Search Engine Information Retrieval: Empirical Research on the usage of Meta Tags to enhance Web Site Visibility and Ranking of e-Commerce Web Sites

Melius WEIDEMAN
Faculty of Business Informatics, Cape Technikon
Cape Town, South Africa. meliusw@yahoo.com

and

Wouter KRITZINGER
Faculty of Business Informatics, Cape Technikon
Cape Town, South Africa. kritzingerw@ctech.ac.za

ABSTRACT

The principal objective of this research project was to determine the extent to which e-Commerce websites make use of metadata to enhance website visibility to search engines. The methods employed were to firstly identify a number of e-Commerce websites, and secondly to inspect, record, and analyze the relevant meta tags used in its coding.

A subset of 200 e-Commerce websites was compiled by extensive Internet searching using standard search engines and portals. Each site was evaluated to confirm its qualification as an e-Commerce site, and its three visibility-related HTML meta tags were inspected. The results prove that a reasonable percentage of e-Commerce web pages make use of the three meta tags in question, but none make use of Dublin Core. An average score of 66.83% was earned overall.

The primary conclusion reached is that initially meta tag usage appears to be reasonably high, considering figures of 60% and above. However, if the percentages of the non-users are extrapolated across the size of the WWW, while assuming that not all web pages are e-Commerce based, a large number of web pages are not availing themselves of one of the most basic visibility features.

Keywords: search engine, information retrieval, meta tags, website visibility, e-commerce, web sites.

1. INTRODUCTION

The amount of data available on the Internet cannot be measured. New and existing authors constantly add more by uploading new and revised web pages to web servers, some on an hourly basis. There is no central body responsible for categorizing, validating or censoring data on the Internet. These factors contribute to the rather chaotic situation Internet users are facing when attempting to retrieve relevant information from the Internet. Although many programs exist to enable the average user to sail the data seas (friendly browsers, free search engines), general consensus exists that navigating the Internet is not a straightforward task [45].

There appears to be a link between an Internet user, search engines, metadata and websites, as these elements are defined in section 2. The purpose of this research project was to inspect and report on the usage of metadata on e-Commerce web sites towards search engine friendliness.

Countless reports in research literature seem to suggest that Internet searching is a difficult, error prone task. For example, Sherman echoed Stoll's sentiments by stating that the Internet had become the world's largest and most complex, chaotic and unstructured search space [34]. Another author has found that retrieval and organization of materials on the web are not standardized, while others claimed that the WWW was not designed to support organized publishing of data, or the retrieval thereof [6], [33]. Information overload has arrived [52].

2. DEFINITIONS AND OTHER RESEARCH

2.1 The User

A skill which eludes many average Internet users, is the finding of relevant information on the Internet in a short time [3]. However, many Internet users rely on search engines to find relevant data for a variety of purposes on a daily basis [46].

However, some studies have shown that early information users were not all eager to become involved in the extraction of knowledge from an electronic source. Over three decades ago Jackson found that engineers are reluctant to use information sources. The United Engineering Information Service failed to elicit financial support from the engineering profession to establish its services [20].

To further complicate matters, Large, Tedd and Hartley warn that information seekers must not be treated as a homogenous group – they differ in many aspects, where their information retrieval experience level has a large differentiating effect [24]. Other authors warn that there is a difference in the way that

WWW searchers and traditional OPAC searchers work when looking for information [21].

For the purpose of this project, an Internet user is defined as a person who uses search engines to find relevant information on the Internet.

2.2 The Search Engine

Search engines provide the average Internet user with a (mostly) free, apparently easy way to find general information on the Internet. They are programs which offer interaction with the Internet through a front end, where the user can type in a search term, or make successive selections of relevant directories. The search engine software then compares the search term against an index file, which contains information about many websites. Matches found are returned to the user via the front end. The index is updated regularly either by human editors or by automated programs (called spiders, robots or crawlers). Both humans and spiders simply collect information of new websites by visiting as many websites as possible, and then building them into the index [49].

The three components of a typical search engine (front end, index file and information collectors) have close parallels in the components of a typical information retrieval system, as defined by Lancaster many years before the Internet and search engines became freely accessible. A search engine front-end maps to the ”user system interface”, the index file maps to the “indexing subsystem” and the information collectors map to the “document selection subsystem” [23].

However, one author claims that search engines are complex, trusted without being understood and that users simply deal with their answers without understanding why they receive those answers [27].

Many authors from both the popular press and other sources have done evaluations, comparisons, measurements and a variety of other tests on a large number of search engines [2], [3], [4], [7], [8], [9], [11], [12], [13], [15], [16], [17], [18], [26], [28], [29], [30], [31], [32], [36], [37], [42], [44], [50]. In these studies, the following search engines were mentioned:

AltaVista, Excite, Galaxy, Harvest, Hotbot, Infoseek, Looksmart, Lycos, Magellan, MetaCrawler, Northern Light, Open Text, PlanetSearch, Savvysearch, Search.com, UKPlus, Webcrawler, WWW Worm and Yahoo!. A list of commonly used search engines and meta search engines with their URLs and a short evaluation, is available on the WWW [48].

Another study confirmed that users make little use of advanced search features. For example, a test on 1 025 910 queries submitted to the Excite search engine revealed that the following features were only used by small percentages of users [39]:

-Inclusion (+)	2%
-Exclusion (–)	0.001%
-Boolean operators (NOT / AND NOT)	0.0003%

Yet another study indicated that searchers spent a relatively short amount of time searching for one topic: the average search session seemed to last between five and 10 minutes only [10]. This could possibly be a reason for the lack of use of advanced operators.

From the work already done on search engines, it has become clear that these programs play an important role in the lives of Internet users. For the purposes of this study, **a search engine is defined as a program that is accessible by any average user, capable of accepting user input which defines the information it produces as output to this user.**

2.3 The Web page

A large number of authors claim that the relationship between users and web pages with relevant information, stored on the Internet, is not always a positive one:

- “Currently, search is simply bad” [35].
- “…some respondents seemed confused about what they were to report when asked to list query terms for their search” [38].
- “I find it difficult to search information on the Internet…” [41].
- “…information seeking is a complex and difficult process for these students, who seek to reduce the task to finding an obvious answer or finding a good website…” [42].
- “Only 33% of the Internet users agree or strongly agree with the statement ‘It is easy to perform subject searches on the Internet.’” [45].
- “… both novice and experienced searchers were overconfident in their performance” [51].

Furthermore, Jones criticizes web robots as using crude methods of gathering data, not distinguishing between opinion and fact [22].

In summary it is clear that web pages also play an important part in the life of the Internet user. For the purposes of this study, **a web page is defined as an entity, stored on the Internet, accessible by any average user, which contains some information.**

2.4 Metadata

During the Middle Ages, indexing and simple classification of manuscripts were done. The indexers involved in this task were surrounded by an aura of mysticism, resulting from the coding schemes and alphabetical keys [19].

The high-powered computer era we currently find ourselves in provided the much-needed technology to empower new developments, where document matching is made through inverted indices, string and positional searches. There was now no technical constraint to prevent an index from including every single term of a given textual document in the index. This has been done some time ago, in the form of typical Bible concordances such as *Strong’s Exhaustive Concordance*, first published in 1890.

During the late 1950s and 1960s, landmark work was done by noted authors in this area of document content presentation. The controversial Uniterm system sparked interest in the UK and the USA, which led to the Cranfield tests by Cleverdon and Keen [5], [38], [43].

Although certain inherent limitations of the study were evident, it did prove the effectiveness of the Uniterm system above the UDC classification [14]. Modern day versions of metadata include HTML meta tags and the DCMI (Dublin Core Metadata

Initiative). More recent work indicates that the use of these metadata elements plays an important part in making web pages more visible. A large number of authors make a case to urge website authors to make use of HTML and the DCMI meta tags respectively. The HTML TITLE, KEYWORD and DESCRIPTION meta tags have been singled out as being most relevant to website visibility [1], [47].

Again it is clear that metadata plays an important part in making web pages more visible to the search engine. For the purposes of this study, **metadata is defined as data about a web page, contained in the web page, used to describe its contents and other features.**

3. RELATIONSHIPS

The four entities identified in section 2 are related to each other in a true quadrilateral fashion. The author intuitively defined a set of relationships between these four. The nature of each relationship is given by some action term(s) in Figure 1, and these terms should be used to complete an implied English sentence.

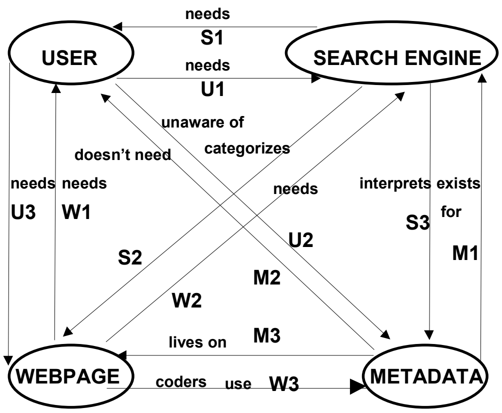


Figure 1

No claim is made in this project as to the accuracy of these definitions. However, the author had reason to believe that at least two of them could be invalid as initially defined: S3 and W3. The empirical nature of this project manifested in the inspection and reporting on the accuracy of W3's definition, with specific reference to e-Commerce websites. The "action sentence" for W3 is defined as: "Web page coders use metadata on web pages to enhance their visibility to search engines". The remaining eleven relationships could be inspected and criticized in detail in future research.

4. METHODOLOGY

4.1 URL set

According to a landmark study done on Internet metrics, a total of 256^4, or about 4.3 billion Internet servers could exist. Furthermore, each server could store one or more web pages, typically an average of 289 [25]. As a result of these large values, all web pages could not be inspected. A subset had to be identified, from which a sample would be drawn. This concept is based on the extraction of documents used in the Cranfield tests. The subset was defined as all those who appear to be involved in the selling of a product or service via online methods.

A series of Internet searches was started to find a sample of 200 e-Commerce based websites. Standard search engines and portals such as Google, Yahoo and Hotbot were used. In each case the web site was traversed in width and depth to ensure that it met the set criteria of an e-Commerce web site: products or services were being sold online.

4.2 Scoring scheme

A decision was taken to inspect the occurrence and usage of those meta tags considered to be relevant to website visibility. This includes the HTML TITLE, KEYWORD and DESCRIPTION meta tags, as well as the following DCMI tags: TITLE, SUBJECT, DESCRIPTION, TYPE and SOURCE. A web page would earn a score of 0 for each one of these meta tags if it were not present at all. A score of 1 would be allocated if the tag was present, but not used effectively (eg. title totally irrelevant, too many keywords, no text inside tag, text contains excessive forms of spam, spelling mistakes in important keywords, etc). A score of 2 was allocated if the element was present, and if it was used effectively. The web page score therefore ranges between 0 and a maximum of (8 X 2) = 16.

The author would use the judgement of Sullivan [40] plus personal experience based on many years of exposure in the field, for the subjective issues of effective use of meta tags just described.

5. RESULTS AND ANALYSIS

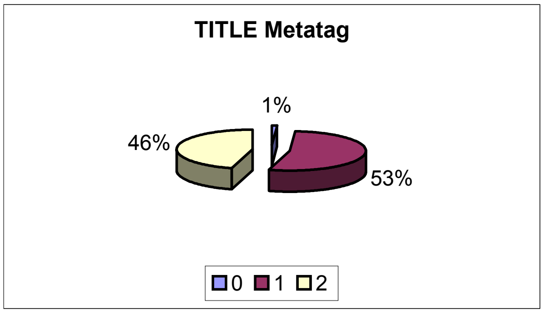
Inspection of the 200 websites' HTML coding produced not one instance of the usage of any one of the DCMI meta tags. As a result, it was decided to omit the five DCMI meta tags from further discussion. That left the three HTML meta tags only to produce scores. The maximum score for a website was therefore reduced to 2+2+2 = 6.

The header section of every URL in the sample was inspected, a judgement was made on the effective usage of each one of the three meta tag elements and the scoring system was applied. The results are ummarized in Table 1.

SCORE	TITLE	KEYW.	DESCR.	TOTAL
0	2	57	69	1
1	107	15	20	41
2	91	128	111	16
3				15
4				15
5				48
6				64

Table 1

5.1 TITLE meta tag



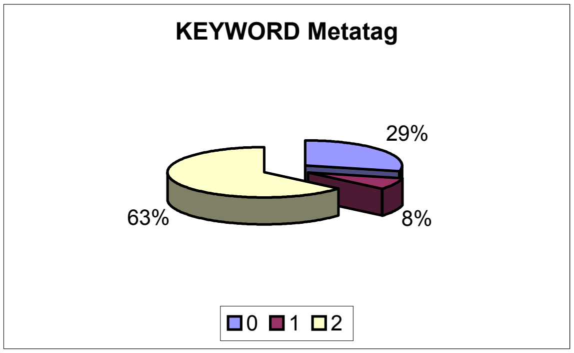
Graph 1

As Table 1 indicates, only 2 sites did not use the TITLE meta tag (score=0), 107 used it ineffectively (score=1) and 91 used it correctly (score=2). These figures are translated to percentages in Graph 1.

The average for this meta tag was 1.445 out of 2, which equates to 72.25%. Since this feature is used to identify a web page in the result listing produced by most search engines, users would be able to view at least the title of the web page, as seen by the designer in 99% of all cases. However, the value of having a correct title displayed could be diminished by the lack of other, more useful indicators like the contents. It remains unknown why ALL web page coders do not make use of this simple feature to enhance web page visibility, or why more than half the sample exhibited an apparent indifference to its use.

5.2 KEYWORD meta tag

From Table 1, 57 sites did not use the KEYWORD meta tag (score=0), 15 used it ineffectively (score=1) and 128 used it correctly (score=2). Graph 2 shows these figures as percentages.

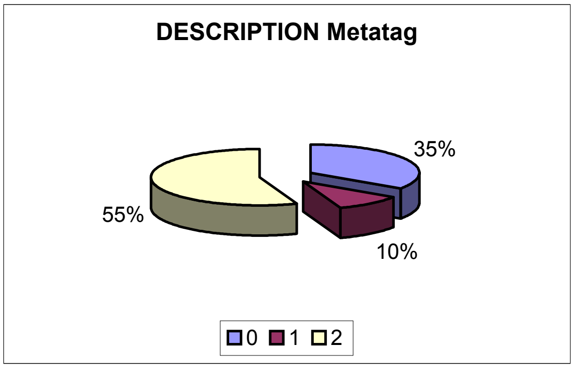


Graph 2

The average for this meta tag was 1.355 out of 2, which is equivalent to 67.75%. This figure is lower than the equivalent figure for the TITLE meta tag, indicating possible indifference by web page authors, or an inability to identify a number of important words which describe the core business of the relevant company. Since its inception however, the KEYWORD meta tag has been abused extensively by spammers in an attempt to increase website ratings with search engines. This was done by, for example, repeating the same important keyword thousands of times, using the same colour text and background for keywords, etc. As a result, only one search engine data base (Inktomi) currently still reads this meta tag when it indexes a web site. Therefore this result does not surprise, considering the low value attached to the use of the KEYWORD meta tag.

5.3 DESCRIPTION meta tag

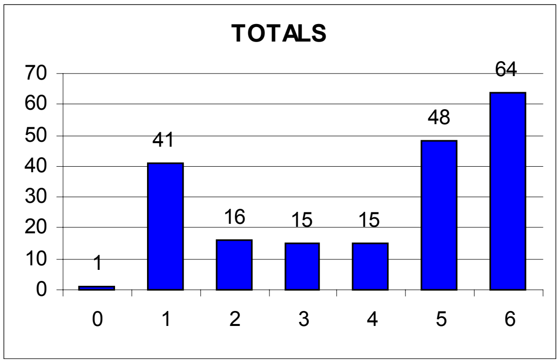
As Table 1 indicates, 69 sites did not use the DESCRIPTION meta tag (score=0), 20 used it ineffectively (score=1) and 111 used it correctly (score=2). These figures are shown as percentages in Graph 3.



Graph 3

The average for this meta tag was 1.21 out of 2, which equates to 60.5%. Once again there appears to be an even lower amount of effective usage of this meta tag, compared to the previous two. Just over ½ of the sampled web pages used it effectively, leaving the other ½ without it, or using it ineffectively. This is a surprising result, since this meta tag is used in a prominent way on the result pages of many search engines.

5.4 Overall Score



Graph 4

As seen on Graph 4, 63.5% of the 200 web sites earned a score of 4 and more. The other 36.5% did not use meta tags efficiently in order to enhance web page visibility.

From this graph it can also be seen that 58 out of 200 web pages (29%) earned a score of 2/6 and lower, while only 64 out of 200 (32%) earned the maximum of 6/6. On the other hand, only 1 web page (0.5%), did not use any one of the three meta tags in any way.

The overall average for website visibility related meta tag usage for all sites was 4.01 out of 6, yielding a value of 66.83%.

6. CONCLUSION AND RECOMMENDATIONS

To summarize some of the basic results listed above:

It appears as if the web page authors that earned a score of 2/6 and lower are aware that meta tags exists but do not bother to use them. It would also appear that the web page authors that earned a score of 3 to 4 out of 6 are also aware that meta tags exist but do not necessarily consider them to be important. The 56% of web page authors that earned a score of 5/6 and higher are fully aware of meta tags and their value to web page visibility.

It is thus concluded that the relationship W3, as defined in section 3, is invalid. Web page authors do not use metadata as a rule to enhance the electronic visibility of web pages.

The figures above indicate a lack of application of a basic but potentially effective methodology to increase web page visibility. It is recommended that company management insist on the implementation of these meta tags on mission critical websites. If this visibility enhancing mechanism is ignored, potential income opportunities are neglected.

At the same time, the other relationship which was identified earlier as being suspect (S3), should be investigated in future research. The three meta tags considered in this project were initially designed to enhance electronic visibility. If search engines fail to recognize meta tags on a large scale, web page authors have little reason to include them in their coding efforts. However, the status quo of this situation can only be determined by an empirical investigation into the recognition of meta tags by search engines.

Finally, the total absence of Dublin Core meta tags indicates either ignorance on the part of designers, or lack of usefulness of these tags. Further research could shed light on this aspect of web site visibility.

7. ACKNOWLEDGEMENTS

The authors would like to acknowledge to National Research Foundation for financial assistance, and the Cape Technikon for financial assistance and leave to complete this project.

8. REFERENCES

[1] Altavista, **Search Help,** [Online], Available: http://help.altavista.com/adv_search/ast_haw_metatags [Site visited on 01/07/2002].

[2] E.A. Brewer, "When everything is searchable", **Communications of the ACM,** 44(3), 2001, pp.54.

[3] T.Y. Chun, "World wide web robots: an overview", **Online & CD-ROM Review,** 23(3), 1999, pp. 135-142.

[4] S.J. Clarke and P. Willett, "Estimating the recall performance of web search engines", **Aslib Proceedings,** 49(7), 1997, pp. 184-189.

[5] C.W. Cleverdon, J. Mills and E.M. Keen, **Factors determining the performance of indexing systems,** Vol. 1, Design, Part 1, Cranfield: Aslib Cranfield Research Project, 1966.

[6] H.S. Coetzee, "The Dublin core initiative as a metadata standard for information retrieval on the Internet", **Meta-info Bulletin**, 8(4), 1999, pp. 9,10.

[7] L.B. Cohen and T.E. Jacobson, "Choosing tools to search the web", **The Teaching Professor,** 13(9), 1999, pp. 1.

[8] S. Collins, "Seek and ye shall find", **.net**, January, 1997, pp. 50-66.

[9] R. Conte, "Guiding lights", **Internet World**, May, 1996, pp. 41-44.

[10] M.D. Cooper, "Usage patterns of a web-based library catalog", **Journal of the American Society for Information Science and Technology,** 52(2), 2001, pp. 137.

[11] M.P. Courtois and M.W. Berry, "Results ranking in web search engines", **Online**, 1999, pp. 39-46.

[12] X. Dong and L.T. Su, "Search engines on the world wide web and information retrieval from the Internet: a review and evaluation", **Online & CD-ROM Review**, 21(2), 1997, pp. 67-80.

[13] P. Dwyer, "Search Engines – The Future", **.net**, July, 1997, pp. 97-100.

[14] D. Ellis, **Progress and problems in information retrieval,** London: Library Association Publishing, 1996, pp. 1,2.

[15] N. Garman, "Meta search engines", **Online**, 23(3), 1999, pp. 74-78.

[16] R. Hock, "Web search engines - features and commands", **Online,** May/June, 1999, pp. 24-28.

[17] S.R.C. Hoff, **AltaVista vs. Excite vs. Hotbot vs. Infoseek: which is the one to rely on?** [Online], Available: http://www4.zdnet.com/pccomp/srchoff/srchoff.html [Site visited on 1/10/2002].

[18] T.K. Huwe, "New search tools for multidisciplinary digital libraries", **Online**, March/April, 1999, pp. 67-74.

[19] P. Ingwersen, **Information retrieval interaction**, London: Taylor Graham., 1992, pp. 61.

[20] E.B. Jackson, "The engineer as reluctant information user – a remedial plan", **Proceedings of the International Conference on Training for Information Work,** Rome, November, 1971, pp. 15-19, 430.

[21] B.J. Jansen and U. Pooch, "A review of web searching studies and a framework for future research", **Journal of the American Society for Information Science and Technology**, 52(3), 2001, pp. 244.

[22] S. Jones, "Indexing the Internet – a job for machine? (or does it take human intervention?)", **NFAIS Newsletter**, April/May, 1996, pp. 80.

[23] F.W. Lancaster, **Information retrieval systems: characteristics, testing and evaluation**, New York, NY: John Wiley, 1978, pp. 13, 72, 331.

[24] A. Large, L.A. Tedd and R.J. Hartley, **Information seeking in the online age: principles and practice**, London: Bowker-Saur, 1999. pp. 5, 29.

[25] S. Lawrence and L. Giles, "Accessibility of information on the web", **Nature**, 400, 1999, pp. 107.

[26] D. Lidsky and R. Kwon, "Searching the Net", **PC Magazine,** December, 1997, pp. 227-255.

[27] C.A. Lynch, "When documents deceive: trust and provenance as new factors for information retrieval in a tangled web", **Journal of the American Society for Information Science and Technology**, 52(1), 2001, pp. 17.

[28] T. Nobles, **What will it be, directory or search engine?** [Online], Available: http://www.phoenixmlm.com/sengine.html. [Site visited on 16/10/2002].

[29] G.R. Notess, "The Infoseek databases", **Online**, August/September, 1995a, pp. 85-87.

[30] G.R. Notess, "Searching the world wide web: Lycos, Webcrawler and more", **Online**, July/August, 1995b, pp. 48-53.

[31] C. Oppenheim, A. Morris, C. McKnight and S. Lowley, "The evaluation of www search engines", **Journal of Documentation**, 56(2), 2000, pp. 190.

[32] A. Page, **The search is over**, [Online]. Available: http://www4.zdnet.com/pccomp/features/fea1096/sub2.html. [Site visited on 16/10/2002].

[33] S.E. Robertson, **Conflicting philosophies,** [Online]. Available: http://www.soi.city.ac.uk/research/cisr/ser/ucla/node3.html [Site visited on 10/04/2003].

[34] C. Sherman, "The future of web search", **Online**, May/June, 1999, pp. 54, 57.

[35] C. Sherman, "The future of web search", **Online**, May/June, 1999, pp. 54.

[36] C. Sherman, "Reference resources on the web", **Online**, January/February, 2000, pp. 52-56.

[37] A. Singh and D. Lidsky, "All-out search", **PC Magazine**, December, 1996, pp. 213-245.

[38] A. Spink, J. Bateman and B.J. Jansen, "Searching the web: a survey of Excite users", **Internet Research: Electronic Networking Applications and Policy,** 9(2), 1999, pp. 122, 125.

[39] A. Spink and others. "Searching the web: the public and their queries", **Journal of the American Society for Information Science and Technology,** 52(3), 2001, pp. 229.

[40] D. Sullivan, **More About Meta Tags,** [Online], Available:http://searchenginewatch.com/subscribers/more/metatags.html. [Site visited on 01/05/2003].

[41] S.A. Sutton, "Conceptual design and development of a metadata framework for educational resources on the Internet", **Journal of the American Society for Information Science,** 50(13), 1999, pp.1191.

[42] G. Taubes, "Indexing the Internet", **Science**, 269, 1995, pp. 1354-1356.

[43] Y.A. Tonta, **Failure analysis in document retrieval systems: a critical review of studies,** [Online], Available: http://yunus.hacettepe.edu.tr/~tonta/yayinlar/phd/bolum-3.htm [Site visited on 18/04/2002].

[44] G. Venditto, "Search engine showdown", **Internet World**, May, 1996, pp. 79-80.

[45] H.J. Voorbij, "Searching scientific information on the Internet: a Dutch academic user survey", **Journal of the American Society for Information Science**, 50(7), 1999, pp. 604-605.

[46] R.M. Wallace, J. Kupperman and J. Krajcik, "Science on the web: students online in a sixth-grade classroom", **The Journal of the Learning Sciences,** 9(1), 2000, pp. 75.

[47] **Washington University**, [Online], Available:http://depts.washington.edu/trio/comp/howto/site/design/metatags.shtml. [Site visited on 01/04/2003].

[48] M. Weideman, **Search engines**, [Online], Available: http://www.mwe.co.za/seaengin.htm. [Site visited on 23/5/2003].

[49] M. Weideman, **Newspaper Articles**, [Online], Available: http://www.mwe.co.za/seaarticles.htm [Site visited on 20/5/2003].

[50] I.R. Winship, "World-wide web searching tools: an evaluation", **Vine,** 99, 1995, pp. 49-54.

[51] D. Wolfram and A. Dimitroff, "Preliminary findings on searcher performance and perceptions of performance in a hypertext bibliographic retrieval system", **Journal of the American Society for Information Science**, 48(12), 1997, pp. 1145.

[52] M.F. Wyle, **Information overload**, [Online]. Available: http://eiger.wyle.org/~mfw/diss/node12.html. [Site visited on 10/04/2003].