Metadata Schema Used in OCLC Sampled Web Pages

Fei Yu

Ph.D. Candidate
Department of Library and Information Science
School of Information Science, University of Pittsburgh
Pittsburgh, Pennsylvania, U.S.A.
E-mail: fey5@pitt.edu

Abstract

The tremendous growth of Web resources has made information organization and retrieval more and more difficult. As one approach to this problem, metadata schemas have been developed to characterize Web resources. However, many questions have been raised about the use of metadata schemas such as which metadata schemas have been used on the Web? How did they describe Web accessible information? What is the distribution of these metadata schemas among Web pages? Do certain schemas dominate the others? To address these issues, this study analyzed 16,383 Web pages with meta tags extracted from 200,000 OCLC sampled Web pages in 2000. It found that only 8.19% Web pages used meta tags; description tags, keyword tags, and Dublin Core tags were the only three schemas used in the Web pages. This article revealed the use of meta tags in terms of their function distribution, syntax characteristics, granularity of the Web pages, and the length distribution and word number distribution of both description and keywords tags.

Keywords: *Metadata schema; Meta tag; Description; Keyword; Granularity; Web resource; Web page; Information retrieval*

Introduction

Since the inception of the World Wide Web in the early 1990s, the vast quantity and variety of resources on the Web have grown exponentially. If measured by the number of hosts, the Web has grown faster than the Internet at large. According to the statistics from Matthew Gray of the Massachusetts Institute of Technology (2002)[1], in June 1993, the number of websites was just 130; however, after 6 months growth, it increased to 623, and in January 1997, the number of websites reached 650,000. In 1999, it was estimated that there were 288,221,000 publicly accessible Web pages on 2,229,000 public websites, which had increased more than 683 percent since 1997 (OCLC Research Office Web Characterization Project, 1999). The Internet Software Consortium (ISC) provides the "Internet Domain Survey Host Count" diagram on its website, which clearly demonstrates the dramatic increase since 1998 at 30,000,000 to 180,000,000 in 2002 and to 360,000,000 in 2005. The increasing rate is almost 20,000,000 per year.[2]

Ease of use and the availability of rich information have made the Web an attractive place for research and education. The World Wide Web is commonly referred to as a "digital library", where with the wealthy information being proliferated in electronic formats, users seem to be able to find everything they need. The user surveys conducted in several academic libraries confirmed the prevalent adoption of the Web by college students, especially for undergraduates. They are "addicted" to the Web: their first response to coursework related information is to go to the Web (Bancroft, etc., 1998; Perkins & Yuan, 2000; Friedlander, 2002).

Because the growth rate of Web resources is prodigious and beyond the ability of human intervention to catalogue them or control access to them, to a degree, Internet searches are far more difficult than searching brick-and-mortar libraries despite technical advances (Luh, J. C., 2000). Researchers find the software tools that have so far emerged to assist searching are inadequate for the task due to the vast size, dynamic and "chaotic repository" nature of the Web (Lynch, 1997; Luh, J. C., 2000). Therefore, the Web is described as "big, unstructured and a bit of mess really" (Maclennan, 1998).

Because of this "chaotic" nature, the issue of Web accessibility has been increasingly occupying the minds of researchers. Since the current barriers encountered in Web accessibility not only stem from design issues, but also from many other factors as well, for decades, generally, researchers approached this problem in four ways: First, to develop sophisticated search tools; Second, to study user online searching behavior; Third, to create and develop standards for cataloguing and classification: such as metadata schema DC, RDF; Forth, to characterize Web resources.

More and more researchers have been aware of the significance of the last approach. To provide high-quality service in conjunction with access to the Web, researchers understand that they need to have a complete understanding that requires a systematic description of Web accessible information, and this understanding of Web resources will be significantly helpful and complementary to other approaches. A number of studies have been devoted to characterizing Web resources (e.g., Lawrence & Giles, 1999; OCLC Office of Research, 1999-2000; O'Neill, et al., 1998-2000; Greenberg, 2003; Alimohammadi, 2003-2004; Craven, 2000-2004). These studies include estimation of Web size, categorization of websites, characteristics of public Web pages (e.g., distribution by subject and language), search engine coverage, and metadata usage by Web resources. However, as O'Neill (1998) points out, still, very little is known about what type of information is available and about the collection of documents that constitute the Web.

Statement of the Problem

With the tremendous growth of Web resources, people are able to enjoy the great convenience of wealthy information accessibility, but at the same time, are challenged by information organization and retrieval. As one approach to the "chaotic" Web, metadata schemas were created and have been developed to organize and characterize Web resources in terms of a systematic description of Web accessible information. However, the resource type issue and how the metadata schemas work on the Web are still not clear in several respects: What resource types are used to categorize Web-accessible information? How are Web pages distributed by resource type? Do certain resource types dominate the others? How are home pages in different subject areas distributed by resource type? How are the resource types of a home page different from the resource types of its internal Web pages? To what extent is the proposed control list of Dublin Core resource types applicable to Web accessible information? Studies are needed for these questions. A better understanding of the content on the Web will assist the development of information organization and retrieval and will tame the "chaotic" nature of the Web.

Research Purpose

With the data from OCLC office of Research's Web Characterization Projects (WCP) datasets (OCLC Office of Research, 2000), this research intends to reveal the metadata schemes adopted by Web pages in 2000. Specifically, the following topics of Web content will mainly be addressed:

- Which metadata schemas were used to describe Web-accessible information;
- Did certain schemas dominate the others;
- Who created the metadata records—by domain name;
- At what level of the Web pages were metadata normally created;
- Were Web pages classified by keywords, and how?

Methodology

WCP (Web Characterization Project) at OCLC office is an ongoing initiative to pursue research in the area of Web characterization. It has used carefully designed methodology to obtain a set of Web page datasets from the Web (O'Neill, McClain, & Lavoie, 1997). Unique public websites in WCP's sample datasets collected for the year 2000 were identified and ready for data analysis. The population was restricted to public websites because non-public websites were not intended for or were not ready for public consumption, and, therefore, would not be likely to provide information that facilitates resource discovery. This project was manually analyzed. There were a total of 16,383 Web pages with tags extracted from more than 200,000 OCLC sampled Web

pages. The extracted Web pages were downloaded to a Microsoft Excel file. According to each research purpose, multiple statistical functions provided by Excel were utilized for data analysis such as sort, calculation, and graphic display.

Definitions of Terms

1. Web page: a collection of information consisting of one or more Web resources, which is intended to be rendered simultaneously, and identified by a single URL (W3C, 1999);

2. Web resource: a resource, identified by a URL, which is a member of the Web Core (W3C, 1999);

3. Web accessible resource or Web accessible information: a Web page available and accessible from public websites;

4. Metadata: data about data; structured encoded data that describe characteristics of information—bearing entities to aid in the identification, discovery, assessment, and management of the described entities (ALA Committee on Cataloging: Description and Access: Available at http://www.ala.org/alcts/organzation/ccs/ccda/tf-meta6.html);

5. Meta tag: an extensible container for use in identifying specialized document meta-information (the HTML 2.0 specification); In general, the tag can be used to embed any information that a Web document author feels is relevant for describing the document (O'Neill, T. E., et. al, 2000);

6. Resource type: category or genre of the content of the resource.

Literature Review

There have been many recent published works and activities in the field of metadata. This review focuses on the types of metadata schemas prominently used on the Web such as which these schemas are, how they are implemented (e.g., their semantics, syntaxes), and what are the problems behind the implementations of these schemas in promoting Web resource description and discovery.

There are literally hundreds of metadata schemas, and the number is growing rapidly, as different communities seek to meet the specific needs of their members. However, the review shows there are generally just three infirmly entrenched standards pertinent in the Web context: the "keyword" and "description" tags, as implemented by the search engines; the prominent general standard "Dublin Core Metadata Initiative"; the Resource Description Framework (Gill, 2002). Among these three standards, tags are the most prevalent on the Web, and along with Dublin Core, they bring up many new standards that are predecessors to breakthrough technologies like RDF (Resource Description Framework), which provides a framework for describing various types of metadata (Stanek,1999). There have also been a number of large-scale deployments of

Dublin Core metadata around the globe. On the official Dublin Core website (http://dublincore.org/projects/), DCMI (Dublin Core Metadata Initiative) projects page lists about 20 projects in North American and Mexico, 38 in Europe, and 12 across Asian countries and Australia. These initiative projects have been conducted for different document preservation purposes and in various scales (e.g., cross countries vs. national scale vs. one library scale).

Meta Tags

HTML permits document authors to control not only how text, graphics, and multimedia materials are displayed, but also the information available about the document itself through the use of meta tags. Weibel (1996) claims that meta tags are the best section of the HTML Specification in which this data can be placed. As the No.1 tip provided for Web page quality enhancement, Maddux (1998) suggests to always use the meta tag because it enables a Web author to exert some control over how some of the major search engines categorize and describe a page. Quite a few researchers in library and information science, computer or other fields discussed meta tags in terms of definition, structure, function, and classification, etc. (e.g., Clark, 2002; Duval, 1996; Futterman, 2001; Sullivan, 2002; and Gill, 2002).

What are Meta Tags?

"Meta" in "meta tags" arises from the term "metadata", which means data about other data. Meta tags may be used to describe the content of a Web page. They are non-displaying or hidden HTML tags that may provide site owners and authors with a degree of control over how a Web page is indexed (Christensen, 1999; Henshaw, 1999; Henshaw & Valauskas, 2001). Any Web page can have a variety of metadata associated with it (Sullivan, 1997).

When examining meta tags, generally, researchers study both their features and functions in the context of HTML Web environment. Looking at features, Ramiscal (2000) states that meta tags can only be understood within the context of HTML because HTML has provided the impetus for the creation of tags to make the task of information gathering easy and quick; through these "tags", the way a document is presented to the viewers is determined by a Web browser intentionally instructed. According to the HTML Library, meta tags are used to embed any useful information not defined by other HTML elements. The nature of these tags lies in the fact that they are used to exert control over the entire document rather than to format or otherwise modify content. Another general feature of meta tags is that they are hidden from any mortal eye: they are machine readable and cannot be seen by the user unless the user views the document's source. Hanks (1998) praises meta tags as a triumph of post-modernism simplicity for the keywords operate as summaries of documents contained in the websites that the search engine has searched. Dillon (2001) ranks meta tags as

the most common form of metadata characterizing the content of fields with a great variety of uses. He lists the advantages of simplicity, machine and human readability, and great expressive power, as HTML has demonstrated in the Web environment.

Looking at functions, researchers identified over 40 different kinds of meta tags currently used for different purposes: to assist search engines better indexing a document, to determine relevance, to refresh a Web page or to redirect a user from one Web page to another, to identify properties of a document (e.g., author, expiration date, keywords, etc.) as well as assign values to those properties, and to allow clients to label and rate online content according to standards developed by the World Wide Web Consortium, etc. (Clark, 2000; Futterman, 2001; Beeline, 2002; Sullivan, 2002). Meta tags might help authors and publishers ensure that their materials are found when appropriate searches are executed (Turner & Brackbill, 1998). Both Sullivan (2000) and Bradley (2002) suggest that meta tags be used as a magnifying glass to help the search engines focus on the most important parts of a page, and that they should be put on every page that is created and published so that search engines will find all of the pages.

Classification of Meta Tags

Meta tags are generally known as two kinds: HTTP-EQUIV tags and meta tags with a NAME attribute (Stanek 1999, Ramiscal 2000, Clark 2000).

HTTP-EQUIV Tags

Meta HTTP-EQUIV tags are the equivalent of HTTP (Hypertext Transfer Protocol) headers that usually control or direct the actions of Web browsers. HTTP-EQUIV tags are designed to affect the Web browser in the same manner as normal (HTTP) headers: when you click on a link for a page, the Web server receives your browser's request via HTTP; once the Web server has made sure that the page you've requested is indeed there, it generates an HTTP response. The initial data in that response is called the "HTTP header block." The header tells the Web browser information that may be useful for displaying this particular document. In a similar fashion, Meta HTTP-EQUIV tags direct the actions of Web browsers, but they are also able to further refine the information provided by the actual (HTTP) headers (Stanek,1999; Ramiscal, 2000).

Normally, HTTP headers are set automatically by Web servers based on responses to requests for resources. However, people can modify existing headers or create their own by using the HTTP-EQUIV tag so that browsers and server behavior can be customized. For example, say the default content type and character set for the server are text/html and ISO-8859-1 (Western, Latin-1). The server sets this information in a HTTP header as follows:

Content-Type: text/html;
CHARSET=ISO-8859-1

With a meta tag, you can override the default content type by setting the default character set to ISO-8859-5 (Cyrillic) as follows:

<META HTTP - EQUIV="Content-Type"
CONTENT="text/html; CHARSET=ISO8859 - 5">

Now instead of seeing the Latin-1 character set, readers see the Cyrillic character set, which works for the Russian translation of readers' favorite play. Content- Type is only one of many similar Meta values. Here are a few more:

HTTP-EQUIV="Content-Disposition" specifies an application handler for the file.
HTTP- EQUIV="Content-Scri-Type" sets the default scripting language.
HTTP-EQUIV="Content-Style-Type" sets the default stylesheet language.
HTTP-EQUIV="Content-Language" declares the natural language for the page.

Meta Tags with a Name Attribute

Meta tags with a Name attribute are used for Meta types that do not correspond to normal HTTP headers. These meta tag sources designate supplemental information that doesn't have a related HTTP header. In each of these tags, NAME identifies the value and CONTENT sets the actual value. The following example sets the author's name,

<META NAME:"Author"
CONTENT="William R. Stanek">

Here are more Meta values that use the NAME attribute to provide further information:

NAME="Copyright" sets the copyright information.
NAME="Generator" sets the authoring tool that created the page.
NAME="Reply-To" sets the contact e-mail address.
NAME="description" sets the specified text.
NAME="keywords" sets the index words for search engines.

For either HTTP-EQUIV tags or meta tags with a Name attribute, the Meta information is always added to the page header inside the <head> and </head> tags, like this:

<html>
<head>
<META HTTP-EQUIV:"Content-Type"
CONTENT="text/html; CHARSET=windows-1252">
<META HTTP-EQUIV="Content-Language"
CONTENT="en-us">
<META NAME="Generator"

```
CONTENT="FrontPage 4.0">
 <title>My Web Page</title>
 </head>
 <body>
...
 </body>
 </html>
```

Because the content of a meta tag directly corresponds to their diverse functions, some researchers categorize meta tags by their functions. For example, Sullivan (2002) classified meta tags that users can see into four types: Meta Robots, Meta Description, Meta Keywords, and Meta everything Else.

Description Tags and Keyword Tags

There are several types of meta tags in HTML: Coopee (2000) estimates that there are over 50, and Lawrence & Giles (1999) count as many as 123, but the meta tags of most interest to library science are those that provide a unique representation of the information source and enable users to locate information (Nowick, 2002). The most important meta tags that are used by search engines for indexing purposes are description tags and keywords tags. For these two types, Rasmical (2000) argues the importance of their placement in every framed page cannot be overemphasized since "they are the trigger mechanisms for 'hits' on a particular website." The keywords and description tags are widely used in conducting information searches via search engines like Hotbot, Infoseek, Altavista, and Yahoo!; additionally, they are the most prevalent meta tags in practical Web design and academic research (Alimohammadi, 2003, 2004; Craven, 2001a, 2001b, 2004).

The description tags are used to describe Web pages, and the "description" attribute allows any text description related to the Web pages. The search engines that recognize this tag will display the text specified here, rather than the first few lines of the text from the actual document when the document shows up in a search result. It is particularly useful for documents with a small amount of text. A number of keywords should be included in the description, and the most important ones should be near the beginning of the description (Craven, 2001b). The following is an example of description tags from the source code of University of Pittsburgh Homepage (2005):

<meta name="description" content="The University of Pittsburgh is among the nation's most distinguished comprehensive universities, with a wide variety of high-quality programs in both the arts and sciences and professional fields." />

It is recommended that the length of description be limited to 20-25 words, or about 150-200 characters of text (Christensen, 1999; Henshaw, 1999). Description could be a word, a sentence, or even a paragraph. As a general rule, it should be reasonably short, concise and to the point; however, it should not be so compressed that it cannot

be an appropriate reflection of the contents (Bradley, 2002).

Keywords help search engines to easily index Web pages and to allow people to find Web pages more quickly (Kyrnin, 2002). Without this tag, search engines will choose words from the title and text of the site (Henshaw, 1999). As for which words should be used as keywords, researchers have their own opinions: for example, only those keywords should be used that the customer might type in a search engine to try to find the website (Sliwa, 1998); acronyms (Richmond, 2002); synonyms or Americanisms (Bradley, 2002); related words or word combinations should be used (Guenther, 1999); furthermore, common misspellings should be included in the keywords tag (Spider Food, 2002); the best advice is selecting keywords based on a thesaurus (Richmond, 2002). As for the number of keywords used, Spider Food (2002) suggests keywords be fewer than 1,000 characters including spaces and commas because most search engines have limits on how many keywords are viewed. It is a good idea to review the keywords and make sure that they are as concise and specific as possible (Kyrnin, 2002). The following is an example of a keyword tag that is also from the source code of University of Pittsburgh Homepage (2005):

<meta name="Keywords" content="University, Pittsburgh, Pitt, College, Learning, Research, Students, Undergraduate, Graduate" />

Meta Tag Use on the Web

The use of meta tags, though it has been improving, is still unsatisfactory. According to O'Neill & MaClain's survey (1998), although nearly three-quarters of sampled sites had at least one installation of a meta tag, only about 45 percent of internal Web pages contained metadata, much of which can be attributed to automatic generation of meta tags by HTML editors. Lawrence and Giles' survey (1999) showed a similar situation, where among 2,500 randomly sampled Web servers, only about 34.2% of servers contained meta tags, which may not be useful for the purpose of describing and discovering Web accessible resources. In Craven's (2000; 2001a; 2001b) successive reports of how and to what extent people and organizations make use of meta tags, his 2000 survey showed, of 628 Web pages registered with Yahoo, 357 (56.8%) contained meta tags and 163 (25.9%) used description tags; the same year, another survey he conducted showed, of 1,937 Web pages surveyed, 35.5% of the Web pages used description tags; in 2001, of 460 Web pages surveyed, 95.1% pages used description tags; however, in another 2001 study, he found that among 1,947 retrieved Web pages, 1,198 pages (61.5%) used meta tags, of which 592 (30.4%) were description tags. Drott (2002) conducted studies in the same vein by examining 60 corporate websites in both 2000 and 2001. He found that the quantities of meta tag use were equal for keywords and description in 2000—35%; however, in 2001, the use of keyword tags

increased to 41.6%, and the use of description tags dropped to 31.6%.

Dublin Core

The Dublin Core Metadata Element Set (DC) is a set of 15 information elements that can be used to describe a wide variety of information resources on the Internet for the purpose of simple cross-disciplinary resource discovery. DC metadata is embedded within Web pages as a static descriptive record complete in itself. The 15 elements are Contributor, Coverage, Creator, Date, Description, Formal, Identifier, Language, Publisher, Relation, Rights, Source, Subject, Title, and Type. The HTML DC tag elements are:

<META NAME = "DC.Type" CONTENT = "Text">
<META NAME = "DC.Type" CONTENT = "Text.Serial">
<META NAME = "DC.Type" CONTENT = "Text.Serial.Journal>

Dublin Core serves as an interchange format between various systems using different metadata standards; it is used for harvesting metadata from data sources within and outside of library domains; it supports the simple creation of library catalog records for a resource within a variety of systems; it also exposes MARC data to other communities and allows for acquisition of resource discovery metadata from non-library creators who are using it.

Despite the significant progress over the last few years in raising awareness and increasing deployment of Dublin Core, it is rarely adopted on the Web. For example, both O'Neill (1998) and Lawrence & Giles (1999) found that only 0.3% of the websites contained DC metadata in their studies. One factor for this low adoption is lack of search engine support. The Search Engine Report (1997) showed that none of the major search engines did anything with them: they did not index them, nor did they provide a way to search within the Dublin Core tag fields. Another factor that has hindered the widespread adoption of DC metadata is the length of time it has taken to reach consensus on an approved interoperability qualifier. Gill (2002) points out, "the intellectual difficulty in reaching agreement on qualifiers is partly the result of well-intentioned attempts to apply Dublin Core far more broadly than what it was originally designed for — simple discovery of 'document-like objects' on the World Wide Web."

It is obvious that description tags, keyword tags, and Dublin Core tags have been underused on the Web. Stanek (1999) states in his article that, as one cornerstone of Web technology, "tags have been neglected lately." Nevertheless, though only a meager 20% to 30% of Web pages use them, this does not mean meta tags are outdated or some new technology is better (Stanek, 1999). Henshaw & Valauskas's (2001) experiment clearly showed that Dublin Core Metadata Initiative and HTML meta tags have improved the rank of selected articles among retrieved results. Therefore, meta tags

deserve more attention in the further study of Web resources.

Research Results

As stated, 16,383 Web pages with meta tags were extracted from 200,000 OCLC sampled Web pages. Only 8.19% of Web pages were found to use meta tags, and this percentage is much lower than the previous studies mentioned in the literature review. Among these Web pages, description tags, keyword tags, and Dublin Core tags were identified as the only three schemas used; 4,469 Web pages contained description tags by a percentage of 27.3, which is surprisingly similar to Craven's result of 25.9% (Craven, 2000); 4,141 Web pages contained keyword tags by a percentage of 25.3.

There were just 2 tags containing Dublin Core elements, which confirmed the current underused situation. The Dublin Core metadata elements were identified as follows:

Attribute	Frequency in Sample
DC.Title	2
DC.CREATOR	2
DC.SUBJECT	2

Among the 16,383 extracted Web pages with meta tags, there were 8,786 HTTP-EQUIV tags and 12,896 meta tags with a Name attribute.

HTTP-EQUIV Tags

- Function Distribution of HTTP-EQUIV Tags

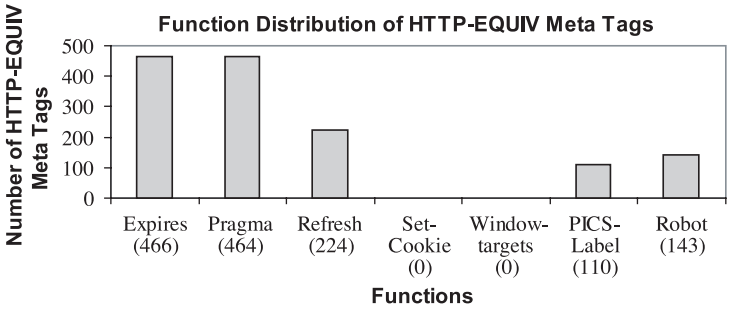

Figure 1 Function Distribution of the HTTP-EQUIV Meta Tags

- Syntax of HTTP-EQUIV Tags

  Generally, there were two syntax types among these HTTP-EQUIV tags:

  (1)HTTP-EQUIV is at the beginning of meta tags. There were three formats in this type in this research:

  ○ *HTTP-EQUIV=Content-Type CONTENT=text/html; harset=iso-8859-1*
  ○ *HTTP-EQUIV=Content-Type CONTENT=text/html; charset=iso-8859-1:*

NAME=Hidden CONTENT=Thanks for peeking at my source code - use it for examples, but don't copy it! Ashley Williams aawwconsulting@hotmail.com

- *HTTP-EQUIV=expires CONTENT=0:HTTP-EQUIV=Pragma CONTENT=no-cache:HTTP-EQUIV=Cache-Control CONTENT=no-cache*

The first format was the most common syntax among 8,786 HTTP-EQUIV meta tags. Consistent with the literature review, *Content- Type* is one of the most frequently used Meta values; the second format had a NAME attribute tag that followed a colon mark. The number of tags of the second format was considerable among the16,838 meta tags. The third formate was a typical HTTP-EQUIV function syntax except that two HTTP-EQUIVs were included in one tag and separated by a colon mark.

(2)HTTP-EQUIV is not at the beginning of meta tags. There were two formats in this research:

- *content=text/html; charset=windows-1252 http-equiv=Content-Type:content=MSHTML 5.00.2614.3500 name=GENERATOR: content=FrontPage.Editor.Document name=ProgId*
- *NAME=GENERATOR CONTENT=Adobe PageMill 3.0 Mac:HTTP-EQUIV=Content-Type CONTENT=text/html; charset=iso-8859-1: NAME=Generator CONTENT=Microsoft Word 98*

The first format was an irregular meta tag syntax that was not led by a HTTP-EQUIV or a NAME attribute tag. The second format was a tag with a Name attribute, in which a HTTP-EQUIV tag and a second Name attribute tag were included.

Meta Tags with a Name Attribute

- Distribution of Meta Tags with a Name Attribute

Of the 16,383 Web pages with meta tags, there were 4,469 description tags, or 27.28%; there were 4,141 keyword tags, or 25.28%.

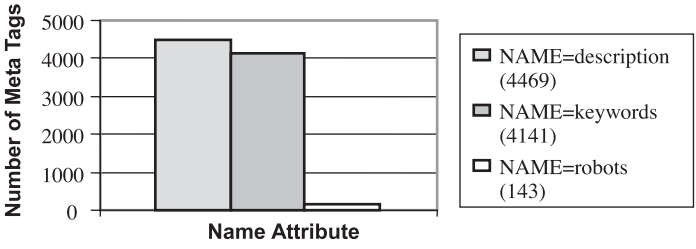

Figure 2 Distribution of Meta Tags with a Name Attribute

- Distribution of Meta Values

The Meta values found in the meta tags were:

NAME=GENERATOR; NAME=AUTHOR; NAME=CLASSIFICATION; NAME=DISTRIBUTION; NAME=TEMPLATE; NAME=FORMATTER.

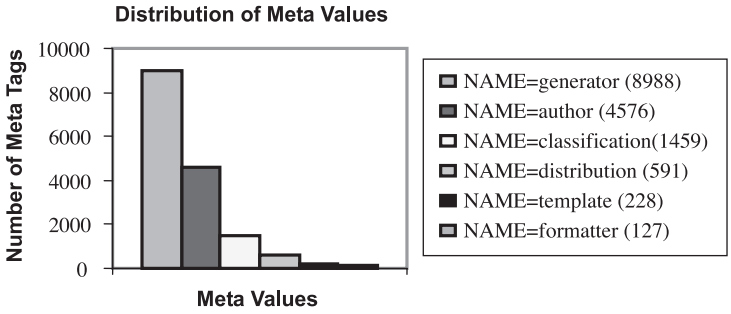

Figure 3 Distribution of Meta Values

- Syntax of Meta Tags with a Name Attribute

Similar to the syntax of HTTP-EQUIV tags, meta tags with a Name attribute had two syntax types in this research:

(1)A Name attribute is at the beginning of meta tags. Among these tags, there were three formats:

One meta tag had only one Name attribute:

○ *NAME=Author CONTENT=A.KalinichevOther tags*

One meta tag had more than one Name attribute:

○ *NAME=GENERATOR CONTENT=Mozilla/3.0Gold (X11; I; SunOS 5.5 sun4m) [Netscape]:NAME=AUTHOR CONTENT=ANP*

One meta tag had more than one Name attribute and a HTTP-EQUIV tag:

○ *NAME=GENERATOR CONTENT=Adobe PageMill 3.0 Mac:HTTP-EQUIV=Content-Type CONTENT=text/html; charset=iso-8859-1: NAME=Generator CONTENT=Microsoft Word 98*

The first and the second format comprised the largest percentage of meta tags with Name attributes; just a few meta tags were found in the third format.

(2)A Name attribute is not at the beginning of meta tags. There were two formats in this research:

HTTP-EQUIV leading meta tags include at lease one Name attribute.

○ *HTTP-EQUIV=Content-Type CONTENT=text/html; charset=iso-8859-1: NAME=Author CONTENT=qusheng Jin:NAME=GENERATOR CONTENT =Mozilla/4.04 [en] (WinNT; I) [Netscape]*

Irregular meta tag syntax that is not led by HTTP-EQUIV or a NAME attribute

○ *content=MSHTML 5.00.2314.1000 name=GENERATOR content=Microsoft FrontPage 4.0 name=GENERATOR:content=text/html; charset=iso-8859-1 http-equiv=Content-Type*

- "Irregular" Meta Tags

Besides HTTP-EQUIV tags and the meta tags with a Name attribute, there were 216 other "irregular" tags that were not led by either <Meta HTTP-EQUIV… > or <Meta Name=…> at the beginning. All these 216 irregular tags started with

"content=…" Among them, there were 168 tags in the following format:

<Meta content=text/html; charset=windows-1252 http-equiv=Content-Type: content=MSHTML 5.00.2614.3500 name=GENERATOR:content=FrontPage.Editor. Document name=ProgId>

Above all, the results showed that the number of meta tags with Name attributes were much more than the number of HTTP-EQUIV tags; description and keyword were still the major meta tag functions while the functions of Expire, pragma, refresh, and PICS-Label were used to a limited degree. The result indicated no use of "set-cookie" and "robots" functions among all the sampled Web pages. As for the syntax, in both HTTP-EQUIV and tags with Name attributes, the syntax was not rigid as those mentioned in the literature review; instead, the syntax revealed in this study was in various formats within each type, and there were also some irregular meta tags.

Granularity: At What Level These Web Pages Were Created

There were a total of 305 domains within 16,383 Web pages. The domain distribution on levels 1-9 could be seen in this sample web page as follows:

WCP00_10/209.61.69.117/position.cgi_positionname_I/T_Specialist/graphics/

Domain first level second third fourth

general/department/group/team/employ.html

fifth sixth seventh eighth ninth

(Each level is defined by a slash after the domain.)

The dataset showed that, there were 241 (79%) domains describing the first level Web pages with meta tags; 151 (49.5%) domains describing the second level; 96 (31.5%) domains describing the third level; 60 (19.7%) domains describing the fourth level; 25 (8.2%) domains describing the fifth level; 17 (5.6%) domains describing the sixth level; 6 (1.9%) domains describing the seventh level; 3(0.98%) domains describing the eighth level; and just one domain describing the ninth level.

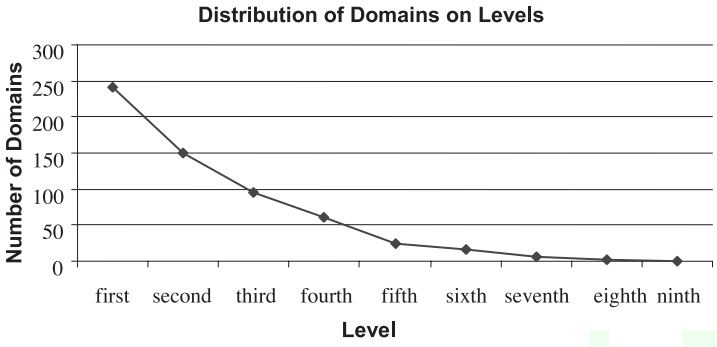

Figure 4 Distribution of Domains on Granularity

The distribution of domains on various levels was extremely negatively skewed, which suggests that most domains just described the Web pages on the first level, and the number of domains decrease as the granularity of Web page description increases.

Another way to measure the granularity of each Web page is to analyze how many attributes are included in the meta tags of each Web page. A rough method is to calculate how many equal "=" marks are contained in the meta tags of each Web page because each attribute is defined after an "=" mark. However, the final results should delete the duplicate equal marks, then divide the number by two. For example:

<Meta name=description content=BrainBug is a provider of digital media communications for business.:name=keywords content=…>

There are two Name attributes in this tag: one is description and the other is keyword. The total number of equal marks is four, but the equal marks after "content" should not be counted because their function is just to show the Meta value. Therefore, for the total number of attributes describing one Web page, the number of equal marks should be divided by 2: 4 / 2 = 2.

The attribute distribution of meta tags per Web page showed the mode was 2, and the distribution was positively skewed. The more attributes each Web page contained, the less the frequency of Web page numbers. Most meta tags preferred 2 to 5 attributes. Therefore, the granularity of each Web page was less than 5.

Attribute Range	Frequency	Cumulative %
2	3732	22.78
3	3007	41.13
5	2291	55.12
7	2110	68.00
10	1342	76.19
4	1232	83.71
6	962	89.58
8	751	94.16
9	553	97.54
More	387	99.90
1	12	99.98
0	4	100.00

Figure 5 Attribute Distribution of Meta Tags per Web Page

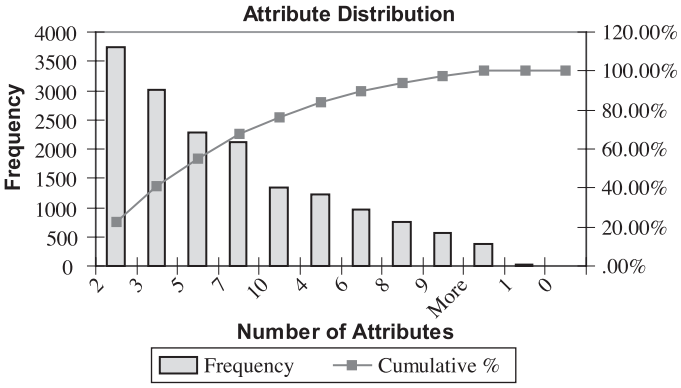

Figure 6 Histogram of Attribute Distribution of Meta Tags per Web page

Description Tags

Description tags were examined in two ways: the length of description tag characters and the number of description words.

• Length of Description Tag Characters

According to the statistical calculation, the mode of the character length distribution was 40, the mean was 119, and the median was 120. A total of 3,768 (85.90%) Web pages contained less than 150 characters in length, and 3,940 (89.8%) contained less than 200. The longest length of the description tag was 229 characters. Compare to Craven's 55% (2000), 85.90% description tags less than 150 characters in this study are much higher. Similarly, 89.8% description tags no more than 200 characters in length are also higher than Craven's 76.7% (2000). However, for the longest description tag, 229 characters in this study are much smaller than 1,789 characters in Craven's (2000) result.

Column1	
Mean	119.1538462
Standard Error	21.22403275
Median	120
Mode	40
Standard Deviation	76.52433834
Sample Variance	5855.974359
Kurtosis	-1.271074691
Skewness	-0.055949588
Range	229
Minimum	0
Maximum	229
Sum	1549
Count	13
Confidence Level(95.0%)	46.24319406

Figure 7 Table of Length of Description Tag Character Distribution

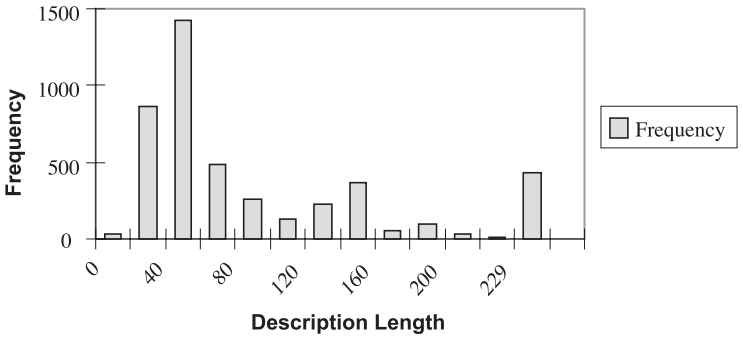

Figure 8 Histogram of Length of Description Tag Character Distribution

- Number of Description Words

The distribution showed that the mode was 5, the mean was 19, and the median was 20. This distribution range was from 0 to 37. Compared to Craven's study (2000), the mean in this study is slightly smaller (19 vs. 25); however, the median is exactly the same (20 vs. 20); the longest description in one Web page is much shorter than Craven's (2000) result (37 vs. 294). This study indicates the description for these 16,383 extracted Web pages are generally fewer and shorter than the results of the previous studies.

Column1	
Mean	19.6666667
Standard Error	4.39064662
Median	20
Mode	5
Standard Deviation	13.1719399
Sample Variance	173.5
Kurtosis	-1.34287304
Skewness	-0.1194575
Range	37
Minimum	0
Maximum	37
Sum	177
Count	9
Confidence Level(95.0%)	10.1248558

Figure 9 Table of Length of Description Word Distribution

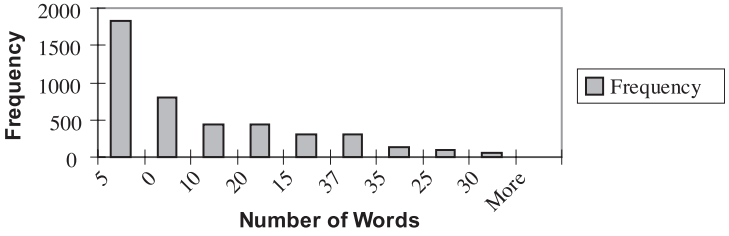

Figure 10 Histogram of Length of Description Word Distribution

Keyword Tags

Keyword tags were also examined in two ways: the length of keyword tag characters and the number of keywords.

- Length of Keyword Tag Characters

The length of character distribution showed that, the mode was 180, the mean was 80, and the median was 80. This distribution range was from 0 to 160, and the standard deviation was 54.77. There were 52 (37.68% of the total Web pages) keyword tags of no more than 150 characters in length, and 124 (89.85%) of no more than 200 characters. The longest keyword tag had 160 characters. The percentage of keyword tags of no more than 200 characters in length was almost the same as the percentage of description tags, which was 89.85% (keyword tags) vs. 89.8% (description tags).

Column1	
Mean	80
Standard Error	18.25741858
Median	80
Mode	180
Standard Deviation	54.77225575
Sample Variance	3000
Kurtosis	-1.2
Skewness	0
Range	160
Minimum	0
Maximum	160
Sum	720
Count	9
Confidence Level(95.0%)	42.10170998

Figure 11 Table of Length Distribution of Keyword Tags

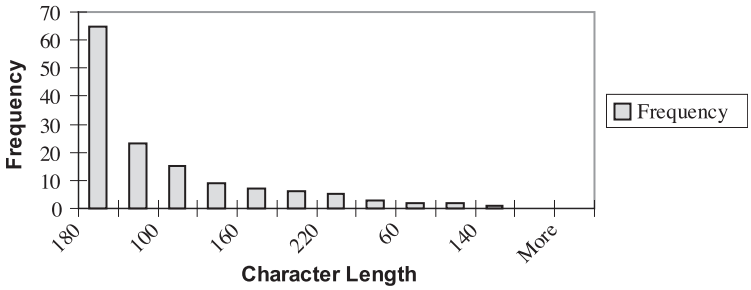

Figure 12 Histogram of Length Distribution of Keyword Tags

- Number of Keywords

The number of keyword distribution showed that the mode was 28, the mean was 20, and the median was 20. This distribution range was from 4 to 36. There were 18.84% keyword tags with the keyword number less than 20; 80.4% keywords tags with the keyword numbers more than 22; 78.2% keywords tags concentrated on the keyword number from 16 to 28.

Column1	
Mean	20
Standard Error	3.651483717
Median	20
Mode	28
Standard Deviation	10.95445115
Sample Variance	120
Kurtosis	-1.2
Skewness	0
Range	32
Minimum	4
Maximum	36
Sum	180
Count	9
Confidence Level(95.0%)	8.420341996

Figure 13 Table of Word Number Distribution of Keyword Tags

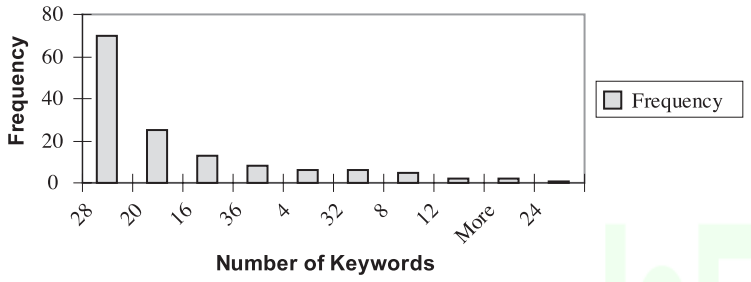

Figure 14 Histogram of Word Number Distribution of Keyword Tags

Discussion

1. It is obvious that many current metadata is attributed to the automatic generation of tags by HTML editors. Although it is not clear that metadata of this kind is particularly useful to facilitate resource discovery and description (O'Neill et al, 2002), it is an area that needs further investigation.

2. This study shows the percentage of description tags and keyword tags on the sampled Web pages were just 27.28% and 25.28%, which sounds low, they were actually the highest counts of any tags in this research. These two attributes are widely favored in both practical Web resource description (discovery) and academic research conducted on various scales because they enjoy full support from search engines. This conclusion is consistent with the results from the previous studies (Search Engine Report, 1997; Henshaw, 1999; Craven, 2000; 2001a; 2001b; 2001c; 2001d; Drott, 2002; Alimohammadi, 2004).

3. To compare the tag use on the Web, the relevant previous studies and results are listed as follows:

Research	Description tag Proportion	Keyword tag Proportion
Search engine report (1997)	12%	11%
Henshaw (1999)	27%	30%
Craven (2000)	25.9%	
Craven (2001a)	35.5-36.5%	
Craven (2001c)	30.4%	
Craven (2001d)	33.3%	
Drott (2002)	35% (2000); 31.6%(2002)	35% (2000); 41.6%(2002)

Generally, it is observed that in each passing year, the use of description and keyword tags increased, and the keyword tags were used more than the description tags. This research perfectly confirmed the first part of this observation; however, it was a departure of the latter part by showing that the description tags were used more than the keyword tags in the year 2000, which was 27.28% (description) vs. 25.28% (keyword). Nevertheless, the general tendency revealed from these successive studies is to increase the use of keyword tags. This is affected by the keyword approach to information storage and retrieval on the Web: since Internet users tend to do their searches by typing keywords and phrases, Web designers tend to use the keyword tag more than description tags (Alimohammadi, 2004).

4. The vast majority of description tags (89.8%) in this study was consistent with the maximum length guideline of 200 characters provided by Infoseek (1999), and 85.9% of description tags in this study were consistent with HotBot's (1998) more restrictive guideline of 150 characters. In contrast to Craven's conclusion (2000), no results from this project show that "there is a need on the part of some authors for a

way of including much longer descriptive information." In this project, 229 was the maximum character length for description tags.

5. The percentage of keyword characters that were less than 200 in length per Web page is similar to that in the description tags (89.85% vs. 89.8%) in this study. Since Craven (2000) observed the same similarity in his own studies, he proposed an assumption of the relationship between keyword density and description length. However, no research has yet proven that authors who create longer descriptions also tend to create more comprehensive lists of keywords. This could be an interesting topic for further research.

6. This study arrives at the same conclusion on the use of Dublin Core as Lawrence & Giles (1999) and O'Neill & MaClain (1998): DC is not well used and has been ignored by the search engines and practical information retrieval. However, the underused situation cannot undermine the significance of DC. It is necessary for researchers to constantly observe its use and development in Web resource description and rediscovery.

7. As for the general granularity, most domains (about 79%) only described their Web pages on the first level; the more granular they were, the fewer domain numbers they had. For example, on the first and second level, there were 79% and 49.5% domains; while on the ninth level granularity, there was only one domain. The domain distribution on the granularity was extremely negatively skewed. As for the granularity in tags, most of them used two to five attributes to describe one Web page; however, after the deduction of duplicate attributes and automatically generated ones, the number of real attributes used was less than three. The granularity study implies that the Web resource description is coarse.

Because time, assistance, and method were limited, this project is only a small initiative of identifying Web accessible information. More research is needed to address the issues of Web resource identity. For example, to classify Web pages by domain was not conducted as planned because the extracted tags did not provide usable domain information. However, domain study is important such as the relationship between tags and the subject/geographical domain of websites. In addition, it is necessary to further analyze the content of description and keyword tags (e.g., density of keywords). Craven's successive (2000, 2001a, b, c, d, 2004) investigations into description and keyword tags provided good reference for further study in terms of research fields, techniques, and data analysis methods, etc.

Notes

1.Matthew Gray is a former MIT graduate student, one of three members of the Student Information Processing Board (SIPB) who set up **www.mit.edu** in the Spring of 1993. Currently,

he is the Chief Software Architect for Newbury Networks in Boston. “Growth of Web Report” is published and updated on his website: http://www.mit.edu/people/mkgray/net

2.“Internet Domain Survey Host Count” is from the source of Internet Software Consortium: www.isc.org.

Reference

Alimohammadi, D. (2003). Meta-tag: a means to control the process of Web index. *Online Information Review, 27*(4), 238-242.

Alimohammadi, D. (2004). Measurement of the presence of keywords and description meta-tags on a selected number of Iranian websites. *Online Information Review, 28*(3), 220-223.

American Society for Information Science and Technology (2002). Retrieved from www.asis.org

Bancroft, A. F., Croft, V. F., Speth, R., & Phillips, D. M. (1998). A forward-looking library use survey: WSU Libraries in the 21st century. *Journal of Academic Librarianship, 24*(3), 216-224.

Beaubien, Rick (2002). *METS Tutorial Presentation*. Retrieved from http://www.loc.gov/standards/mets/presentations.html

Beeline (2002). *Tips and tricks: meta-tags*, Retrieved from http://bton.com/tb16/metatags.html

Bradley, P. (2002). *Meta-tags—what, where, when, why?*. Retrieved from www.philb.com/metatag.htm .

Charles, F. Thomas (2002). Who will Create The Metadata For the Internet. *First Monday, 3*(12). Retrieved from http://www.firstmonday.dk/issues/issue3_12/thomas/index.html

Christensen, D. (1999). Golden retrievers. *School Library Journal, 45*(11), 38-41.

Clark, S. (2000). *Back to basics: META tags*. Retrieved from Webdevelopercom web site: http://www.webdeveloper.com/html/html_metatags_parts.html

Coopee, T. (2000). How to climb the search engine rankings. *InfoWorld, 22*(24), 61-64.

Craven, T. C. (2000). Features of description meta tags in Public Home Pages. *Journal of Information Science, 26*(5), 303-311

Craven, T. C. (2001a). Changes in meta tag descriptions over time. *First Monday, 6*(10). Retrieved from http://www.firstmonday.org/issues/issue6_10/craven/index.html

Craven, T. C. (2001b). Description meta-tags in locally linked Web pages. *Aslib Proceedings, 53*(6), 203-16.

Craven, T. C. (2001c). Description meta-tags in pages returned on different search engines. *Canadian Journal of Information and Library Science, 26*(1), 1-17.

Craven, T. C. (2001d). Description meta tags in public home and linked pages, *LIBRES: Library and Information Science Research Electronic Journal, 11*(2). Retrieved from http://libres.curtin.edu.au/LIBRE11N2/

Craven, T. C. (2004). Variations in use of meta tag keywords by Web pages in different languages. *Journal of Information Science, 30*(3), 268-279.

Dillon, M. (2001). *Metadata for Web resources: How metadata works on the Web*. Retrieved from http://lcweb.loc.gov/catdir/bibcontrol/dillon_paper.html

Drott, M. C. (2002). Indexing aids at corporate websites: the use of Robots.txt and Meta Tags. *Information Processing & Management, 38*, 209-219.

Dublin Core—Tagging the Web for Better Search and Retrieval(n.d.). Retrieved from http://

webreference.com/xml/column24/index.html

Duval, K. B. (1996). Building Web pages: An update. *Library Software Review, 15*(3), 158-162.

E-Commerce-indecs. (2000). Retrieved from http://indecs.org

Friedlander, A. (2002). Digital preservation looks forward. *Information Outlook, 6*(9), 12-14.

Futterman, D. (2001). How to help your intranet search engine do a better job. *Online*, May/June 2001, 36-40.

Gill, Tony (2002). *Metadata and the World Wide Web 2000*. Retrieved from http://www.getty.edu/research/institue/standards/intrometadata/2_articles/gill/content.html

Greenberg, Jane (2000). *Metadata for a Digital Library of Educational Resources*. NY: Haworth Press.

Guenther, K. (1999). Publicity through better website design. *Computers in Libraries, 19*(8), 62-64, 66-67.

Hanks, P. (chief ed.) (1998). *New Oxford Dictionary of English*. UK, Oxford: Clarendon Press, p.1162.

Hensen, S. L. (2001). Archival cataloging and the Internet: the implications and impact of EAD. *Journal of Internet Cataloging*, (3/4), 74-95.

Henshaw, R. (1999). The first Monday metadata project. *Libri, 49*(3), 125-31.

Henshaw, R. & Valauskas, E. J. (2001). Metadata as a catalyst: experiments with metadata and search engines in the Internet journal. *Libri, 51*(2), 86-101.

How to use the HTML Meta Tag. (2002). Retrieved from http://www.upenn.edu/computing/web/webdev/meta/index.html

Infoseek, Help: Submitting Tips (1999). Retrieved from http://www.go.com/AddURL?pg-submitTips.html

Kyrnin, J. (2002). *Magic with meta-tags*. Retrieved from http://html.miningco.com/library/weekly/aa083099.htm

Lawrence, S. & Giles, C. L. (1999, July). Accessibility of information on the Web. *Nature*, 107-109.

Lynch, Clifford (1997, March). Searching the Internet. *Scientific American*, 52-56.

Lynch, Clifford.(2002). *The Dublin Core Descriptive Metadata Program: Strategic Implications for Libraries and Networked Information Access*. Retrieved from http://www.arl.org/newsltr/196/dublin.html

Luh, J. C. (2000). Metadata to the rescue? *Internet World, 6*, p38

Maclennan, A. (1998). Interesting times. *Library Review, 47*(2), 106-109.

Maddux, D. C. (1998). The World Wide Web: Some Simple Solutions to Common Design Problems. *Education Technology*, 24-28.

Metadata. (2002, September). *HCI Journal*. Retrieved from http://www.hci.com.au/hcisite2/journal/Metadata.htm

Meta-tag optimisation tutorial. (2002). Retrieved from http://spider-food.net/meta-tags.html.

Miller, E. (2002). An Introduction to the Resource Description Framework. *D-Lib Magazine, 4*. Retrieved from http://www.dlib.org/dlib/may98/miller/05miller.html

Nowick, E.A. (2002). Use of meta-tags for Internet documents. *Journal of Internet Cataloging, 5*(1), 69-75.

O'Neill, E. T., Lavoie, B. F., & McClain, P. D. (1998). *Web characterization project: An analysis of metadata usage on the Web*. Retrieved from http://www.oclc.org/oclc/research/publications/

review98/oneill_etal/metadata.htm

O'Neill, E. T., et. al (2000). *Web characterization project: An analysis of metadata usage on the Web*. Retrieved from www.oclc.org/oclc/research/publications/review98/oneill_etal/metadata.htm

ONIX For Books(2002). Retrieved from http://www.editeur.org/onix.html

Perkins, G. H., & Yuan, H. (2000). Genesis of a Web based satisfaction survey in an academic library: the Western Kentucky University Libraries' experience. *Library Administration and Management, 14*(3), 159-166.

Qin, J. & Wesley, K. (1998). Web indexing with meta fields: a survey of Web objects in polymer chemistry. *Information Technology and Libraries, 17*(3), 149-56.

Richmond, A. (2002). *Meta-tagging for search engines*. Retrieved from www.wdvl.com/Search/Meta/Tag.html,

Ramiscal, N. G. (2000). The nature and functions of meta-tags: Covert infringement of trademarks and other issues. *Academic search Elite, 18*(1).

Sliwa, C. (1998). Not all search engines are created equal: cases in point. *Computerworld, 32*, 37-38.

Stanek, W. R. (1999). tags target your pages. *PC Magazine, 18*(13), 253-256.

Sullivan, Danny (2002). *How to use HTML meta tags*. Retrieved from http://searchenginewatch.internet.com/webmasters/meta.html

Thelwall, M. (2000). Use of meta-tags for Internet documents. *Journal of Internet Cataloging, 5*(1), 69-75.

The New Meta Tags are Coming—or Are They? (1997). Retrieved from http://searchenginewatch.internet.com/sereport/97/12-metatags.html

The Head Element and Related Elements. (n.d.). Retrieved from http://www.w3.org/MarkUp/html3/dochead.html

Turner, P. T. & 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, 42*(4), 258-270.

Web Characterization Project: June 1999 Web statistics. (1999). Retrieved from http://www.oclc.org/research/projects/archive/wcp/default.htm

Websites: Concepts, issues, and definitions. (2000). Retrieved from http://www.oclc.org/oclc/research/projects/Webstats/definitions.htm

Weibel, S. (1996). *Metadata: The Foundation of Resource Description*. Retrieved from http://digitalarchive.oclc.org/da/ViewObject.jsp?objid=0000003349

Weibel, S., Kunze, J., Lagoze C., &Wolf, M. (1998). *Dublin Core Metadata for Resource Discovery*. Retrieved from http://www.ietf.org/rfc/rfc2413.txt