Received 28 May 2010
Revised 22 June 2011
Accepted 12 August 2011

How are XML-based Marc 21 and Dublin Core records indexed and ranked by general search engines in dynamic online environments?

A. Hossein Farajpahlou and Faeze Tabatabai
Department of Library and Information Science, Shahid Chamran University, Ahvaz, Iran

Abstract

Purpose – The aim of this paper is to examine the indexing quality and ranking of XML content objects containing Dublin Core and MARC 21 metadata elements in dynamic online information environments by general search engines such as Google and Yahoo!

Design/methodology/approach – In total, 100 XML content objects were divided into two groups: those with DCXML elements and those with MARCXML elements. Both groups were published on the web site www.marcdcmi.ir in late July 2009 and were online until June 2010. The web site was introduced to Google and Yahoo! search engines. The indexing quality of metadata elements embedded in the content objects in a dynamic online information environment and their indexing and ranking capabilities were compared and examined.

Findings – Google search engine was able to retrieve fully all the content objects through their Dublin Core and MARC 21 metadata elements; Yahoo! search engine, however, did not respond at all. Results of the study showed that all Dublin Core and MARC 21 metadata elements were indexed by Google search engine. No difference was observed between indexing quality and ranking of DCXML metadata elements with that of MARCXML. The results of the study revealed that neither the XML-based Dublin Core Metadata Initiative nor MARC 21 demonstrate any preference regarding access in dynamic online information environments through Google search engine.

Practical implications – The findings can provide useful information for search engine designers.

Originality/value – The present study was conducted for the first time in dynamic environments using XML-based metadata elements. It can provide grounds for further studies of this kind.

Keywords Dublin Core, MARC 21, Indexing, Ranking, Dynamic online environments, Classification, Search engines

Paper type Research paper

Aslib Proceedings: New Information Perspectives
Vol. 63 No. 6, 2011
pp. 586-592
© Emerald Group Publishing Limited
0001-253X
DOI 10.1108/00012531111187234

1. Introduction

In line with recent developments in information and communication technology, we are witnessing an increasing growth and improvement in different dynamic online information databases. Such databases contain content objects and up-to-date scientific sources in different branches of knowledge. Therefore, librarians and information professionals have always recognized the significance of knowledge and information classification. As a result, numerous research programs have been conducted on the development of metadata initiatives and standards based on the needs of various domains. In other words, the need for the application of metadata standards is now unavoidably associated with ongoing developments in digital libraries and dynamic online information databases.

MARC metadata format is one of the most important metadata schemes which have been made compatible with dynamic environments, with the capacity to identify, classify, and retrieve web resources and content objects. Dublin Core (DC) metadata initiative is another main and international metadata initiative which was originally created for the application in identification, retrieval and classification of web content objects. However, in order to take full advantage of metadata schemes in the web environment, a few challenges should be resolved. One important issue is the ability of general search engines such as Google to index metadata elements and retrieve content objects using their embedded metadata elements.

Another major issue in identification and effective retrieval of content objects that contain metadata elements relates to their semantic environment and the interoperability between different platforms and applications. The use of extensible markup language (XML) in DC and MARC 21 is because of the high capacity of XML in increasing the interoperability. One advantage of using XML in these two metadata schemes is that the indexing software of search engines can index XML-based elements in static information environments (Taheri and Hariri, 2011). However, retrieval of information in dynamic online information environments remains an issue that should be addressed. Hence, the present study examines the indexing quality and ranking of content objects consisting of the XML-based MARC 21 and DC metadata elements in dynamic online information environments by two general search engines, Yahoo! and Google.

As most of the information and scientific content objects are in dynamic online information environments, the importance of this research lies in unveiling the quality of indexing of content objects containing DCXML and MARCXML elements in dynamic online information environments by Yahoo! and Google. The findings can show the efficiency of each of these two popular search engines in indexing XML-based metadata elements.

2. Research questions

This research sought answers to the following seven questions:

(1) What is the indexing quality of content objects containing XML-based DC metadata elements in dynamic online information environments as performed by Yahoo! and Google search engines?

(2) What is the indexing quality of content objects containing XML-based MARC 21 metadata elements in dynamic online information environments as performed by Yahoo! and Google search engines?

(3) What is the difference between the indexing quality of three main elements (title, author and subject) of content objects containing DC and XML-based MARC 21 metadata elements in dynamic online information environments as performed by Yahoo! and Google search engines?

(4) What is the difference in the ranking procedure of content objects containing XML-based MARC 21 and DC metadata elements in dynamic online information environments as performed by Yahoo! and Google search engines?

(5) What is the reaction of Yahoo! and Google search engines to content objects of dynamic online information environments containing XML-based metadata elements with flat structure (DC) and hierarchical structure (MARC 21)?

(6) What is the reaction of Yahoo! and Google search engines to metadata initiatives with language-based tags (DC) and without language-based tags (MARC 21)?

(7) Which one of MARC 21 and Dublin Core metadata initiatives is more suitable for classification of XML-based content objects in dynamic online information environments in regard to access through Google and Yahoo! search engines?

3. Literature review

Turner and Brackbill (1998) looked at the ways in which access to (hypertext markup language (HTM) documents could be improved using HTML meta-tags. They found out that assigning the "description" meta-tag alone was not able to improve the retrieval of the document in general search engines; however, the "keywords" meta-tag did improve the access.

Sokvitne (2000) conducted research on the web sites of 20 Australian large educational and government organizations aimed at identifying the ability to retrieve key elements such as title, publisher, author, and subject in DC. The results of the study revealed that because of inconsistencies in the content records' formats, elements such as author, publisher and co-author which could be useful in searching and retrieving objects, remained useless. Since the "subject" was not used properly and the "title" content was the same as the HTML title's tag content, these elements are not effective in the retrieval process.

Henshaw and Valauskas (2001) conducted experimental research on some selected pages of *First Monday*'s electronic magazine. Two groups of pages were included in this research including a control group with no metadata element and a test group with DC metadata elements as well as HTML keywords and description meta-tags. Results of the study revealed that metadata alone did not have any impact on increasing the probability of indexing the resources and giving them top ranks in search engines' results.

Zhang and Dimitroff (2004) in a study entitled "Internet search engines' response to metadata Dublin Core implementation" examined the function of seven main search engines which were categorized into two groups: a target group and a control group. The target group consisted of the subject element of the DC metadata scheme as well as the HTML "keyword" element. The control group lacked any such elements. The results showed that there was a significant difference between the two groups in terms of visibility for search engines; i.e. six out of the seven search engines responded positively to metadata elements.

Quevedo-Torrero (2004) looked for ways to improve search quality and retrieval of web pages by inserting keywords in HTML meta-tags as metadata. The research used a selection of search results rendered by search engines like Google and AltaVista. Some strategies were formulated and suggested for improvement in ranking of search results by using HTML meta-tags as metadata, and clustering web pages according to their link structures.

Zhang and Dimitroff (2005a) examined the effect of web page content features on their visibility and inclusion in search engines' results. This research aimed at finding answers to the question: "how could the ranking of a page or a site in a search engine result be improved in view of authors or developers of pages or web sites?" The results revealed that repetition of keywords in the title as well as in the full text body improves the visibility of pages in search engines' results. Factors like colour and font size appeared to have no effect on the visibility.

Zhang and Dimitroff (2005b) conducted another experimental study to examine the effect of implementing metadata on the visibility of web pages in search engines' results. For this purpose they introduced 40 test web pages to 19 search engines. The results of the study showed that metadata is an appropriate and effective mechanism for increasing the visibility and ranking of web pages. Moreover, keywords extracted from web pages, especially from the title and full-text body, proved to be very effective in ranking.

Mohamed (2006) investigated the effect of metadata usage on the ranking and retrieval of web pages. This research was conducted in two parts. In part one, the effect of metadata initiative on the access to content objects was considered and examined. In part two, by adding metadata elements to web pages, the extent of their indexing was measured as well as the effect of metadata on page ranking. The results showed that description elements and keywords have a significant role in page ranking.

Also, a couple of relevant studies have been conducted in Iran. Safari (2005) in research on 16 articles that were published on the web version of the *Iranian International Journal of Science* studied the effect of DC metadata elements (four out of 15 elements) on the ranking of web sources by Google, AltaVista and Lycos. His results showed no significant differences between the ranking of pages that contained DC metadata elements and those pages that did.

Taheri and Hariri (2011) conducted a comparative study on the indexing quality and ranking of content objects containing DCXML and MARCXML metadata elements by general search engines. His findings showed that there was no significant difference between the indexing quality of content objects containing DCXML and MARCXML elements as performed by Google and Yahoo! Also, there was no significant difference between content objects ranking containing the two metadata initiatives in Google search engine; however, there was a significant difference in the ranking status of content objects containing the two metadata initiatives in Yahoo! search engine.

4. Methodology

A total of 100 content objects (i.e. e-books) were selected from a California digital library source set. They were selected using the URL www.archive.org and focusing on the subject "theory of knowledge". The e-books were divided into two groups. The first group contained DCXML elements, and the second group contained MARCXML elements. Both groups were mounted on www.marcdcmi.ir and introduced to Yahoo! and Google search engines from late July 2009 till June 2010. The data were collected in April 2010. The mentioned web site was introduced to Google search engine by "Webmaster Tools" through "XML Sitemap" option and "Suggest a site". Introduction to Yahoo! was done using "Yahoo! Search URL Status Review Form" and "ROR & Text Sitemap" with the same condition. Google search engine could retrieve all the content objects fully by DCXML and MARCXML metadata elements; however, Yahoo! search engine, despite many follow-ups, did not respond at all. Therefore, we had to rely only on Google results.

The data were collected by means of a checklist which was devised on the basis of, and according to, research questions and requirements. Searches were done in Google using the query: ["keyphrase" site:marcdcmi.ir"] and the results were analyzed using the checklist. The data that were collected by means of the checklist were transferred to worksheets in which " + " and " − " signs were assigned as indications of "being indexed" or "not being indexed" respectively. Each of these positions received "1" and "0" values respectively for calculation purposes. The sum of these values were then used in analyses and answering the research questions.

5. Findings

As mentioned above, Yahoo! search engine never responded to metadata elements retrieval during the study and it was omitted from the research. What we present here is only the findings of the test on Google.

Table I answers the first two research questions. The findings indicate that Google has been able to index DCXML elements (nine elements) as well as MARCXML elements (ten elements). Therefore, XML-based content objects which were embedded in the research dynamic online information environment proved to be retrievable. In fact, the indexing quality of the selected elements by Google search engine is suitable.

Table II illustrates the indexing quality of Google search engine in regard to title, author and subject elements both in DCXML and MARCXML. The content of Table II answers the third research question. Findings show that Google search engine is able to index title, author, and subject content elements in DCXML and MARCXML. Therefore, there is no difference between these elements in this regard.

Table III is used to answer the fourth research question regarding the rank quality. Out of 50 content objects containing MARCXML metadata elements, Google placed only 25 objects higher than content objects containing DCXML elements. In other words, the ratio of XML-based content objects containing metadata elements is equally 25 out of 50 for both metadata schemes.

Table I.
Indexing quality of DCXML and MARCXML elements in dynamic online information environments by Google

Metadata initiative (in Google)	The number of the studied elements	The number of content objects	Indexing percentage
Dublin Core	9	50	100
MARC 21	10	50	100

Table II.
Indexing quality of title, author and subject elements of DCXML and MARCXML embedded in content objects in dynamic information environments by Google

Metadata initiative (in Google)	The studied main elements	The number of content objects	The obtained point by content objects
Dublin Core	Title	50	50
MARC 21	Title	50	50
Dublin Core	Author	50	50
MARC 21	Author	50	50
Dublin Core	Subject	50	50
MARC 21	Subject	50	50

To answer the fifth and sixth questions, the contents of Table II are useful. The data in Table II show that Google search engine indexing software does not discriminate between content objects with a flat structure and with a language-based tag and those with a hierarchical structure and without a language-based tag.

Finally, the seventh research question aims to determine more suitable metadata initiatives for organization of the XML-based content objects in dynamic online information environments in terms of accessibility by Google search engine. Answering this question, one would conclude that none of the XML-based Dublin Core metadata initiative or MARC 21 shows any preference over the other in this regard. In other words, both metadata schemes are appropriate for organization of XML-based content objects in dynamic online information environments, as far as accessibility by Google is concerned.

6. Discussion and conclusion

Review of the literature tells us that a few researchers such as Zhang and Dimitroff (2005b), Mohamed (2006) and Safari (2005) studied "ranking" in relation to the application of metadata elements. Taheri and Hariri (2011) tackled exactly the same issue, but in regards to static environments, and Henshaw and Valauskas (2001) focused on indexing. Our findings lend support to the findings of Taheri and Hariri (2011), as both showed that there is no significant difference between the quality of indexing and ranking of general search engines in regards to content objects containing XML-based DC and MARC 21 metadata elements.

Based on the findings, one could conclude that XML, as the syntax ground for implementing the metadata elements of DC and MARC 21, in comparison with HTML, can be effective both in static and dynamic environments. This might justify the preference of XML over HTML because it maximizes the interoperability between search engines and the metadata initiatives. Therefore, both metadata initiatives can be regarded as appropriate for making different content objects accessible in dynamic online environments via Google. On the other hand, none of the two metadata initiatives proved to have clear superiority over the other in terms of indexing capabilities.

Regarding the ranking of the content objects under study, it was found that Google does not discriminate between the two metadata initiatives. This means that Google follows a similar pattern and policy in indexing and ranking of the content objects containing these two metadata schemas.

Also, from the answer to the third research question it was clear that the structure, whether flat or hierarchical, does not impact on the quality of indexing of the content objects. Google does not treat differently metadata initiatives with and without

Table III. Ranking of XML-based content objects containing DC and MARC21 in dynamic online information environments by Google

Metadata initiative (in Google)	Total number of content objects	The point of content objects placed higher
Dublin Core	50	25
MARC 21	50	25

language-based tags (DC against MARC 21 respectively). That means Google indexes both types of content objects in XML-based dynamic online information environments.

Therefore, all in all, it can be concluded that both DC and MARC 21 are suitable for organization of XML-based content objects in dynamic and static online information environments. Designers of the relevant software, therefore, could benefit from the results of the present study to improve the quality and reliability of search engines they develop.

References

Henshaw, R. and Valauskas, E.J. (2001), "Metadata as a catalyst: experiments with metadata and search engines in the internet journal, *First Monday*", *Libri*, Vol. 51 No. 2, pp. 86-101, available at: www.librijournal.org/pdf/1999-3pp125-131.pdf (accessed 5 November 2009).

Mohamed, K.A.F. (2006), "The impact of metadata in web resources discoverin", *Online Information Review*, Vol. 30 No. 2, pp. 155-67.

Quevedo-Torrero, J.U. (2004), "Improving web retrieval by mining the HTML tags for keywords and exploring the hyperlink structures of web pages", doctoral dissertation, University of Houston, Houston, TX, available at: http://lib.umi.com/dissertations/fullcit/3156028 (accessed 23 October 2009).

Safari, M. (2005), "Search engines and resource discovery on the web", *Webology*, Vol. 2 No. 2, available at: www.webology.org/2005/v2n2/a13.html (accessed 13 November 2009).

Sokvitne, L. (2000), "An evaluation of the effectiveness of current Dublin Core metadata for retrieval", available at: www.vala.org.au/vala2000/2000pdf/Sokvitne.PDF (accessed 13 November 2009).

Taheri, S.M. and Hariri, N. (2011), "A comparative survey on the indexing and ranking of the content objects including the MARCXML and Dublin Core's metadata elements by general search engines", *The Electronic Library* (forthcoming).

Turner, T.P. and Brackbill, L. (1998), "Rising to the top: evaluating the use of the HTML meta tag to improve retrieval of world wide web documents through internet search engines", *Library Resources and Technical Services*, Vol. 42 No. 4, pp. 258-71, available at: http://cat.inist.fr/?aModele=afficheN&cpsidt=1748620 (accessed 25 September 2009).

Zhang, J. and Dimitroff, A. (2004), "Internet search engines' response to metadata Dublin Core implementation", *Journal of Information Science*, Vol. 30 No. 4, pp. 310-20, available at: http://portal.acm.org/citation.cfm?id=1142111 (accessed 15 November 2009).

Zhang, J. and Dimitroff, A. (2005a), "The impact of webpage content characteristics on webpage visibility in search engines' results (Part I)", *Information Processing & Management*, Vol. 41 No. 3, pp. 665-90, available at: www.elsevier.com/locate/infoproman (accessed 15 November 2009).

Zhang, J. and Dimitroff, A. (2005b), "The impact of metadata implementation on webpage visibility in search engines' results (Part II)", *Information Processing and Management*, Vol. 41 No. 3, pp. 691-715.

Corresponding author
A. Hossein Farajpahlou can be contacted at: farajpahlou@scu.ac.ir