Paulina E. Vinyard. An Analysis of Embedded Metadata Usage on the World Wide Web. A Master's Paper for the M.S. in L.S. degree. April, 2001. 33 pages. Advisor: Jane Greenberg.

This study examined the use of HTML meta tags for embedding metadata in World Wide Web (Web) resources. While many metadata initiatives that organize and facilitate resource discovery have been successful in organizational and corporate environments, little work has been done to determine whether or not Web resource creators are using metadata. This study examined the HTML source code for 299 web pages to provide a snapshot of how embedded metadata is being used on the Web and used a survey to explore the reasons why resource creators do or do not use metadata. The results found that 71% of the pages examined contained metadata, but very little of it conformed to any metadata standard. The survey results indicated that the majority of respondents did not grasp the larger context of metadata outside of its use on the web, and that they were unfamiliar with the concept of a metadata standard. These findings suggest that to improve the amount and quality of metadata used in web pages, resource creators should be given a context larger than the web in order to understand its value. Additionally, search engines need to get involved by providing metadata field-searching capabilities.

Headings:

Metadata

Dublin Core

World Wide Web

Online searching

AN ANALYSIS OF EMBEDDED METADATA USAGE ON THE WORLD WIDE WEB

by
Paulina E. Vinyard

A Master's paper submitted to the faculty
of the School of Information and Library Science
of the University of North Carolina at Chapel Hill
in partial fulfillment of the requirements
for the degree of Master of Science in
Library Science.

Chapel Hill, North Carolina

April, 2001

Approved by:

____
Advisor

TABLE OF CONTENTS

Chapter

Chapter 1

INTRODUCTION

Finding information on the Internet has become an increasingly difficult task as the number of available resources has multiplied exponentially over the years. This rapid growth, combined with commercial search engines' emphasis on keyword searches and term frequencies, has created an environment in which a single search query often returns hundreds or thousands of search results, more than any searcher has the time or patience to comb through. In the last several years, this problem of information overload has triggered many efforts to organize and facilitate the discovery of information on the Internet. Many of these efforts have been focused on the use of metadata.

On its most basic level, metadata is "data about data." A metadata record consists of a set of attributes, or elements, necessary to describe a particular resource. These elements include information about the resource's form and content such as its author, date of publication, or other similar details that can help information seekers and providers with a variety of tasks. These include identification of a resource that meets a particular information need, evaluation of its suitability for use, and the tracking of the characteristics of a resource for maintenance or usage over time. If search engines were developed to take advantage of metadata assigned to information resources in the Internet environment, this would facilitate searching by attributes such as title or author which are not currently possible.

While many metadata initiatives to organize and facilitate the discovery of information have enjoyed success within various environments, particularly organizational and corporate, the use of metadata for general resource discovery does not seem to have been widely implemented on the Internet. The search algorithms used by most of the commercial search engines likely do not take metadata into account at all (since their algorithms are proprietary, it is impossible to know this for certain). The few that do indicate that they search metadata look only at two rather generic fields, keywords and description, and do not allow users to search these fields specifically. No commercial search engines currently allow field searching with a template that identifies the elements of Dublin Core metadata, a 15-element metadata set intended to be usable by non-specialists and to be generalizable across all fields of knowledge. This schema is particularly well-suited for Internet resource description and discovery because it would allow users to limit searches by using specific standardized fields such as “creator”, “title”, “date”, etc, similar to the way in which standalone literature databases and library online catalogs are already searched.

Not only are search engines not taking advantage of metadata, but little is known about whether or not it is being used by the creators of Internet resources. The purpose of this study is twofold: 1) to provide a snapshot of how embedded metadata (HTML meta tags) is being used on the World Wide Web and 2) to explore the reasons why resource creators are or are not using metadata and what factors might influence them to use it if they are not.

Chapter 2

LITERATURE REVIEW

Although the concept of metadata predates both the Internet and the World Wide Web, the growth of electronic publishing, digital libraries and the resulting information overload have caused an explosion of interest in metadata standards and practices all over the world. Several metadata initiatives are evaluated in Greenberg (2000). These include BIBLINK: Linking Publishers and National Bibliographic Services [http://hosted.ukoln.ac.uk/biblink/], DESIRE (Development of a European Service for Information on Research and Education) [http://www.desire.org/], Nordic Metadata [http://www.lib.helsinki.fi/meta/], OCLC CORC (Cooperative Online Resource Catalog) [http://www.oclc.org/oclc/corc/], and ROADS (Resource Organization and Discovery in Subject-based services) [http://www.ilrt.bris.ac.uk/roads/].

Projects like these have either developed their own metadata schemas or used schemas developed by other organizations. One of several metadata schemas developed in the last few years is the Dublin Core, an international, cross-disciplinary metadata standard comprised of fifteen elements, developed for the description of a wide range of networked resources (Hillman 2000). Eric Jul concludes in his 1997 survey of the status of the field of Internet resource cataloging that the Dublin Core "represents significant effort building consensus among disparate user groups for a basic set of data elements that facilitate network resource discovery and retrieval . . . Once [a metadata transfer syntax is] in place, producers of Internet resources will be able to express and associate metadata with the resource." Weibel (1999) agrees with this in an assessment of the Dublin Core metadata initiative written two years later, calling the Dublin Core "the leading candidate for achieving the goal of simple resource description for Internet resources."

While the general consensus seems to be that distributed resource description for Internet resources using the Dublin Core is the best way to improve the ability to find information on the Internet, Thomas and Griffin (1998) question who is going to do the intensive work of creating the vast amounts of metadata that this solution would require. Although they believe the adoption of a common metadata standard such as the Dublin Core offers promise, they are concerned that supporters of such a step have not considered the challenge of how to persuade Internet resource creators to participate.

An evaluation of the United States Environmental Protection Agency's (EPA) metadata system performed by Shauna Stephenson, a Master's student in the School of Information and Library Science at the University of North Carolina at Chapel Hill, lends support to Thomas and Griffin's concern. Stephenson (1999) found that the success of the EPA's system in improving public access to the agency's information resources hinges upon a commitment by data owners to generate metadata for their documents. At the time the evaluation was performed, she noted that the organization was experiencing difficulty in generating support among employees for metadata creation, and that as a result, metadata generation was only consistently carried out for top-level pages on the EPA's web site. It seems reasonable that this lack of commitment to the use of metadata on the part of EPA employees, for whom resource description is a part of their jobs, can be extended to resource creators on the Internet as a whole who have no such obligation.

In March 1997, The Vancouver Webpages VWBot traversed an unknown number of web pages; a breakdown by count of the meta attributes it found is available at http://www.vancouver-webpages.com/META/bycount.shtml. About 2/3 of the data was collected from traversing www.yahoo.com to depth 4 and the rest from regular runs of searchBC, a search engine that confines itself to sites in the bc.ca domain, sites in the .com, .net and .org domains which are listed by InterNIC as being located in BC, and sites in the .ca domain which are not part of another Provincial domain. It is interesting to note that the *keywords* attribute was only present on about 12% of the web pages visited, and *description* on about 10%. This page gives a good snapshot of the meta attributes in use at the time and an idea of its overall prevalence on the Web. In addition to this data, Qin and Wesley (1998) surveyed 1037 Web objects in polymer science between October 1996 and March 1997 and found that many HTML documents in the field did not include any metadata, and that errors in coding, imprecision of data description, and inadequate use of metadata were prevalent in those that did. They concluded that Internet resource creators are largely unaware of the existence of metadata schemas and why and how they should implement these mechanisms.

Unfortunately, three years later, there is still little information available about metadata standards for Internet resource creators who are not part of the information science community. Aside from the name attributes *description, keywords,* and *robots* that are supported by some search engines and *PICS-Label*, which is supported by a W3C Recommendation (PICS 1.1 Label Distribution -- Label Syntax and Communication

Protocols, available at http://www.w3.org/TR/REC-PICS-labels), there does not seem to be any commonly accepted standard for metadata associated with web pages. HTML META tags are addressed in the W3C HTML 4.01 specification (Raggett, et al., 1999) and also mentioned briefly in most web sites dealing with the construction of web pages or their optimization for search engines, but these resources do not really explain the importance of adhering to a metadata standard. While the HTML 4.01 Specification does recommend that resource creators should refer to a profile where metadata properties and their legal values are defined and designate this profile by using the profile attribute of the HEAD element, it does not do a good job of explaining why resource creators should take the time to do this.

Information gathered in 1997 indicates that metadata creation for web pages was not in widespread practice at that time. While many metadata initiatives have enjoyed success in organizational and corporate environments since that time, little has been published on the use of metadata for resource discovery on the Internet. It is not known whether or not metadata is being used by the creators of Web resources any more today than it was in 1997, nor have the questions raised by Thomas and Griffin been answered. For these reasons, it seems important to develop a current picture of metadata usage on the World Wide Web and to explore the reasons why resource creators are or are not using it.

Chapter 3

METHODOLOGY

This study utilized a multi-method approach to examine the research questions. First, it analyzed a sample of web pages viewed between February 24, 2001 and March 12, 2001 in order to develop a snapshot of how embedded metadata (in the form of HTML meta tags) is being used on the World Wide Web. Second, a survey was used to gain insight into the reasons why resource creators do or do not use metadata and the factors that might influence them to use it if they do not.

To develop the sample of web pages, a series of five searches was performed on each of three search engines. While the enormity and ephemeral nature of the Web make it difficult to achieve a representative sample, the search terms “astronomy”, “buckyball”, “e-commerce”, “Italian greyhound”, and “vehicle registration” were chosen because they seemed likely to lend themselves to a broad sample of the various types of pages found on the Web. “Astronomy” and “buckyball” were intended to yield a sample of scientifically-oriented web pages (in both commercial and educational domains), “e-commerce” was intended to yield a sample of business-oriented web pages, “Italian greyhound” was intended to yield a sample of personal web pages, and “vehicle registration” was intended to yield a sample of government web pages. The search engines chosen for this study were AltaVista (http://www.altavista.com), Hotbot (http://www.hotbot.com), and Google (http://www.google.com). AltaVista and Hotbot were selected because they index HTML meta keywords and description tags, according to both the Search Engine Watch article "Search Engine Features for Webmasters" (http://www.searchenginewatch.com/webmasters/features.html) and their own site documentation (http://doc.altavista.com/help/search/customizing_results.html, http://hotbot.lycos.com/help/addurl/#2)[1]. The third search engine, Google, was chosen because it received a high ranking and was designated an Editor's Pick in a recent review done by PC Magazine (Sirapyan, 2000). According to Search Engine Watch, the appearance of search terms in meta tags does not boost the page ranking in the results returned by any of these three search engines, so the choice of search engines that index metadata should not have introduced any bias toward pages that include meta tags into the search results.

The first twenty hits returned by each search were included in the sample unless they fell into one of the following categories:

- Pages that returned a "404 Not Found" error when visited;
- Pages in languages other than English or Spanish (since the researcher could not determine to whom the survey for the second part of the study should be sent);

[1] While neither of these sites specifically states that these search engines index metadata, it is strongly implied on both. AltaVista's "Choose How AltaVista Displays Search Results" page (http://doc.altavista.com/help/search/customizing_results.html) says "the description of a Web page is entered in the page's HTML code by the Webmaster specifically to aid your Web searches. If the Webmaster does not include a description, AltaVista uses the first 150 bytes on the page as the description." Further down on the page is information about an option to have AltaVista highlight the search term in the search results. If a page has few or no highlights, "the Web developer may have included your search terms in HTML code as keywords." Both of these statements imply the indexing of meta tags. Hotbot's "Submit a Web Site & Webmaster's FAQ" (http://hotbot.lycos.com/help/addurl/#2) includes a question about how a webmaster can include description and keywords metadata when submitting a site, and states that the meta tags most search engines use are author, description, and keywords.

- Pages returned by more than one search engine, each time they were encountered beyond the first.

All omitted pages were replaced by the next consecutive page on the search results list, in order to ensure that twenty pages were examined for each search. Since this was the case, the order in which search engines were queried was varied for each search term in order to prevent the results from being biased by any one search engine's algorithm.

Once the sample was developed, the HTML source code of each page included was viewed and the meta name or http-equiv attributes used (if any) were recorded. Additionally, the meta tag name-value pairs from each page were recorded in their entirety so that they would be available for qualitative examination. These tags were not associated in any way with the pages from which they were taken. This study did not record title metadata that was included in a web page through the use of HTML title tags because having title tags is required by the W3C HTML 4.01 specification (Raggett, 1999). Additionally, the focus of this study is on the use or non-use of HTML meta tags specifically.

For the second part of this study, a web survey was used to gather data about resource creators' use or non-use of metadata and the factors that might influence them to use it if they are not. A message containing the survey cover letter and directing the recipient to the survey web site was emailed to the webmaster of each page visited in the process of data gathering for the first part of this study. Copies of the survey cover letter and the web survey can be found in Appendices A and B. To determine the "webmaster" of each page to whom the survey should be sent, this study used the following rubric, with preference given to definitions appearing higher on the list:

1. Use an email address that specifically states that it is a means of contacting the webmaster. If the means of contacting the webmaster is through a form linked from the page in question, use the form.
2. Use an email address belonging to the person who maintains the page (as identified by the page).
3. Use an email address belonging to the person who created the page (as identified by the page).
4. Follow the links from the page that go to other pages in the same site (here defined to be other pages under the same base URL and at the same hierarchical level in the file structure as the page in question) and repeat the above steps.

If no webmaster could be determined (or an email address for said individual could not be obtained), no email was sent. In cases where more than one individual was determined to be the webmaster, the email was sent to all individuals so identified.

Because the sample of web pages developed using this methodology included only English-language pages (due to the choice of English-language search terms and the language limitations of the researcher), it cannot be truly representative of the entire Web. The subject-bias introduced by the chosen search terms provides additional limitations. For future studies, a different sampling methodology should be used in order to ensure a more random, and thus more representative, sample of the variety of pages available on the World Wide Web. The sampling methodology using randomly-generated IP addresses proposed by O’Neill, McClain, and Lavoie (1997) would be a good starting point (unfortunately, implementation of this methodology was not feasible within the time constraints needed for it to benefit this study). Although the sample used for this study was not random, it was chosen in such a way as to minimize factors that may have biased the study in favor of the presence or absence of metadata (for example, the choice of search engines that did not boost rankings of pages for which the search term appeared in meta tags), so it seems likely to be at least somewhat representative of the variety of pages available on the Web as a whole. Regardless, it is possible that a repetition of this study would return different results. The reader should view the results of this study in light of these limitations.

Chapter 4

RESULTS AND DISCUSSION

Analysis of the data collected in this study provides a snapshot of the use of embedded metadata in HTML pages available on the World Wide Web. To gather this data, twenty pages were examined for all of the searches described in Chapter 3 except for the Google search for “astronomy”, where only nineteen were examined due to a counting error. A remarkably high percentage of these pages contained meta tags. Of the 299 pages examined, 212 (70.90%) contained at least one meta tag. Given that the sampling method used in this study aimed to achieve a representative sample of the different kinds of pages found on the Web, this data might suggest that between 65.7% and 76.1% of web pages have metadata of some sort ($p<.05$). Of the 212 sampled pages with meta tags, 33 had only meta tags (*content-type*, *generator*, and/or *progID*) that are known to be generated automatically by popular HTML editors such as Netscape, Claris Home Page, Microsoft FrontPage and Microsoft Word. A breakdown of the occurrence of these tags can be seen in Table 1.

Table 1

Number of Sites with only Automatically Generated Meta Tags

	Only *Content-type*	**Only *Generator***	**Only *Generator* and *Content-type***	**Only *Generator*, *Content-type*, and *ProgID***
Number of sites	**16**	**8**	**8**	**1**

Combinations of the above three meta tags not appearing in the table did not occur in the data set.

It is possible that more than these 33 pages contained only automatically-generated metadata, since aside from the above fields it is hard to tell if the information was entered by the site-builder or generated automatically from information an editor gleaned from elsewhere. For example, an HTML editor could extract the value for an *author* tag from the name of the individual or company to whom the editor is registered. Keeping this caveat in mind, approximately 59.53% of the pages looked at for this study had meta tags that likely involved some effort on the part of the web site builder. Eighty-four percent (84.43%) of the pages with meta tags had tags that probably involved effort on the part of the site builder. Again assuming that the sample selected for this study is representative of the population of HTML pages, this data suggests that between 53.93% and 65.13% of HTML pages available on the World Wide Web contain metadata generated by a human being (p<.05). Appendix C contains a chart which lists the various meta tags found in the sample selected for this study and the percentage of pages on which they occurred.

In contrast with the just over 70% of pages in this study that contained metadata, only 7 (2.34%) of the 299 pages examined showed evidence of using the Dublin Core metadata schema or a variant of it. Table 2 contains a breakdown of the Dublin Core meta tags used and the percentage of pages on which they occurred. Only five of these pages made extensive use of the Dublin Core (in fact, two of the pages used it exclusively), and interestingly, these five were the only ones that used it correctly, though even their use was not flawless. All five pages used Dublin Core Qualifiers for at least one element in accordance with the principles governing them, but not all of the qualifiers they used are approved by the Dublin Core Usage Committee (“Dublin Core

Qualifiers," 2000). The non-approved qualifiers are probably part of a standard developed internally by the organizations responsible for the web pages in question, and since the appropriate method of designating the qualifier was used, it would be possible for an indexing algorithm to strip off the unknown qualifiers and include the information with the appropriate main element. Two of the five web sites that used the Dublin Core extensively also made use of encoding scheme qualifiers. Problems noted with these five pages' uses of the Dublin Core include inconsistency in the date formats used (particularly when an encoding scheme qualifier is not used), lack of use of a controlled vocabulary for the *Type* element, and the presence of an element without a value associated with it on one of the pages.

Table 2

Usage of Dublin Core Meta Tags

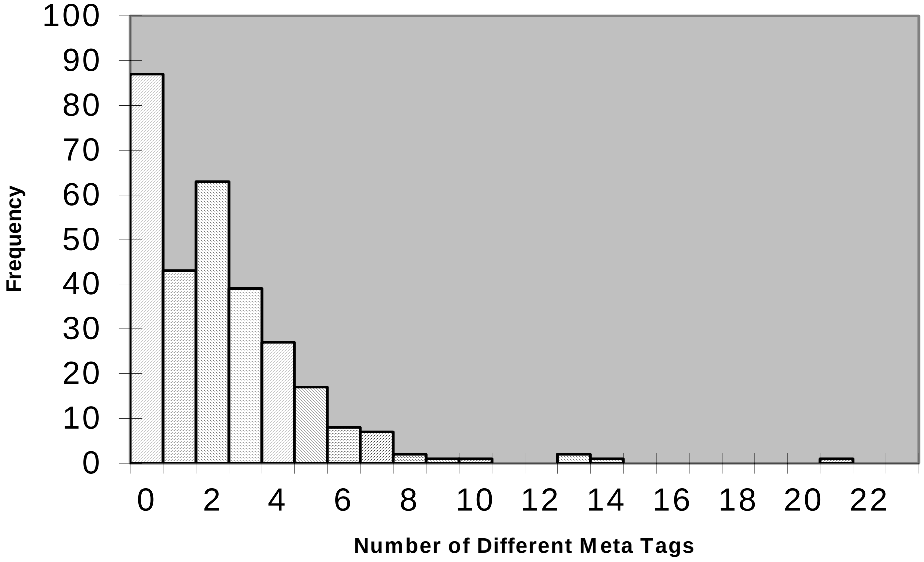
Dublin Core META Tag	**Number of Sites Including Tag**	**Percentage of Sites Including Tag**	**Percentage of Sites Including DC Including Tag**
DC.Date	5	1.67%	71.43%
DC.Description	5	1.67%	71.43%
DC.Identifier	5	1.67%	71.43%
DC.Publisher	5	1.67%	71.43%
DC.Subject	5	1.67%	71.43%
DC.Title	5	1.67%	71.43%
DC.Type	5	1.67%	71.43%
DC.Creator	4	1.34%	57.14%
DC.Format	4	1.34%	57.14%
DC.Language	4	1.34%	57.14%
DC.Source	2	0.67%	28.57%
DC.Contributor	1	0.33%	14.29%
DC.Coverage	1	0.33%	14.29%
DC.Relation	1	0.33%	14.29%
DC.Rights	0	0.00%	0.00%

The final two pages that attempted to use Dublin Core metadata did not do so successfully. One of the two sites used only the DC.Source element, and the value associated with the element does not adhere to the definition provided by the Dublin Core Element Set, Version 1.1 (2000). The definition calls for “a reference to a resource from which the present resource is derived” and recommends that best practice is to “reference the resource by means of a string or number conforming to a formal identification system.” The page in question supplied simply “Magazines,” which really provides very little information. The other page that attempted to use the Dublin Core did not even use elements defined as part of the Dublin Core Element set (they have been included as examples of Dublin Core metadata because the element names were preceded by “dc”, which is the standard way to signify Dublin Core). The meta tags used were *dcFileID*, *dcModified*, and *dcPublisher*, only the last of which is one of the fifteen Dublin Core elements. The values associated with these elements do not conform to any controlled vocabulary or encoding scheme (unless an internal one), and would be quite useless to a search engine or a searcher who tried to search on one of these fields. Unless this is some kind of internal metadata standard, it seems likely that the resource creator has had some exposure to the Dublin Core, but does not fully understand how it is intended to be used.

Figure 1 is a histogram of the distribution of the number of meta tags per page with distinct name or http-equiv attributes. Disregarding the 87 web pages that had no metadata and the largest outlier, a page with 21 different tags, the mean number of meta tags per page with different name or http-equiv attributes was 3.09, with a standard deviation of 2.15 and median and mode of 2. If the study sample is considered to be representative of the web as a whole, computing the One-Sample *t* confidence interval at the 95% confidence level gives a margin of error for the mean of ±.29, implying that the true mean number of distinct meta tags for all pages that contain metadata is close to 3.

However, since the data is right-skewed, as Figure 1 shows, it is possible that the median and mode of 2 is a more accurate number. Regardless, the results from this sample seem to demonstrate that web pages with metadata contain an average of 2 or 3 meta tags with distinct name or http-equiv attributes.

Figure 1
Number of Distinct Meta Name and Http-equiv Attributes Per Web Page



Not all of the meta tags present on the pages examined for this study are there for the purpose of improving information retrieval or ranking on search engine results lists. Many meta tags seem to be used to instruct web browsers to display pages in certain ways or to instruct a search engine’s spider as to whether and how to index the page. For example, the *refresh* tag is often used to forward users to a different page, a technique deprecated by the W3C HTML 4.01 Specification because it makes pages inaccessible to some users (Raggett, 1999). Tags like *robots* and *revisit-after* are examples of tags used to provide instructions to search engines. Additionally, the *generator* tag does not seem to serve a purpose other than to allow companies that produce web authoring software to assess their market penetration.

Of the 185 surveys emailed to the webmasters of the pages examined for the first part of this study, 29 usable responses were received, a response rate of 16%. The survey was intended to explore the reasons why resource creators do or do not use metadata and how they might be persuaded to use it if they are not. The responses received provide interesting qualitative insight into these issues.

Understanding of what metadata is varied greatly among the respondents. They were asked to rate their understanding of metadata on a scale of 1 to 5, with 1 being "never heard of it" and 5 being "I have a clear understanding." Of the 29 respondents, 7 had never heard of metadata and 8 felt that they had a clear understanding. Fourteen respondents rated their understanding somewhere in the middle of that continuum – 6 on the low end, 4 in the middle, and 4 on the high end. Interestingly, most of the respondents who rated their understanding of metadata highly provided definitions of it specifically in the context of web pages and search engines. Some of these responses were, "the data about the data that is supplied by the Web, i.e. – where the data originated and who is maintaining it", "data from the head of an HTML page that can be gathered for generation of reports and searchable databases", and "a way of helping search engines understand the content of a web page." Some of those who rated their understanding of metadata at 3 or 4 on the scale provided more general definitions, however: "a method of summarizing key document data", "data which are not the primary product of a repository, but which are used to organize, describe, and/or retrieve information from the primary product", and "fields of information added to datarecords that describe their contents.” These responses indicate that while many of the respondents have a fairly good understanding of how metadata is used on the Web, most of them do not grasp its larger context. Instead, they see it mainly as a tool to improve search engine listings and to control the way a page is handled by a web browser.

Perhaps related to this general lack of knowledge about metadata’s uses outside of the web environment is the fact that there does not seem to be any schema consistently used to create the metadata associated with web pages. Ten of the 18 survey respondents who had used metadata for a web page said that they had not adhered to any standard for its creation. Of the 8 survey respondents who said they had used a metadata standard, it is unclear whether any of them actually have. Two of them gave responses indicating that they assumed they had used a standard. One of these two said he assumed his HTML editor conforms to “some sort of defined standard for meta tags” and said he thought the standard might be ISO. The other said he appropriated from “various other pages, magazines and books” and assumed that one or two standards had been met that way. Three of the respondents who said they had used a standard indicated that they adhered to W3C standards or HTML 4.0 standards. However, both the W3C HTML 4.01 Specification and the HTML 4.0 Specification that it supercedes state, “This specification does not define a set of legal meta data properties. The meaning of a property and the set of legal values for that property should be defined in a reference lexicon called a profile. For example, a profile designed to help search engines index documents might define properties such as "author", "copyright", "keywords", etc.” (Raggett 1999, 1998). It seems that these respondents do not understand the difference between the rules for the structure of HTML meta tags and a schema that defines metadata attributes and the information that should be associated with them. The final three respondents who said they adhered to a metadata standard (plus one of those who said he follows HTML standards) said they used an in-house standard. Without the ability to examine these standards, it is hard to know whether or not they are true metadata schemas, or whether these respondents are as confused as the others.

In examining knowledge of the Dublin Core, it was discovered that only six of the 29 respondents had heard of it and none of them had used it for a web site. One of these six respondents had used it to catalog slides for a library using a template supplied by the library; from his response, it sounds like it had never occurred to him to use Dublin Core on the web. Four respondents felt that there was no real need to use Dublin Core metadata for web pages, and one of these four additionally stated that he felt that creating metadata for web pages was more work than he had time to do, and that he felt metadata standards should be embedded in site development tools. The last respondent who had heard of Dublin Core said that his company preferred to develop proprietary solutions that were fitted to its needs.

One limitation of the survey was the small number of responses received. While 16% is a reasonably good response rate, the actual number of responses received was too low to permit the statistical analysis of the relationships between education, web experience, and understanding of metadata that the researcher had hoped to carry out. Nonetheless, the qualitative information gleaned provides a foundation on which further research can be based.

Chapter 5

CONCLUSION AND FUTURE WORK

This study found that 212 (70.90%) of the 299 web pages in the sample had meta tags. Just under 16% (33) of the pages with meta tags had only tags that were probably automatically generated. The sample mean number of meta tags per page with different name or http-equiv attributes was found to be 3.09, with standard deviation 2.15 and median and mode of 2. Only 7 (2.34%) of the 299 pages examined had Dublin Core metadata or a variant of it.

If one accepts the study sample as representative of the variety of pages available on the World Wide Web, given the limitations of the sampling methodology as explained in Chapter 3, it is likely that between 65.7% and 76.1% of all web pages have metadata of some sort (p<.05). It is also likely that between 53.93 % and 65.13% of HTML pages available on the World Wide Web contain metadata generated by a human being (p<.05). The results from this sample also seem to demonstrate that web pages with meta tags contain an average of two or three meta tags with distinct name or http-equiv attributes.

The survey responses, while small in number, generated several interesting insights into the respondents' understanding and use of metadata. While many of them had a fairly good understanding of how metadata is used on the Web, most of them did not grasp its larger context. Perhaps as a result of this, the majority of the respondents had not adhered to any standard for the creation of metadata for their web pages, and in fact most of them seemed not to understand what a metadata standard is. In keeping with this lack of knowledge of metadata standards, only six of the respondents had heard of the Dublin Core, and none had ever used it for a web page.

These findings suggest a need to educate resource creators about standards like the Dublin Core and the uses of metadata outside of simply improving placement in search engine results. However, providing them with knowledge about metadata standards is probably not enough; it seems likely that as long as search engines do not index most metadata or allow field searching, resource creators will feel no real incentive to conform to any standard. The observed relationship between the fact that many search engines index metadata in the *keywords* and *description* fields (though they do not allow searches limited to these fields) and the fact that *keywords* and *description* were the most frequently occurring meta tags in this study seems to support this connection. If and when search engines do begin to index metadata fields and offer searches by these fields, it would be ideal if most of them choose to adhere to the same metadata standard to save resource creators the work of having to code their metadata in several different schemas.

The variety of levels of knowledge about metadata among the survey respondents is a particularly interesting finding of this study. One possible explanation for the broader definitions of metadata provided by respondents who are less sure of their understanding of it is that these individuals became familiar with the concept of metadata outside the web environment, and thus realize that its use on the Internet is only one application. Those who define it strictly in terms of its use by search engines and web browsers are probably very familiar with its use in that context but have probably not been exposed to it outside of the web-development environment. Therefore, they don’t realize how many other environments and applications it can be used in, nor do they realize its potential to drastically improve Internet searching if indexing of metadata fields were implemented by search engines. Adding to the confusion, it appears that some respondents who have used metadata outside of the web environment don’t realize that that is what they have been working with; for example, the respondent who had used the Dublin Core in a library setting rated his understanding of metadata at 2, and defined it tentatively as “keywords to assist search algorithms.” While this definition could encompass both the Dublin Core and the generic metadata most often used in web pages, nothing in this individual’s response indicates that he has connected the two cognitively.

Based on these inferences, it seems that one approach that might increase both the amount and quality of metadata used in web pages is to educate resource creators about its value. This could be done by including more contextual information about what metadata is and about its potential uses along with the technical instruction provided by many web sites on the use of metadata. Providing users with a glimpse of the bigger picture might help them grasp the full extent of metadata’s potential to improve information retrieval on the Internet, improving their willingness to implement the Dublin Core or another standard on their sites, rather than dismissing it as “too complex” or because they do not feel it necessary to provide so much information. Getting search engines involved by convincing them to provide metadata field-searching capabilities is another important step. It is also essential to provide Web resource creators with a clear understanding of the difference between the rules for the structure of meta tags and a metadata schema that defines metadata attributes and the information that should be associated with them. Metadata is far less useful for information retrieval if it does not conform to a standard set of attributes and have the information associated with those attributes in a format that can be parsed by an indexing technology.

In conclusion, this study's most valuable contribution is probably a base number (71%) of web pages with metadata against which a number produced by a more thorough study can be compared. It has also provided information researchers with some basic insight into web resource creators' use of metadata and attitudes towards it.

It would be useful for a future study to examine more than just embedded metadata on the World Wide Web. While embedded metadata can only be used for collections made up of HTML files on the Web, this does not mean that there is not metadata for these resources stored in other locations – for example, library OPACs that utilize the MARC 856 field (Electronic Location and Access) or OCLC's CORC database. Metadata stored in locations like these (separate from the resource being described) was not examined in the course of this study, and it is possible that this kind of metadata could be more prevalent than the embedded form. A more comprehensive future study could explore this issue and try to determine what kind of impact the storage of metadata outside the document described could have on resource discovery on the Internet. It would also be interesting to see what kind of search mechanisms, if any, are available to take advantage of externally-stored metadata.

A final piece of data worth tracking in a future study is the domain (government, commercial, educational, personal home page, etc.) into which each site in the sample falls. Domain diversity is something that the sampling method employed by this study intended to achieve, but since the category of each page visited was not recorded during data collection, it is hard to say whether or not this goal was met. Additionally, it would be interesting to see if and how metadata usage varies across these domains.

Appendix A

SURVEY COVER LETTER

Dear Webmaster:

I am a graduate student writing to request your participation in a metadata study because one of the web pages selected (**«URL»)** identified you as a webmaster. This study is being carried out under the supervision of Dr. Jane Greenberg to fulfill the master's paper requirement for a Master of Science in Library Science degree from the University of North Carolina (UNC).

By UNC policy, you must be at least 18 years of age to participate in my study. If you meet this criteria and choose to participate, please fill out the anonymous survey located at http://www.ils.unc.edu/~vinyp/mp/survey.shtml . It should take no more than 10 minutes of your time. This questionnaire is intended to assess your knowledge of and experience with metadata, particularly as it pertains to the accessibility of information on the World Wide Web. If you choose to participate, please base your answers to the questions about your use of metadata on a specific web site on the URL mentioned above. Even if you are not familiar with the term "metadata," I am still interested in your response to my survey!

Participation in this study is completely voluntary, and no risks are anticipated to respondents. You may refuse to answer any question, and all information you provide will be completely anonymous and confidential. The Academic Affairs Institutional Review Board (AA-IRB) of the University of North Carolina at Chapel Hill has approved this study.

Please contact Dr. Greenberg or me if you have any questions about the study or the survey itself, and contact the UNC-CH AA-IRB if you have any questions or concerns about your rights as a participant in this research. Contact information for each of us is available below.

Sincerely,
Paulina Vinyard, MSLS Student
School of Library and Information Science
University of North Carolina at Chapel Hill
919-969-9085
vinyp@ils.unc.edu

Advisor:
Dr. Jane Greenberg
(919) 962-7024
janeg@ils.unc.edu

Academic Affairs Institutional Review Board
Dr. Barbara D. Goldman, Chair
CB# 4100, 201 Bynum Hall
The Univ. of North Carolina at Chapel Hill
Chapel Hill, North Carolina 27599-4100
(919) 962-7761, or Email: aa-irb@unc.edu

Appendix B

SURVEY

Survey of Metadata use on the World Wide Web

Thank you for agreeing to participate in my metadata study! This questionnaire is intended to assess your knowledge of and experience with metadata, particularly as it pertains to the accessibility of information on the World Wide Web. It should take no more than 10 minutes of your time and all information you provide will be completely anonymous and confidential. The results will be used for my masters' paper for the degree of Master of Science in Library Science from the University of North Carolina at Chapel Hill. Please feel free to contact me at vinyp@ils.unc.edu or my advisor, Dr. Jane Greenberg, at janeg@ils.unc.edu, if you have any questions.

By filling out this survey and clicking on the "submit" button, you are indicating that you are at least 18 years of age and that you consent to participate in this study.

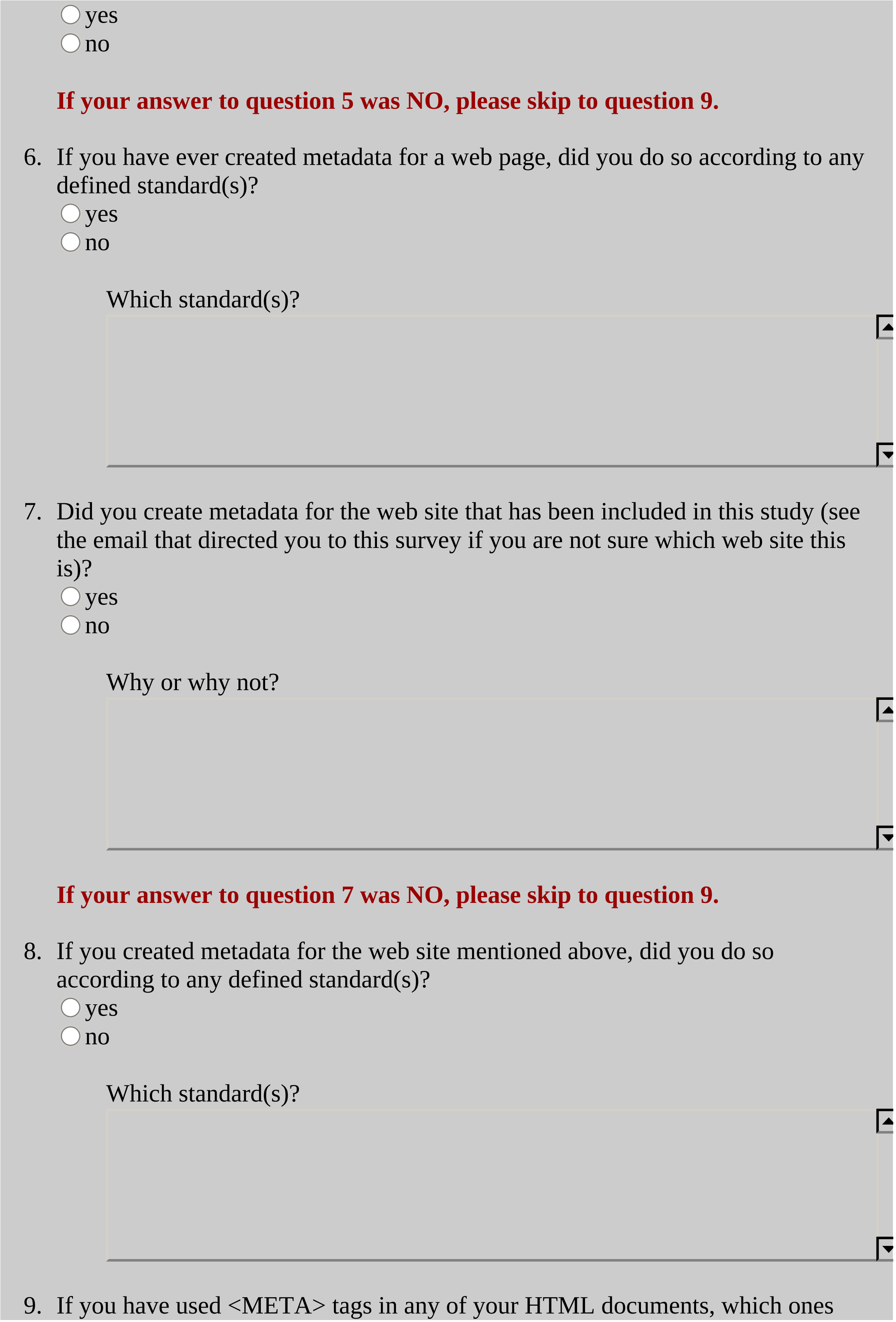
1. How long have you been building web sites?

2. If you have any degrees more advanced than a high school diploma, please list them and specify your major or subject of study for each degree.

3. On a scale of 1 to 5, how well do you feel you understand what metadata is?
   ○ 1 Never heard of it
   ○ 2
   ○ 3
   ○ 4
   ○ 5 I have a clear understanding

4. If you have heard of metadata, briefly explain your understanding of what it is.

5. Have you ever created metadata for any web page you have worked on?



○ yes
○ no

If your answer to question 5 was NO, please skip to question 9.

6. If you have ever created metadata for a web page, did you do so according to any defined standard(s)?
○ yes
○ no

Which standard(s)?

7. Did you create metadata for the web site that has been included in this study (see the email that directed you to this survey if you are not sure which web site this is)?
○ yes
○ no

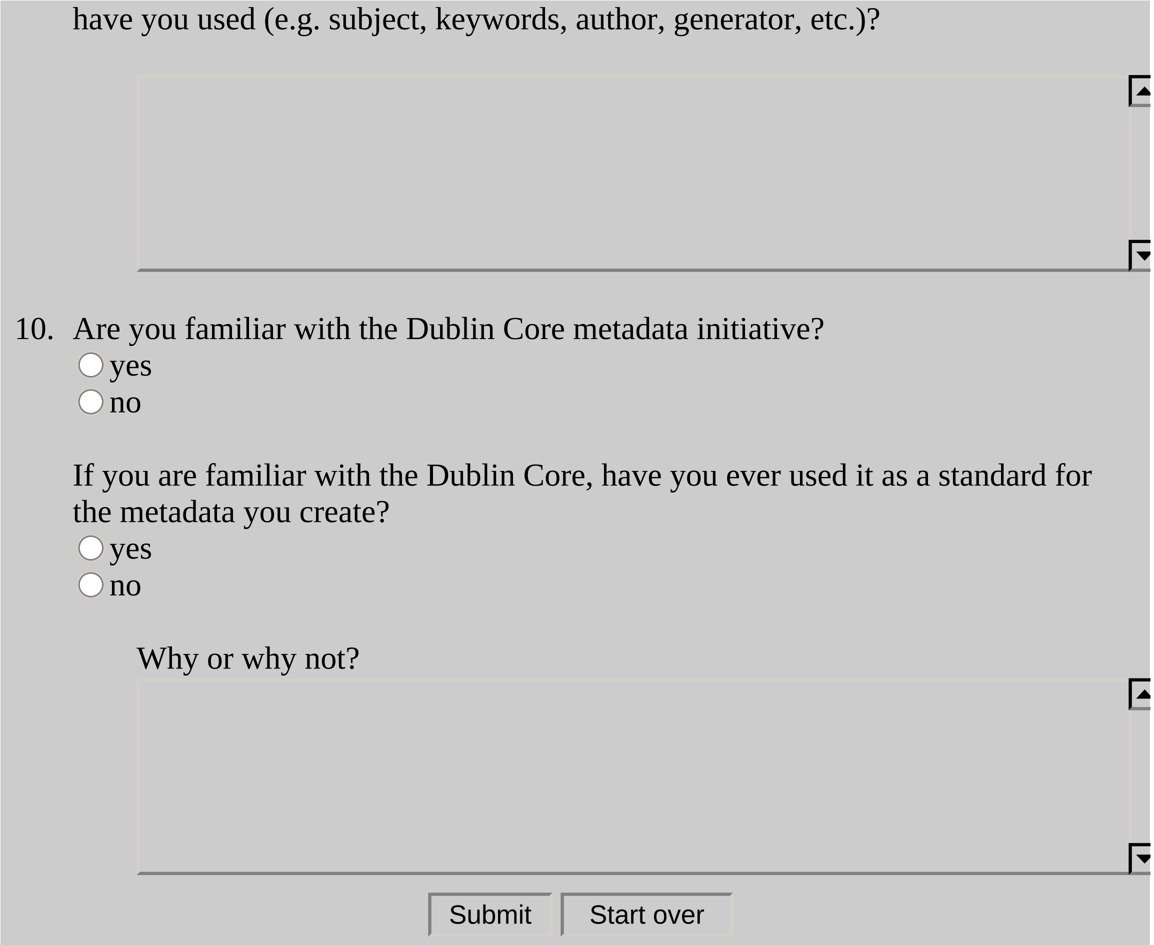
Why or why not?

If your answer to question 7 was NO, please skip to question 9.

8. If you created metadata for the web site mentioned above, did you do so according to any defined standard(s)?
○ yes
○ no

Which standard(s)?

9. If you have used <META> tags in any of your HTML documents, which ones



have you used (e.g. subject, keywords, author, generator, etc.)?

10. Are you familiar with the Dublin Core metadata initiative?
○ yes
○ no

If you are familiar with the Dublin Core, have you ever used it as a standard for the metadata you create?
○ yes
○ no

Why or why not?

This page created and maintained by Paulina Vinyard, Masters Student, School of Information and Library Science, University of North Carolina at Chapel Hill.

Last updated April 23, 2001 -- Background color changed for printing purposes.

Appendix C

META TAG USAGE IN THE STUDY SAMPLE

META Tag	Number of Sites Including Tag	Percentage of Sites Including Tag
Keywords	144	48.16%
Description	143	47.83%
Content-Type	73	24.41%
Generator	55	18.39%
Author	33	11.04%
Robots	24	8.03%
PICS-Label	12	4.01%
Copyright	10	3.34%
ProgId	10	3.34%
Revisit-after	10	3.34%
Pragma	8	2.68%
Distribution	7	2.34%
Rating	7	2.34%
Content-Language	6	2.01%
Microsoft Border	6	2.01%
Classification	4	1.34%
Refresh	4	1.34%
Title	4	1.34%
Expires	3	1.00%
Language	3	1.00%
Date	2	0.67%
Microsoft Theme	2	0.67%
Owner	2	0.67%
Publisher	2	0.67%
Reply-To	2	0.67%
Review	2	0.67%
Template	2	0.67%
zgitemplate	2	0.67%
Abstract	1	0.33%
Astronomy Now Online	1	0.33%
cache-control	1	0.33%
Contents	1	0.33%
Coverage	1	0.33%
creation_date	1	0.33%
dcFileID	1	0.33%
dcModifier	1	0.33%

META Tag	Number of Sites Including Tag	Percentage of Sites Including Tag
dcPublisher	1	0.33%
Element_ID	1	0.33%
entword	1	0.33%
Geography	1	0.33%
Guide	1	0.33%
Longname	1	0.33%
Louisiana Office of Motor Vehicles	1	0.33%
MS.LOCALE	1	0.33%
Netinsert	1	0.33%
objecttype	1	0.33%
Originator	1	0.33%
Page-Enter	1	0.33%
Page-Exit	1	0.33%
QuickSite Border	1	0.33%
Resource-Type	1	0.33%
Save	1	0.33%
Section	1	0.33%
Security	1	0.33%
Service	1	0.33%
Set-Cookie	1	0.33%
Site	1	0.33%
Subject	1	0.33%
Subsection	1	0.33%
Summary	1	0.33%
Update	1	0.33%
VPSiteProject	1	0.33%
VW96.objecttype	1	0.33%
Watch Your Car Program	1	0.33%
web_author_id	1	0.33%
Windows-target	1	0.33%

References

Dublin Core Metadata Initiative. (1999, July 2). Dublin Core Metadata Element Set, Version 1.1: Reference Description. DCMI Recommendation: http://dublincore.org/documents/1999/07/02/dces/

Dublin Core Metadata Initiative. (2000, July 11). Dublin Core Qualifiers. DCMI Recommendation: http://www.dublincore.org/documents/2000/07/11/dcmes-qualifiers/

Greenberg, J. (2000, September 29). A Comparison of Web Resource Access Experiments: Planning for the New Millennium. Conference on Bibliographic Control in the New Millennium. Washington, D.C.: Library of Congress: http://lcweb.loc.gov/catdir/bibcontrol/greenberg_paper.html.

Hillmann, D. (2000, July 16). Using Dublin Core (Dublin Core Metadata Initiative Working Draft): http://purl.org/dc/documents/wd/usageguide-20000716.htm.

Jul, E. (1997). Cataloging Internet Resources: Survey and Prospectus. Bulletin of the American Society for Information Science 24(1): http://www.asis.org/Bulletin/Oct-97/jul.htm.

Krauskopf, T., Miller, J., Resnick, P., and Treese, W. (1996, October 31). PICS Label Distribution Label Syntax and Communication Protocols, Version 1.1. W3C Recommendation: http://www.w3.org/TR/REC-PICS-labels/.

META attributes by count.(1997, June 4). Vancouver Webpages: http://www.vancouver-webpages.com/META/bycount.shtml.

O'Neill, E.T., McClain, P.D., and Lavoie, B.F. (1997). A Methodology for Sampling the World Wide Web. Annual Review of OCLC Research: http://www.oclc.org/oclc/research/publications/review97/main_frameset.htm.

Qin, J. and Wesley, K. (1998). Web indexing with meta fields: a survey of web objects in polymer chemistry. Information Technology and Libraries, 17(3), 149-156.

Raggett, D., Le Hors, A., and Jacobs, I. (1998, April 24). HTML 4.0 Specification. W3C Recommendation: http://www.w3.org/TR/1998/REC-html40-19980424/.

Raggett, D., Le Hors, A., and Jacobs, I. (1999, December 24). HTML 4.01 Specification. W3C Recommendation: http://www.w3.org/TR/html4/.

Sirapyan, N. (2000, December 5). "In search of..." PC Magazine. 187-198. Also available: http://www.zdnet.com/pcmag/stories/reviews/0,6755,2652815,00.html.

Stephenson, S. (1999). An assessment of the effectiveness of metadata as a tool for electronic resource discovery. A Master's paper for the M.S. in L.S. degree. University of North Carolina at Chapel Hill. Unpublished.

Sullivan, D. (2000, December 20). Search Engine Features for Webmasters. Search Engine Watch: http://www.searchenginewatch.com/webmasters/features.html.

Thomas, C. F. and Griffin, L. S. (1998, December). Who will create the metadata for the Internet? First Monday: Peer-reviewed Journal on the Internet. Available: http://www.firstmonday.org/issues/issue3_12/thomas/index.html

Weibel, S. (1999). The State of the Dublin Core Metadata Initiative. D-Lib Magazine 5(4): http://www.dlib.org/dlib/april99/04weibel.html.