Distributed Metadata System and Retrieval Error Ratio

Cheng-juei Wu
Department of Library & Information Science
Fu-Jen University, Taiwan

Abstract

The Dublin Core is a resource description format, which can assist information retrieval of digital documents on the Internet. Firstly, we briefly describe the Distributed Metadata System (DIMES for short), an experimental system of the Dublin Core. DIMES is located on the author's homepage (URL: http://dimes.lins.fju.edu.tw/eng/metadata). Then, in order to test the power of metadata on information retrieval, the author designed and conducted an experiment on a group of graduate students using the Dublin Core as the cataloging metadata. The experimental results show that, on average, the retrieval error ratio (RER) is only 2.9% for the DIMES system, which utilizes the Dublin Core to describe the documents on Web, in contrast with 20.7% for the seven famous search engines. The very low RER indicates that the cataloging information of the Dublin Core is good and enough for users to make judgments of document relevance before retrieving the documents.

Keywords: Dublin Core, Retrieval Error Ratio, Metadata, Information Retrieval

I. INTRODUCTION

The Dublin Core was created in the workshop sponsored by the OCLC (Online Computer Library Center) and the NCSA (National Center for Supercomputing Applications) in 1995. As stated in the first workshop report [1], it is a simple resource description metadata, which allows each information provider to describe their works. Several principles, which become its characteristics and distinguish the Dublin Core from the other metadata, have been established since the first workshop. They are Intrinsicality, Extensibility, Optionality, Repeatability, and Modifiability.

At the present time, the Dublin Core has fifteen elements (or fields) and is in the process of standardization. According to the user guide for the Simple Dublin Core [2] (the Dublin Core without qualifiers), these fifteen elements are Title, Creator, Subject, Description, Publisher, Contributor, Date, Type, Format, Identifier, Source, Language, Relation, Coverage, and Rights. As can be easily seen, some elements come from the traditional library catalog while others are closely related to digital resources on the Web or the Internet. In addition, there are three kinds of qualifiers (Lang, Scheme, and Subelement) which can help us to refine the content of each element [3]. By properly using these three qualifiers, the Dublin Core can be a very powerful, yet flexible, cataloging tool for professional catalogers, such as librarians.

As for the implementation, the Dublin Core utilizes HTML 4.0 format at the present time [4]. For example, <META NAME = “DC.subject” SCHEME=“LCSH” LANG=“EN” CONTENT = “Computer Cataloging of Network Resources”>. However, it will use RDF (Resource Description Framework) [5] for its semantic model and XML (Extensible Markup Language) [6] for its syntactic model in the future.

Metadata, such as the Dublin Core, is used as a tool to describe the documents (or resources), and its functions are similar to the catalogs in libraries. Both metadata and the catalogs supply far more information about the resources than the indexes automatically created by computers. Therefore, the users can make a better judgment on whether or not to retrieve a document based on metadata.

In recent years, to overcome the information overloading problems of search engines, metadata (especially the Dublin Core) has drawn lots of attentions among the researchers in the information retrieval field. Many experimental systems of the Dublin Core have been built around the world and one of them is the Distributed Metadata System (DIMES, URL: http://dimes.lins.fju.edu.tw/eng/) created by the author. Thus, the next question which must be answered is how to evaluate the effectiveness of metadata systems. The author believes that the questionable measures of recall and precision are not the good candidates since they tend to fail in the large-scale information systems.

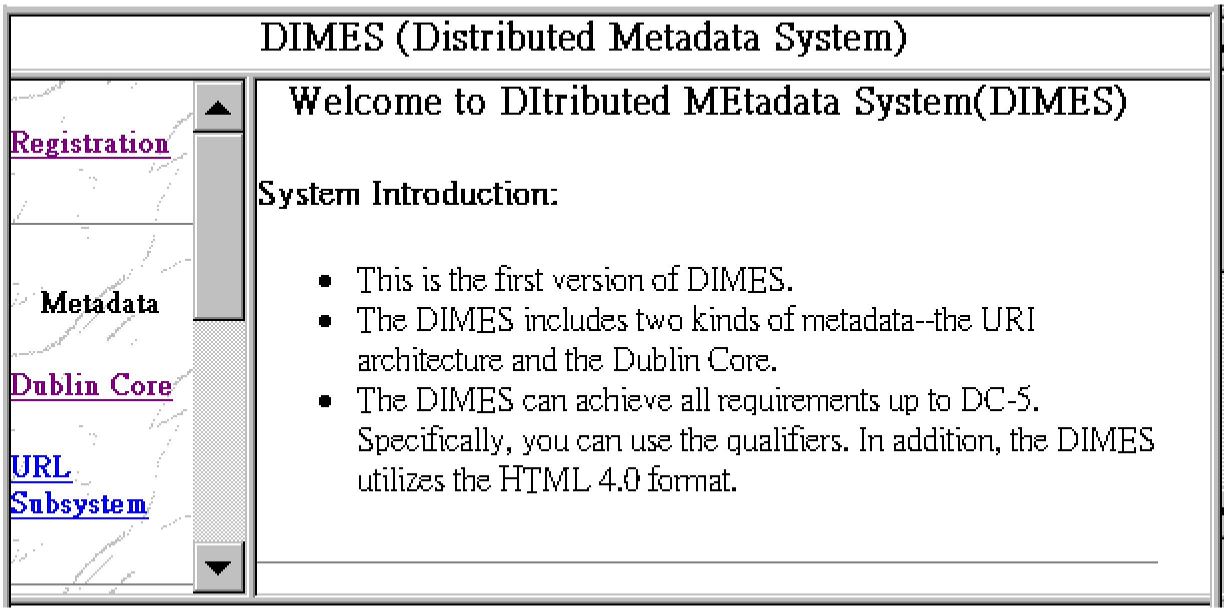
To find a proper measure to evaluate the effectiveness of metadata systems, we need to carefully examine the process of how the users interact with various kinds of retrieval systems including the library automatic systems, the commercial databases, and the search engines. Though these systems look very different, there is a similarity among them. After the users give queries, they all return some information for each hit item. Some systems, such as most of the search engines, only give very little information which typically includes the file names, the URLs, the created time, and some sentences at the beginning of the documents. Other systems, such as the library automatic systems, supply lots of information. Then the users judge the relevance of each hit item and their interests based on those information supplied by the systems. The users will make lots of wrong judgments if the cataloging information is not proper or is insufficient. In addition, they waste their energy and time since it sometimes requires money to get the documents and evaluating documents needs tremendous time. Therefore, we believe that the effectiveness of a retrieval system relies heavily on the cataloging information supplied by the system.

In this work, we introduce a new measure of retrieval effectiveness, the Retrieval Error Ratio (RER), which focuses on the cataloging information. Section II of this paper is a brief introduction of

DIMES, which is an experimental system of the Dublin Core. A brief description of the experimental process and the experimental results are given in Section III. Lastly, Section IV contains the conclusions.

II. BRIEF INTRODUCTION OF THE DUBLIN CORE

The DIMES (DIstributed Metadata System, URL: http://dimes.lins.fju.edu.tw/eng/ metadata, Fig. 1) is an experimental system of the Dublin Core, which is a simple resource description metadata created by the OCLC and the NCSA in 1995. It is an open system since it allows both querying and cataloging. At the present time, DIMES has the following three DC related sub-systems: Registration, Dublin Core, and Query. Other features are adopting URN as the identifications of resources (or documents) and attaching relational values to the element Subject to indicate how important it is to the resource.



Welcome to DItributed MEtadata System(DIMES)

System Introduction:

- This is the first version of DIMES.
- The DIMES includes two kinds of metadata--the URI architecture and the Dublin Core.
- The DIMES can achieve all requirements up to DC-5. Specifically, you can use the qualifiers. In addition, the DIMES utilizes the HTML 4.0 format.

Figure 1. The homepage of DIMES

Registration Sub-system

The homepage (URL: http://dimes.lins.fju.edu.tw/eng/metadata/ register/register_main.html) of this sub-system is Figure 2. Since DIMES allows the users to do the cataloging, it requires the users to register first so that their identifications can attached to their cataloging data. Your identification must be unique, which can be the combination of any Chinese characters, English letters, and numbers, but can not be identical to any other. It has the following functions:

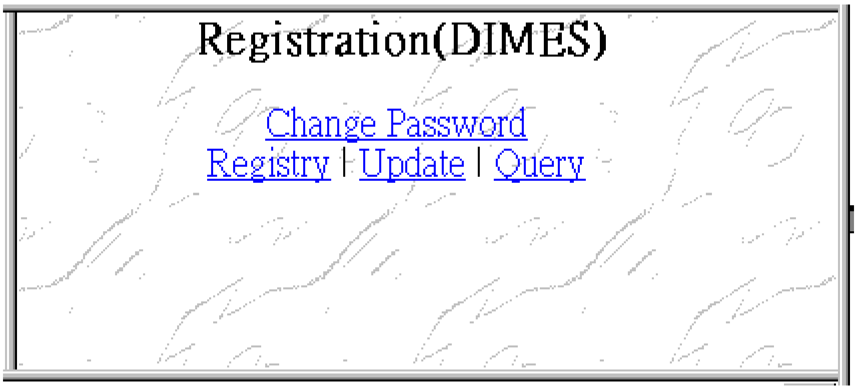

Registration(DIMES)

Figure 2. The homepage of the registration sub-system

1. Registration: The user must give identification and password. Though the other personal information is optional, it is used by the system administer to contact you. In additional, your personal information is protected by your password.
2. Personal Information Update: You can update your personal information (including name, title, address, e-mail, FAX, and telephone) one at a time.
3. Password Change: You can change your password any time you wish.
4. Personal Information Query: It can protect your personal information from peeking by requiring your password.

Dublin Core Sub-system

The homepage (URL: http://dimes.lins.fju.edu.tw/dublin/dublin_main.html) of this sub-system is Figure 3. There are fifteen elements at the present time. The system utilizes the HTML 4.0 format to present the data of the Dublin Core according to the suggestions from the fifth workshop, for example, <meta name= "DC.Subject" SCHEME=”LCSH” LANG=”EN” CONTENT=”Computer Cataloging of Network Resources”>. It has the following functions:

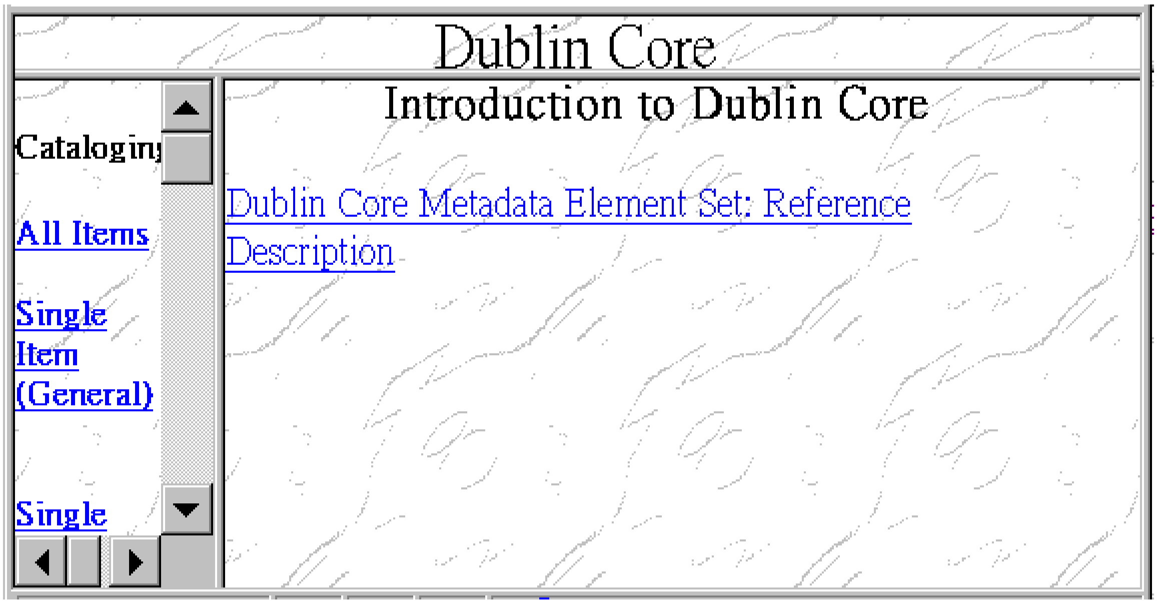
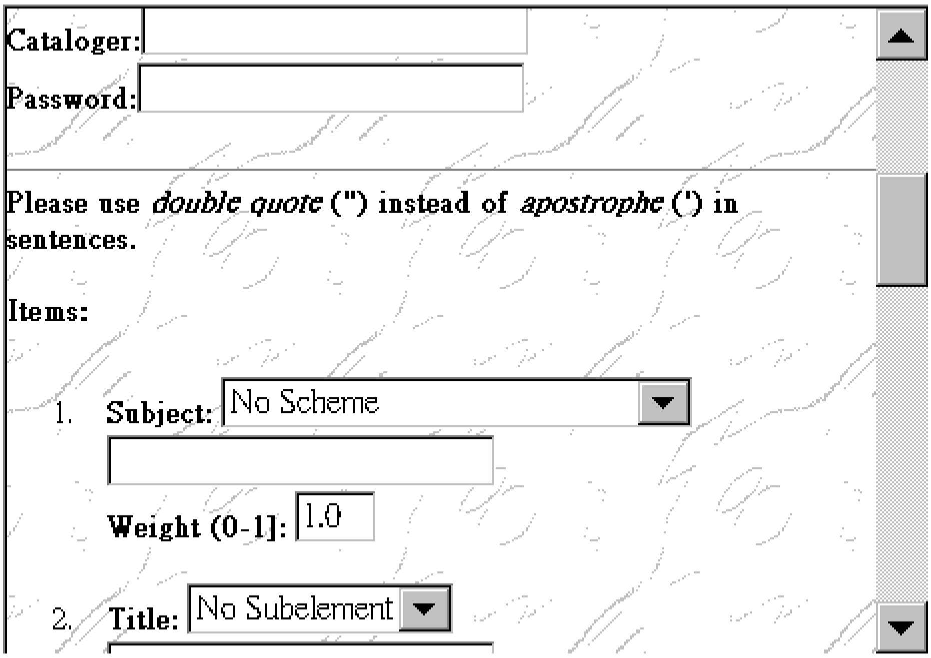

Dublin Core
Introduction to Dublin Core
Cataloging
All Items
Single Item (General)
Single
Dublin Core Metadata Element Set: Reference Description

Figure 3. The homepage of the Dublin Core sub-system

1. All items cataloging (Figure 4): All of the fifteen items are listed one by one. Please leave any unused items empty since all the Dublin Core elements are optional.



Cataloger:

Password:

Please use *double quote* (") instead of *apostrophe* (') in sentences.

Items:

1. Subject: No Scheme

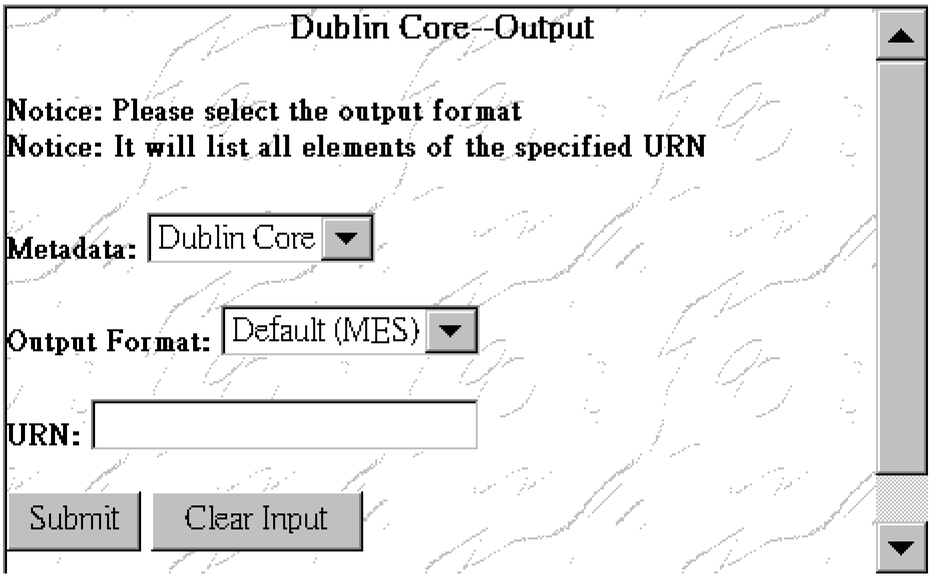
   Weight (0-1]: 1.0

2. Title: No Subelement

Figure 4. The function of all items cataloging

2. Single item cataloging (General): You can select any one of fourteen elements, except the element Subject, to handle those repeating items since all elements of the Dublin Core are repeatable.

3. Single item cataloging (Relational): This function is used when there are two or more subjects applied. The relation values should be between 0 and 1 inclusive.
4. Update: You can update your own cataloging data (but not others) since the system will check the password. Thus, to protect your data, please register before you begin the cataloging.
5. Delete: This only applies to the element Description.
6. Single element query (General): This can find all documents which have the specified keyword or string in the specific element.
7. Single element query (Relational): This allows retrieval of documents which match both of the specified string and the minimal relational value.
8. Compound-field query: You can apply the Boolean logic to at most three elements.
9. Keyword query: This allows you to find documents which have at least one element matching the specified string.
10. URN query: Please refer to the relative description in the Query sub-system below.
11. Single document retrieval (Figure 5): This shows all the cataloging information of the specified document, and you can choose either the default format or the HTML 4.0 format. If you use DIMES to catalog a Web homepage, please choose the HTML format and copy the data back to that homepage.



Dublin Core--Output

Notice: Please select the output format
Notice: It will list all elements of the specified URN

Metadata: Dublin Core

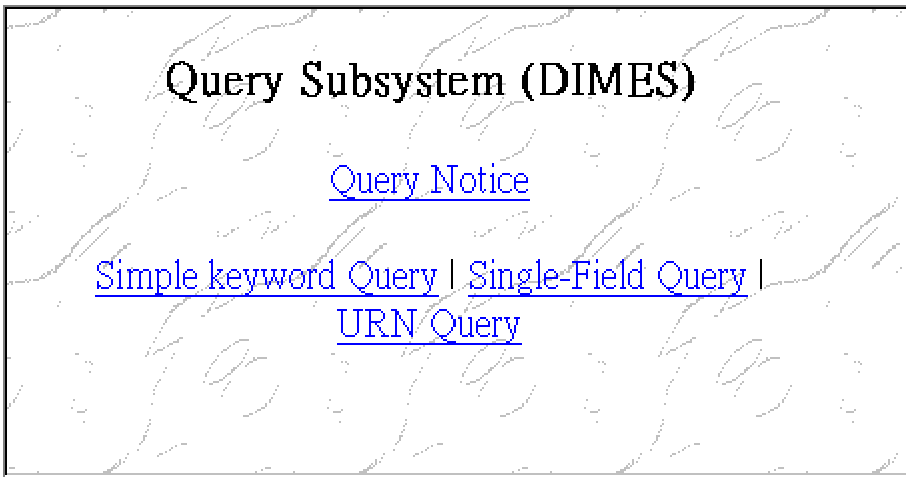
Output Format: Default (MES)

URN:

Figure 5. The function of single document retrieval

Query Sub-system

The homepage (URL: http:// dimes.lins.fju.edu.tw /DIMES/query/ search_main.html) of this sub-system is Figure 6. DIMES is designed to handle many kinds of metadata; therefore, this query sub-system is for the entire system, not just for a specific metadata such as the Dublin Core. It has the following functions:



Query Subsystem (DIMES)

Query Notice

Simple keyword Query | Single-Field Query | URN Query

Figure 6. The function of single document retrieval

1. Keyword query: This allows you to find the documents that have at least one element matching the specified string in the specified metadata.
2. Single element query: This can find all the documents which have the specified keyword or string in the specific element of the specified metadata.
3. URN query: This allows you to find all URNs which have match the specified string

III. EXPERIMENTAL RESULTS

Since the Dublin Core supplies more information than automatic generated indexes, we assume that the number of wrong judgments will be decreased. How much improvement can we expect from the Dublin Core? The following experiment was designed make a comparison between the retrieval effectiveness of the Dublin Core systems and the search engines. Seven graduate students, three females and fore males, participated in this experiment. Each of them learned the Dublin Core at the first time. Before the experiment, they received two kinds of training: one was an hour introduction of the Dublin Core, the other was two hours of demonstration of DIMES, an experimental system created by the author.

The experiment proceeded as follows:

1. Two students formed a team. One acted as a user while the other acted as a librarian.
2. The user gave a topic and the librarian selected a search engine to find the relevant documents on this topic.
3. The librarian only adopted the first twenty items returned by the search engine. In addition, he (or she) recorded the information associated with those twenty items and retrieved them via their URLs as well.

4. The librarian created the Dublin Core records for selected twenty documents using DIMES.

The evaluation was analyzed using the three steps below, and their order should be strictly followed. At each step, the librarian recorded the results.

1. The user was required to answer whether a document was relevant to his (or her) topic or not based on the information returned from the selected search engine.
2. The user did the same judgments based on the Dublin Core records created by the librarian.
3. The user read each document and judged its relevance.

In the three-step evaluation of relevance above, the last set of results should be the most accurate judgments of relevance; therefore, it can be used as the standard answers to evaluate the correctness of the other two sets of results. For convenience, we refer the first set of results based on the information of the search engines as the control group, while the second set of data based on the information of DIMES as the test group. In theory, the test group should perform better than the control group since the Dublin Core supplies more information than the automatically generated indexes, in general. In other words, the difference between the test group and the standard answers should be smaller than the one between the control group and the standard answers.

There are two types of errors when we compare the control and the test groups with the standard answers. A Type I error occurs when a document is rated relevant in the control or test groups but is re-rated as an irrelevant document after the users read it. On contrary, a Type II error occurs when a document rates irrelevant in the control or test groups but is re-rated relevant after reading the document.

Table 1. The experimental results.

		I	II	III	IV	V	VI	VII
Topic		Unicode	圖書館利用教育	Distribution Searching	Information Filter	Asia Financial Crisis	Z39.50	Metadata
Search engine		HOTBOT	GAIS	OCTOPUS	LYCOS	EXCITE	INFOSEEK	YAHOO
The number of relevant documents in the last set of results		14	12	0	2	17	14	17
The number of relevant documents in the test group		14	14	0	2	16	14	18
The number of relevant documents in the control group		13	17	8	2	9	15	19
The Test Group	Type I Error	0	2	0	0	0	0	1
	Type II Error	0	0	0	0	1	0	0
	Total Error (RER)	0	2 (10%)	0	0	1 (5%)	0	1 (5%)
The Control Group	Type I Error	2	5	8	0	0	1	2
	Type II Error	3	0	0	0	8	0	0
	Total Error (RER)	5 (25%)	5 (25%)	8 (40%)	0	8 (40%)	1 (5%)	2 (10%)

Additional explanations of the Table 1 are as follows:

1. The number of relevant documents is calculated based on twenty documents in total.
2. To avoid improper influence of previous readings, the users were required to do the judgments according to the following order: the control group -> the test group -> the documents.
3. RER is the total number of errors, which is the sum of Type I and Type II errors, divided by the number of evaluated documents of each topic, twenty documents in this case. The formula is as follows:

$$RER(\%) = \frac{TC}{N} \times 100$$

where TC is the total errors and N is the number of documents in evaluation. Unlike recall and precision, RER uses the evaluated documents, not the whole collection or the number of hit items, as the basis for calculation.

IV. CONCLUSIONS

The users will make a lot of wrong judgments if the cataloging information is not proper or is insufficient. The cost of wrong judgments can be either wasted time, missing some relevant documents, or both. Thus, the author invent a new measure, RER, to focus on this key issue. RER is a measure used to evaluate how well the cataloging information helps the users make correct judgments of document relevance before reading documents. Specifically, RER is composed of two kinds of errors, Type I and Type II errors. Type I errors refer that the users waste of time and effort to retrieve irrelevant documents while Type II errors mean that the users will miss some relevant documents. Additionally, unlike recall and precision, RER uses only the evaluated documents, not the whole collection or the number of hit items, as the basis for calculation.

In this work, an experiment to explore the retrieval effectiveness of the Dublin Core was conducted. Specifically, the aim of this experiment was to compare the retrieval effectiveness of the Dublin Core with that of the search engines. Seven graduate students participated in this experiment and the selected topics were academic-related. The experimental results showed that the Dublin Core had only 2.9% of RER while the seven famous search engines had nearly 21% of RER on average. The very low RER indicates that the cataloging information of the Dublin Core is good enough for the users to make judgments of document relevance before retrieving the documents. However, since the participants of this experiment are graduate students and their topics are academic related topics, more investigations are needed before firm conclusions can be claimed.

REFERENCES

[1] Stuart Weibel, Jean Godby, Eric Miller, and Ron Daniel, “OCLC/NCSA Metadata Workshop Report,” <http://www.oclc.org:5047/oclc/research/publications/weibel/metadata/dublin_core_report.html>, 1995.

[2] B. Rajapatirana and D. Hillman, “A USER GUIDE FOR SIMPLE DUBLIN CORE,” <http://128.253.70.110/DC5/UserGuide5.html>, 1998.

[3] S. Weibel, R. Iannella, and W. Cathro, “The 4th Dublin Core Metadata Workshop Report,” D-Lib Magazine, <http://www.dlib.org/dlib/june97/metadata/ 06weibel.html>, 1997.

[4] B. Rajapatirana, ”The 5th Dublin Core Metadata Workshop: a report and observations,” <http://www.nla.gov.au/nla/staffpaper/helsinki.html>, 1997.

[5] O. Lassila and R. R. Swick, "Resource Description Framework (RDF) Model and Syntax Specification," <http://www.w3.org/TR/WD-rdf-syntax>, 1998.

[6] D. Connolly and J. Bosak, "Extensible Markup Language (XML)," <http://www.w3.org/XML/>, 1998.