Rising to the Top: Evaluating the Use of the HTML META Tag to Improve Retrieval of World Wide Web Documents through Internet Search Engines

Thomas P. Turner and Lise Brackbill

We evaluate the effectiveness of using the HTML META tag to improve retrieval of World Wide Web documents through Internet search engines. Twenty documents were created in five subject areas: agricultural trade, farm business statistics, poultry statistics, vegetable statistics, and cotton statistics. Four pages were created in each subject area: one with no META tags, one with a META tag using the keywords attribute, one with a META tag using the description attribute, and one with META tags using both the keywords and description attributes. Searches were performed in AltaVista and Infoseek to find terms common to all pages as well as for each keyword term contained in the META tag. Analysis of the searches suggests that the use of the keywords attribute in a META tag substantially improves accessibility while use of the description attribute alone does not. These results suggest that HTML document authors should consider using keywords attribute META tags. We also suggest that more search engines index the META tag to improve resource discovery.

The problem of finding materials on the World Wide Web has been discussed in library and information science journals, computer literature, and the popular media. Internet search engines have been developed to aid in finding materials; however, their performances vary considerably. Numerous researchers have evaluated these tools and have detailed their strengths and weaknesses. Melee's Indexing Coverage Analysis (MICA) report, issued weekly, details the number of pages indexed by various Internet search engines; in addition, the speed of the systems is evaluated (Melee 1998). Other authors have analyzed particular aspects of Internet search engines, such as their retrieval precision (Leighton and Srivastava 1997), their usability (Pollock and Hockley 1997), and their indexing methods (Srinivasan, Ruiz, and Lam 1996). Some researchers have offered advice to the authors of Hypertext

THOMAS P. TURNER (tpt2@cornell.edu) is Metadata Librarian, Albert R. Mann Library, Cornell University, Ithaca, New York. LISE BRACKBILL (liseb@nethost.multnomah.lib.or.us) is Technology Trainer, Multnomah County Library, Portland, Oregon. Manuscript received February 4, 1998; accepted for publication April 17, 1998.

Markup Language (HTML) documents about improving retrieval of their materials. The current research was designed to determine how useful one method, the HTML META tag, is in improving accessibility via Internet search engines; here we focus on indexing rather than on search engine performance.

Metadata and the HTML META Tag

Much has been written about the importance of metadata for understanding and using electronic resources. This literature sheds light on the types of issues that the HTML META tag (see figure 1) is

```
<HTML>
<HEAD>
<TITLE>Poultry, production and value</TITLE>
<META NAME="keywords" CONTENT="USDA, Mann Library, poultry produc-
tion and value, agriculture, livestock, dairy, poultry, agricul-
tural economics, business, trade, commodities, statistics">
<META NAME="Description" CONTENT="This full-text file presents the
annual estimates of production and value for commercial broilers,
eggs, turkeys raised, and chickens sold by states and U.S.  This
report is a supplement to Broiler hatchery, Chickens and eggs, and
Turkey hatchery.">
</HEAD>
<BODY>
<A HREF="http://www.mannlib.cornell.edu/gateway.html">Mann Library
Home Page</A> :
<A HREF="http://www.mannlib.cornell.edu/catalog/catalog.html">Gate-
way</A>
<H1><img src="http://www.mannlib.cornell.edu/icons/world.gif" ALT
="[World]">
Poultry, production and value.</H1>
<HR><P>
<form action="http://www.mannlib.cornell.edu/cgi-bin/connect.cgi"
method="post"><input type="submit" name="Connect" VALUE="Connect">
<input NAME="theID" TYPE="hidden" VALUE="728"></form><br clear=all>
<HR>
<H3>Description</H3>
This full-text file presents the annual estimates of production and
value for commercial broilers, eggs, turkeys raised, and chickens
sold by states and U.S.  This report is a supplement to Broiler
hatchery, Chickens and eggs, and Turkey hatchery.<P>
Resource type: Full text<P>
Update Frequency: Annually<P>
Summary Holdings: 1995-<P>
Publisher: Washington, DC : National Agricultural Statistics Ser-
vice,<P>
<H3>Access Notes</H3>
No access restrictions apply. <P>
<HR>
<B>Crossroads</B>...from here you can:<P>
<DL><DD>Go to other titles of similar subject:
<DD><UL>
<LI><A HREF="http://www.mannlib.cor-
nell.edu/cgi-bin/subj.cgi?ag-econ">Agriculture - Agricultural Eco-
nomics</A>
<LI><A HREF="http://www.mannlib.cor-
nell.edu/cgi-bin/subj.cgi?ag-live">Agriculture - Livestock, Dairy
and Poultry</A>
<LI><A HREF="http://www.mannlib.cor-
nell.edu/cgi-bin/subj.cgi?bus-tra">Business and Economics - Trade
and Commodities</A>
</UL></DL><P>
</BODY></HTML>
```

Figure 1. Example HTML Document with Embedded META Tags.

intended to address. Metadata is commonly defined as data about data. A more complete definition notes that metadata provides "a user (human or machine) with a means to discover that the resource exists and how it might be obtained or accessed. It can cover many aspects, such as subject content, creators, publishers, quality, structure, history, access rights and restrictions, relationship to other works or appropriate audience" (Efthimiadis and Carlyle 1997, 5). Metadata is important for what it enables; its strength is not description but the support it provides for resource discovery and data use (Lynch 1998). Metadata also prevents ambiguity about data (Lide 1995). Weibel (1995) describes metadata as the centerpiece of information gathering. He argues that new types of metadata need to be developed to facilitate document discovery and suggests the Dublin Core element description set as a solution for metadata problems.

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 the META tag. Several authors have suggested that the HTML META tag can be used to enhance information retrieval, especially through Internet search engines. AltaVista Search Network (1997) documentation suggests that authors use the keywords and description attributes of the META tag to improve retrieval and control the description of the document that appears on a search results page. Bremser (1997) offers more detailed advice to Web authors about using different aspects of the META tag.

The META tag has also been seen as a way of providing additional types of metadata about documents. Miller (1995) discusses the potential use of the META tag to contain formatted information defined by the Dublin Core element set. The Dublin Core provides a means of creating basic metadata about a resource in a simple manner and is not formally connected to the HTML META tag. However, the META tag is the best section of the HTML specification in which this data can be placed (Weibel 1997). Many of these authors envision resources that are "self-declaring" because the items provide important information about themselves to human catalogers and automated indexers.

The HTML META tag resides within the header and can have the attributes CONTENT, HTTP-EQUIV, or NAME. It is intended to provide "a place to put meta-information that is not defined by the other HEAD elements. This allows an author to more richly describe the document content for indexing and cataloging purposes" (Graham 1995, 147). In this research, we are most concerned with two attributes: CONTENT and NAME. The NAME attribute requires that a CONTENT attribute also be present. Although the NAME attribute can take the values of author, document type, distribution, keywords, and description among other values, most of the Internet search engines that currently support use of the META tag recognize only those NAME attributes defined as keywords or description. The keywords attribute provides important terms associated with a document, while the description attribute briefly details it and is often used as a summary on the results page generated by Internet search engine queries. This example of a META tag from the header of the USDA report "Agriculture and trade: Europe" illustrates the use of both the keywords and description attributes:

```
<META NAME="Keywords" CONTENT="USDA, Mann Library, agriculture, Europe, agricultural economics, international agriculture, business, economics, trade, commodities, statistics">

<META NAME="Description" CONTENT="Database contains macroeconomic data on Western Europe, budget and price data, and time-series data on supply and utilization of agricultural commodities for the EC-12 and the European Free Trade Association.">
```

Several authors have voiced some concerns about the potential misuse and failure of the META tag. Kuhn (1996) notes that although the META tag can be used for certain information, there is not enough agreement about the types of information that can be implemented. He is especially concerned about information related to authors of documents, abstracts, and document content beyond keywords assigned by authors. One concern about the use of the META tag involves the various opinions about the nomenclature for the NAME attribute. Currently some search engines recognize NAME designated as keywords and description, but other options, such as history, access restrictions, and audience, are ignored. Without consensus about the nomenclature among the HTML standard developers, authors using HTML, and Internet indexing services, the META tag will never be widely implemented (Pfaffenberger 1995). Current court cases will also set precedents for the use of the META tag. Using names in a META tag that have nothing to do with the content of a site has been called into legal question by companies whose names appear in documents with which they have no connection (Kaplan 1997).

Problem Statement

The Albert R. Mann Library at Cornell University works in conjunction with the Economic Research Service, the National Agricultural Statistics Service, and the World Agricultural Outlook Board of the United States Department of Agriculture to produce the USDA Economics and Statistics System (http://usda.mannlib.cornell.edu/usda/). This system provides access to over 300 statistical reports and data sets in various agricultural commodity and business areas. Like other large content providers, producers want these materials to be discovered by Internet users who might not have been previously aware of the service. Users would find these materials relevant whether they were searching for agricultural economics materials in general or specific commodity figures, such as watermelon production statistics.

The META tag might help publishers ensure that their materials are found when appropriate searches are executed. Although the META tag is being put to use by many World Wide Web publishers, its effectiveness has not been evaluated. In this study, we examine the following questions related to the use of the HTML META tag:

1. Do pages that use the META tag have higher retrieval ranks than pages that do not?
2. Is one method of META tag authoring more effective than other methods?
3. Do pages that use both of the META tag attributes have better retrieval ranks than pages that use only one attribute?

To answer these questions, it is necessary to understand how search engines deal with the META tag.

Method

At the time of this research, Mann Library provided access to many USDA reports and data sets, as well as other networked electronic resources, through the Mann Library Gateway (http://www.library.cornell.edu/). All resources previously available through the Mann Library Gateway are now available through the Cornell University Library Gateway (http://www. library.cornell.edu). The gateway is a searchable database of electronic resources that allows users to connect via a hyperlink to resources that match their queries. Searches yield dynamically generated HTML pages with lists of appropriate records. A gateway record lists the title of the work, a description, the publisher, the publication date, update frequency, type of material, summary holdings, access information, and general subject categories. Users can connect to the resource by clicking on a hyperlink from the record. For this experiment, static gateway-like HTML documents were created to test how access to this type of metadata record could be improved.

Twenty HTML documents were created in five subject areas: agricultural

TABLE 1
LIST OF USDA ECONOMIC AND STATISTICS SYSTEM DOCUMENTS USED

Agricultural trade between Asia/Near East countries and the United States
Agricultural trade of former Soviet Republics
Agricultural trade policies "Redbook"
Agriculture and trade: Europe
Cotton and wool outlook
Cotton and wool yearbook
Cotton ginnings
Cotton/citrus production
Farm business balance sheet
Farm operating and financial characteristics
Farm production expenditures
Farm sector balance sheet
Poultry outlook
Poultry slaughter
Poultry yearbook
Poultry production and value
Vegetable yearbook
Vegetables and specialties
Vegetables annual summary
Vegetables

trade, farm business statistics, poultry statistics, vegetable statistics, and cotton statistics. Static pages were used for two reasons: dynamically generated pages are created from a script that would not allow for different HTML markup and, more importantly, many Internet search engines do not index dynamically generated pages. Four pages were created in each subject area: one with no META tags, one with a META tag using the keywords attribute, one with a META tag using the description attribute, and one with META tags using both the keywords and description attributes. The report and data set titles used are listed in table 1. These pages contained all information present for the titles in the gateway record, including a working hyperlink to the resource. These documents were placed on a separate server and were not linked to the gateway. As far as an Internet user finding these pages is concerned, they are functioning gateway records. However, we were able to alter the HTML markup for research control purposes, and the unique location allowed us to determine quickly where these particular documents resided on the ranked list of search results.

Keyword and description attribute terms were chosen from descriptive information available through the documentation for that report or data set as well as from cataloging records created at Mann Library. The number of keyword terms chosen for the keywords attributes ranged from 10 for "Poultry Slaughter," to 40 for "Vegetables." The average number of keyword terms assigned using the keywords attribute was 18.4 and the mode was 12. Keywords ranged from specific terms, such as "watermelon," to abstract or general terms, such as "business." Descriptions matched summaries routinely provided in cataloging records for each item and contained keywords associated with the title.

All 20 pages were submitted to the three Internet search engines that support the use of the META tag—AltaVista, HotBot, and Infoseek—in late December 1996. Of these three services, only AltaVista and Infoseek indexed the pages. Infoseek indexed the pages within two days of submission. AltaVista indexed the documents after a month's delay and after several submission attempts over a six-week period. HotBot failed to index the pages after several requests. From the perspective of a content provider, we suggest that the process for submitting pages and the speed of indexing sites by Internet search engines might be improved.

Once the pages were indexed, searches were performed in AltaVista and Infoseek to find terms common to all pages as well as for terms paired with each keyword and description term contained in the META tags. The only search conducted for all 20 pages was the search for "Mann and agriculture." Searches followed a set format in which "Mann," "statistics," or "USDA" were combined with the keyword terms contained on the page. For example, the searches "Mann and poultry," "statistics and poultry," and "USDA and poultry" were each performed. "Mann and [keyword]" was searched to test a specific term and name with a variety of specific and general keywords. "Statistics and [key-

TABLE 2
SAMPLE QUERIES USED TO SEARCH FOR USDA ECONOMIC AND STATISTICS SYSTEM DOCUMENTS

Mann and [Keyword]	USDA and [Keyword]	Statistics and [Keyword]
Mann USDA	USDA Mann Library	Statistics Mann Library
Mann statistics	USDA statistics	Statistics USDA
Mann poultry production	USDA poultry production	Statistics poultry production
Mann value	USDA value	Statistics value
Mann agriculture	USDA agriculture	Statistics agriculture
Mann livestock	USDA livestock	Statistics livestock
Mann dairy	USDA dairy	Statistics dairy
Mann poultry	USDA poultry	Statistics poultry
Mann agricultural economics	USDA agricultural economics	Statistics agricultural economics
Mann business	USDA business	Statistics business
Mann trade	USDA trade	Statistics trade
Mann commodities	USDA commodities	Statistics commodities

word]" was searched to test a general term search with a range of specific and general keywords. "USDA and [keyword]" was searched to test a commonly expected general name with a number of specific and general keywords. Table 2 lists a sample of queries. In total, 579 search combination results were recorded.

All Infoseek searches were completed in January and early February 1997. All AltaVista searches were completed during late March and April 1997. The length of time required to complete the searches was due partly to delays in indexing and partly to the time required to complete all searches. More efficient automated means of checking the ranks of pages, such as software like Webposition Analyzer, were not available at the time the searches were conducted. These delays are not expected to have had a significant impact on the results observed because searches repeated at intervals during this process did not reveal any changes in ranking. Once search results were achieved, the first 200 results were examined to determine which pages fell within those retrieved-item lists. If a page was found, the rank of the page within that list was recorded.

To evaluate the effectiveness of using the HTML META tag to improve retrieval of HTML documents searched on Internet search engines, the ranks of pages retrieved by both AltaVista and Infoseek were recorded. For each set of markup comparisons, searches were examined in which the terms appeared on both of the pages analyzed either in the text or in the description or keywords attributes of the META tag. To be considered for analysis in a given comparison, the search must either have retrieved both pages being compared or have been expected to retrieve both pages.

By basing queries on terms known to be present in these documents, several concerns must be noted. In his analysis of the ASLIB Cranfield Research Project, Swanson (1965) argues that it should not be assumed that nonsource documents behave in the same manner as source documents. Swanson also notes that the opportunity to find unexpected results might be lessened if the focus remains on a relatively small set of documents and search terms. In addition, terms searched might not reflect the wide range of terms used in actual searches from a diverse user population (Furnas et al. 1987). Searches performed in this research were designed to determine how search engines

TABLE 3
NUMBER OF RANK SCORES COMPARED IN MANN-WHITNEY U TESTS BY META TAG COMPARISONS

Attributes Compared	Number of Rank Scores Compared (sample size)
AltaVista and Infoseek Ranks Combined	
None versus keywords	198
None versus description	96
None versus both keywords and description	194
Keywords versus both keywords and description	506
Description versus keywords	166
Description versus both keywords and description	174
AltaVista Ranks Only	
None versus keywords	96
None versus description	44
None versus both keywords and description	96
Keywords versus both keywords and description	252
Description versus keywords	84
Description versus both keywords and description	88
Infoseek Ranks Only	
None versus keywords	102
None versus description	52
None versus both keywords and description	98
Keywords versus both keywords and description	254
Description versus keywords	82
Description versus both keywords and description	86

Notes: No fewer than 22 searches were run for each set of comparisons. Each search generated ranks for each of the pages being compared. As a result, the number of searches is half that of the number of ranks compared. For example, 42 searches resulted in 84 ranks being generated: 42 ranks for pages with the description attribute and 42 ranks for pages with the keywords attribute.

index page text in relation to META tag text rather than to simulate user behavior. The results generated reflect an ideal scenario. Page ranks might be lower with more diverse search terms and search engine retrieval might be less effective under those conditions.

The first 200 documents retrieved were examined and the ranks of source pages within those first 200 were recorded. In cases in which more than 200 sites were retrieved, the first 200 were examined as an arbitrary cutoff point supported by all the search engines used. Harman (1993, 371) reported using 200 as a retrieval threshold, although she concluded that for the purposes detailed at the first Text Retrieval Conference, this point was too low. If a search could have resulted in retrieving a page but it was not in the top 200 sites, we gave that page the rank of 201 for that search. As a result, rankings ranged from 1 (highest) to 201 (not retrieved) and the number of searches analyzed in each comparison varied.

The ranks of the differently coded pages were compared using the Mann-Whitney U test. The U statistic measures "the number of times that the rank of a score in one group precedes the rank of a score in the other group" (Kiess 1989, 468). This provides for a comparison of two sets of ranked scores to determine whether or not the sets can be expected to fall within the same distribution. If the ranks are part of the same statistical distribution of ranks, then the addition of

TABLE 4
MANN-WHITNEY U TEST RESULTS FOR META TAG COMPARISONS

Attributes Compared	Significance Level (p=)
AltaVista and Infoseek Ranks Combined	
None versus keywords	.0000*
None versus description	.3900
None versus both keywords and description	.0000*
Keywords versus both keywords and description	.3913
Description versus keywords	.0000*
Description versus both keywords and description	.0000*
AltaVista Ranks Only	
None versus keywords	.0000*
None versus description	.3806
None versus both keywords and description	.0000*
Keywords versus both keywords and description	.2898
Description versus keywords	.0000*
Description versus both keywords and description	.0000*
Infoseek Ranks Only	
None versus keywords	.0000*
None versus description	.3605
None versus both keywords and description	.0000*
Keywords versus both keywords and description	.4439
Description versus keywords	.0000*
Description versus both keywords and description	.0000*

*Indicates significance at the .01 level.

META tags has probably not affected the retrieval rank. If the U test shows that the ranks are most likely not from the same distribution, then the type of META tag markup used has probably affected the rank.

The Mann-Whitney U test was chosen because it tests rankings of at least ordinal-level data. The rankings received from the search engines were considered ordinal-level data because AltaVista and Infoseek have different algorithms for ranking materials. In addition, the degree of relevance attributed to a site by a search engine is not directly correlated to its rank. For instance, one search might yield a result in which a site ranked first is considered 100% relevant to the search query, while in another search, a site given the first ranked position is considered 75% relevant to the query. As a result, the distance between ranks is not consistent for all searches analyzed. Ranks were recorded rather than relevance percentages because not all search engines provided relevance percentage information.

The Mann-Whitney U test was run to compare several sets of search result rankings for pages with: no META tag and keywords attribute META tags; no META tag and description attribute META tags; no META tag and both keywords and description attribute META tags; keywords attribute META tags and both keywords and description attributes META tags; description attribute META tags and keywords attribute META tags; and description attribute META tags and both keywords and description attribute META tags. In each test the number of times that the rank of one set of scores exceeded the rank of another set of scores was tallied. The results are summarized in tables 3 and 4.

Table 3 lists the number of searches whose rankings were compared in each pair.

TABLE 5
MEDIAN, MODE, AND RANGE FOR META TAG COMPARISONS
(ALTAVISTA AND INFOSEEK COMBINED)

Attributes Compared	Page Type	Median	Mode	Range
None versus keywords	None	59	7	199
	Keywords	8	3	181
None versus description	None	39.5	3	197
	Description	43	4*	197
None versus both keywords and description	None	79	15	198
	Both	14	2*	193
Keywords versus both keywords and description	Keywords	14	2	197
	Both	20	1	196
Description versus keywords	Description	65	14	198
	Keywords	13	3	109
Description versus both keywords and description	Description	57	4	198
	Both	12	2	192

*Multiple modes exist. Smallest mode is shown.

No fewer than 22 searches (yielding 44 rankings) were observed for each set of comparisons. Table 4 lists the results of the U test for each comparison by recording whether a statistically significant difference was noted at the .01 level. Sets of ranks are considered statistically significant in their differences if the smaller U score observed is less than or equal to the critical U value for that sample size.

The U tests were run for all searches in AltaVista and Infoseek combined as well as for AltaVista and Infoseek results separately. Results are reported for AltaVista and Infoseek ranks combined to determine how well META tags work regardless of the search engine used. AltaVista and Infoseek were also considered separately to determine whether one search engine's performance skewed the combined rankings. These results show how Internet search engines deal with META tag data rather than how well Internet search engines work to retrieve known items.

In addition to the Mann-Whitney U test, the medians, modes, and ranges of ranks were generated to test the practical significance of the findings from the Mann-Whitney U test. If the U test suggests a statistically significant difference but the median, mode, and range are similar, then the U test difference may be said to have less practical significance. However, if the U test suggests a statistically significant difference and the median, mode, and range are different, then the U test can be said to be reflect a practical as well as statistically significant difference. The median, mode, and range values are summarized in tables 5, 6, and 7.

DATA ANALYSIS: COMPARISONS OF PRESENCE AND ABSENCE OF META TAG

The use of the HTML META tag was expected to improve the ranking of a document when searched using an Internet search engine. We tested this assertion by measuring the rankings of three sets of comparisons: pages with no META tag to those with keywords attribute META tags, pages with no META tag to those with description attribute META tags, and pages with no META tag to those with both keywords and description attributes META tags. The inclusion of the keywords attribute, with or without the description attribute also present, consistently improved the accessibility of HTML documents. Unexpectedly, however, the inclusion of only the description

TABLE 6
MEDIAN, MODE, AND RANGE FOR META TAG COMPARISONS
(ALTAVISTA ONLY)

Attributes Compared	Page Type	Median	Mode	Range
None versus keywords	None	153.5	156	197
	Keywords	10	5	181
None versus description	None	127	148	196
	Description	134	65*	197
None versus both keywords and description	None	153	156	198
	Both	25.5	12	199
Keywords versus both keywords and description	Keywords	23.5	4	181
	Both	34.5	3	193
Description versus keywords	Description	197.5	198	195
	Keywords	17.5	5*	199
Description versus both keywords and description	Description	196.5	198	199
	Both	34.5	4	199

attribute did not improve ranking of a page over a page with no META tag.

Pages containing the keywords attribute META tag were consistently ranked by Internet search engines as more relevant than pages lacking a META tag. Page rankings were found to have statistically significant differences at the .01 level. The U test results of pages employing only the keywords attribute versus pages lacking a META tag reflect a comparison of 198 search ranks. The results detailed in table 5 suggest that pages that use the keywords attribute META tag are consistently ranked as more relevant than those without the META tag. In addition, the disparity among the median, mode, and range for these two groups (table 5) is consistent with the findings of the Mann-Whitney test. Similar results were obtained when comparing the ranks of pages with no META tags and of pages with both the description and keywords attributes META tags. This comparison of 194 search results (tables 3 and 4) revealed that having both attributes in the META tag resulted in higher rankings than having no META tag present. As was the case with the keywords attribute only results, the median, mode, and range (table 5) were sufficiently different for each set to reinforce the results of the Mann-Whitney test.

Surprisingly, the comparison of the ranks of pages with no META tag and of pages with only the description attribute META tag revealed no significant difference. The ranks of 96 searches were compared, with a U test result that failed to reach the .01 level (table 4). In addition, median, mode, and range were very similar for both sets of ranks (table 5). This suggests that the effect of the description attribute was negligible.

The U test and median, mode, and range were generated for AltaVista and Infoseek scores separately to verify that the similarities noted in all scores combined were not the result of the bias of one particular search engine. The U test for AltaVista and for Infoseek (table 4) did not reveal a statistically significant difference between these sets of ranks. Moreover, the median, mode, and range for both AltaVista (table 6) and Infoseek (table 7) validated the U score results. Although it is possible that the AltaVista scores, which reflected a higher median for ranks of pages without a META tag than for pages with the description attribute alone, may have influenced the median for the scores combined, it is not likely that the results from either AltaVista or Infoseek biased the combined totals.

TABLE 7
MEDIAN, MODE, AND RANGE FOR META TAG COMPARISONS
(INFOSEEK ONLY)

Attributes Compared	Page Type	Median	Mode	Range
None versus keywords	None	26	2*	199
	Keywords	5	1	119
None versus description	None	18	3*	79
	Description	19	4*	82
None versus both keywords and description	None	37	3*	198
	Both	9	2	168
Keywords versus both keywords and description	Keywords	12	4	197
	Both	11	2	196
Description versus keywords	Description	27	19	198
	Keywords	8	1	103
Description versus both keywords and description	Description	22	19	197
	Both	4	2	40

*Multiple modes exist. Smallest mode is shown.

The differences noted here might be due to the ways in which Internet search engines index and weigh the text from a page in comparison with the text in a description attribute. According to information that we received from AltaVista Support, if the description attribute is indexed, the page text is not (Alta Vista Support 1998). This might be the case with other search engines as well and would naturally favor pages with no META tag over those with only the description attribute META tag because multiple occurrences of the same terms might be more likely in the text of a page than in the description attribute.

DATA ANALYSIS: COMPARISONS OF DIFFERENT META TAG ATTRIBUTES

The Mann-Whitney U test was also performed on the ranks of pages using various attributes within the META tag. U scores were generated to compare these rankings: pages with only the keywords attribute META tag to those with both keywords and description attributes META tags; pages with only the description attribute META tag to those with both keywords and description attributes META tags; and pages with only the keywords attribute META tags to those with only the description attribute META tag. It was assumed that using both the keywords and description attributes would improve retrieval in relation to those pages with only one type of META tag. In addition, it was assumed that the keywords attribute META tag would result in better retrieval than the description attribute META tag.

No statistically significant differences were found between the ranks of pages with both the keywords and description attributes and those of pages containing only the keywords attribute in the META tag. We expected that pages with META tags containing both keywords and description attributes would result in higher rankings than those containing only the keywords attribute. Although the U test did not uncover a statistically significant difference between the sets of scores (table 4), the median rank for the keywords attribute only pages was somewhat better than the median rank for pages with both the keywords and description attributes (table 5).

The ranks of pages using only the description attribute META tag were compared to ranks of pages using only the keywords attribute and to ranks of pages using both the keywords and description attributes. In both cases, the U test recorded a statistically significant difference (table 4). Additionally, the median ranks observed in both cases validated the U score results (table 5). The evidence suggests that using the keywords attribute, either with or without the description attribute, improves retrieval rank over using only the description attribute. This might be the result of the weight given to page text versus the weight given to description attribute text, or it might reflect the decision by search engine designers not to index the page text when a description attribute is present.

Conclusion and Future Work

Enabling users to find materials on the World Wide Web is an important problem faced by librarians, search engine designers, and Internet publishers and content providers. As new standards in metadata emerge, such as the Dublin Core (Weibel 1995) and Extensible Markup Language, or XML (Flynn 1998), it will be possible to embed richer types of metadata into documents, thereby improving automated indexing processes. The goal of this research was to determine how useful one current method, the HTML META tag, is in improving accessibility via Internet search engines. Because only AltaVista, Infoseek, and HotBot currently recognize and use the META tag, we suggest that search engine designers enable their services to accept META tag data because it does benefit retrieval rank. Concern over improper uses of the META tag do not justify failure to index them when appropriately used.

This research serves as a snapshot of how current forms of embedded metadata are processed by Internet search engines. Newer technologies and methods for embedding and indexing World Wide Web documents are evolving that will alter the view presented here. The searches used to test indexing methods reflect an idealized situation because terms searched were known to be present in the documents sought. Retrieval rates based on more realistic search scenarios might reveal lower rankings. It was found that using the keywords attribute of the HTML META tag, with or without the description attribute, consistently improved the retrieval rank of a Web document. Mann-Whitney U test comparisons of rankings of pages with the keywords attribute versus those with no META tags revealed a statistically significant difference at the .01 level. However, using only the description attribute in the META tag did not appear to improve retrieval over not using a META tag. The Mann-Whitney U test did not reveal a statistically significant improvement in retrieval rank between pages using only the description attribute and pages with no META tag. Furthermore, the U test shows that pages with only the description attribute were given consistently less-relevant ranks compared to pages employing either the keywords attribute alone or pages containing both keywords and description attributes.

This discrepancy may reflect different ways in which Internet search engines index and weigh the text from a page and the text in a description attribute and the failure to index page text when the description attribute is present. This process favors pages with no META tag over those with only the description attribute META tag because there can be multiple uses of the same terms in the text of a page while the description attribute is likely to use a given term fewer times. It is important to bear in mind that the description attribute is designed to provide a display summary of the resource rather than to improve retrieval of a document. We suggest that search engine designers consider indexing the full text of a page regardless of the presence of a description attribute or improve the relevance assessment of text in the description attribute. The Mann-Whitney U test did not uncover a statistically significant difference between the ranks of pages with only the keywords attribute and pages with both the keywords and description attributes. We suggest that World Wide Web authors use at least keywords attribute META tags in their documents.

More research needs to be done to determine the types of keywords that are most effectively used in HTML META tags. In this research, we relied upon idealized search situations because all the terms searched were known to be present in the documents indexed. Further research is needed to determine whether META tag data assists users in finding known documents when a more diverse group of search terms is used. The method used here might have resulted in better retrieval rates than more realistic searches would generate. Not surprisingly, this research also suggests that because searches for abstract or general terms retrieve large result sets, the inclusion of a META tag might not have practical significance in improving.the retrieval rank of a page. The inclusion of many abstract terms in a META tag might not raise retrieval rank.

It is clear, however, that the inclusion of specific terms in combination with more abstract ones will have an impact on retrieval rank when less general terms are searched. The question of the usefulness of abstract and specific keyword terms requires more study. Finding new ways of embedding significant metadata into documents will enhance the experiences of both content providers and users of Internet search engines.

Works Cited

AltaVista Search Network. 1997. The META tag: Controlling how your Web page is indexed by AltaVista. Online. Available: http://www.altavista.digital.com/av/content/addurl _ meta.htm.

AltaVista Support. 1998. E-mail message to the authors. 4 June.

Babbie, Earl. 1992. *The practice of social research*. 6th ed. Belmont, Calif.: Wadsworth Publishing Co.

Bremser, Wayne. 1997. Gain fame with META tags. *Internet world* 8, no. 10: 94–96.

Efthimiadis, Efthimis N., and Allyson Carlyle. 1997. Introduction to special section: Organizing Internet resources: Metadata and the Web. *Bulletin of the American Society for Information Science* 24, no. 1: 4–5.

Flynn, Peter. 1998. Frequently asked questions about the Extensible Markup Language. Version 1.3. Online. Available: http://www.ucc.ie/xml/. 1 June.

Furnas, G. W., T. K. Landauer, L. M. Gomez, and S.T. Dumais. 1987. The vocabulary problem in human-system communication. *Communications of the ACM* 30: 964–71.

Gibbons, Jean Dickinson. 1993. *Nonarametric statistics: An introduction*. Quantitative applications in the social sciences, no. 90. Newbury Park: Sage Publications.

Graham, Ian S. 1995. *The HTML sourcebook: A complete guide to HTML 3.0*. 2d ed. New York: John Wiley and Sons.

Harman, D.K., ed. 1993. *The first Text Retrieval Conference (TREC-1)*. Gaithersburg, Md.: U.S. Department of Commerce, National Institute of Standards and Technology.

Kaplan, Carl S. 1997. Cyber law journal: Legal roadblocks starting to deter META-tag hijacking. New York Times. 16 Oct. Online. Available: http://search.nytimes.com/library/cyber/law/101697law.html.

Kiess, Harold O. 1989. *Statistical concepts for the behavioral sciences*. Boston: Allyn and Bacon.

Kuhn, Heinrich C. 1996. A proposal for structured indexing with the meta-tag. Online. Available: http://www.gwdg.de/~hkuhn1/wwwcat/ mtprop01.html.

Leighton, H. Vernon, and Jaideep Srivastava. 1997. Precision among World Wide Web search services (search engines): AltaVista, Excite, Hotbot, Infoseek, Lycos. 16 June. Online. Available: http://www.winona.msus.edu/library/webind2/ webind2.htm.

Lide, David R. 1995. Metadata: A description. *Library hi tech* 13, nos. 1/2: 33–34.

Lynch, Clifford. 1998. The Dublin Core Descriptive Metadata Program: Strategic implications for libraries and networked information access. ARL: A bimonthly newsletter of research library issues and actions no. 196: 5–10. Available: http://www.arl.org/newsltr/196/dublin.html.

Melee's Indexing Coverage Analysis. 1998. The MICA report. 9 Aug. Online. Available: http://www.melee.com/mica/index.html.

Miller, Eric J. 1995. Network centric computing: Issues of document description in HTML. Online. Available: http://www.oclc.org/oclc/research/publications/review95/part1/html.htm.

Pfaffenberger, Brian. 1995. *Web search strategies*. New York: MIS.

Phillips, John L., Jr. 1992. *How to think about statistics*. Revised ed. New York: W.H. Freeman and Co.

Pollock, Annabel, and Andrew Hockley. 1997. What's wrong with Internet searching. *D-Lib magazine*. March. Online. Available: http://www.dlib.org/dlib/march97/bt/03pollock.html.

Srinivasan, Padmini, Miguel E. Ruiz, and Wai Lam. 1996. An investigation of indexing on the WWW. In *Global complexity: Information, chaos, and control: Proceedings of the 59th ASIS annual meeting, Baltimore, Md. October 19–24, 1996*, ed. Steve Hardin: 79–83. White Plains, N.Y.: Knowledge Industry Publications.

Swanson, Don R. 1965. The evidence underlying the Cranfield results. *Library quarterly* 35: 1–20.

Weibel, Stuart. 1995. Metadata: The foundation of resource description. *D-Lib magazine*. July. Online. Available: http://www.dlib.org/dlib/July95/07weibel.html.

———. 1997. The Dublin Core: A simple content description model for electronic resources. *Bulletin of the American Society for Information Science* 24, no. 1: 9–11.