Webology, Volume 2, Number 2, August, 2005

Search Engines and Resource Discovery on the Web: Is Dublin Core an Impact Factor?

Mehdi Safari
Encyclopedia Islamica Foundation, Tehran, Iran

Received June 5, 2005; Accepted August 13, 2005

Abstract

This study evaluates the effectiveness of the Dublin Core metadata elements on the retrieval of web pages in a suite of six search engines, AlltheWeb, AltaVista, Google, Excite, Lycos, and WebCrawler. The effectiveness of four elements, including title, creator, subject and contributor, that concentrate on resource discovery was experimentally evaluated. Searches were made of the keywords extracted from web pages of the Iranian International Journal of Science, before and after metadata implementation. In each search, the ranking of the first specific reference to the exact web page was recorded. The comparison of results and statistical analysis did not reveal a significant difference between control and experimental groups in the retrieval ranks of the web pages.

Keywords

Metadata, Dublin Core, Resource discovery, World Wide Web, Search engines

Introduction

Granted that the current World Wide Web contains tremendous amount of information provided by millions of users all over the world, it should be admitted that the problem of discovering the relevant resources is not easy. The Web has enabled users to electronically publish information accessible to millions of people relatively easily, but as the quantity of its information grows, the ability of those people to find relevant materials has decreased dramatically and can be compared to looking for "a needle in the haystack."

To solve the problem of discovering web resources, search engines have been developed that can provide the users with a large body of results by a click. While the value of these tools should not be underestimated, they have many shortcomings as information retrieval systems. With all of their power to provide access to an enormous array of information, it has been shown that they are finding it difficult to cope with the explosion of web resources (Bharat & Broder, 1998; Lawrence & Giles, 1998; Bar-Ilan, 1998/99; Lawrence & Giles, 1999) and accordingly cannot be considered as perfect tools because of their low coverage. Their performance volatility (Rousseau, 1998/99; Snyder & Rosenbaum, 1999), fluctuations and changes in the results set over time (Peterson, 1997; Bar-Ilan, 1998/99; Rousseau, 1999; Mettrop & Nieuwenhuysen, 2001) and a generally low retrieval effectiveness (Gordon & Pathak, 1999) are some major shortcomings and deficiencies which are well documented in the literature. It seems that deficiencies of search engines in retrieving the relevant resources mostly originate from their strategy of indexing the Web. The way in which they index the Web, that indiscriminately harvest whatever they can find and then do selective indexing on those contents, coupled with the enormous mass of web resources results in overly large retrieval sets with low relevancy. It has made it clear that without enforcement of a more rigorous indexing strategy through some level of meta control, search engines effectiveness and efficiency in resource discovery will deteriorate. Since the content of the information resources has not the right and efficient information for them to be indexed effectively, some kind of descriptive information to impose pre-defined meaning on the Web content is essential.

The metadata movement for resource discovery on the Web

The high dynamics of web resources (Lawrence & Giles, 1999), both in size and content, as well as their unique characteristics (Heery, 1996), has posed many challenges for using the traditional procedures of resource organization and discovery, such as cataloging rules, in the Web environment. The challenges in the way of deploying cataloging rules for digital resources (Beacom, 2000; Huthwaite, 2001; Lagoze, 2000; Weiss and Carstens, 2001), have led to favoring "metadata" as the best means of describing and discovering resources on the Web.

Metadata is a heavily loaded term for which many definitions have been offered. It, in general, may be defined as structured data about data (Burnett, Ng & Park, 1999, p.1212). More specifically, it is a structured set of elements that describes the information resource for the purpose of identification, discovery and use of information (Lee-Smeltzer, 2000, p.206). To encompass the main perspectives on metadata and accurately reflect the current status of its studies, Burnett et al. (1999) define metadata as "data that characterizes source data, describes their relationships, and supports the discovery and effective use of source data" (p. 1212).

Metadata is a recent coinage though not a recent concept. The above definitions about metadata are usually followed by the observation that libraries have been producing, standardizing and maintaining metadata for a long time; because descriptive data such as standard bibliographic information, and indexing and cataloging information are all structured data that describe the attributes and contents of an information resource to facilitate their discovery and use, hence metadata. However, while the concept of metadata is a familiar one for information professionals, in today's jargon, this data is considered to "[be] structured so that it can become machine-understandable as well as machine-readable [â€¦] and has largely been identified with issues of Internet resource discovery" (Day, 1999). As Milstead and Feldman (1999) point out, this term "â€¦ is generally applied to electronic resources (though it doesn't have to be) and refers to "data" in the broadest sense--datasets, textual information, graphics, music, and anything else that is likely to appear electronically. While the concept includes indexing and cataloging information (information for "resource discovery" in Webspeak), it can go far beyond conventional document representations, such as MARC records."

Today metadata activities are unprecedented. Because of the exponential growth of information resources on the Web, they expand beyond the traditional library environment to deal with the problem of effective resource description and discovery. The accelerated growth in the related literature on the topic of metadata and the rapid decrease of the word cataloging (Ercegovac, 1999) as well as several metadata standards with different levels of richness and complexity originated from different communities (Heery, 1996; Dempsey & Heery, 1997; Burnett, Ng & Park, 1999), reveals the unprecedented movement towards metadata for resource discovery on the Web.

Dublin Core Metadata Initiative: a simple metadata for the Web

Within the diverse resource discovery activities of the mid 90's, ranging from unstructured indexing of full-text resources by search engines to richly-structured data like Machine Readable Cataloging (MARC) and Text Encoding Initiative (TEI) records, Dublin Core metadata standard arose as a means to mediate these extremes. it originated from a workshop sponsored by Online Computer Library Center (OCLC) and the National Center for Supercomputing Applications (NCSA) in March 1995 to "form an international consensus on the semantics of a simple description record for networked resources" (Weibel, Iannella & Cathro, 1997). It was believed that resource discovery is the most pressing need that metadata can satisfy (Weibel et al., 1995). Therefore, only descriptive data elements required to support resource discovery were considered and data elements covering other characteristics of the resource such as terms and conditions, archival status, and other types of metadata were not included (Dempsey & Weibel, 1996).

The primary deliverable from the OCLC/NCSA workshop was a set of thirteen metadata elements, named the Dublin Core Metadata Element Set (or Dublin Core, for short) by the workshop participants. The Dublin Core was proposed as the minimum number of metadata elements required to facilitate the resource discovery in a networked environment such as the Internet (Weibel et al., 1995); and until the third workshop, the elements were increased to 15 (Weibel & Miller, 1997). This metadata elements set includes Title, Creator, Subject, Description, Publisher, Contributor, Date, Type, Format, Identifier, Source, Language, Relation, Coverage, and Rights (Dublin Core, 1999).

The functions of the Dublin Core elements can be categorized into four classes, according to the four elementary uses of the bibliographic data. The IFLA statement on the purpose of bibliographic records identifies four 'generic tasks' the users perform and these records should support (IFLA Study Group on the Functional Requirements for Bibliographic Records, 1998): To find, to identify, to select, and to acquire or obtain the resource. Dublin Core metadata elements support these four generic tasks as follows (Dublin Core Metadata and the Cataloging Rules, 1998):

- *FIND*: Elements that are likely to be primary search categories for discovery or finding of electronic resources include Title, Creator, Contributor, and Subject. The elements that are likely to be secondary or restricting features of a search are Language, Coverage, and Format.
- *IDENTIFY*: the elements that are related to the "instantiation" of the resource include Date, Type, Format, and Identifier.
- *SELECT*: the elements intended to provide some information for a user to make a selection among multiple search results include Description and Coverage.
- *OBTAIN*: In a networked environment, obtaining a resource should be fully supported by the inclusion of an accurate address in the Identifier element.

The implementation of Dublin Core elements on the Web requires a formal syntax. In 1996, a consensus concerning embedding metadata in HTML was reached at the *W3C Distributed Indexing and Searching Workshop* (Dempsey & Weibel, 1996). Because of the changes in HTML as well as a general need for greater formalization of the syntax, an Internet Draft authored by John Kunze (1999) was released after the debates at the sixth Dublin Core workshop, which explains how to encode Dublin Core elements in HTML. Current implementation of Dublin Core on the Web is often based on metadata embedded in HTML metatags.

Metadata Effectiveness: problem statement

With any metadata schema, there is a question of effectiveness. Does metadata provide a basis for increased effectiveness of retrieval by search engines? While there have been many studies done to evaluate search engines from different points of view, few studies have been done to test the effectiveness of metadata on resource discovery by search engines. Turner and Brackbill (1998) did a research on how the embedded metadata (HTML metatags) effects retrieval of web pages. The use of keywords metatag was shown to cause the significant improvement in the retrievability of a web page. However, another type of metatag (description metatag) exhibited no improvement in retrieval. Henshaw and Valauskas (2001) studied the effectiveness of Dublin Core metadata together with HTML keywords and description metatags on enhancing information retrieval in a suit of specific search engines. Results suggested that metadata did not play a significant role in increasing the likelihood of a web page being indexed or highly ranked by search engines.

This study aims to examine the following questions related to the use of four Dublin Core metadata elements, which are likely to be primary search categories for resource discovery:

- Do Dublin Core elements (Title, Subject, Creator and Contributor) improve the retrieval rank of a web page?
- Is retrieval performance of the major search engines improved after embedding metadata into the web pages?

Methodology

The web pages tested in this study, are a group of articles published in the home page of the *Iranian International Journal of Science* (freely available at: http://www.fos.ut.ac.ir/~journal/iijs.html). At the time of this research, the total number of the articles published online by this journal was 16 articles (see Appendix A). The articles were submitted to the major search engines (see table 1). Among these search engines, AOL Search, HotBot and Iwon failed to index the submitted articles and AlltheWeb, AltaVista, Google, Lycos, MSN Search, Excite and WebCrawler indexed the articles. Table 1 shows the results of searches of the titles in the search engines. The presence of the articles in the databases of the search engines is indicated by "+".

As table 1 indicates, MSN indexed only 4 articles and excluded from the study. Therefore, the maximum number of articles indexed by the maximum number of search engines is 10 and 6

respectively. These articles are shown by "*" sign. Table 2 shows the search engines that have indexed the articles and are tested in this study.

Table 1. The presence of articles in the database of Internet search engines

Article	Allthe Web	Alta Vista	AOL Search	Google	HotBot	Iwon	Lycos	MSN Search	Excite	Web Crawler
1	-	+	-	+	-	-	-	+	-	+
2	-	+	-	+	-	-	-	-	-	+
3	-	+	-	-	-	-	-	-	-	+
4	-	+	-	+	-	-	-	+	-	+
5*	+	+	-	+	-	-	+	-	+	+
6	-	+	-	+	-	-	-	-	+	+
7*	+	+	-	+	-	-	+	-	+	+
8*	+	+	-	+	-	-	+	+	+	+
9*	+	+	-	+	-	-	+	-	+	+
10*	+	+	-	+	-	-	+	+	+	+
11*	+	+	-	+	-	-	+	-	+	+
12*	+	+	-	+	-	-	+	-	+	+
13*	+	+	-	+	-	-	+	-	+	+
14*	+	+	-	+	-	-	+	-	+	+
15*	+	+	-	+	-	-	+	-	+	+
16	+	+	-	+	-	-	+	-	-	-

Table 2. Search engines tested in this study

Search Engine	URL
AlltheWeb	www.alltheweb.com
AltaVista	www.altavista.com
Excite	www.excite.com
Google	www.google.com
Lycos	www.lycos.com
WebCrawler	www.webcrawler.com

In the next step, keywords were extracted from the articles. Keyword extraction was performed in accordance with their corresponding Dublin Core elements. All of the keywords in the titles were extracted as the value of element "title"; the keywords provided by the authors in each article were considered as the value of element "subject"; the first creator of each article was considered as the value of element "creator" and the other ones (if any) as the value of element "contributor". In the case of some articles in which there was a complete overlap between subject and title keywords (all of the keywords assigned as subject keyword by authors existed in the title), a keyword was extracted from the abstract as the value of element subject to avoid any common keywords between title and subject elements. Totally 82 keywords were extracted from articles.

Using the simple search of search engines and regarding the nature of searching in each of engines, the keywords were searched. The phrases were searched with double quotes, so that the entire phrase was searched rather than each word of the phrase. The retrieval rank of a web page in a search engine results list was used to measure performance of the web pages. The higher the retrieval rank of a web page in a search engine results list, the better its performance and vice versa.

As at the first *Text Retrieval Conference*, using 200 results was reported as a retrieval threshold (Turner & Brackbill, 1998, p.264), the first 200 results of each search were examined as an arbitrary cutoff point and the ranking of the first specific reference to the exact web page within those first 200 hits was recorded. If a search could have resulted in retrieving a page but it was not in the top 200 results, that keyword was given the rank of 201 for that search. Rankings, therefore, ranged from 1 (highest) to 201 (not retrieved).

In the next step, the web pages were randomly divided into two groups: experimental group and control group. It resulted in totally 43 keywords in experimental group and 39 keywords in the

control group. The metadata elements were embedded into the web pages of experimental group through HTML metatags. Figure 1 shows the metadata elements of one of the web pages.

To ensure that search engines revisit the pages, a numeric character (digit 1) was added to the beginning of the titles of the pages in HTML title tag. Since search engines take the title of each page from this tag and show it to the user in the list of the retrieved results, showing the changed title (a title with digit 1) was considered as an indicator for ensuring that the search engines had revisited the pages through their continuous crawling and refreshing. It took 4 months for search engines to revisit all pages and among them, Google was the first one and AltaVista was the last one that revisited all pages. From the perspective of a content provider and regarding the high dynamics of the Web, it is highly suggested that the speed of revisiting web pages and updating the databases by Internet search engines might be improved.

Figure 1. Dublin Core elements in HTML metatags

```
<title> 1 Theoretical and Experimental Investigation on Back-scattered
Low Energy Gamma Radiation from Different Metals <title/>
<META NAME="DC.Title" CONTENT ="Theoretical and Experimental
Investigation on Back-scattered Low Energy Gamma Radiation from
Different Metals">
<META NAME="DC.Creator" CONTENT="A. Pazirandeh">
<META NAME="DC.Subject" CONTENT="Compton scattering, Rayleigh
scattering, Double scattering, Albedo spectrum, Coherent and incoherent
scattering, photo-electric effect">
<META NAME= "DC.CONTRIBUTOR" CONTENT="N. Sobhkhiz">
```

The web pages were controlled in order not to be changed during the study. The only change was metadata implementation in the experimental group. Once search engines revisited the pages, the searches were repeated with the same search terms and exactly in the same fashion and the results were recorded for future comparison. Appendix B and C indicate the first (R1) and second (R2) ranking of the keywords in the experimental and control groups. To determine the differences between the first and the second ranks, the first rank of each keyword was subtracted from its second rank. The differences achieved are indicated in R3 column. The negative numbers indicate losing ground and positive ones indicate gaining ground in the ranking of keywords.

Analysis of Findings

To determine whether metadata elements have affected the retrieval rank of the web pages, the achieved ranks (R3) of the keywords in the experimental and control groups were compared. The comparison was made by running Mann-Whitney U test. The U statistic "is used to test the significance of differences in central tendency between independent groups when the scores are ranks or when ranks have been substituted for the original scores" (Willemsen, 1974, p.193). This provides for a comparison of two sets of ranked scores and tests rankings of at least rank-order or ordinal level data. Since the results provided by the search engines are considered ordinal level data (Turner & Brackbill, 1998, p.265), Mann Whitney U test can be used to determine the significance of differences of ranks between two independent groups in this study. This test was run by SPSS (Statistical Package for Social Sciences) software.

In order to answer the question that whether Dublin Core elements improve the retrieval rank of a web page, the R3 of the keywords in experimental and control groups were compared. R3 of two groups were compared to determine if there is a significant difference for two groups with respect to the web pages retrieval ranks before and after the metadata implementation. Table 3 and 4 represent the descriptive statistics and Mann-Whitney U test statistics of the comparison.

Table 3. Descriptive Statistics of experimental and control groups

Groups	N	Minimum	Maximum	R3 Mean	Std. Deviation
Experimental	258	-187	200	9.736	52.74
Control	234	-200	200	9.974	51.05

Table 4. Mann-Whitney U Test Statistics of experimental and control groups

Mann-Whitney U	29437.000
Z	-.530
Asymp. Sig. (2-tailed)	.596

The significance level (p) or sig. for all tests is 0.01. Adopting .01 or 1 per cent as the significance level at or below which the difference of ranks between control and experimental groups are unlikely due to the chance, we can determine whether these differences are statistically significant or not. That is, if P .01 then we can conclude that the differences between two groups are statistically significant. As table 4 indicates, the significance level (P=596) is greater than .01 (.596>.01). The R3 Mean of the web pages with metadata elements and those without metadata elements are 9.73 and 9.94 respectively. Therefore, there is no statistically significant difference between experimental and control groups with respect to their retrieval rank improvement. In other words, using Dublin Core elements (Title, Subject, Creator and contributor) did not affected the retrieval rank of the web pages.

Is retrieval performance of the major search engines improved after embedding metadata into the web pages? To answer the second question of this study, each search engine was considered separately. As search engines use different algorithms for indexing and ranking the web pages, to determine the significance of the differences of ranks between experimental and control groups, the U test was run for each search engine.

Tables 5 and 6 show the statistical results for AlltheWeb. From table 6, the differences between two groups are not significant, (P=.410 [>0.01]). In other words, there is no statistically significant difference between the experimental and control groups and therefore, the retrieval performance of AlltheWeb is not improved after metadata implementation. The R3 Mean for the web pages with metadata and those without metadata are -2.55 and -2.84 respectively (see table 5). This, therefore, shows that the web pages with metadata did not achieve better rankings than the web pages without metadata in AlltheWeb.

Table 5. Descriptive Statistics for AlltheWeb

Groups	N	Minimum	Maximum	R3 Mean	Std. Deviation
Experimental	43	-94	86	-2.55	27.44
Control	39	-94	74	-2.84	21.21

Table 6. Mann-Whitney U Test Statistics for AlltheWeb

Mann-Whitney U	785.500
Z	-.825
Asymp. Sig. (2-tailed)	.410

The statistical results for AltaVista are presented in the tables 7 and 8. It is evident that the differences between the pages in the experimental and control groups with respect to their retrieval rankings are not statistically significant because of P=.519 (>0.01). It suggests that the variances of the rankings for two groups are not substantially different. The R3 Mean ranks for the experimental group and the control group are -2.51 and -7.10 respectively. Therefore, we cannot assume that the metadata implementation has enhanced web pages retrievability and ranking in AltaVista.

Table 7. Descriptive Statistics for AltaVista

Groups	N	Minimum	Maximum	R3 Mean	Std. Deviation
Experimental	43	-39	2	-2.51	7.31
Control	39	-198	2	-7.10	31.97

Table 8. Mann-Whitney U Test Statistics for AltaVista

Mann-Whitney U	776.50
Z	-.644
Asymp. Sig. (2-tailed)	.519

Tables 9 and 10 demonstrate the statistical results for Excite. From tables, the web pages in the experimental group (R3 Mean =27) did not achieve the better ranking than the web pages in the control group (R3 Mean=32.38) because of P=.729 (>0.01). It suggests that the variances of the rankings for two groups are not substantially different. The web pages with metadata elements, therefore, did not achieve better performance than those without metadata elements with respect to their retrieval ranking in Excite.

Table 9. Descriptive Statistics for Excite

Groups	N	Minimum	Maximum	R3 Mean	Std. Deviation
Experimental	43	-187	200	27	85.04
Control	39	-9	200	32.38	69.26

Table 10. Mann-Whitney U Test Statistics for Excite

Mann-Whitney U	806
Z	-.346
Asymp. Sig. (2-tailed)	.729

Tables 11 and 12 show the statistical results for Google. From table 12, the differences between the rankings of the web pages in the experimental group and the control group are not statistically significant (P=188 [>0.01]). The R3 Mean for the web pages with metadata elements and those without metadata elements are 12.25 and 7.64 respectively. Therefore, we cannot assume that metadata elements have caused better performance for the web pages with respect to their retrieval ranks in Google.

Table 11. Descriptive Statistics for Google

Groups	N	Minimum	Maximum	R3 Mean	Std. Deviation
Experimental	43	-90	189	12.25	32.71
Control	39	-21	159	7.64	29.33

Table 12. Mann-Whitney U Test Statistics for Google

Mann-Whitney U	707.50
Z	-1.31
Asymp. Sig. (2-tailed)	.188

The statistical results for Lycos are demonstrated in the tables 13 and 14. It is evident that the differences between the web pages in the experimental group (R3 Mean=1.69) and the control group (R3 Mean=-2.25) with respect to their retrieval rankings are not statistically significant because of P=.434 (>0.01). It suggests that the variances of the rankings for two groups are not substantially different. Therefore, we cannot assume that the metadata implementation has enhanced web pages retrievability and ranking in Lycos.

Table 13. Descriptive Statistics for Lycos

Groups	N	Minimum	Maximum	R3 Mean	Std. Deviation
Experimental	43	-70	84	1.69	22.05
Control	39	-92	75	-2.25	20.81

Table 14. Mann-Whitney U Test Statistics for Lycos

Mann-Whitney U	762
Z	-.783
Asymp. Sig. (2-tailed)	.434

Tables 15 and 16 show the statistical results for WebCrawler. From tables, the differences of rankings between the web pages in the experimental group (R3 Mean=22.53) and the control group (R3 Mean= 32.02) are not statistically significant (P=.547 [>0.01]). Metadata elements, therefore, have not caused better performance for the web pages with respect to their ranking and retreivability in WebCrawler.

Table 15. Descriptive Statistics for WebCrawler

Groups	N	Minimum	Maximum	R3 Mean	Std. Deviation
Experimental	43	-187	200	22.53	81.08
Control	39	-200	200	32.02	82.31

Table 16. Mann-Whitney U Test Statistics for WebCrawler

Mann-Whitney U	781.50
Z	-.602
Asymp. Sig. (2-tailed)	.547

Table 16 shows the significance level (P) of search engines. The significance level of every search engine is more than .01 and consequently, it suggests no statistically significant difference between the ranks of the pages with Dublin Core elements and the pages without Dublin Core elements. In other words, it is unreasonable to assume that the use of four Dublin Core metadata elements has led to an improvement in the retrieving and ranking of the web pages through six search engines: AlltheWeb, AltaVista, Excite, Google, Lycos and WebCrawler.

Table 17. Mann-Whitney U Test Results for Search Engines

Search engines	Significance Level (P=)
AlltheWeb	.410
AltaVista	.519
Excite	.729
Google	.188
Lycos	.434
WebCrawler	.547

Conclusion

The current strategy of search engines to indiscriminately harvest whatever they can find and then do full text indexing on those contents is an unsustainable for resource discovery and generally results in low relevancy in retrieval. Therefore, providing descriptive data to impose some level of meta control on the content is necessary for the current Web to be more effective and efficient. One of the best solutions to web resource discovery is the embedding of descriptive metadata in Web for harvesting by web index services. It consequently has led to developing and maintaining descriptive metadata schemas. Among the current metadata standards, Dublin Core has the potential of being adapted as an international standard for resource description and discovery on the Web and as a *lingua franca* for metadata. The goal of this study was to determine whether the Dubline Core implementation could improve web resource discovery via search engines. The effectiveness of four Dublin Core elements that concentrate on resource discovery was evaluated including Title, Subject, Creator, and Contributor. Two questions were considered in this study: Do Dublin Core elements improve the retrieval rank of a web page? and Is retrieval performance of the major search engines improved after embedding metadata into the web pages? Towards these aims, the articles published online by the *Iranian International Journal of Science* in the form of HTML pages (16 articles at the time of study) were considered as testing web pages and were submitted to 10 major search engines. The maximum number of search engines that indexed the maximum number of articles was 6 and 10 respectively. Keywords extracted from the indexed web pages were searched in the 6 search engines and their retrieval ranks were recorded for future comparisons. Dividing the web pages into two experimental and control groups, the metadata elements were embedded into the web pages of

experimental group. After that search engines revisited the pages through their continuous crawling and refreshing, the searches were repeated exactly in the same way and the results were recorded. Mann-Whitney U Test was employed to compare the results and examine two questions.

Based on the statistical analysis discussed in the previous section, and regarding the first question of the present study, it was found that using Dublin Core elements did not improve the retrieval rank of the web pages. Mann-Whitney U test comparisons of rankings of pages with metadata elements versus those without metadata elements did not reveal a statistically significant difference at the .01 level. The lack of a significant difference between two groups of web pages shows that four Dublin Core elements do not affect the retreivability and ranking of web pages and consequently is not an impact factor for resource discovery on the current Web. To answer the second question of the study, the retrieval performance of 6 search engines (AlltheWeb, AltaVista, Google, Lycos, Excite and WebCrawler) before and after metadata implementation was examined. Final statistical analysis revealed that the difference of ranks of the pages with metadata and those without metadata in each search engine was not significant and thus the retrieval performance of none of the search engines improved after metadata use. It shows that Dublin Core metadata, as a well-known metadata schema, is not widely accepted and used by search engine designers and the spiders do not consider its elements while ranking the web pages.

Resource discovery is impossible without resource description and adequate resource description assures effective discovery (Dillon, 2001). It is believed that the greatest potential for improvements to the resource discovery on the Web lies in the use of metadata. Undoubtedly, there is value in the current search engines as the main resource discovery tools on the Web, which operate without the aid of descriptive metadata. However, for them to be more effective and efficient metadata has to matter and they have to move beyond the full text indexing of the Web. Creating the metadata schemas for web resources is essential, but not sufficient. For a metadata schema to be an impact factor in resource discovery, it has to be widely accepted and deployed both by content providers and by web indexing services in a systematic way. As Lynch (2001, p.14) asks, if web indexing services do not use metadata, who will go to the expense and trouble of creating and maintaining it?

Acknowledgment

The Author would like to express his special gratitude to Prof. Abbas Horri for his encouragements and assistances. The author also would like to give thanks to Mr. Keyvan Salehi who helped with the statistical analysis and Mr. Darush Alimohammadi for his ongoing and useful debates on the subject.

References

- Bar-Ilan, J. (1998/99). Search engine results over time: A case study on search engine Stability. *Cybermetrics*, 2/3, (1). Retrieved January 8, 2004, from http://www.cindoc.csic.es/cybermetrics/articles/v2i1p1.html
- Beacom, M. (2000). Crossing a digital divide: AACR2 and unaddressed problems of networked resources. Retrieved November 27, 2004, from http://www.loc.gov/catdir/bibcontrol/patton_paper.html
- Bharat, K. & Broder, A. (1998). A technique for measuring the relative size and overlap of public web search engines. *Computer Networks and ISDN Systems*, 30, 379-388.
- Burnett, K., Ng, K. & Park, S. (1999). A comparison of the two traditions of metadata developments. *Journal of the American Society for Information Science*, 50(13), 1209-1217.
- Day, M. (1999). Metadata and electronic information. Retrieved December 1, 2004, from http://www.ukoln.ac.uk/metadata/presentations/circe/birmingham.html
- Dempsey, L. & Heery, R. (1997). A review of metadata: A survey of content resource description formats. Retrieved December 1, 2004, from http://www.ukoln.ac.kr/metadata/DESIRE/overview/revÂ¬_it.html
- Dempsey, L. & Weibel, S. L. (1996). The Warwick metadata workshop: A framework for the development of resource description. *D-Lib Magazine*, (July/August). Retrieved July 3, 2004, from http://www.dlib.org/dlib/july96/07weibel.html
- Dillon, M. (2001). Metadata for web resources: how metadata works on the Web. Retrieved July 3, 2004, from http://www.loc.gov/catdir/bibcontrol/dillon-paper.html
- Dublin Core metadata and the cataloging rules (1998). Retrieved December 1, 2004, from http://www.libraries.psu.edu/iasweb/personal/jca/dublin/dcreport.html

- Dublin Core metadata element set, version 1, 1: reference description (1999). Retrieved November 26, 2004, from http://dublincore.org/documents/1999/07/02/dces
- Ercegovac, Z. (1999). Introduction. *Journal of American Society for Information Science*, 50(13), 1165-1168.
- Gordon M. & Pathak P. (1999). Finding information on the World Wide Web: the retrieval effectiveness of search engines. Information Processing and Management, 35(2), 141-180.
- Heery, R. (1996). Review of metadata formats. *Program*, 30(4), 345-373.
- Henshaw, R. & Valauskas, E. J. (2001). Metadata as a catalyst: experiments with metadata and search engines in the internet journal, First Monday. *Libri*, 51(2), 86-101.
- Huthwaite, A. (2001). AACR2 and its place in the digital world: near-term solutions and long-term direction. Retrieved November 27, 2004, from http://www.loc.gov/catdir/bibcontrol/huthwaite_paper.html
- IFLA study group on the functional requirements for bibliographic records (1998). Functional requirements for bibliographic records: Final report. Retrieved December 1, 2004, from http://www.ifla.org/VII/s13/frbr/frbr.htm
- Kunze, J. (1999). Encoding Dublin Core metadata in HTML. Retrieved December 4, 2004, from http://www.fuqs.org/rfcs/rfcs2731/html
- Lagoze, C. (2000). Business unusual: how "event-awareness" may breathe life into the catalog? Retrieved November 27, 2004, from http://www.cs.cornell.edu/lagoze/papers/lagozelc.pdf
- Lawrence, S. & Giles, C. L. (1998). Searching the World Wide Web. *Science*, 280, 98-100.
- Lawrence, S. & Giles, C. L. (1999). Accessibility and distribution of information on the Web. *Nature*, 400, 107-110.
- Lee-Smeltzer, K. (2000). Finding the needle: controlled vocabularies, resource discovery and dublin core. *Library Collections, Acquisitions & Technical Services*, 24, 205-215.
- Lynch, C. A. (2001). When documents deceive: trust and provence as new factors for information retrieval in a tangled web. *Journal of the American Society for Information Science and Technology* , 52(1), 12-17.
- Mettrop, W. & Nieuwenhuysen, P. (2001). Internet search engines - fluctuations in document accessibility. *Journal of documentation*, 57 (5), 623-651.
- Milstead, J. & Feldman, S. (1999). Metadata: Cataloging by any other name. originally published in *ONLINE*, January 1999, Retrieved December 19, 2004 from: http://www.cbuc.es/5digital/1.pdf
- Peterson, R. E. (1997). Eight Internet search engines compared. *First Monday*, 2(2). Retrieved December 20, 2004, from http://www.firstmonday.dk/issues/issue2_2/peterson/index.html
- Rousseau, R. (1998/99). Daily time series of common single word searches in AltaVista and NorthernLight. *Cybermetrics*, 2/3(1). Retrieved December 15, 2004, from http://www.cindoc.csic.es/cybermetrics/articles/v2i1p2.html
- Rousseau, R. (1999). Time evolution of the number of hits in keyword searches on the Internet. Post Conference Seminar - Cybermetrics'99 at the *Seventh International Conference on Scientometrics and Informetrics*, July 9, 1999, Colima, Mexico. Summary Retrieved December 27, 2004, from http://www.cindoc.csic.es/cybermetrics/cybermetrics99.html
- Snyder, H. & Rosenbaum, H. (1999). Can search engines be used as tools for web-link analysis? A critical view. *Journal of Documentation*, 55(4), 375-384.
- Turner, T. P. & Brackbill, L. (1998). Rising to the top: evaluating the use of HTML metatag to improve retrieval of World Wide Web documents through Internet search engines. *Library Resources and Technical Services*, 42 (4), 258-271.
- Weibel, S., Godby, J., Miller, E. & Daniel, R. (1995). *OCLC/NCSA metadata workshop report*. Retrieved November 26, 2004, from http://www.ifla.org/documents/libraries/cataloging/oclcmeta.htm
- Weibel, S., Iannella, R. & Cathro, W. (1997). The 4th Dublin Core metadata report: DC-4, March 3-5, 1997, National Library of Australia, Canberra. *D-Lib Magazine*, (June). Retrieved November 26, 2004, from http://www.dlib.org/dlib/june97/metadata/06weibel.html
- Weibel, S. & Miller, E. (1997). Image description on the Internet: A summary of the CNI/OCLC image metadata workshop, September 24-25, 1996, Dublin, Ohio. *D-Lib Magazine*, (January). Retrieved November 26, 2004, from http://www.dlib.org/dlib/january97/oclc/01weibel.html
- Weiss, A. K. & Carstens, T. V. (2001). The year's work in cataloging, 1999. *Library Resources and Technical Services* , 45(1), 47-53.

- Willemsen, E.W. (1974). *Understanding statistical reasoning: how to evaluate research literature in the behavioral sciences*. San Francisco: W.H. Freeman and Company.

Appendix A: The articles' titles of the Iranian International Journal of Science

No.	Title
1	Some Production Comparisons of Two Celluloytic Fungi
2	Third Virial Coefficient and Compressibility Factors for Dense Spherical Gases Using the HFD-C Potential
3	Digenetic Studies, a Key to Reveal the Timing of Oil Migration: an Example from the Tirrawarra Sandston Reservoir, Southern Cooper Basin, Australia
4	Optimal Control of an Inhomogeneous Problem by Using Measure Theory
5	Probe Diagnostics of Confined Plasma Produced by 13.56 MHz R.F Plasma Source
6	Occurrence and Distribution of Aquatic saprolegniaceae in Northwest and South of Tehran
7	Effect of Pectic Acid and b-Glocan on Prolactin Secretion by Ovine Pituitary Explants
8	Deformational Behavior of Quartz and Feldspar in Quartzites within Shear Zones in the Adelaide Hills Area, South Australia
9	Construction of some Join Spaces Boolean Algebras
10	Characterization of Certain Infinitely Divisible Distributions
11	Theoretical and Experimental Investigation on Back-Scattered Low Energy Gamma Radiation from Different Metals
12	Cytogenetic Biomonitoring of Workers Occupationally Exposed to Aromatic Solvents
13	Notes on the Distribution, Climate and Flora of the Oil Field Areas, South-West of Iran
14	Isotopic Signature of the Diagenetic Fluids and Cement in the Tortachilla Limestone, South Australia
15	Correlating marine Palynomorph Variations with Sequence Boundaries of Upper Jurassic Sediments in a Basin of Northern Switzerland
16	On Approximately Convex Functions

Appendix B: The rankings of the keywords in the experimental group

		AlltheWeb			AltaVista			Excite			Google			Lycos			WebCrawle		
Element	**Keyword**	R1	R2	R3	R1	R2	R3	R1	R2	R3	R1	R2	R3	R1	R2	R3	R1	R2	R3
Title	Energy Gamma Radiation	40	36	+4	29	29	0	201	201	0	178	136	+42	40	36	+4	201	201	0
Title	Metals	201	201	0	201	201	0	201	201	0	201	201	0	201	201	0	201	201	0
Title	Pectic Acid	1	4	-3	6	7	-1	1	1	0	7	6	+1	1	4	-3	1	1	0
Title	b-Glucan	5	9	-4	26	30	-4	15	30	-15	74	53	+21	5	9	-4	15	29	-14
Title	Prolactin Secretion	3	23	-20	10	10	0	14	201	-187	160	92	+68	3	23	-20	14	201	-187
Title	Ovine Pituitary Explants	2	1	+1	1	1	0	1	1	0	1	1	0	2	1	+1	1	1	0
Title	Deformational Behavior	1	1	0	8	6	+2	201	1	+200	1	1	0	1	1	0	201	1	+200
Title	Quartz	201	201	0	201	201	0	201	201	0	201	201	0	201	201	0	201	201	0
Title	Feldspar	201	201	0	201	201	0	201	201	0	201	153	+48	201	201	0	201	201	0
Title	Quartzites	96	82	+14	162	201	-39	201	55	+ 146	201	12	+189	97	82	+15	201	57	+144
Title	Shear Zones	201	201	0	165	187	-22	201	40	+161	201	201	0	201	201	0	201	30	+171
Title	Adelaide Hills Area	101	15	+86	1	6	-5	201	39	+162	201	201	0	99	15	+84	201	45	+156

Title	South Australia	201	201	0	201	201	0	201	201	0	201	201	0	201	201	0	201	201	0
Title	Probe Diagnostics	102	175	-73	11	10	+1	201	201	0	88	54	+34	127	175	+48	201	201	0
Title	Confined Plasma	201	201	0	23	26	-3	201	201	0	166	103	+63	201	201	0	201	201	0
Title	R.F Plasma Source	1	1	0	1	1	0	201	7	+194	59	44	+15	1	1	0	201	8	0
Title	Isotopic Signature	100	194	-94	201	201	0	201	201	0	201	201	0	100	144	-44	201	201	0
Title	Diagenetic fluids	69	10	+59	201	201	0	22	201	-179	55	29	+26	70	10	+60	22	201	-179
Title	Cement	201	201	0	201	201	0	201	201	0	201	201	0	201	201	0	201	201	0
Title	Tortachilla Limestone	7	1	+6	201	201	0	1	1	0	4	3	+1	7	1	+6	1	1	0
Title	South Australia	201	201	0	201	201	0	201	201	0	201	201	0	201	201	0	201	201	0
Subject	Compton scattering	201	201	0	201	201	0	201	201	0	201	201	0	201	201	0	201	201	0
Subject	Reyleig scattering	201	201	0	201	201	0	201	201	0	201	201	0	201	201	0	201	201	0
Subject	Double scattering	82	160	-78	22	27	-5	201	57	+144	49	29	+20	83	153	-70	201	58	+143
Subject	Albedo spectrum	12	7	+5	1	1	0	201	1	+200	10	5	+5	12	9	+3	201	1	+200
Subject	Coherent and incoherent scattering	54	62	-8	18	26	-8	201	55	+146	39	38	+1	54	58	-4	201	55	+146
Subject	Photo-electric effect	201	201	0	201	201	0	201	201	0	201	201	0	201	201	0	201	201	0
Subject	Plant extracts	201	201	0	201	201	0	201	201	0	201	201	0	201	201	0	201	201	0
Subject	Intracrystalline deformation	4	3	+1	3	3	0	6	10	-4	14	11	+3	4	3	+1	6	10	-4
Subject	R.F Plasma reactor	3	8	-5	4	4	0	201	15	+186	27	25	+2	4	8	-4	201	15	+186
Subject	Stable isotopes	201	201	0	201	201	0	201	201	0	201	201	0	201	201	0	201	201	0
Subject	Meteoric cement	1	1	0	1	3	-2	5	2	+3	5	6	+1	1	1	0	5	2	+3
Subject	Diagenesis	201	201	0	201	201	0	201	201	0	201	201	0	201	201	0	201	201	0
Subject	Isotopic composition	201	201	0	201	201	0	201	201	0	201	201	0	201	201	0	201	201	0
Creator	A. Pazirandeh	2	2	0	201	201	0	3	9	-6	201	201	0	2	2	0	3	9	-6
Creator	Houri Sepehri	1	1	0	1	2	-1	1	1	0	1	1	0	1	1	0	1	1	0
Creator	Ali Yassaghi	2	3	-1	1	1	0	2	2	0	4	7	-3	2	3	-1	2	2	0
Creator	M. Khorassani	1	2	-1	2	2	0	9	1	+8	1	1	0	1	2	-1	9	1	+8
Creator	Hossain Rahimpour-Bonab	5	6	-1	4	5	-1	3	2	+1	2	2	0	5	6	-1	3	2	+1
Contributor	N. Sobhkhiz	1	1	0	1	1	0	1	1	0	1	1	0	1	1	0	1	1	0
Contributor	Roya Zoraghi	2	2	0	4	5	-1	3	2	+1	4	5	-1	2	2	0	2	1	+1
Contributor	Ali Haeri Rouhani	1	1	0	1	1	0	1	1	0	1	1	0	1	1	0	1	1	0
Contributor	Yvonne Bone	49	47	+2	7	26	-19	201	201	0	30	39	-9	50	47	+3	201	201	0

R1: The first retrieval rank
R2: The second retrieval rank
R3: The difference achieved

Appendix C: The rankings of the keywords in the control group

		AlltheWeb			**AltaVista**			**Excite**			**Google**			**Lycos**			**WebCrawler**		
Element	**Keyword**	R1	R2	R3	R1	R2	R3	R1	R2	R3	R1	R2	R3	R1	R2	R3	R1	R2	R3
Title	Boolean	201	201	0	74	76	-2	201	201	0	201	201	0	201	201	0	201	201	0

	Algebras																		
Title	Join Spaces	1	1	0	5	5	0	201	8	+193	16	12	+4	2	1	+1	201	8	+193
Title	Infinitely Divisible Distributions	28	21	+7	32	35	-3	201	38	+163	32	28	+4	28	21	+7	201	30	+171
Title	Cytogenetic Biomonitoring	25	36	-11	2	5	-3	201	201	0	12	21	-9	25	32	-7	201	201	0
Title	Aromatic solvents	201	201	0	22	20	+2	201	201	0	201	201	0	201	201	0	201	201	0
Title	Marine Palynomorph	22	20	+2	2	1	+1	25	7	+18	3	4	-1	22	17	+5	25	7	+18
Title	Sequence boundaries	201	201	0	40	38	+2	201	201	0	129	82	+47	201	201	0	201	201	0
Title	Upper Jurassic Sediments	6	56	-50	3	201	-198	201	201	0	18	16	+2	6	52	-46	201	201	0
Title	Northern Switzerland	107	201	-94	201	201	0	201	201	0	46	30	+16	109	201	-92	201	24	+177
Title	Climate	201	201	0	201	201	0	201	201	0	201	201	0	201	201	0	201	201	0
Title	Flora	201	201	0	201	201	0	201	201	0	201	201	0	201	201	0	201	201	0
Title	Oil Field Areas	81	7	+74	1	1	0	201	29	+172	28	10	+18	82	7	+75	201	29	+172
Title	South-West of Iran	66	73	-7	9	8	+1	201	37	+164	201	42	+159	66	73	-7	201	38	+163
Subject	Hypergroup	30	48	-18	10	8	+2	201	201	0	21	19	+2	32	48	-16	201	201	0
Subject	Algebraic hyperstructure	1	1	0	1	1	0	2	1	+1	1	1	0	1	1	0	2	1	+1
Subject	Infinite divisibility	201	201	0	70	75	-5	201	201	0	42	55	-13	201	201	0	1	201	-200
Subject	Strictly stable distributions	2	2	0	1	1	0	201	2	+199	3	4	-1	2	2	0	201	2	+199
Subject	Cauchy distribution	201	201	0	201	201	0	201	201	0	201	201	0	201	201	0	201	201	0
Subject	Normal distribution	201	201	0	201	201	0	201	201	0	201	201	0	201	201	0	201	201	0
Subject	Characteristics function	201	201	0	105	120	-15	201	201	0	199	123	+76	201	201	0	201	201	0
Subject	Unimodality	201	201	0	48	79	-31	201	201	0	125	146	-21	201	201	0	201	201	0
Subject	Chromosomal Aberrations	201	201	0	183	201	-18	201	201	0	201	201	0	201	201	0	201	201	0
Subject	Lymphocytes	201	201	0	201	201	0	201	201	0	201	201	0	201	201	0	201	201	0
Subject	Occupational Exposure	201	201	0	201	201	0	201	201	0	201	201	0	201	201	0	201	201	0
Subject	Organic Solvents	201	201	0	201	201	0	201	201	0	201	201	0	201	201	0	201	201	0
Subject	Dinoflagellates	201	201	0	201	201	0	201	201	0	201	201	0	201	201	0	201	201	0
Subject	Rhodano-Swabian basin	1	1	0	1	1	0	1	1	0	1	1	0	1	1	0	1	1	0
Subject	Floristic composition	201	201	0	119	123	-4	201	201	0	201	201	0	201	201	0	201	201	0
Subject	Saharo-Sindian region	3	4	-1	1	1	0	201	1	+200	2	1	+1	3	2	+1	201	1	+200
Subject	Plant geography	201	201	0	201	201	0	201	201	0	201	201	0	201	201	0	201	201	0
Subject	SW. Iran	47	53	-6	17	15	+2	201	50	+151	66	47	+19	47	51	-4	201	48	+153
Subject	Khuzistan	23	22	+1	21	25	-4	201	201	0	34	40	-6	23	22	+1	201	201	0
Creator	Ali Reza Ashrafi	5	6	-1	7	11	-4	7	16	-9	10	11	-1	5	6	-1	7	16	-9
Creator	M. Hossein Alamatsaz	1	1	0	1	1	0	1	1	0	1	1	0	1	1	0	1	1	0
Creator	Hossein Mozdarani	6	9	-3	5	5	0	4	5	-1	8	6	+2	6	9	-3	4	5	-1
Creator	Ebrahim Ghasemi-Nejad	7	9	-3	3	2	+1	7	5	+2	2	2	0	7	8	-1	7	5	+2

Creator	Ebrahim Alaie	1	1	0	1	2	-1	1	1	0	1	1	0	1	1	0	1	1	0
Creator	Shirazeh Arghami	1	1	0	1	1	0	1	1	0	1	1	0	1	1	0	1	1	0
Creator	A. Ghahreman	1	2	-1	3	3	0	12	2	+10	3	3	0	1	2	-1	12	2	+10

R1: The first retrieval rank
R2: The second retrieval rank
R3: The difference achieved

Bibliographic information of this paper for citing:

Safari, M. (2005). "Search Engines and Resource Discovery on the Web: Is Dublin Core an Impact Factor?" *Webology*, **2** (2), Article 13. Available at: http://www.webology.org/2005/v2n2/a13.html

---

This article has been cited by other articles.
