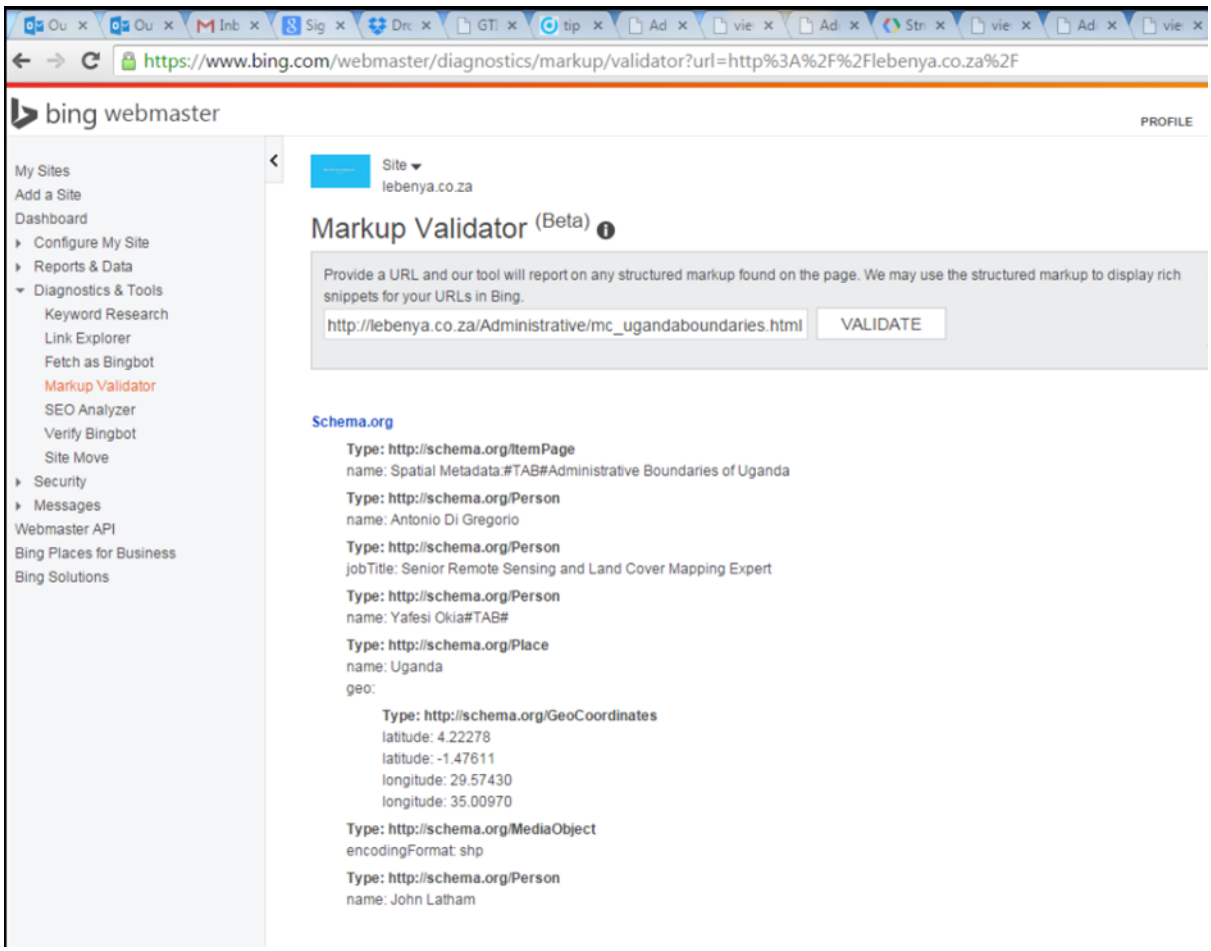

Figure 22: HTML page marked-up with Schema.org validated by Bing

7.3.2 Markup validation in Google webmaster tools

In contrast to the Bing Markup validator tool which did not recognise any markups with HTML web pages marked-up with Dublin Core vocabulary using the ‘meta’ tag, the Google Data Testing Tool identified markups in HTML web pages marked-up with Dublin Core vocabulary (refer Figure 23).

© University of Pretoria

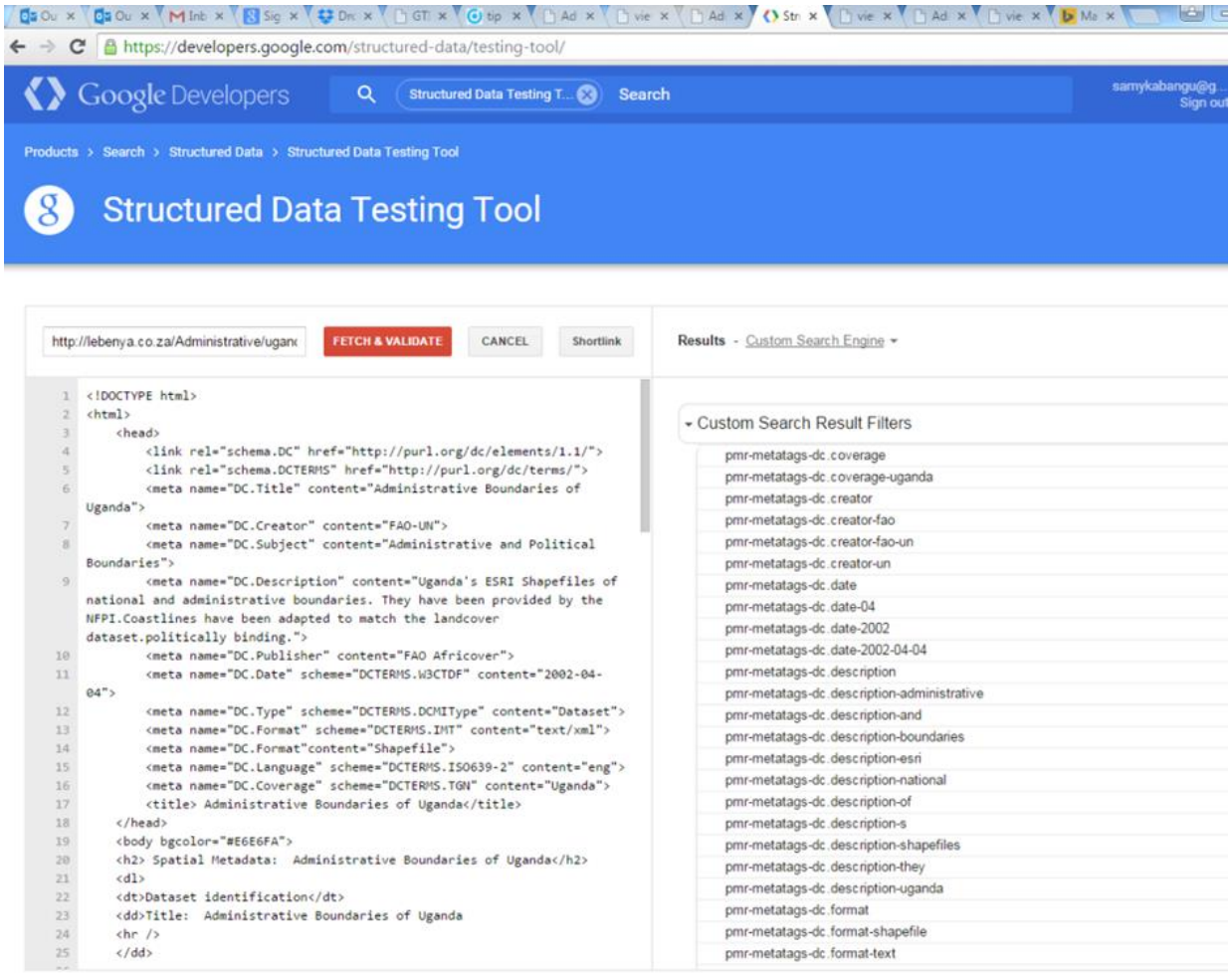

Figure 23: Structured data testing tool (Google)

Finally, the Google Data Testing Tool recognised all HTML web pages marked-up with selected types of Schema.org vocabulary using microdata (see Figure 24).

© University of Pretoria

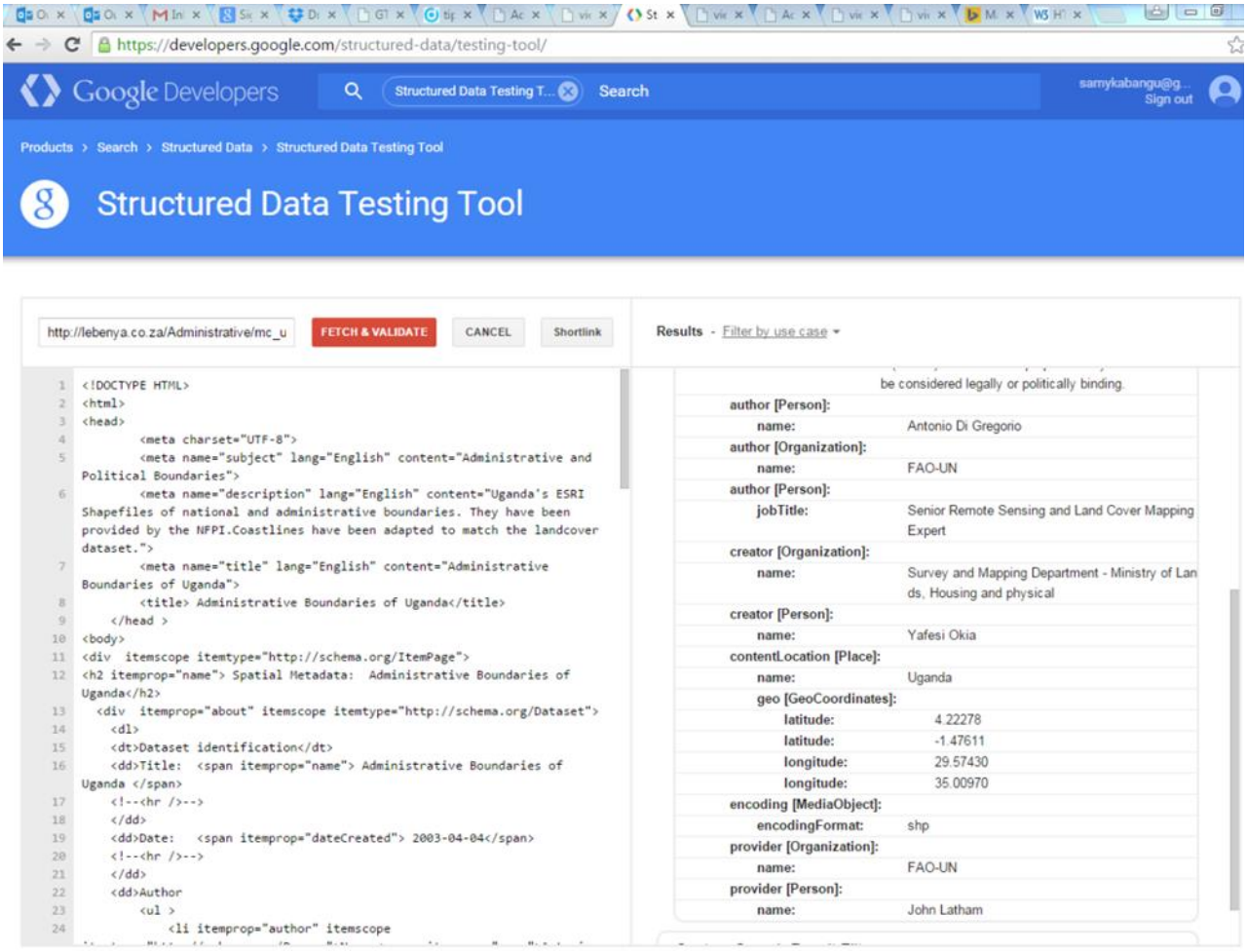

Figure 24: HTML web pages marked-up with Schema.org vocabulary validated by Google

7.4 Comparison of the retrieval effectiveness between the two sets of HTML web pages: Dublin Core vs. Schema.org

The Bing and Google webmaster tools applications provide reports on page traffic, search engine optimisation (SEO) recommendations, crawl information, search keywords and measures that quantify the number of times a web page appeared in search results pages of the web search engine. The latter is also referred to as impressions. The average search appearance position of a web page on the search engine results pages is also provided among other information obtained from search engines' webmaster tools applications. Measures that are relevant to the study (second experiment) described in this dissertation are impressions and average search appearance positions of retrieved HTML web pages. Both Bing and Google web search engines provide a similar definition of the two measures. The measure which quantifies 'Appeared in search' (in

© University of Pretoria

Bing web search engine, http://www.bing.com/webmaster/help/page-traffic-report-69c2c352) and impressions (in Google web search engine, https://support.google.com/analytics/answer/1308626?hl=en), refer to the number of times URL(s) of web page(s) in a website are visible on the search results pages of web search engines. The measure which quantifies 'Average Search Appearance Position' (in Bing) or 'Average Position' (in Google) provides the average (ranking) position of a web page on the search results pages of web search engines in response to queries submitted in the search boxes by users. The lower the position number, the higher the ranking position of a given web page on the search results page. Hence for the purpose of this research, 'Appeared in search (impressions)' and 'Average Search Appearance Position (Average Position)' are assumed to provide a measure of web page visibility. In the context of this research, such a measure (web page visibility) is used as a means of evaluating the retrieval effectiveness of HTML web pages with spatial metadata content marked-up with Dublin Core and Schema.org vocabularies. Furthermore, given that the appearance of the submitted HTML web pages on search engine results pages is of a great concern in this study, it can be stated that the 'recall' as a measure for evaluating the relevance of web search engine results, is also evaluated (refer section 3.7 of chapter 3 – Related work). This is because all the submitted HTML web pages with spatial metadata content are considered relevant. Recall provides the ratio of the number of relevant retrieved web pages to the total number of all web pages uploaded onto the web server.

A summary description of the measures used for comparing the results is provided in Table 13.

Measure	Description
Appeared in Search/Impressions	The number of times (frequency) a URL (web pages) from a website appeared on the search results pages of a web search engine.
Average Position/Average Search Appearance Position	The average position of a given web page (URL) on the search results pages of a web search engine. The lower the average position of a web page, the higher its visibility on search results pages of a web search engine.
Recall	The ratio of all relevant web pages retrieved (displayed on the search results pages of a web search engine) to the total number of web pages uploaded onto the web server.

Table 13: Description of measures for web search engines performance evaluation

© University of Pretoria

7.5 Analysis and comparison of results

The comparison is performed based on the results about web pages appearance obtained from the Bing and Google web search engines' webmaster tools applications. For each of the two search engines appearance results, web pages marked-up with Dublin Core elements are compared with web pages marked-up with properties of selected types contained in Schema.org. Furthermore, appearance results of HTML web pages marked-up with Dublin Core and submitted to Bing, are compared with the appearance results of HTML web pages marked-up with Dublin Core and submitted to Google. The same applies to Schema.org whereby the appearance results of web pages marked-up with Schema.org vocabulary and submitted to Bing are compared with appearance results of HTML web pages marked-up with Schema.org vocabulary and submitted to Google (see Table 14 and Table 15).

Comparison within each web search engine				
Bing Results			Google Results	
Dublin Core Vs Schema.org			Dublin Core Vs Schema.org	
Web pages marked-up with Dublin Core elements	Web pages marked-up with Schema.org properties		Web pages marked-up with Dublin Core elements	Web pages marked-up with Schema.org properties

Table 14: Comparison within each web search engine

Comparison between the two search engines				
Dublin Core Results			Schema.org Results	
Bing Vs Google			Bing Vs Google	
Web pages marked-up with Dublin Core submitted to Bing	Web pages marked-up with Dublin Core submitted to Google		Web pages marked-up with Schema.org submitted to Bing	Web pages marked-up with Schema.org submitted to Google

Table 15: Comparison between the two search engines

The results obtained from HTML web pages marked-up with Dublin Core vocabulary were collected for a period of 87 days starting from the 06th of July 2014 to the 01st of October 2014 (refer Figure 25 and Figure 26) after submission of the HTML web pages. The 58 HTML web pages, the index page and a sitemap (in XML format) containing URL links to the 58 HTML web pages, were all registered with the webmaster tools of the two web search engines (Bing and Google) on the same day.

© University of Pretoria

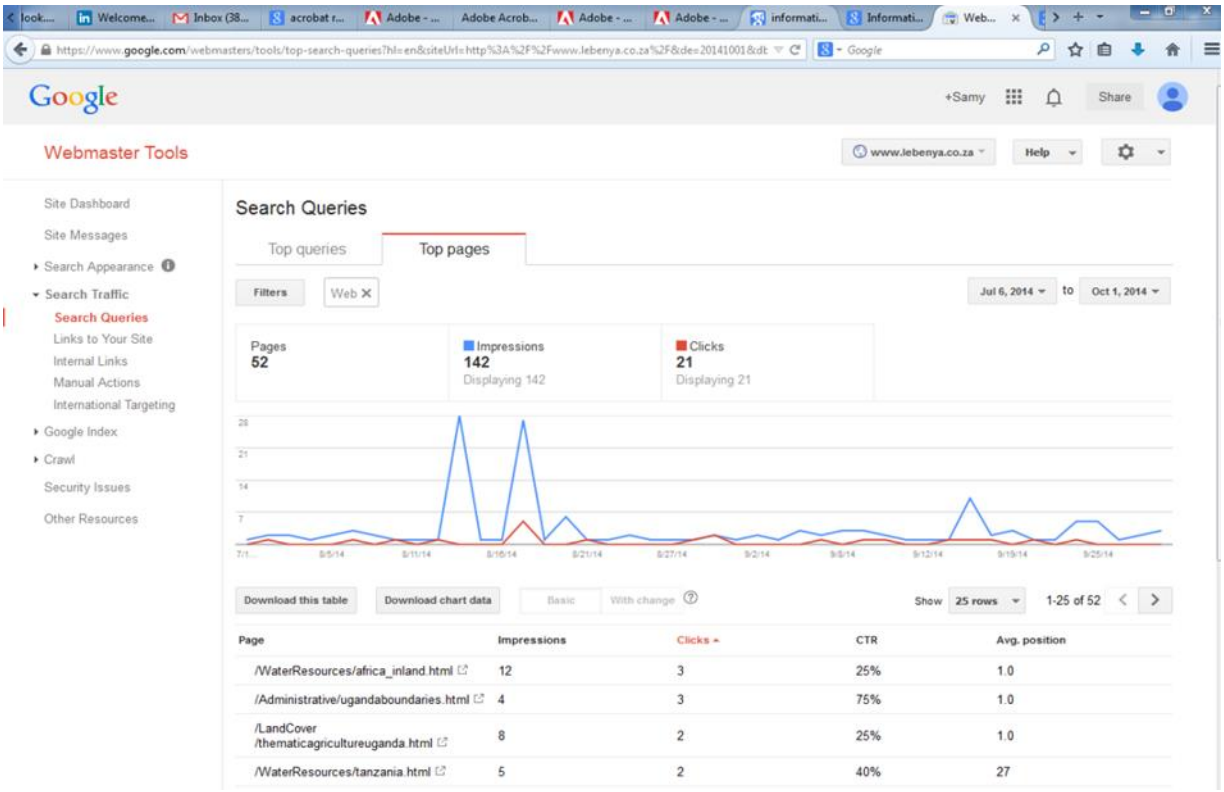

Figure 25: HTML web pages marked-up with Dublin Core listed on Google Webmaster Tools

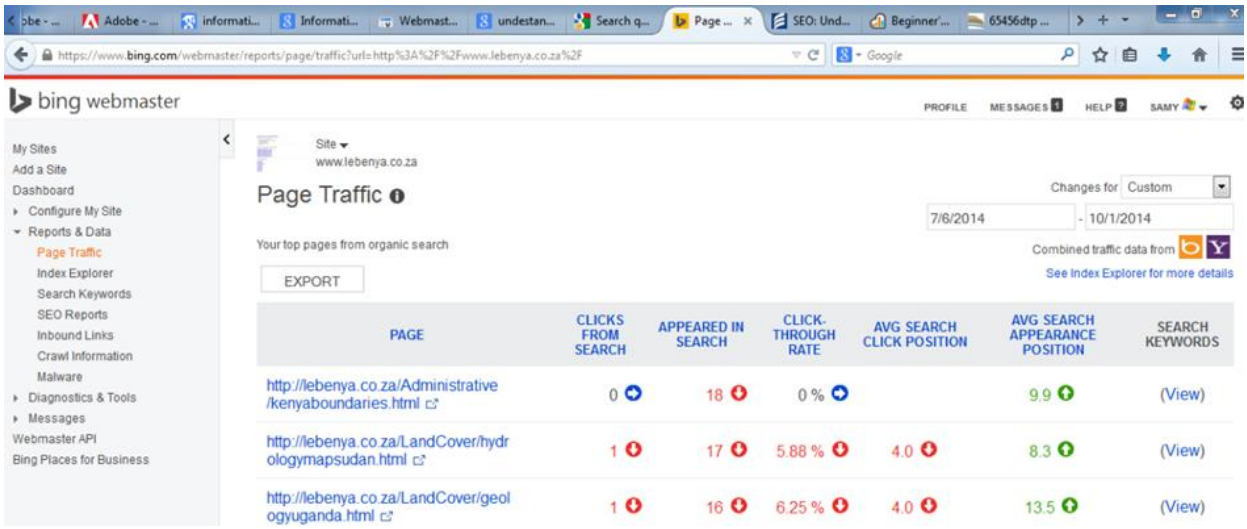

Figure 26: HTML web pages marked-up with Dublin Core listed on Bing Webmaster Tools

© University of Pretoria

Results obtained from HTML web pages marked-up with Schema.org vocabulary were also collected for a period of 87 days but starting from the 23rd of November 2014 to the 18th of February 2015 (refer Figure 27 and Figure 28). The 58 HTML web pages and the index page were all registered with the webmaster tools of the two web search engines (Bing and Google) on the same day. Furthermore, a sitemap (in XML format) containing URL links to the 58 HTML web pages, was submitted (on the same day) to each of the two web search engines (Bing and Google).

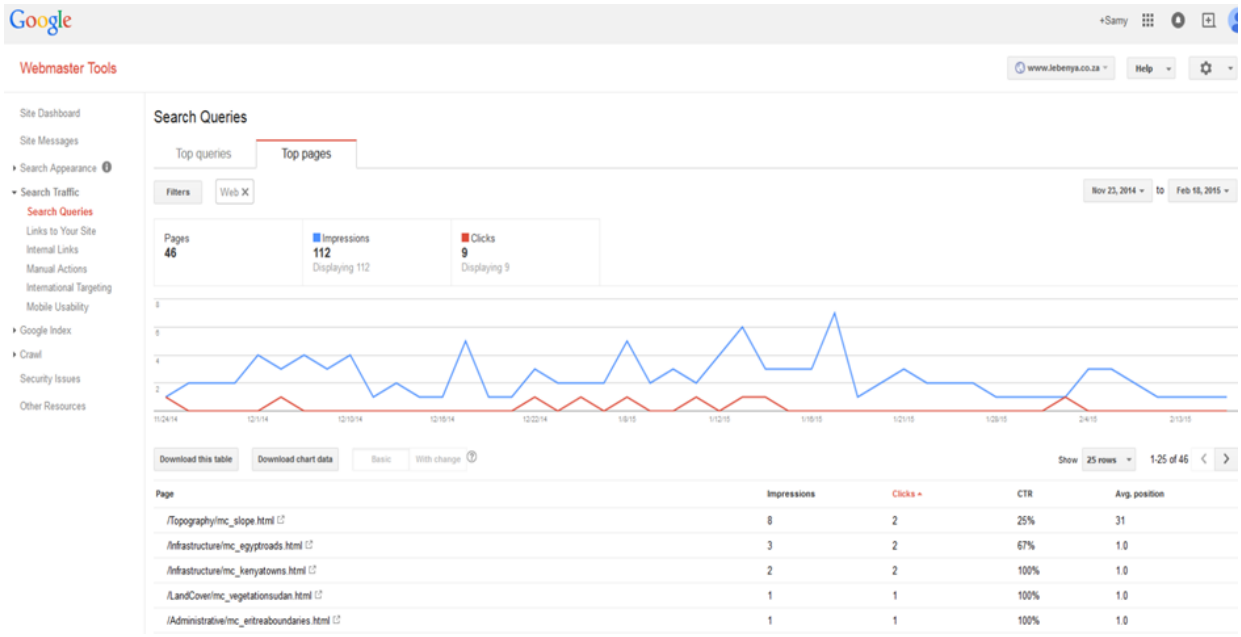

Figure 27: HTML web pages marked-up with Schema.org listed on Google Webmaster Tools

© University of Pretoria

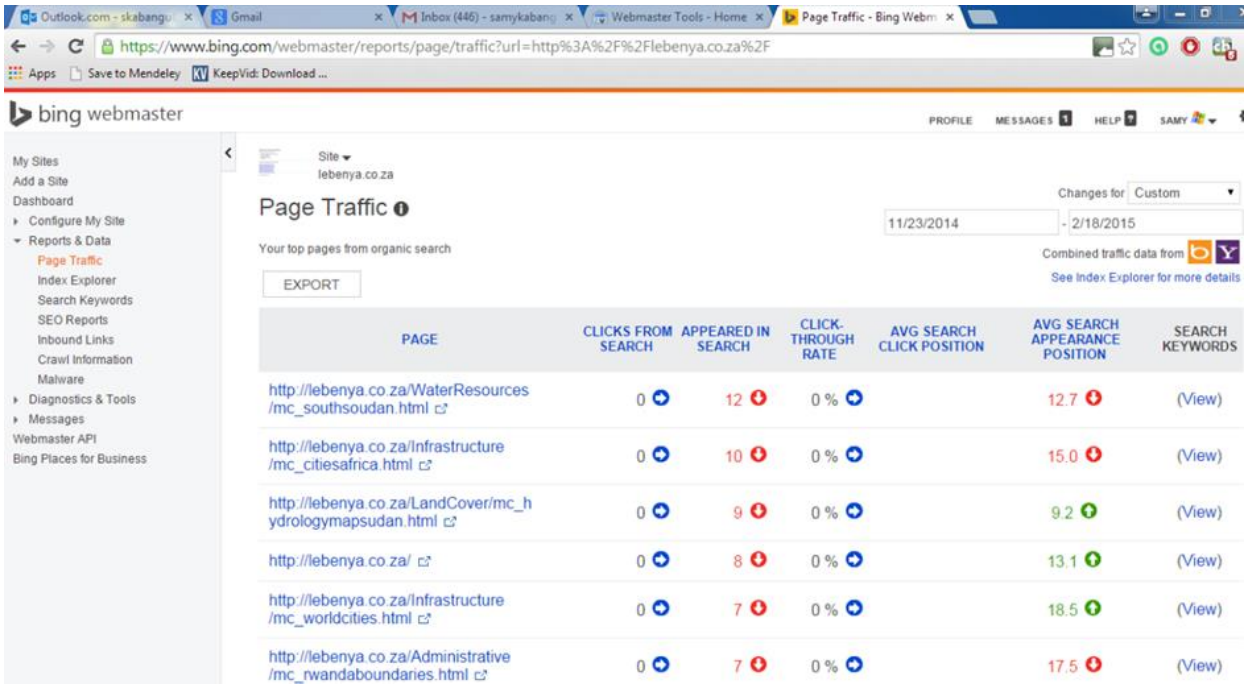

Figure 28: HTML web pages marked-up with Schema.org listed on Bing Webmaster Tools

7.6 Comparison between the two search engines based on metadata vocabularies (Dublin Core and Schema.org)

The main aim of this comparison is to assess the effect of metadata vocabularies namely, Dublin Core and Schema.org on the retrieval effectiveness of the submitted HTML web pages (with spatial metadata content) by web search engines (Bing and Google). For each metadata vocabulary, results obtained from Bing webmaster tools and Google webmaster tools applications are compared. The analysis of the results is discussed in subsequent sections.

7.6.1 Dublin Core results: Bing versus Google

The analysis described in this section reports on the comparison of results obtained from the Bing and Google web search engines' webmaster tools applications based on HTML web pages marked-up with Dublin Core vocabulary.

Out of a total of 58 HTML web pages marked-up with Dublin Core metadata vocabulary, uploaded onto the online web server and registered with both Bing and Google web search engines, 49 HTML web pages were visible on the search results pages of Google and 15 HTML web pages were visible on the search results pages of Bing.

© University of Pretoria

Dublin Core Results	**Bing**	**Google**
HTML web pages Retrieved (Impressions)	15	49
Recall	0.26	0.84

Table 16: Dublin Core results (Bing versus Google)

Google achieved a recall of 0.84 which is higher than 0.26 for Bing (refer Table 16). Hence, it can be said that based on the time period for collecting the results and the recall measures computed thereafter, HTML web pages marked-up with Dublin Core are effectively retrieved by Google web search engine as compared to Bing search engine. Such an argument could also be supported by the fact that Bing web search engine via its 'Markup Validator' tool did not recognise any markups (or tags) in HTML web pages marked-up with Dublin Core vocabulary. Google detected HTML web pages marked-up with Dublin Core vocabulary through its 'Structured Data Testing Tool' application. However, one could also argue that Google achieved a high recall because of its popularity among web search engines' users as compared to Bing.

To test whether there is a statistically significant difference in the average positions of web pages on the search results pages of web search engines between the two samples of results, an appropriate test statistics had to be employed. Being cognisant of the fact that the results (average position) from the two samples are dependent and paired, the Wilcoxon Signed Ranks test was chosen. It should be noted that not all the submitted 58 pages (as classified under the two vocabularies: Dublin and Schema.org) appeared on the search results pages of the two search engines for all the comparison cases considered. As such, any HTML page that did not appear on the results pages of any of the two considered search engines was assigned the value of 1000 as average position. This was done to allow the 'one to one' matching or pairing of average positions between HTML web pages. Hence, there was no need to test for normality. Descriptive statistics of the average positions of web pages marked-up with Dublin Core as retrieved by both search engines respectively are provided in Table 17 (Please note that the maximum value of 1000 as average position was assigned to HTML web pages that did not appear on the search results pages of the two considered web search engines).

© University of Pretoria

Descriptive Statistics

	N	Mean	Std. Deviation	Minimum	Maximum
DC Bing AvgPosition	58	744.5034	436.37868	1.00	1000.00
DC Google AvgPosition	58	163.1345	363.52145	1.00	1000.00

Table 17: Descriptive Statistics – Average positions of web pages marked up with Dublin Core (Bing versus Google)

Ranks

		N	Mean Rank	Sum of Ranks
DC Google AvgPosition - DC Bing AvgPosition	Negative Ranks	47[a]	27.06	1272.00
	Positive Ranks	4[b]	13.50	54.00
	Ties	7[c]		
	Total	58		

a. DC Google AvgPosition < DC Bing AvgPosition

b. DC Google AvgPosition > DC Bing AvgPosition

c. DC Google AvgPosition = DC Bing AvgPosition

Table 18: Description of ranks – Average positions of web pages marked-up with Dublin Core (Bing versus Google)

© University of Pretoria

Test Statistics[a]

	DC Google AvgPosition - DC Bing AvgPosition
Z	-5.906[b]
Asymp. Sig. (2-tailed)	.000

a. Wilcoxon Signed Ranks Test

b. Based on positive ranks.

Table 19: Test Statistics – Average positions of web pages marked up with Dublin Core (Bing versus Google)

The results from the statistical tests performed using SPSS 23 statistical software package as described in the Table 18 and Table 19, suggest that at 5% level of significance, the null hypothesis stating that the average positions between the two samples are equal, be rejected. Such a decision can be confirmed by the p-value (0.00) being less than the significance level (0.05). As such, it can be observed that HTML web pages marked-up with Dublin Core appeared on higher positions (lower average position) on Google web search engine as compared to Bing (refer Table 18 and Table 19).

7.6.2 Schema.org results: Bing versus Google

A comparison of results obtained from the Bing and Google web search engines' webmaster tools applications based on HTML web pages marked-up with Schema.org vocabulary is discussed in this section.

A total number of 45 HTML web pages marked-up with Schema.org vocabulary were retrieved on the search results pages of Google web search engine. On the search results pages of Bing web search engine, 39 HTML web pages marked-up with Schema.org vocabulary were retrieved (refer Table 20).

© University of Pretoria

Schema.org Results	**Bing**	**Google**
HTML web pages Retrieved (Impressions)	39	45
Recall	0.67	0.78

Table 20: Schema.org results (Bing versus Google)

Google achieved a recall measure of 0.78 higher than the one achieved by Bing, i.e. 0.67 (refer Table 20). Similarly to the observation made in section 7.6.1, it can be said that based on the time period it took to collect the results and the computed recall measures thereafter, HTML web pages marked-up with Schema.org are effectively retrieved by Google web search engine as compared to Bing. However, in this instance, both web search engines (Bing and Google) detected HTML web pages marked-up with Schema.org vocabulary.

To test whether there is a statistically significant difference in the average positions of web pages on the search results pages of web search engines between the two samples of results, an appropriate test statistics had to be employed. Given the fact that the results (average position) from the two samples are dependent (related) and paired, the Wilcoxon Signed Ranks test was chosen. It should be noted that not all the submitted 58 pages (as classified under the two vocabularies: Dublin Core and Schema.org) appeared on the search results pages of the two search engines for all the comparison cases considered. Therefore, to allow for the 'one to one' matching or pairing of average positions between HTML web pages, any HTML page that did not appear on the results pages of any of the two considered search engines was assigned the value of 1000 as average position. Hence, there was no need to test for normality. Descriptive statistics of the average positions of web pages marked-up with Schema.org as retrieved by both search engines respectively are provided in Table 21 (Please note that the maximum value of 1000 as average position was assigned to HTML web pages that did not appear on the search results pages of the two considered web search engines).

© University of Pretoria

Descriptive Statistics

	N	Mean	Std. Deviation	Minimum	Maximum
Schema Bing AvgPosition	58	338.1828	466.41361	2.00	1000.00
Schema Google AvgPosition	58	209.8793	407.14617	1.00	1000.00

Table 21: Descriptive Statistics – Average positions of web pages marked up with Schema.org (Bing versus Google)

Ranks

		N	Mean Rank	Sum of Ranks
Schema Google AvgPosition - Schema Bing AvgPosition	Negative Ranks	43[a]	26.16	1125.00
	Positive Ranks	10[b]	30.60	306.00
	Ties	5[c]		
	Total	58		

a. Schema Google AvgPosition < Schema Bing AvgPosition

b. Schema Google AvgPosition > Schema Bing AvgPosition

c. Schema Google AvgPosition = Schema Bing AvgPosition

Table 22: Description of ranks – Average positions of web pages marked up with Schema.org (Bing versus Google)

© University of Pretoria

Test Statistics[a]

	Schema Google AvgPosition - Schema Bing AvgPosition
Z	-3.631[b]
Asymp. Sig. (2-tailed)	.000

a. Wilcoxon Signed Ranks Test

b. Based on positive ranks.

Table 23: Test Statistics – Average positions of web pages marked up with Schema.org (Bing versus Google)

The results from the statistical tests performed using SPSS 23 statistical software package as described in the Table 22 and Table 23, suggest that at 5% level of significance, the null hypothesis stating that the average positions between the two samples are equal, be rejected. Such a decision can be confirmed by the p-value (0.00) being less than the significance level (0.05). As such, it can be observed that HTML web pages marked-up with Schema.org appeared on higher positions (lower average position) on Google web search engine as compared to Bing (refer Table 22 and Table 23).

7.7 Comparison within each web search engine based on web search engine results (Bing and Google)

This comparison aims to assess for each web search engines (Bing and Google) the effect that metadata vocabularies namely, Dublin Core and Schema.org have on the retrieval effectiveness of the submitted HTML web pages with spatial metadata content. In other words, this evaluation assesses for a given search engine (Bing or Google), the impact that metadata vocabularies have on the retrieval effectiveness of HTML web pages. The comparison is performed within the search engine. The results obtained with HTML web pages marked-up with Dublin Core vocabulary were collected from the 06th of July 2014 to the 01st of October 2014 while results obtained with HTML web pages marked-up with Schema.org vocabulary were collected from the 23rd of November 2014 to the 18th of February 2015.

© University of Pretoria

7.7.1 Bing Results: Dublin Core versus Schema.org

On the search results pages of Bing, 39 HTML web pages marked-up with Schema.org vocabulary were retrieved and only 15 HTML web pages marked-up with Dublin Core were retrieved. These web pages were retrieved out of a total number of 58 pages uploaded onto a web server. This gives Bing recall measures of 0.67 for HTML web pages marked-up with Schema.org vocabulary and 0.26 for HTML web pages marked-up with Dublin Core vocabulary (refer Table 24). Hence, it can be said that within the time period it took to collect the results with Bing web search engine, HTML web pages marked-up with Schema.org vocabulary were more effectively retrieved as compared to those marked-up with Dublin Core vocabulary.

Bing Results	**Dublin Core**	**Schema.org**
HTML web pages Retrieved (Impressions)	15	39
Recall	0.26	0.67

Table 24: Bing Results(Dublin Core versus Schema.org)

Given that the content of HTML web pages remained the same for both metadata vocabularies (however, differing on the mark-ups employed), one could argue that HTML web pages marked-up with Schema.org vocabulary achieved a higher recall measure because web pages' contents were already cached by Bing web search engine when such contents were crawled with earlier submitted HTML web pages marked-up with Dublin Core vocabulary. However, all the 58 HTML web pages marked-up with Dublin Core vocabulary were replaced on the web server with HTML web pages marked-up with Schema.org vocabulary and were renamed appropriately in order to distinguish them from the previously submitted HTML web pages marked-up with Dublin Core (see retrieved URLs and corresponding average positions in appendix H and in appendix I).

To test whether there is a statistically significant difference in the average positions of web pages on the search results pages of the web search engine (Bing) between the two samples of results, an appropriate test statistics had to be employed. Given the fact that the results (average positions) from the two samples are dependent (related) and paired, the Wilcoxon Signed Ranks test was chosen. It should be noted that not all the submitted 58 pages (as classified under the two vocabularies: Dublin Core and Schema.org) appeared on the search results pages of Bing to

© University of Pretoria

allow for all the comparison cases considered. That is, the 'one to one' matching or pairing of average positions between HTML web pages. As such, any HTML web page that did not appear on the results pages of the considered search engine (Bing) was assigned the value of 1000 as average position. Hence, there was no need to test for normality. Descriptive statistics of the average positions of web pages marked-up with Dublin Core and Schema.org as retrieved by Bing, are provided in Table 25 (Please note that the maximum value of 1000 as average position was assigned to HTML web pages that did not appear on the search results pages of Bing web search engine).

Descriptive Statistics

	N	Mean	Std. Deviation	Minimum	Maximum
Bing DC AvgPosition	58	744.5034	436.37868	1.00	1000.00
Bing Schema AvgPosition	58	338.1828	466.41361	2.00	1000.00

Table 25: Descriptive Statistics results - Bing Results (Dublin Core versus Schema.org)

The null hypothesis was set to be that the average positions between the two samples are equal. The results from the statistical tests performed using SPSS 23 statistical software package as described in Table 26 and Table 27, suggest that at 5% level of significance, the null hypothesis stating that the average positions between the two samples are equal, be rejected. Such a decision can be confirmed by the p-value (0.001) being less than the significance level (0.05). Hence, it can be concluded that HTML web pages marked-up with Schema.org vocabulary and retrieved with Bing achieved a higher visibility (lower average positions) as compared to HTML web pages marked-up with Dublin Core and retrieved with Bing (refer Table 26 and Table 27).

© University of Pretoria

Ranks

		N	Mean Rank	Sum of Ranks
Bing Schema AvgPosition - Bing DC AvgPosition	Negative Ranks	35[a]	21.59	755.50
	Positive Ranks	8[b]	23.81	190.50
	Ties	15[c]		
	Total	58		

a. Bing Schema AvgPosition < Bing DC AvgPosition

b. Bing Schema AvgPosition > Bing DC AvgPosition

c. Bing Schema AvgPosition = Bing DC AvgPosition

Table 26: Description of ranks - Bing Results (Dublin Core versus Schema.org)

Test Statistics[a]

	Bing Schema AvgPosition - Bing DC AvgPosition
Z	-3.411[b]
Asymp. Sig. (2-tailed)	.001

a. Wilcoxon Signed Ranks Test

b. Based on positive ranks.

Table 27: Test Statistics - Bing Results (Dublin Core versus Schema.org)

7.7.2 Google Results: Dublin Core versus Schema.org

When comparing results obtained from Google between the two vocabularies, 45 HTML web pages marked-up with Schema.org vocabulary and 49 HTML webpages marked-up with Dublin Core were retrieved on the Google search results page. Recall measures are: 0.78 for HTML

© University of Pretoria

web pages marked-up with Schema.org vocabulary and 0.84 for HTML web pages marked-up with Dublin Core vocabulary (refer Table 28).

Google Results	**Dublin Core**	**Schema.org**
HTML web pages Retrieved (Impressions)	49	45
Recall	0.84	0.78

Table 28: Google Results (Dublin Core versus Schema.org)

In contrast to the results described in section 7.7.1 where HTML web pages marked-up with Dublin Core achieved a lower recall with Bing web search engine, in this instance, HTML web pages marked-up with Dublin Core achieved a higher recall measure with Google as compared to those marked-up with Schema.org.

A further statistical analysis involving the comparison of average positions of the retrieved web pages (HTML web page marked-up with Schema.org and HTML web pages marked-up with Dublin Core) on Google's search results pages was performed. The aim was to confirm whether there is a statistically significant difference between values of the two samples (retrieved HTML web pages marked-up with Schema.org and retrieved HTML web pages marked-up with Dublin Core) in terms of their average positions on search results pages. Given the fact that the results (average positions) from the two samples are dependent (related) and paired, the Wilcoxon Signed Ranks test was employed. It should be noted that not all the submitted 58 pages (as classified under the two vocabularies: Dublin Core and Schema.org) appeared on the search results pages of Google to allow for all the comparison cases considered. That is, the 'one to one' matching or pairing of average positions between HTML web pages. As such, any HTML web page that did not appear on the results pages of the considered search engine (Google) was assigned the value of 1000 as average position. Hence, there was no need to test for normality. Descriptive statistics of the average positions of web pages marked-up with Dublin Core and Schema.org as retrieved by Google, are provided in Table 29 (Please note that the maximum value of 1000 as average position was assigned to HTML web pages that did not appear on the search results pages of Google web search engine).

© University of Pretoria

Descriptive Statistics

	N	Mean	Std. Deviation	Minimum	Maximum
Google DC AvgPosition	58	163.1345	363.52145	1.00	1000.00
Google Schema AvgPosition	58	227.1034	419.10995	1.00	1000.00

Table 29: Descriptive Statistics results - Bing Results (Dublin Core versus Schema.org)

The null hypothesis was set to be that the average positions between the two samples are equal. The results from the statistical tests performed using SPSS 23 statistical software package as described in Table 30 and Table 31, suggest that at 5% level of significance, the null hypothesis stating that the average positions between the two samples are equal, be accepted. Such a decision can be confirmed by the p-value (0.458) being greater than the significance level (0.05). Hence, it can be concluded that HTML web pages marked-up with Schema.org vocabulary and retrieved with Google achieved the same level of visibility as compared to HTML web pages marked-up with Dublin Core and retrieved with Google (refer Table 30 and Table 31).

© University of Pretoria

Ranks

		N	Mean Rank	Sum of Ranks
Google Schema AvgPosition - Google DC AvgPosition	Negative Ranks	12[a]	13.25	159.00
	Positive Ranks	15[b]	14.60	219.00
	Ties	31[c]		
	Total	58		

a. Google Schema AvgPosition < Google DC AvgPosition

b. Google Schema AvgPosition > Google DC AvgPosition

c. Google Schema AvgPosition = Google DC AvgPosition

Table 30: Description of ranks - Bing Results (Dublin Core versus Schema.org)

Test Statistics[a]

	Google Schema AvgPosition - Google DC AvgPosition
Z	-0.743[b]
Asymp. Sig. (2-tailed)	0.458

a. Wilcoxon Signed Ranks Test

b. Based on negative ranks.

Table 31: Test Statistics - Bing Results (Dublin Core versus Schema.org)

© University of Pretoria

7.8 Summary of results and chapter conclusion

The summary of the results discussed in this chapter is provided in Table 32.

	Dublin Core		Schema.org	
Search Engines	**Web pages retrieved**	**Recall**	**Web pages retrieved**	**Recall**
Google	49	0.84	45	0.78
Bing	15	0.26	39	0.67

Table 32: Retrieval effectiveness: summary of results

7.8.1 Results by metadata vocabulary

In terms of evaluating the effect that each of the metadata vocabularies (Dublin Core and Schema.org) has on the retrieval effectiveness of the submitted HTML web pages by search engines, Google achieved a high recall as compared to Bing in retrieving HTML web pages marked-up with Dublin Core vocabulary. This is further confirmed in the detailed summary table (refer Table 33) where 37 web pages (out of 49 pages) marked-up with Dublin Core vocabulary were only retrieved by Google. However, web pages marked-up with Schema.org vocabulary were effectively retrieved by both search engines (31 web pages retrieved) as compared to web pages marked-up with Dublin Core vocabulary (only 12 web pages retrieved). This latter observation suggests that the two search engines were more effective at retrieving web pages marked-up with Schema.org vocabulary as compared to Dublin Core vocabulary.

Vocabulary	Found only in Google	Found only in Bing	Found in Both	None
Dublin Core	37	3	12	6
Schema.org	14	8	31	5

Table 33: Retrieval effectiveness: detailed summary of results by vocabularies

Based on the computed descriptive statistics and the statistical test (Wilcoxon Signed Ranks test) performed, web pages marked-up with Dublin Core vocabulary appeared at higher positions

© University of Pretoria

(more visible) on Google's search results pages when compared to web pages retrieved on Bing's search results pages. With respect to Schema.org vocabulary, Google achieved a high recall as compared to Bing in retrieving HTML web pages marked-up with Schema.org vocabulary. Based on the computed descriptive statistics and the statistical test (Wilcoxon Signed Ranks) performed, HTML web pages marked-up with Schema.org appeared at higher positions (more visible) on Google's search results pages as compared to HTML web pages retrieved on Bing's search results pages. The observed difference between the average positions of the submitted HTML web pages on Google and Bing search results pages was also statistically significant. Overall, HTML web pages marked-up with either of the two metadata vocabularies (Schema.org or Dublin Core) are effectively retrieved by Google as compared to Bing. Furthermore, the difference in recall measures is reduced when HTML web pages are marked-up with Schema.org vocabulary (refer Table 32).

7.8.2 Results by web search engine

Another way of comparing the results consisted of evaluating the retrieval effectiveness of the submitted HTML web pages for each of the web search engines. That is, the retrieval effectiveness is compared between HTML web pages marked-up with Dublin Core vocabulary and HTML web pages marked-up with Schema.org vocabulary within each web search engine.

With respect to Bing's results, HTML web pages marked-up with Schema.org vocabulary achieved a high recall as compared to HTML web pages marked-up with Dublin Core. This is confirmed in the detailed summary table (Table 34) where 29 web pages (out of 39 pages) marked-up with Schema.org vocabulary were only retrieved by Bing. However, web pages marked-up with both vocabularies (Schema.org and Dublin Core) were effectively retrieved by Google (37 web pages in total) as compared to Bing (only 10 web pages retrieved). This latter observation suggests that Google was more effective at retrieving web pages marked-up with both vocabularies (Schema.org and Dublin Core) as compared to Bing.

© University of Pretoria

Search Engine	Found only with Dublin Core	Found only with Schema.org	Found with Both (Dublin Core & Schema.org	None
Google	12	8	37	1
Bing	5	29	10	14

Table 34: Retrieval effectiveness: detailed summary of results by search engines

Based on the computed descriptive statistics and the statistical test (Wilcoxon Signed Ranks test) performed, the average positions of HTML web pages marked-up with Schema.org on Bing's search results are lower than the ones of HTML web pages marked-up with Dublin Core. This suggests that HTML web pages marked-up with Schema.org have a better visibility on Bing's search results pages as compared to those (HTML pages) marked-up with Dublin Core. Furthermore, based on the statistical analysis performed, the difference between the average positions (of the two sets of results) was also found to be statistically significant.

With respect to Google's results, HTML web pages marked-up with Dublin Core vocabulary achieved a high recall measure as compared to HTML web pages marked-up with Schema.org. However, this difference in recall measures between the two sets of HTML web pages is not very large. A further statistical analysis test (Wilcoxon Signed Ranks test) suggests that the average positions of web pages from the two sets of results (HTML web pages marked-up with Dublin Core and HTML web pages marked-up with Schema.org) as retrieved by Google are equal. This implies that the difference of the average positions of HTML web pages marked-up with Dublin Core and those marked-up with Schema.org (as retrieved by Google), is not statistically significant (refer Table 30 and Table 31).

© University of Pretoria

Chapter 8

8 Conclusion

8.1 Introduction

This research emanates from the need to contribute towards finding possible solutions to the current challenges that impede the easy discovery of geospatial data on the web. Such a problem becomes even more evident when data requesters do not belong to any particular geo-information community, or do not know of a specific geoportal that serve the data they need. Given that current geoportals and associated web services that enable the discovery of geospatial data are often invisible to web search engines commonly used to find various kind of digital resources online, HTML pages with spatial metadata content have been proposed in this study as a means for discovering geospatial data published on the web. This dissertation reports on the evaluation of the retrieval effectiveness of these HTML pages (by Bing and Google) and the process involved in their construction. The later includes the computer laboratory experiment and the Bing webmaster tools experiment that resulted in the collection and analysis of search terms that users employed when searching for geospatial data on the web. A taxonomy of these search terms which provides an understanding of how users search for geospatial data on the web was constructed. The chain of mappings from the taxonomy of search terms to the *ISO 19115:2003, Geographic Information – Metadata* standard, to the Dublin Core metadata standard and finally to the Schema.org vocabulary, enabled the insertion of spatial metadata content (informed by the taxonomy of search terms) into HTML pages and their marking-up with Dublin Core and Schema.org vocabularies. A summary of the main findings, research contribution and possible future work based on the study reported in this dissertation, are provided in subsequent sections.

8.2 Summary of main findings

The main findings of this dissertation are summarised per the three objectives as set in the introduction chapter and attained in chapters 5, 6, and 7.

© University of Pretoria

8.2.1 Analysis of search terms and taxonomy construction

Search terms obtained from the two experiments performed, were analysed and assigned to four categories of the taxonomy of search terms. The four categories are: 'Application Domain', 'Location', 'Feature Type' and 'Data Model'. For both sets of results (search terms collected from the computer laboratory experiment and from the Bing webmaster tools experiment), the 'Application Domain' and 'Location' categories had the highest proportion (or number) of search terms. The 'Feature Type' and 'Data Model' categories had the lowest proportions of search terms in both experiments. The observed similarities led to conclude that the taxonomy of search terms was consistent. It was also found that the frequency of search terms (quantified using the mode as descriptive statistics) in a search query to be three after analysing the results of the two experiments separately. Such an observation suggests that search queries used for finding geospatial data on the web are frequently comprised of three terms. With the 'Application Domain' and 'Location' categories having the highest number of search terms, the three terms are more likely to fall within these two mentioned categories. Lastly, it was observed that the group of professionals had the highest ratios of the number of participants to the number of queries suggesting that professionals found relevant information (geospatial data) using fewer queries as compared to the first years and the third years. Such results corroborate the findings from the literature that suggest that users' expertise in the knowledge of search influences success when searching online information using web search engines.

8.2.2 Mappings between standards (ISO 19115:2003, Geographic Information – Metadata & Dublin Core) and Schema.org

The second objective of this dissertation was attained by creating two mappings (crosswalks) solely based on the definitions or descriptions of the elements involved. The two created crosswalks are between:

- The categories of the taxonomy of search terms and selected elements of *ISO 19115: 2003, Geographic Information – Metadata* standard.
- Selected elements of Dublin Core vocabulary and proprieties of selected types of Schema.org vocabulary.

The existing mapping between Dublin Core vocabulary and *ISO 19115:2003, Geographic Information – Metadata* standard provided the basis for incorporating spatial metadata contents

© University of Pretoria

into HTML web pages. While the created mapping between the taxonomy of search terms and *ISO 19115:2003, Geographic Information – Metadata* standard, provided a means for understanding how keywords employed by users when searching for geospatial data online, relate to appropriate elements of *ISO 19115:2003, Geographic Information – Metadata* standard. Based on these mappings (crosswalks), two sets of HTML web pages with spatial metadata contents were constructed: one set marked-up with Dublin Core vocabulary and the other set marked-up with Schema.org vocabulary. HTML web pages constructed in this manner serve as proxies for finding geospatial data on the web using general purpose web search engines.

8.2.3 Comparison of the retrieval effectiveness of HTML web pages with spatial metadata content

The retrieval effectiveness of HTML web pages marked-up with Dublin vocabulary Core was evaluated and compared to that of HTML web pages marked-up with Schema.org vocabulary. Analyses were performed in two instances: by metadata vocabulary whereby the results obtained were compared between the two web search engines (Bing versus Google) and within each search engine whereby the results obtained were compared between the metadata vocabularies (Dublin Core versus Schema.org).

In the first instance, HTML web pages marked-up within Dublin Core were retrieved effectively by Google (with a recall measure of 0.84) as compared to Bing (with a recall measure of 0.26). HTML web pages marked-up with Schema.org were effectively retrieved by Google (with a recall measure of 0.78) as compared to Bing (with a recall measure of 0.67). HTML web pages marked-up with Schema.org vocabulary appeared at higher positions on Google as compared to HTML web pages retrieved by Bing. Such a difference was found to be statistically significant. HTML web pages marked-up with Dublin Core appeared at higher positions on the results pages of Google as compared to Bing. The difference between the average positions of the two sets of HTML web pages was also found to be statistically significant.

In the second instance, Bing achieved a recall measure of 0.67 when retrieving HTML web pages marked-up with Schema.org and a recall measure of 0.26 when retrieving HTML web pages marked-up with Dublin Core. HTML web pages marked-up with Schema.org vocabulary appeared at higher positions on Bing as compared to HTML web pages marked-up with Dublin Core. This difference between the average positions of HTML marked-up with Dublin Core and

© University of Pretoria

HTML web pages marked-up with Schema.org (as retrieved by Bing) was found to be statistically significant. With respect to Google, HTML web pages marked-up with Dublin Core were effectively retrieved (with a recall measure of 0.84) as compared to HTML web pages marked-up with Schema.org (with a recall measure of 0.78). Such a difference can be said to be small as compared to the difference in recall measures obtained from the Bing results (0.26 and 0.67). The average positions of both sets: HTML web pages marked-up with Schema.org and HTML web pages marked-up with Dublin Core (as retrieved by Google) were found to be the same (statistically).

The overall outcome of this comparison is to conclude that in terms of web search engines, Google was more effective in retrieving the considered HTML web pages as compared to Bing. In terms of metadata vocabularies, HTML web pages marked-up with Schema.org were more effectively retrieved as compared to HTML web pages marked-up with Dublin Core when Bing was employed. However, the latter difference was found to be statistically insignificant when Google web search engine was employed.

8.3 Research contribution to science and technology

Besides the state of the art body of knowledge presented in the literature review and related work chapters, the study performed throughout this dissertation contributes towards two fields: geoinformatics and computer or information science. In the field of geoinformatics, this dissertation demonstrates how through a chain of mappings, a geospatial standard namely *ISO 19115:2003 , Geographic Information – Metadata*, can be used in conjunction with Dublin Core and Schema.org vocabularies, to construct HTML web pages with spatial metadata content for enhancing the discovery of geospatial data on the web using general purpose web search engines.

In the field of computer science or information science, namely the domain of online information retrieval, this dissertation:

- Elaborates on search engine optimisation (SEO) techniques applied in embedding spatial metadata into HTML web pages for easy discovery of geospatial data on the web.
- Proposes an approach which saw the construction of a taxonomy of search terms. Such a taxonomy contribute towards understanding search terms users employed when searching for geospatial data online using web search engines and how such search terms can be

© University of Pretoria

incorporated into HTML web pages that contain contents for describing geospatial data (also known as spatial metadata).

- Presents results on retrieval effectiveness of online resources (HTML web pages) describing geospatial data). In so doing, search engine performance evaluation is also performed.

8.4 Possible future work

After the attainment of the objectives assigned in this research, new arising questions requiring further research were identified. Such possible future works are elaborated in subsequent paragraphs.

In the construction of the taxonomy, search terms were manually allocated to categories based on subjective judgements. Future work could explore automatic categorisation assisted by human supervision. Furthermore, given that only descriptive statistics and inferential statictics were performed on small samples of search queries, advanced inferential statistical analyses could be performed on bigger search query sets in future work. Possibilities for broadening the taxonomy to have more levels so as to include more search terms can also be explored as part of the future work emanating from this research.

With respect to the mapping between selected elements of Dublin Core and Schema.org which was based on crosswalks, ontology-based semantic interoperability can be considered in future work. This would involve establishing semantic relationships between the classes (elements or types) of the two metadata vocabularies using relevant ontologies and Resource Description Framework (RDF). The development of a flexible metadata service that would be permeable to web search engines could also be considered as future work.

Extended periods for collecting the results pertaining to the evaluation of the retrieval effectiveness of web search engines can also be considered when planning for future work.

© University of Pretoria

References

Aalders HJGL (2005). An Introduction to Metadata for Geographic Information, In Moellering H, Aalders HJGL, Crane A, (Eds). World Spatial Metadata Standards, 1st Edition, Elsevier Ltd., pp. 3–27

Agichtein E, Brill E, Durmais Sand Ragno R (2006). Learning user interaction models for predicting web search results preferences, SICIR' 06, August 6-11, 2006, Seattle, Washington, USA

Akinci H and Comert C (2007). Geoportals and their role in spatial data infrastructures. Department of Geodesy and Photogrammetry Engineering, Turkey. Available at: http://www.researchgate.net/publication/228948121_Geoportals_and_their_role_in_spatial_data_infrastructures [Accessed October 1, 2014]

Ali R and Sufyan Beg MM (2011). An overview of web search evaluation methods, Computers and Electrical Engineering, vol. 37, pp. 835-848

Aloteibi S (2014). Analyzing geographic query reformulation: An exploratory study, Journal of the association for information science and technology, vol. 65(1), pp. 13-24, 2014

Ambiah N and Lukose D (2012), Enriching web pages with semantic information, International Conference on Dublin Core and Metadata Applications, DC-2012, Kuching, Sarawak, Malaysia

Beall J (2008). The Death of Metadata, The Serials Librarian: From the printed Page to the Digital Age, vol. 51(2), pp. 55-74

Bordogna G, Ghisalberti G and Psaila G (2012). Geographic information retrieval: Modeling uncertainty of user's context, vol. 196, pp. 105-124

Broder A (2002). A taxonomy of web search, SIGIR Forum, vol. 36 (2), pp. 3-10

Can F, Nuray R and Sevdik AB (2004). Automatic performance evaluation of web search engines, Information Processing and Management, vol. 40, pp. 495-514

Cahill K and Chalut R (2009). Optimal Results: What Libraries Need to know About Google and Search Engine Optimization, The Reference Libraries, vol. 50 (3), pp. 234-247

Carmel E, Crawford S and Chen H (1992). Browsing in hypertext: A cognitive study, pp. 865–884, Paper presented at the IEEE transactions on systems, man and cybernetics, 5–10 October, Chicago IL.

CEN Workshop Agreement (2003). Available at: ftp://cenftp1.cenorm.be/PUBLIC/CWAs/e-Europe/MMI-DC/cwa14857-00-2003-Nov.pdf , [Accessed 9 November 2013]

© University of Pretoria

Chuang SL and Chien LF (2002), Towards automatic generation of query taxonomy: a hierarchical query clustering approach, In International Conference on Data Mining (ICDM), pp. 75-82

Chrisman NR (1999). What Does 'GIS' Mean?, Transaction in GIS, vol. 3, no. 2, pp. 156-186

Clough P (2005). Extracting metadata for spatially-aware information retrieval on the internet, Proceedings of the 2005 workshop on Geographic Information retrieval, Pages 25 – 30

Croft B, Metzler D and Strohman T (2009). Search Engines: Information Retrieval in Practice, 1st edition, Addison-Wesley Publishing Company, USA

De Longueville B, (2010). Community-based geoportals: The next generation Concepts and methods, Computers, Environment and Urban Systems, vol. 34, pp. 299-308

DCMI (2001). Dublin Core Metadata glossary, Final draft. Available at: http://library.csun.edu/mwoodley/dublincoreglossary.html. [Accessed 18 November 2014]

Dublin Core Metadata Initiative (DCMI) (2012), Available at: http://dublincore.org/, [Accessed 15 November 2013]

Dueker KJ and Kjerne D (1989). Multipurpose Cadastre: Terms and Definitions, *American Society for Phogrammetry and Remote Sensing and American Congress on Surveying and Mapping*, Falls Church VA

Foley R (2009). Integrated spatial data infrastructure, Maynooth, Elsevier Ltd.

Gan Q, Attenberg J, Marowetz A and Suel T (2008). Analysis of geographic queries in search engine log, LocWEB 2008, April 22, Beijing China

Giacomeli P (2006). Integration of GIS and Simulation Models, In GIS for Sustainable development, [Campagna M. (ed.)], Taylor & Francis Group, USA, chapter 11, pp. 181-192

Google (2014), Google search engine optimization starter guide, Available at: http://static.googleusercontent.com/media/www.google.com/en//webmasters/docs/search-engine-optimization-starter-guide.pdf, [Accessed 1st November 2014]

Google Webmaster tools (2014), Meta tags that Google understands, Available at: https://support.google.com/webmasters/answer/79812?hl=en [Accessed 22nd November 2014]

Govorov M (2008). Training Material, STANDARDS, SPECIFICATIONS AND METADATA FOR GEOGRAPHIC INFORMATION (GII-03), Civil servants leaning programme, Distance learning of Geographic information infrastructure, Vilnius, Available at:

© University of Pretoria

http://www.geoportal.lt/download/gii_mokymai/GII_03_mokomoji_medziaga/En/Paskaitu_konspektai/GII-03_training_material.pdf , [Accessed 1 July 2014]

Green D and Bossomaier T (2002). Online GIS and Spatial Metadata, Taylor & Francis, London

Harper CA (2010). Dublin Core Metadata Initiative: Beyond the element set, Information standards quarterly, Winter 2010, vol. 22(1), ISSN 1041-0031

Hawking D (2006). Web search engines - Part1, IEEE Computer Society, vol. 39 (6), pp.86-88

Holscher C and Strube G (2000). Web search behavior of Internet experts and newbies, Computer Networks, vol. 33, pp. 337–346

ISO 15836:2009 (2009). The Dublin Core metadata element set', International Organization for Standardisation (ISO), Geneva, Switzerland.

ISO 19115: 2003(2003). Geographic Information – Metadata, International Organization for Standardisation (ISO), Geneva, Switzerland

Katumba S, Coetzee S and De la Rey A (2012). An assessment of spatial data availability for renewable energy planning in sub-Saharan Africa. GISSA Ukubuzana 2012, ISBN 978-0-620-52913-6, Ekurhuleni, South Africa, 2-4 October 2012.

Katumba S and Coetzee S (2013). Spatial data discovery using general purpose web search engines, 26th International Cartographic Conference, Dresden, Germany, 25 - 30 August 2013

Katumba S and Coetzee S (2015). Enhancing the online discovery of geospatial data through taxonomy, folksonomy and semantic annotations, South African Journal of Geomatics, 4(3), pp. 339-350

Krutil J, Kudelka M and Snasel V (2012). Web page classification based on Schema.org collection, Fourth International Conference on Computational Aspects of Social Networks (CASoN).

Kumar S and Prakash JN (2009). Precision and relative recall of search engines: A comparative study of Google and Yahoo, Singapore Journal of Library & Information Management, vol. 38

Limam L, Coquil D, Kosch H and Brunie L (2010). Extracting user interests from search query logs: A clustering approach, Workshops on Database and Expert Systems Applications, pp. 5-9, Washington, DC, USA

Liu X, Song Y, Liu S and Wang H (2012). Automatic Taxonomy Construction from Keywords, In Proceedings of the 18th International conference on Knowledge discovery and data mining, pp. 1433-1441, New York, USA

© University of Pretoria

Longley AP, Goodchild M, Maguire JD and Rhind WD (2005). Geographic Information Systems and Science (2nd ed.), Chichester, John Wiley & Sons Ltd.

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

Lopez-Pellicer JF, Florczyk JA, Nogueras-Iso J, Muro-Medrano P and Zarazaga JF (2010). Exposing CSW catalogues as Linked Data, In Painho M, Santos MY and Pundt H, (Eds), Geospatial Thinking, Lecture Notes in Geoinformation and Cartography (LNG&C), Springer-Verlag, vol. 0, pp. 183-200.

Maguire D and Longley P (2005). The emergence of geoportals and their role in spatial data infrastructure, Computers, Environment and Urban Systems, vol. 29, pp. 3-14

Malaga RA (2010). Search Engine Optimization-Black and White Hat Approaches, Advances in computers, vol. 78, pp. 1-39

Masser I (2005). GIS worlds: Creating spatial data infrastructure, Redlands, ESRI Press.

McGee M (2010). By the numbers: Twitter vs Facebook vs Google Buzz", Search Engine Land, available at: http://searchengineland.com/by-the-numbers-twitter-vs-facebook-vs-google-buzz-36709 [accessed 01 March 2015]

Muhleisen H, Bizer C (2012). Web Data Commons-Extracting structured data from two large web corpora. In Proceedings of the www2012 workshop on Linked Data on the web, April 16, 2012, Lyon, France

Mustafa R, Naawaz MS, Lali MI (2015). Search engine optimization techniques to get high score in SERP's using recommended guidelines, Science International (Lahore), vol. 27, no. 6, pp. 1013-5316

Nebert D (ed.) (2004). Developing spatial data infrastructures: The SDI cookbook, version 2.0

Nogueras-Iso J, Zarazaga-Soria FJ, Lacasta J, Bejar RB and Muro-Medrano PR (2004). Metadata standard interoperability: application in the geo-graphic information domain, Computers, Environment and Urban Systems, vol. 28, pp.611–634

Onaifo D and Rasmussen D (2013). Increasing libraries' content findability on the web with search engine optimization, Library Hi Tech, vol. 31, no. 1, pp. 87-108

Ochoa E (2012). An analysis of the application of selected search engines optimization (SEO) techniques and their effectiveness on Google's search ranking algorithm, MSc dissertation, California State University. 2012

© University of Pretoria

Open Geospatial Consortium (OGC) (2007). "OpenGIS® Catalogue Services Specification", in Nebert D, Whiteside A, Vretanos PP (Eds), available at http://www.opengeospatial.org/standards/cat, [accessed 01 May 2013]

Pew Internet Project (2011). Search and email still top the list of most popular online activities: two activities nearly universal among adult internet users, Pew Research Center's Internet & American Life Project, available at: http://pewinternet.org/,/media/Files/ Reports/2011/PIP_Search-and-Email.pdf [accessed 02 March 2015]

Rauch E, Bukatin M and Baker (2003). A confidence-based framework for disambiguating geographic terms, Proceedings of the HLT-NAACL 2003 workshop on Analysis of geographic references, Volume 1, Pages 50-54

Rose DE and Levinson D (2004). Understanding user goals in Web search, World Wide Web conference (WWW 2004), pp. 13–19, 17–22 May, New York, NY, USA

Sanchez D and Moreno A (2004). Automatic generation of taxonomies from the WWW. *In proceedings of the 5th International Conference on Practical Aspects of Knowledge Management (PAKM 2004) LNAI*, vol. 3336, Vienna, Austria

Saracevic Tand Kantor PB (1997). Studying the value of Library and Information Services. Part II. Methodology and Taxonomy, Journal of the American Society for Information Science, vol. 48 (6), pp. 543-563

Sellen AJ, Murphy R, and Shaw K L (2002). How knowledge workers use the Web, pp. 227–234. Paper presented at the conference on human factors in computing systems (CHI '02), 20–25 April, Minneapolis, Minnesota, USA.

Shen X, Tan B and Zhai C (2005). Implicit User Modeling for Personalized Search, CIKM'05, October 31–November 5, Bremen, Germany

Sontag D, Collins-Thompson K and Bennet PN (2012). Probabilistic Models for Personalizing Web Search, In WSDM'12, February 8–12, Seattle, Washington, USA

Spiteri L and Richard N (1999). Evaluation of internet search engines: Methodological issues and assumptions, Proceedings of the 27th Annual Conference of the Canadian Association for Information Science (CAIS)

Sugiyama K and Hatano K (2004). Adaptive Web Search Based on User Profile Constructed without Any Effort from Users, WWW2004, May 17–22, 2004, New York, New York, USA

Sullivan D (2009). The Microsoft-Yahoo Search Deal, In Simple terms, Available at: http://searchengineland.com/microsoft-yahoo-search-deal-simplified-23299 [Accessed 22 November 2014]

© University of Pretoria

Swati P, Pawar BV, Patil Ajay S (2013). Search Engine Optimization: A study, Research Journal of Computer and Information Technology Sciences, Vol. 1 (1), pp. 10–13

Thatcher A (2008). Web search strategies: The influence of Web experience and task type, Information, Processing and Management, vol. 44, pp. 1308–1329

Tait MG (2005). Implementing geoportals: applications of distributed GIS, Computers, Environment and Urban Systems, vol. 29, pp. 33–47

Vaughan L (2004). New measurements for search evaluation proposed and tested, Information processing and Management, vol. 40, pp 677-691

Vockner B and Mittlböck M (2014). Geo-Enrichment and Semantic Enhancement of Metadata Sets to Augment Discovery in Geoportals, vol. 3, pp. 345-367

Wang F, Li Y and Zhang Y (2011). An empirical study on the Search Engine Optimization Technique and Its Outcomes, Proceedisngs of the 2nd International Conference on Artificial Intelligence, Management Science and Electronic Commerce (AIMSEC), 2011

Weibel SL (2005). An Introduction to Metadata for Geographic Information, in Moellering H., Aalders HJGL, Crane A, (Eds). World Spatial Metadata Standards, 1st Edition: Elsevier Ltd., pp.493-513

White RW, Dumais ST and 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, pp. 132-141, New York, USA

Zhang J and Dimitroff A (2004). Internet search engines' reponse to metadata Dublin Core implementation, Journal of Information Science, vol. 30 (4), pp. 310-320

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

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

Zhang L, Sun X and Zhug H (2014). Density-based spatial keyword querying, Future Generation computer Systems, vol. 3 (2), pp 211-221

© University of Pretoria

Appendix A – Instructions to participants in the user experiment

Geoinformatics

Research Experiment

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Participant consent:
by signing this form, I agree to voluntarily participate to this experiment.
Name:…….. Signature:……

Objectives:

- Monitor participants' web searching behaviours in terms of their keywords selection.

Instructions:

1. Using Firefox web browser, open a web search engine of your choice (Google, Bing or Yahoo).
2. Search for spatial data to address a particular need (or to solve a given problem) of your choice for specific countries or cities in Africa. One should be able to visualize (and/or analyse) such spatial data on a GIS desktop software (ArcGIS or QGIS).
3. Continue to search up until you find spatial datasets that satisfy your need (or approximately so).
4. Once satisfied, please make sure to record the keywords that led you to the web site from which you found your most suitable spatial datasets. Please record that web site as well. (type these details into a word document).
5. With Firefox web browser opened, press **CTRL+H** to open the History tab (on the left side of the screen), right click on **Today** icon, copy it and paste the history results on a second word document.
6. Please write at least a paragraph on your spatial data web searching experience (on the first word document).
7. **Please remember to attach both word documents and email them to** skabangu@hotmail.com

Thank you!

© University of Pretoria

Appendix B – ISO 19115:2003, Geographic Information – Metadata (Core Elements)

Package	Name of the element	Definition	Maximum Occurrence	Data Type	Domain
MD_Metadata					
	fileIdentifier	unique identifier for this metadata file	1	CharacterString	Free text
	metadataStandardName	name of the metadata standard (including profile name) used	1	CharacterString	Free text
	metadataStandardVersion	version (profile) of the metadata standard used	1	CharacterString	Free text
	language	language used for documenting metadata	1	CharacterString	ISO 639-2, other parts may be Used
	characterSet	full name of the character coding standard used for the metadata set	1	Class	MD_CharacterSetCode <<CodeList>>
	contact	party responsible for the metadata information	N (multiple occurrences)	Class	CI_ResponsibleParty |<<DataType>>
	dateStamp	date that the metadata was created	1	Class	Date
MD_Identification		basic information required to uniquely identify a resource or resources		Aggregated Class (MD_Metadata) <<Abstract>>	
	citation	citation data for the resource(s)	1	Class	CI_Citation <<DataType>>
	abstract	brief narrative summary of the content of the resource(s)	1	CharacterString	Free text
	purpose	summary of the intentions with which the resource(s) was developed	1	CharacterString	Free text
	pointOfContact	identification of, and means of communication with, person(s) and organization(s) associated with the resource(s)	N	Class	CI_ResponsibleParty <<DataType>>
	Role name:	provides category keywords, their	N	Association	MD_Keywords

© University of Pretoria

	descriptiveKeywords	type, and reference source			
	spatialRepresentationType	method used to spatially represent geographic information	N	Class	MD_SpatialRepresentation TypeCode <<CodeList>>
MD_DataIdentification					
	spatialResolution	factor which provides a general understanding of the density of spatial data in the dataset	N	Class	MD_Resolution <<Union>>
	language	language(s) used within the dataset	N	CharacterString	ISO 639-2, other parts may be used
	characterSet	full name of the character coding standard used for the dataset	N	Class	MD_CharacterSetCode <<CodeList>>
	topicCategory	main theme(s) of the dataset	N	Class	MD_TopicCategoryCode <<Enumeration>>
	Extent (GeographicBoundingBox Or GeographicDescription)	extent information including the bounding box, bounding polygon, vertical, and temporal extent of the dataset	N	Class	EX_Extent <<DataType>>
CI_Citation					
	title	name by which the cited resource is known	1	CharacterString	Free text
	date	reference date for the cited resource	N	Class	CI_Date <<DataType>>
CI_ResponsibleParty					
	individualName	name of the responsible person surname, given name, title separated by a delimiter	1	CharacterString	Free text
	organisationName	name of the responsible organization	1	CharacterString	Free text
	positionName	role or position of the responsible person	1	CharacterString	Free text
	contactInfo	address of the responsible party		Class	CI_Contact <<DataType>>
	role	function performed by the	1	Class	CI_RoleCode

© University of Pretoria

		responsible party			<<CodeList>>
MD_Keywords		keywords, their type and reference source		Aggregated Class (MD_Identification)	
	keyword	commonly used word(s) or formalised word(s) or phrase(s) used to describe the subject	N	CharacterString	Free text
	Type	subject matter used to group similar keywords	1	Class	MD_KeywordTypeCode <<CodeList>>
	ThesaurusName	name of the formally registered thesaurus or a similar authoritative source of keywords	1	Class	CI_Citation <<DataType>>
MD_SpatialRepresentation		digital mechanism used to represent spatial information		Aggregated Class (MD_Metadata) <<Abstract>	
MD_Distribution					
	distributionFormat	provides a description of the format of the data to be distributed	N	Association	MD_Format
	distributor	provides information about the distributor	N	Association	MD_Distributor
	transferOptions	provides information about technical means and media by which a resource is obtained from the distributor	N	Association	MD_DigitalTransferOptions
MD_Format		description of the computer language construct that specifies the representation of data objects in a record, file, message, storage device or transmission channel			
	name	name of the data transfer format(s)	1	CharacterString	Free text
	version	version of the format (date, number, etc.)	1	CharacterString	Free text
MD_Distributor		information about the distributor		Aggregated Class (MD_Distribution and MD_Format)	
	distributorContact	party from whom the resource may be obtained. This list need not be exhaustive	1	Class	CI_ResponsibleParty <<DataType>>

© University of Pretoria

	distributorFormat	provides information about the format used by the distributor	N	Association	MD_Format
MD_DigitalTransferOptions		technical means and media by which a resource is obtained from the distributor		Aggregated Class (MD_Distribution and MD_Distributor)	
	unitsOfDistribution	tiles, layers, geographic areas, etc., in which data is available	1	CharacterString	Free text
	transferSize	estimated size of a unit in the specified transfer format, expressed in megabytes. The transfer size is > 0.0	1	Real	> 0,0
	onLine	information about online sources from which the resource can be obtained	N	Class	CI_OnlineResource <<DataType>>
	offLine	information about offline media on which the resource can be obtained	N	Class	MD_Medium <<DataType>>
CI_OnlineResource		information about on-line sources from which the dataset, specification, or community profile name and extended metadata elements can be obtained		Class <<DataType>>	
	linkage	location (address) for on-line access using a Uniform Resource Locator address or similar addressing scheme such as http://www.statkart.no/isotc211	1	Class	URL (IETF RFC1738 IETF RFC 2056)
	protocol	connection protocol to be used	1	CharacterString	Free text
	applicationProfile	name of an application profile that can be used with the online resource	1	CharacterString	Free text
	name	name of the online resource	1	CharacterString	Free text
	description	detailed text description of what the online resource is/does	1	CharacterString	Free text
	function	code for function performed by the online resource	1	Class	CI_OnLineFunctionCode <<CodeList>>
MD_ReferenceSyst		information about the reference		Aggregated Class	

© University of Pretoria

em		system		(MD_Metadata)	
	referenceSystemIdentifier	name of reference system	1	Class	RS_Identifier
RS_Identifier		identifier used for reference systems		Specified Class (MD_Identifier)	
	codeSpace	name or identifier of the person or organization responsible for namespace	1	CharacterString	Free text
	version	version identifier for the namespace	1	CharacterString	Free text
DQ_DataQuality		quality information for the data specified by a data quality scope		Aggregated Class (MD_Metadata)	
	lineage	non-quantitative quality information about the lineage of the data specified by the scope	1	Association	LI_Lineage
LI_Lineage		information about the events or source data used in constructing the data specified by the scope or lack of knowledge about lineage		Aggregated Class (DQ_DataQuality)	
	statement	general explanation of the data producer's knowledge about the lineage of a dataset	1	CharacterString	Free text
	dataSource	information about the source data used in creating the data specified by the scope	N	Association	LI_Source
LI_Source	description	detailed description of the level of the source data	1	CharacterString	Free Text

© University of Pretoria

Appendix C – Dublin Core Metadata Elements

Name	Label/URI	Definition	Comment
Title	Title http://purl.org/dc/elements/1.1/title	name given to the resource	
creator	Creator http://purl.org/dc/terms/creator	entity primarily responsible for making the resource	Examples of Creator include a person, an organization, or a service. Typically, the name of a Creator should be used to indicate the entity.
subject	Subject http://purl.org/dc/terms/subject	topic of the resource	Typically, the subject will be represented using keywords, key phrases, or classification codes. Recommended best practice is to use a controlled vocabulary. To describe the spatial or temporal topic of the resource, use the Coverage element.
description	Description http://purl.org/dc/elements/1.1/description	account of the content of the resource	Description may include, but is not limited to, an abstract, a table of contents, a graphical representation, or a free-text account of the resource.
publisher	Publisher http://purl.org/dc/elements/1.1/publisher	entity responsible for making the resource available	Examples of Publisher include a person, an organization, or a service. Typically, the name of a Publisher should be used to indicate the entity.
Contributor	Contributor http://purl.org/dc/terms/contributor	entity responsible for making contributions to the resource	Examples of a Contributor include a person, an organization, or a service. Typically, the name of a Contributor should be used to indicate the entity.
Date	Date http://purl.org/dc/elements/1.1/date	point or period of time associated with an event in the lifecycle of the resource	Date may be used to express temporal information at any level of granularity. Recommended best practice is to use an encoding schema such as the W3CDTF profile of ISO 8601 [W3CDTF].

© University of Pretoria

Type	Type http://purl.org/dc/elements/1.1/type	nature or genre of the resource	Recommended best practice is to use a controlled vocabulary such as the DCMI Type Vocabulary [DCMI-TYPE]. To describe the file format, physical medium, or dimensions of the resource, use the Format element.
Format	Format http://purl.org/dc/elements/1.1/format	file format, physical medium, or dimensions of the resource	Examples of dimensions include size and duration. Recommended best practice is to use a controlled vocabulary such as the list of Internet Media Types [MIME].
identifier	Identifier http://purl.org/dc/elements/1.1/identifier	unambiguous reference to the resource within a given context	Recommended best practice is to identify the resource by means of a string conforming to a formal identification system.
Source	Source http://purl.org/dc/elements/1.1/source	related resource from which the present resource is derived	The described resource may be derived from the related resource in whole or in part. Recommended best practice is to identify the related resource by means of a string conforming to a formal identification system.

© University of Pretoria

Appendix D – Selected Schema.org Types and Properties

Properties inherited from type CreativeWork			
Property	Expected type / Data type	Description	
About	Thing	the subject matter of the content	
accountablePerson	Person	specifies the Person that is legally accountable for the CreativeWork.	
associatedMedia	MediaObject	a media object that encodes this CreativeWork. This property is a synonym for encoding	
Author	Person or Organization	The author of this content	
Contributor	Person or Organization	a secondary contributor to the CreativeWork.	
Creator	Person or Organization	the creator/author of this CreativeWork or UserComments. This is the same as the Author property for CreativeWork	
dateCreated	Date	the date on which the CreativeWork was created	
dateModified	Date	the date on which the CreativeWork was most recently modified	
datePublished	Date	Date of first broadcast/publication	
Editor	Person	specifies the Person who edited the CreativeWork	
Encoding	MediaObject	a media object that encodes this CreativeWork. This property is a synonym for associatedMedia. Supercedes encodings	

© University of Pretoria

inLanguage	Text	the language of the content. please use one of the language codes from the IETF BCP 47 standard	
Keywords	Text	keywords or tags used to describe this content. Multiple entries in a keywords list are typically delimited by commas	
Provider	Person or Organization	the service provider, service operator, or service performer; the goods producer. Another party (a seller) may offer those services or goods on behalf of the provider. A provider may also serve as the seller. Supercedes carrier.	
Publisher	Organization	the publisher of the creative work	
sourceOrganization	Organization	the Organization on whose behalf the creator was working	
Properties borrowed from type MediaObject			
encodingFormat	Text	mp3, mpeg4, etc.	
Properties borrowed from type Organisation			
Email	Text	email address	
Telephone	Text	the telephone number	
Properties borrowed from type Person			
jobTitle	Text	the job title of the person (for example	
Email	Text	email address	
Telephone	Text	the telephone number	
Properties borrowed from type Place			
Address	PostalAddress	physical address of the item	

© University of Pretoria

Geo	GeoCoordinates	the geo coordinates of the place	
Properties borrowed from type PostalAddress			
addressCountry	Country	the country. For example, USA. You can also provide the two-letter ISO 3166-1 alpha-2 country code	
addressLocality	Text	the locality. For example, Mountain View	
addressRegion	Text	The region. For example, CA	
postOfficeBoxNumber	Text	The post office box number for PO box addresses	
postalCode	Text	The postal code. For example, 94043	
streetAddress	Text	The street address. For example, 1600 Amphitheatre Pkwy	
Properties borrowed from type GeoCoordinates			
Latitude	Text or Number	the latitude of a location. For example 37.42242	
Longitude	Text or Number	the longitude of a location. For example -122.08585	

© University of Pretoria

Appendix E – Mapping between the taxonomy of search terms and metadata standards

Taxonomy of search terms categories	**Corresponding ISO 19115 elements** (adapted from the long name)	**Corresponding Dublin Core elements**
Subject (Theme)	Title (Name by which the cited resource is known)	TITLE (A name given to the resource. Typically, a Title will be a name by which the resource is formally known)
	Topic Category (Main theme(s) of the dataset	SUBJECT (The topic of the content of the resource. Typically, a Subject will be expressed as keywords, key phrases, or classification codes that describe a topic of the resource)
	Abstract (Brief narrative summary of the content of the resource(s)	DESCRIPTION (An account of the resource. Description may include but is not limited to; an abstract, table of contents, reference to graphical representation of content or a free-text account of the content)
	Lineage (Detailed description of the level of the source data)	SOURCE (A reference to a resource from which the present

© University of Pretoria

		resource is derived. The present resource may be derived from the Source resource in whole or part.)
Location (Spatial Extent)	Geographic Description (Geographic location of the dataset (by four coordinates, or geographic identifier)	COVERAGE(The extent or scope of the content of the resource. Coverage will typically include spatial location, geographic coordinates, place name or jurisdiction such as named administrative entity)
Geographic Feature Type	Spatial Representation Type (Method used to spatially represent geographic information)	TYPE (The nature or genre of the content of the resource. Type includes terms describing general categories, functions, or aggregation levels for content)
Data Model	Resource Format (Provides a description of the format of the resource(s))	FORMAT (The physical or digital manifestation of the resource. Typically Format may include the media-type or dimensions of the resource. Format may be used to determine the software,

© University of Pretoria

		hardware or other equipment needed to display or operate the resource)
	Responsible party	CREATOR
	Reference date	DATE
	On-line resource	INDENTIFIER
	Language	LANGUAGE

© University of Pretoria

Appendix F – ISO 19115:2003 spatial metadata standard elements' description

Packages	Name of the element	Definition	Maximum Occurrence	Data Type	Domain
MD_Metadata> MD_DataIdentification.citation >CI_Citation.title	Dataset title (M)	name by which the cited resource is known	1	CharacterString	Free text
MD_Metadata> MD_DataIdentification.topicCategory	Dataset topic category (M)	main theme(s) of the dataset	N	Class	MD_TopicCategoryCode <<Enumeration >>
MD_Metadata> MD_DataIdentification.extent> EX_Extent>EX_GeographicExtent > EX_GeographicBoundingBox or EX_GeographicDescription	Geographic location of the dataset by four coordinates (C)	geographic position of the dataset	1	Specified Class (EX_GeographicExtent)	westBoundLongitude, eastBoundLongitude, southBoundLatitude, northBoundLatitude
	Geographic location of the dataset by geographic identifier (C)	description of the geographic area using identifiers	1	Class	geographicIdentifier : MD_Identifier (Code)
MD_Metadata> MD_DataIdentification.abstract	Abstract describing the dataset (M)	brief narrative summary of the content of the resource(s)	1	CharacterString	Free text
MD_Metadata>	Lineage (statement)	general explanation of	1	CharacterString	Free text

© University of Pretoria

DQ_DataQuality.lineage> LI_Lineage	(O)	the data producer's knowledge about the lineage of a dataset			
MD_Metadata > MD_DataIdentification.spatialRepresentationType	Spatial representation type (O)	method used to spatially represent geographic information	N	Class	MD_SpatialRepresentationTypeCode <<CodeList>>
MD_Metadata > MD_Distribution > MD_Format.name	Distribution format (O)	name of the data transfer format(s)	1	CharacterString	Free text

© University of Pretoria

Appendix G – HTML web page marked-up with Schema.org vocabulary

```
<!DOCTYPE HTML>
<html>
        <head>
                <meta charset="UTF-8">
                <meta name="subject" lang="English" content="Road Infrastructures">
                <meta name="description" lang="English" content="Egypt's Roads ESRI Shapefiles">
                <meta name="title" lang="English" content="Roads of Egypt">
                <title> Roads of Egypt</title>
        </head >
<body>
<div  itemscope itemtype="http://schema.org/ItemPage">
<h2 itemprop="name"> Spatial Metadata:          Roads of Egypt</h2>
 <div  itemprop="about" itemscope itemtype="http://schema.org/Dataset">
  <dl>
        <dt>Dataset identification</dt>
        <dd>Title:        <span itemprop="name"> Roads of Egypt</span>
        </dd>
        <dd>Date:        <span itemprop="dateCreated"> 2002-04-04</span>
        </dd>
        <dd>Author
                <ul>
                        <li itemprop="author" itemscope itemtype="http://schema.org/Person">Name:<span  itemprop="name">Antonio Di
Gregorio </span></li>
                        <li itemprop="author" itemscope itemtype="http://schema.org/Person"> Organisation: <span itemprop="name">FAO-
UN </span></li>
                        <li itemprop="author" itemscope itemtype="http://schema.org/Person">Position name:<span
itemprop="jobTitle">Senior Remote Sensing and Land Cover Mapping Expert</span></li>
                </ul>
```


© University of Pretoria

```
</dd>
<dd>Owner
    <ul>
        <li itemprop="creator"  itemscope itemtype="http://schema.org/Person">Organisation:<span itemprop="name">Soil and Water Research Institute - Ministry of Agriculture</span></li>
        <li itemprop="creator" itemscope itemtype="http://schema.org/Person">Contact name:<span itemprop="name">Nabil El Mowelhi</span></li>
    </ul>
</dd>
<dt>Data content of the dataset</dt>
<dd>Theme:<span itemprop="description">       Roads Infrastructures  </span></dd>
<dd>Abstract:<span itemprop="description">The full resolution dataset of roads have been produced from visual interpretation of digitally enhanced
LANDSAT TM images (Bands 4,3,2) acquired mainly in the year 1999.</span></dd>
<dd>Lineage: <span itemprop="description" > The Roads dataset is part of the Multipurpose Africover Database for the Environmental Resources (MADE)
    which was produced by the Africover Project to establish a digital georeferenced database on land cover at a 1:200,000 scale (1:100,000 for small countries and specific areas), and a geographic
    referential for the whole of Africa.</span>
</dd>
<dd itemprop="spatial"  itemscope itemtype="http://schema.org/Place">Geographic area name:<span itemprop="name">Egypt</span></dd>
<dd itemprop="contentLocation"  itemscope itemtype="http://schema.org/Place">Geographic area name:<span itemprop="name">Egypt</span></br>
<span itemprop="geo" itemscope itemtype="http://schema.org/GeoCoordinates">Geographic extent
    <ul>
        <li>Bounding Box:
            <dl>
                <dt>North bound: </dt>
                    <dd> <span itemprop="latitude">31.64690</span> </dd>
```


© University of Pretoria

```
                                    <dt>South bound: </dt>
                                        <dd><span itemprop="latitude">21.99420</span> </dd>
                                    <dt>West bound: </dt>
                                        <dd><span itemprop="longitude"> 24.70680</span></dd>
                                    <dt>East bound: </dt>
                                        <dd><span itemprop="longitude">36.89580</span>  </dd>
                                </dl>
                        </li>
                </ul>
        </span>
        </dd>
  <dd>Temporal extent:</dd>
        <dt>Spatial data structure of the dataset</dt>
        <dd itemprop="encoding"  itemscope itemtype="http://schema.org/MediaObject">Spatial data type (Format): ESRI Shapefile <meta
itemprop="encodingFormat" content="shp" /></dd>
        <dt>Spatial reference of the dataset</dt>
        <dd>Geographic coordinate system: WGS 1984</dd>
        <dt>Availability and distribution of the dataset</dt>
        <dd  itemprop="provider" itemscope itemtype="http://schema.org/Person">Organisation:<span itemprop="name"> FAO - UN
</span></dd>
        <dd>Distribution media:          Online</dd>
        <dd>Distribution format: ESRI Shapefile, Digital map (in vector or shapefile) </dd>
        <dd itemprop="provider" itemscope itemtype="http://schema.org/Person">Contact person:<span itemprop="name">    John Latham
</span></dd>
        </dl>
        <hr />
                <div>
                        <b> Note: </b> This web page is only designed for a master's research experiment undertaken at the <a
href="http://web.up.ac.za/default.asp?ipkCategoryID=16053">Centre for Geoinformation Science (CGIS) </a>  at the University of Pretoria.
                        <br>
```


© University of Pretoria

```
                                Spatial Metadata information displayed here has been adapted from the <a
href="http://www.fao.org/geonetwork/">FAO Africover Geonework web portal.</a>
                        <br>
                        <b>Contacts:</b>
                        <ul>
                                <li>Dr.Serena Coetzee:<a href="mailto:serena.coetzee@up.ac.za">(serena.coetzee@up.ac.za)</a></li>
                                <li> Samy Katumba: <a href="mailto:skabangu@hotmail.com">(skabangu@hotmail.com)</a></li>
                        </ul>
                </div>
        </div>
</div>
</body>
</html>
```


© University of Pretoria

Appendix H – HTML web pages retrieved by Bing

Web pages marked up with Dublin Core Submitted to Bing	Avg. Position		Web pages marked up with Schema.org submitted to Bing	Avg. Position
http://lebenya.co.za/Administrative/drcongoboundaries.html	1		http://lebenya.co.za/Administrative/mc_eritreaboundaries.html	4.5
http://lebenya.co.za/Administrative/kenyaboundaries.html	9.9		http://lebenya.co.za/Administrative/mc_kenyaboundaries.html	6
http://lebenya.co.za/Administrative/namibiaboundaries.html	16		http://lebenya.co.za/Administrative/mc_namibiaboundaries.html	10
http://lebenya.co.za/Administrative/ugandaboundaries.html	25.3		http://lebenya.co.za/Administrative/mc_ugandaboundaries.html	4
http://lebenya.co.za/Administrative/somaliaboundaries.html	11.9		http://lebenya.co.za/Administrative/mc_somaliaboundaries.html	30
http://lebenya.co.za/Administrative/egyptboundaries.html	5		http://lebenya.co.za/Administrative/mc_rwandaboundaries.html	17.5
http://lebenya.co.za/Infrastructure/burunditowns.html	13.2		http://lebenya.co.za/Infrastructure/mc_drcongotowns.html	18.1
http://lebenya.co.za/Infrastructure/eritrearoads.html	7		http://lebenya.co.za/Infrastructure/mc_ghanaroads.html	18.7
http://lebenya.co.za/Infrastructure/roadsafrica.html	8		http://lebenya.co.za/Infrastructure/mc_rwandaroads.html	4

© University of Pretoria

http://lebenya.co.za/Infrastructure/somaliatowns.html	4		http://lebenya.co.za/Infrastructure/mc_ugandatowns.html	10.6
http://lebenya.co.za/LandCover/geologicalmapsudan.html	17		http://lebenya.co.za/Infrastructure/mc_citiesafrica.html	15
http://lebenya.co.za/LandCover/geologyuganda.html	13.5		http://lebenya.co.za/Infrastructure/mc_eritrearoads.html	7
http://lebenya.co.za/LandCover/groundwaternamibia.html	19		http://lebenya.co.za/Infrastructure/mc_eritreatowns.html	5.5
http://lebenya.co.za/LandCover/hydrologymapsudan.html	8.3		http://lebenya.co.za/Infrastructure/mc_kenyatowns.html	20
http://lebenya.co.za/WaterResources/drcongo.html	22.1		http://lebenya.co.za/Infrastructure/mc_worldcities.html	18.5
			http://lebenya.co.za/Infrastructure/mc_worldrailroads.html	11.5
			http://lebenya.co.za/Infrastructure/mc_cityandpopulationafrica.html	10
			http://lebenya.co.za/Infrastructure/mc_tanzaniaroads.html	7.8
			http://lebenya.co.za/Infrastructure/mc_worldroads.html	18
			http://lebenya.co.za/LandCover/mc_agriculturalproductnigeria.html	8.3
			http://lebenya.co.za/LandCover/mc_geologicalmapsudan.html	11

© University of Pretoria

		http://lebenya.co.za/LandCover/mc_geologyuganda.html	19.8
		http://lebenya.co.za/LandCover/mc_groundwaternamibia.html	2
		http://lebenya.co.za/LandCover/mc_hydrologymapsudan.html	9.2
		http://lebenya.co.za/LandCover/mc_landsuitforcropproduction.html	162
		http://lebenya.co.za/LandCover/mc_landusemauritius.html	3
		http://lebenya.co.za/LandCover/mc_landusetanzania.html	13.5
		http://lebenya.co.za/LandCover/mc_thematicagriculturerwanda.html	6
		http://lebenya.co.za/WaterResources/mc_africa.html	12
		http://lebenya.co.za/WaterResources/mc_africa_hydrological.html	17
		http://lebenya.co.za/WaterResources/mc_africa_inland.html	9.8
		http://lebenya.co.za/WaterResources/mc_africa_irrigation.html	16
		http://lebenya.co.za/WaterResources/mc_drcongo.html	9.6
		http://lebenya.co.za/WaterResources/mc_eritrea.html	18

© University of Pretoria

		http://lebenya.co.za/WaterResources/mc_irrigation.html	6
		http://lebenya.co.za/WaterResources/mc_rwanda.html	7
		http://lebenya.co.za/WaterResources/mc_southsoudan.html	12.7
		http://lebenya.co.za/WaterResources/mc_sudan.html	9
		http://lebenya.co.za/WaterResources/mc_tanzania.html	26

© University of Pretoria

Appendix I - HTML web pages retrieved by Google

Web pages marked up Dublin Core Submitted to Google	Avg. position		**Web pages marked up with Schema.org Submitted to Google**	Avg. position
http://lebenya.co.za/Administrative/drcongoboundaries.html	1		http://lebenya.co.za/Administrative/mc_drcongoboundaries.html	1
http://lebenya.co.za/Administrative/eritreaboundaries.html	1		http://lebenya.co.za/Administrative/mc_eritreaboundaries.html	1
http://lebenya.co.za/Administrative/kenyaboundaries.html	1		http://lebenya.co.za/Administrative/mc_kenyaboundaries.html	1
http://lebenya.co.za/Administrative/namibiaboundaries.html	3		http://lebenya.co.za/Administrative/mc_ugandaboundaries.html	1
http://lebenya.co.za/Administrative/ugandaboundaries.html	1		http://lebenya.co.za/Administrative/mc_somaliaboundaries.html	1
http://lebenya.co.za/Administrative/somaliaboundaries.html	1		http://lebenya.co.za/Administrative/mc_rwandaboundaries.html	1
http://lebenya.co.za/Administrative/egyptboundaries.html	3.3		http://lebenya.co.za/Infrastructure/mc_burunditowns.html	1
http://lebenya.co.za/Administrative/rwandaboundaries.html	2.5		http://lebenya.co.za/Infrastructure/mc_ghanaroads.html	1
http://lebenya.co.za/Infrastructure/drcongotowns.html	1		http://lebenya.co.za/Infrastructure/mc_rwandaroads.html	1
http://lebenya.co.za/Infrastructure/burunditowns.h	1		http://lebenya.co.za/Infrastructure/mc_ugandatowns.	1

© University of Pretoria

tml			html	
http://lebenya.co.za/Infrastructure/ghanaroads.html	1		http://lebenya.co.za/Infrastructure/mc_citiesafrica.html	1
http://lebenya.co.za/Infrastructure/rwandaroads.html	1		http://lebenya.co.za/Infrastructure/mc_drcongoroads.html	1
http://lebenya.co.za/Infrastructure/ugandatowns.html	1		http://lebenya.co.za/Infrastructure/mc_egyptroads.html	1
http://lebenya.co.za/Infrastructure/burundiroads.html	1		http://lebenya.co.za/Infrastructure/mc_egypttowns.html	1
http://lebenya.co.za/Infrastructure/citiesafrica.html	1		http://lebenya.co.za/Infrastructure/mc_eritreatowns.html	1
http://lebenya.co.za/Infrastructure/egyptroads.html	1		http://lebenya.co.za/Infrastructure/mc_kenyatowns.html	1
http://lebenya.co.za/Infrastructure/eritrearoads.html	1		http://lebenya.co.za/Infrastructure/mc_roadsafrica.html	1
http://lebenya.co.za/Infrastructure/eritreatowns.html	1		http://lebenya.co.za/Infrastructure/mc_tanzaniatowns.html	1
http://lebenya.co.za/Infrastructure/kenyaroads.html	1		http://lebenya.co.za/Infrastructure/mc_tanzaniaroads.html	1
http://lebenya.co.za/Infrastructure/kenyatowns.html	1		http://lebenya.co.za/Infrastructure/mc_worldroads.html	23
http://lebenya.co.za/Infrastructure/somaliatowns.h	1		http://lebenya.co.za/LandCover/mc_agriculturalprod	1

© University of Pretoria

tml			uctnigeria.html	
http://lebenya.co.za/Infrastructure/tanzaniatowns.html	1		http://lebenya.co.za/LandCover/mc_agroecologicalzonestanzania.html	1
http://lebenya.co.za/Infrastructure/worldcities.html	1		http://lebenya.co.za/LandCover/mc_geologicalmapsudan.html	1
http://lebenya.co.za/Infrastructure/worldrailroads.html	1		http://lebenya.co.za/LandCover/mc_geologyuganda.html	24
http://lebenya.co.za/Infrastructure/tanzaniaroads.html	1		http://lebenya.co.za/LandCover/mc_groundwaterbotshwana.html	1
http://lebenya.co.za/Infrastructure/worldroads.html	1		http://lebenya.co.za/LandCover/mc_groundwaternamibia.html	1
http://lebenya.co.za/LandCover/agriculturalproductnigeria.html	1		http://lebenya.co.za/LandCover/mc_hydrologymapsudan.html	1
http://lebenya.co.za/LandCover/agroecologicalzonestanzania.html	1		http://lebenya.co.za/LandCover/mc_landsuitforcroppproduction.html	1
http://lebenya.co.za/LandCover/geologicalmapsudan.html	1		http://lebenya.co.za/LandCover/mc_landusemauritius.html	1
http://lebenya.co.za/LandCover/geologyuganda.html	24		http://lebenya.co.za/LandCover/mc_landusetanzania.html	1
http://lebenya.co.za/LandCover/groundwaterbotshwana.html	1		http://lebenya.co.za/LandCover/mc_thematicagriculturerwanda.html	1
http://lebenya.co.za/LandCover/hydrologymapsud	1		http://lebenya.co.za/LandCover/mc_thematicagricult	21

© University of Pretoria

an.html			ureuganda.html	
http://lebenya.co.za/LandCover/landsuitforcropproduction.html	1		http://lebenya.co.za/LandCover/mc_vegetationsudan.html	1
http://lebenya.co.za/LandCover/landusemauritius.html	1		http://lebenya.co.za/Topography/mc_slope.html	31
http://lebenya.co.za/LandCover/landusetanzania.html	1		http://lebenya.co.za/WaterResources/mc_africa.html	1
http://lebenya.co.za/LandCover/thematicagriculturerwanda.html	1		http://lebenya.co.za/WaterResources/mc_africa_inland.html	1
http://lebenya.co.za/LandCover/thematicagriculturerwanda.html	1		http://lebenya.co.za/WaterResources/mc_africa_irrigation.html	1
http://lebenya.co.za/LandCover/vegetationsudan.html	1		http://lebenya.co.za/WaterResources/mc_burundi.html	1
http://lebenya.co.za/WaterResources/africa.html	110		http://lebenya.co.za/WaterResources/mc_drcongo.html	1
http://lebenya.co.za/WaterResources/africa_hydrological.html	1		http://lebenya.co.za/WaterResources/mc_egypt.html	1
http://lebenya.co.za/WaterResources/africa_inland.html	1		http://lebenya.co.za/WaterResources/mc_eritrea.html	33
http://lebenya.co.za/WaterResources/africa_irrigation.html	1		http://lebenya.co.za/WaterResources/mc_irrigation.html	1
http://lebenya.co.za/WaterResources/burundi.html	1		http://lebenya.co.za/WaterResources/mc_rwanda.htm	1

© University of Pretoria

			l	
http://lebenya.co.za/WaterResources/eritrea.html	1		http://lebenya.co.za/WaterResources/mc_southsouda n.html	1
http://lebenya.co.za/WaterResources/irrigation.ht ml	250		http://lebenya.co.za/WaterResources/mc_tanzania.ht ml	1
http://lebenya.co.za/WaterResources/kenya.html	1			
http://lebenya.co.za/WaterResources/rwanda.html	1			
http://lebenya.co.za/WaterResources/sudan.html	1			
http://lebenya.co.za/WaterResources/tanzania.htm l	27			

© University of Pretoria

Appendix J – Queries obtained from participants in the user experiment

Search Terms (from User experiment)
shapefile of road network in ghana
spatial data set of road network in ghana
land-use data for Africa
road maps for pretoria
global+bathymetry
global+wave+energy+estemates
global+wave+energy+spatial+data
global+wave+energy+spatial+data+download
satellite+wave+height+data
wave+energy+database
wave+energy+data
wave+energy+database
Wave+gis+data
africa+wave+energy+data
Africa%20Cities%20spatial%20data
African+boundaries+shapefile
African+Cities+shapefiles
African+Cities+spatial+data
datasets on railways in south africa

© University of Pretoria

digital data on Railways in South Africa
digital data on railways in south africa
map datasets on railways in south africa
railways in south africa+shapefiles
road maps for pretoria
shapefile of road network in ghana
spatial data set of road network in ghana
satelite+wave
Spatial data search
spatial data
spatial+data+downloads
spatial%20data%20airbourne
spatial%20data%20airborne
download+spatial+data+airborne
download+spatial+data+for+sa
spatial%20data
spatial+data+airborne&sa
South African spatial data
digital data systems
South African ecosystems and watesites shapefiles
waste sites shapefiles
waste sites shapefiles south africa

© University of Pretoria

wetlands shapefiles
free gis population data
free gis population data, Africa
population change gis data, Africa
population change, Africa
land-use data for africa
land+use+data+sets+limpopo
land%20use%20types%20limpopo
shape+files+landuse
shape+files+landuse+South+A&gs_l
vector+layers+Landuse
Ghana land cover shape file
South African vegetation cover spatial data
south+africa+vegetation
south+africa+vegetation+cover+spatial+data
south%20african%20vegetation%20cover%20spatial%20data
south+africa+vegetation
vegitation%20cover%20in%20sa
vector data set land cover type of Mafikeng
vector format for South African power plans
spatial+data+infrastructure
spatial+data+mining

© University of Pretoria

electricity distribution in rwanda
spatial data of electricity distribution in rwanda
esri+shapefiles+schools
electricity distribution Senegal shapefile
Cairo roads spatial data free
Rwanda spatial data sets
rwanda spatial data files
Download free shapefile maps
download shapefiles
south+africa+data+set
demarcation+board+south+africa
shapefiles
shapefiles somalia
spatial data africa shapefile
spatial data mining
free+digital+maps+south+africa
Botswana Map
botswana raster files
botswana raster layer
botswana shp
south+africa+shapefiles+download+(unemployment)
unempolyment%20in%20pretoria

© University of Pretoria

south+african+shapefiles
unemployment%20places%20in%20south%20africa(map%20presentation)%20shapefiles
unemployment+places+in+pretoria
download east african rift system shape file
east african rift system formation
east african rift system formation in kenya
east african rift system modis
east african rift system shape file
East Africa rift system kenya shape file
botswana dem
botswana dem download
botswana dem free
botswana dem layer
botswana digital elevation model
botswana elevation dem
botswana elevation dem download
Botswana Elevation extremes
botswana elevation raster
mount kilimanjaro elevation dataset download
mount k shapefile download
mount kenya shapefile download
mount kilimanjaro contours

© University of Pretoria

mount kilimanjaro contours shapefile
mount kilimanjaro contours shapefile download
mount kilimanjaro dem shapefile download
mount kilimanjaro elevation data set download
mount kilimanjaro elevation shapefile download
tanzania elevation .shp
tanzania elevation data
mount kilimanjaro shapefile download
mount kilimanjaro spatial dataset
anzania elevation shapefile
south+africa+river+spatial+data
south+africa+rivers
nile river model
Spatial data for nile river
botswana streams and rivers
botswana streams and rivers shp
botswana water bodies map download
mount kilimanjaro rainfall dataset
mount kilimanjaro rainfall shapefile
england+ROAD+SHAPE+FILES
south african roads and shapefiles
Burudi roads shapefiles

© University of Pretoria

overty in ethiopia
poverty in ethiopia facts
poverty in ethiopia maps
poverty in ethiopia maps modus
spatial data mining
spatial data mining shapefile
spatial data for nigeria
spatial data for population in nigeria
spatial data for population in south africa
spatial data for population in zimbabwe
spatial data on the population in nigeria
spatial data on the population in nigeria (shape files)
spatial datafor population in egypt
spatial data for population in nigeria
free spatial data, africa, population
free spatial data, africa, population, qgis
free spatial data, africa, population,raster data
free spatial data, africa, population,raster data downloads
free spatial data, africa, poulation
spatial data, human population density, africa
spatial data, human population density, africa. Shapefile
spatial data, shapefiles, human population density, africa

© University of Pretoria

spatial data of uganda on population
spatial data for population density in a
spatial data for population density in africa
spatial data for population in egypt
spatial data for population in egypt for 2000
spatial data for population in nami
spatial data for population in namibia
spatial data for population in south
spatial data for population in
spatial data for population in angola
spatial data for population in angola(shapefiles)
spatial data for population in egypt
dowload+raster+data+population
2Fpopulation%2Fdata
dowload+raster+data+population
download+spatial+data+on+population
spatial+population+data
spatial+population+data+download
spatial data on the water resources in ethopia (shapefiles)
spatial data on the water resources in nigeria (shapefiles)
spatial databases of the water resources in Ethiopia (shapefiles)
spatial datasets of the water resources in Ethiopia (shapefiles)

© University of Pretoria

spatial data for rivers in
spatial data for rivers in egypt
spatial data for rivers in egypt(shapefiles)
spatial data for rivers in kenya
spatial data for rivers in malawi
spatial data for rivers in taga
spatial data for sql server 2008
spatial data of botwsana in africa
Uganda GIS Maps,
spatial data in SA
spatial data
spatial data fil
spatial data file
spatial data files in angola
spatial data files in zambia
spatial data for in namibia
spatial data for statistics in nigeria
spatial data for statistics in south africa
spatial data formats
spatial data research
dowload+raster+data
2Fdata%2Fsets

© University of Pretoria

2Fmapsdata%2Fdownload
dowload+raster+data
download+spatial+data
spatial+data+vectors
download+vectors
African%20spatial%20data%20downloads
Esri+spatial+data+downloads
Spatial data sources for TanzaniaTanzania
gis mapping
shapefiles
shapefiles south africa
shapefiles download free
shapefiles download
shapefiles of south africa
shapefiles of india
shapefiles of spatial data
spatial data to address a need for south africa shp
download+gis+data+south+africa
spatial%20data%20downloads%20(south%20africa)%
spatial+data+sa
spatial+data+sa+for+universities
spatial+data+sa+for+university

© University of Pretoria

shapefile angola
shapefile burundi
mountains of asia
angola mata data
diva gis shapefiles
geodata design
spatial data infrastructure south africa
spatial dataset
spatial+data+infrastructure+of+india
spatial+data+shapefiles+download
spatial data of buildings in pretoria, rastar
spatial data for building in namibia
spatial data for building in south africa
city lodge
city of tshwane
spatial data on the city of tshwane
tshwane municipality
farms mpumalanga spatial data
shapefiles of south africa schools
shapefiles download: provincial boundaries rsa
shapefiles of government schools in johannesburg
shapefiles of pretoria schools

© University of Pretoria

shapefiles of pretoria schools spatial data
shapefiles of south africa railroads
shapefiles of south africa roads
shapefiles of south africa schools
spatialdata%20for%20parks%20in%20gauteng
spatial+data+downloads+(sa)+for+tshwane+schools+shapefiles
spatial+data+downloads+(south+africa)+for+tshwane+schools+shapefiles
spatial+data+for+national+parks+ingauteng
spatial+data+for+national+parks+in+sa
spatial+data+for+parks+in+gauteng+(shape+files)
spatial+data+for+resorts+in
spatial+data+for+resourts
spatial+data+for+resourts+in+sa
spatial+data+sa+for+drainage+systems
spatial+data+sa+for+municipalities
spatial+data+sa+for+municipalities+shape+files
spatial+data+sa+for+municipalities+shape+files+downloads
spatial+data+sa+for+railways+shape+files
spatial+data+sa+for+thswane+shapefiles+downloads
+data+sa+for+thswane+shapefiles+downloads
spatial+datadownloads+sa+for+tshwane+schools+shapefiles
spatial%20data%20of%20road%20networks%20around%20pretoria

© University of Pretoria

spatial data of the biosphere in SA
high hazardz zones in africa
spatial data on high hazardz zones in africa
spatial data for mountains in
spatial data for mountains in namibia
shapefiles of mountains in tanzania
spatial data to address mountains in asi
spatial data to address mountains in asia
mountains in tanzania
spatial data to address mountains in thailand
free%20download%20of%20shape%20files%20spatial%20data%20for%20mountain%20ranges%20in%20south%20africa
spatial data for rivers in egypt (share files)
spatial data for water in namibia
spatial data rivers
spatial data rivers in egypt
shapefiles on water sources in tanzania
shp files on water sources in tanzania
spatial%20data%20sa%20for%20river%20networks
city+of+tshwane+spatial+data+for+rivers+in+shape
city+of+tshwane+spatial+data+for+rivers+in+shape+files
spatial%20data%20of%20rivers%20in%20south%20africa
spatial data roads in angola

© University of Pretoria

spatial data roads in angola (share files)
road networks in pretoria 2012
Road networks in tshwane municipality
shapefile of roads in south africa
spatial data on the road networks in pretoria
roadnetworks on city of tshwane
roadnetworks on city of tshwane shape files
shapefiles of south africa roads
shp files on weather stations in tanzania
+spatial+data+of+south+africa
geological+spatial+data+of+south+africa+shapefiles+download

© University of Pretoria

Appendix K – Queries obtained from Bing webmaster tools experiment (as collected from the log of Bing search engine)

Search Keywords Number of terms per query
wgs nairobi kenya shapefile 4 1 0 0 6 0 = = = = =
uganda administrative boundaries shapefile 4 1 0 0 2 0 = = = = =
geological map of the sudan 5 1 1 100 8 8 = = = = =
www.lefenya.co.za 1 4 0 0 1.5 0 = = = = =
eugene rurangwa fao 3 2 0 0 6.5 0 = = = = =
political map of namibia site:.za 5 1 0 0 26 0 = = = = =
gemstone geology map in sudan 5 1 0 0 52 0 = = = = =
kenya admin boundaries 3 1 0 0 2 0 = = = = =
rwanda map of roads 4 1 0 0 4 0 = = = = =
sudan geological map 3 1 0 0 15 0 = = = = =
utilization of land title in tanzania -pdf 7 1 0 0 8 0 = = = = =
tanzania infrastructures shapefiles 3 1 0 0 8 0 = = = = =
namibia administrative data shapefile 4 1 0 0 4 0 = = = = =
++serena.coetzee@up.ac.za 1 8 0 0 13 0 = = = = =
+serena.coetzee@up.ac.za 1 4 0 0 13 0 = = = = =
#NAME? 4 0 0 40.5 0 = = = = =
shapefile africa countries wgs 1984 5 3 0 0 5.3 0 = = = = =
hydrology maps sudan 3 8 0 0 7.4 0 - = = - =

© University of Pretoria

spatial planning tanzania 3 2 0 0 19 0 = = = = =
biggest inland water bodies in africa 6 1 0 0 3 0 = = = = =
ghana map site:.za 3 1 0 0 19 0 = = = = =
hydrological map of sudan wikipedia 5 1 0 0 12 0 = = = = =
vector street maps site:.za 4 1 0 0 40 0 = = = = =
lebenya 1 8 0 0 4.2 0 + = = - =
africas area and population 4 1 0 0 9 0 = = = = =
administrative bounderies in kenya 4 1 0 0 9 0 = = = = =
congo dr site:.za 3 1 0 0 48 0 = = = = =
field crop production in botswana 5 1 0 0 162 0 = = = = =
shapefile rivers africa 3 1 0 0 9 0 = = = = =
uganda towns shapefile 3 1 1 100 4 4 = = = = =
hydrological maps site:.za 3 7 0 0 13.6 0 - = = - =
"http://theacrc.info" site:za 1 3 0 0 18.6 0 = = = = =
geology south west uganda 4 3 0 0 12.6 0 = = = = =
africa water bodies shapefile 4 2 0 0 3 0 = = = = =
congo dr map site:.za 5 2 0 0 8 0 = = = = =
administrative boundaries kenya 3 1 0 0 7 0 = = = = =
rwanda administrative map 3 1 0 0 27 0 = = = = =
property boundary maps online site:.za 5 2 0 0 37 0 = = = - =
ghana grographical extent 3 1 0 0 8 0 = = = = =
namibia adminstrative boundaries 3 1 0 0 2 0 = = = = =

© University of Pretoria

riversofafrica 1 1 0 0 7 0 = = = = =
rivers of tanzania 3 1 0 0 27 0 = = = = =
dr congo map site:.za 4 2 0 0 4.5 0 = = = = =
www. sudan mc.com 3 2 0 0 8.5 0 = = = = =
kenya towns 2 1 0 0 20 0 = = = = =
cropping patterns in east africa 5 1 0 0 5 0 = = = = =
yafesi wp- /wp- 3 4 0 0 5.6 0 + = = - =
rail roads of the world 5 1 0 0 2 0 = = = = =
geology of uganda 3 2 0 0 23.5 0 = = = = =
r fao vail blue ley 4 1 0 0 2 0 = = = = =
water bodies shapefile 3 1 0 0 27 0 = = = = =
un dataset shapefile cities 4 2 0 0 15.5 0 = = = = =
bountries oferitrea 2 1 0 0 1 0 = = = = =
boundaries of rwanda 3 1 0 0 18 0 = = = = =
somali map site:.za 3 1 0 0 30 0 = = = = =
mil-v-89039 1 1 0 0 23 0 = = = = =
uganda administrative boundaries 3 1 0 0 5 0 = = = = =
rwanda bounding box 3 4 0 0 3 0 = = = = =
bounding box africa 3 2 0 0 11.5 0 = = = = =
shapefile of africa rivers 4 2 0 0 8.5 0 = = = = =
water distribution -china "hotmail.com" 4 2 0 0 3.5 0 = = = = =
drc map site:.za 3 1 0 0 27 0 = = = = =

© University of Pretoria

megadatauganda 1 1 0 0 27 0 = = = = =
list of rivers in dr congo 6 1 0 0 9 0 = = = = =
map of nigeria and agricultural produce 6 1 0 0 19 0 = = = = =
what east african nation has the most inland bodies of water? 2 0 0 10.5 0 = = = = =
hydrological map site:.za 3 0 0 6.3 0 = = = = =
map of africa eritrea site:.za 3 0 0 11 0 = = = = =
nfpi for africover 2 0 0 7.5 0 = = = = =
"south sudan" shapefile 1 0 0 29 0 = = = = =
aaglmru 1 0 0 3 0 = = = = =
african figures site:.za 1 0 0 20 0 = = = = =
geological survey maps online site:.za 2 0 0 27.5 0 = = = = =
tekleyohannes wp- /wp- 2 0 0 5 0 = = = = =
what is cropping patterns for africa 2 0 0 7 0 = = = = =
agricultural product in nigeria site:.za 1 0 0 1 0 = = = = =
africa inland lakes shapefiles 1 0 0 3 0 = = = = =
rivers of rwanda 1 0 0 7 0 = = = = =
geological maps online site:.za 1 0 0 6 0 = = = = =
hydrology south africa site:.za 1 0 0 25 0 = = = = =
hydrological data site:.za 2 0 0 25 0 = = = - =
namibia shapefile mme 1 0 0 10 0 = = = = =
namibia shapefile 1 0 0 6 0 = = = = =
shapefile african rivers 1 0 0 23 0 = = = = =

© University of Pretoria

statial data site:.za 1 0 0 36 0 = = = = =
tanzania leben 1 0 0 4 0 = = = = =
map of nigeria showing distributions of common cultivated crops 1 0 0 5 0 = = = = =
tanzania road map site:.za 1 0 0 9 0 = = = = =
rwanda administrative boundaries 1 0 0 4 0 = = = = =
yafesi okia linkedin 1 0 0 10 0 = = = = =
+land land resources use site:.za 1 0 0 50 0 = = = = =

© University of Pretoria