Andre S. Burton. Meta Tag Usage and Credibility Factors in Alternative Medicine Websites. A Master's paper for the M.S. in I.S. degree. April 2004. 29 pages. Advisor: Paul Solomon

Clearly, the wide range of health information sources on the World Wide Web has the potential to lead to distribution of inaccurate medical information from unqualified sources bringing a great risk. Given the growing number of Internet users that access health-related information, the need for a more standard means to validate web site content is apparent. This paper examines how source, information, timeliness, accessibility, and design factors impact web document credibility on a narrower health topic - Alternative Medicine. It also examines the contrasts of different levels of credibility with metadata usage as well as the relationships between metadata usage measures. These preliminary results and examinations give an overview of how metadata is currently being used in this subject area.

Headings:

Metadata

Meta Tags

Alternative Medicine

Health Websites

META TAG USAGE AND CREDIBILITY FACTORS IN ALTERNATIVE MEDICINE WEBSITES

by
Andre S. Burton

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
Information Science.

Chapel Hill, North Carolina

April 2004

Approved by:

_____

Paul Solomon

Table of Contents

I. Introduction

Searching for documents containing query terms is conceptually simple. A search for health information may return information that covers topics ranging from supplementary treatments and acupuncture methods to leukemia depending on the terms employed. General inquiries on public search engines often return a plethora of information that can be difficult to filter through, oftentimes discouraging use. [1]

The returned websites may also vary in quality and it is difficult to quantify: "making quality and authority judgments for most users is a difficult task due to the absence of a quality control mechanism for the web" [13]. Within the world of print media, journals have long been the traditional form of acceptance and validation for field-specific research and theories. Journals, through the use of selected peer review groups, lend their credibility, and thus pass the approval of the institution along to the researchers and readers who rely on the publications to keep them apprised of emerging high-quality research in respective fields. However, because they are influenced by institutional guidelines, even peer review boards operating in the same fields use different criteria to evaluate the 'worthiness' of research articles. These criteria may vary in terms of the scales and measures employed; additionally, the respective weighting of these criteria may also vary from person to person due to personal experiences, interests, and other types of individual bias.

Without intimate knowledge of a topic -- which most web users lack -- quality assessments are harder to make. Rieh identifies medical information as a topic most

web users have less experience with: “most of the subjects were not familiar with a medical domain in general, and therefore, had difficulties in judging the goodness of information” [13].

Credibility, on the other hand, is less specific to an information topic, and even those with a less esoteric knowledge of a field can detect a flagrantly incredible website. The characteristics of credible information are more easily ascertained by a layperson, and these features are closely related to source reputation, site organization, information reliability, and the responsible dissemination of data, namely, through the use of statistics that support the information reported there-in, quality disclosures, legal disclosures, and money-back guarantees or warranties of any type.

Clearly, misleading health information has the potential to distribute inaccurate medical information from unqualified sources bringing a great risk. “An estimated 43% of Internet users go on-line to gather health information on over 34,000 health-related sites” [3], and, though most users may seek to glean information related to a known diagnosis or treatment or to self-diagnose minor ailments so that they can seek professional treatment, others may attempt to use the information to self-diagnose and even self-medicate serious medical conditions. Importantly, as a measure of self-protection, consumers of health information identify credibility as an important aspect of useful information. It is assumed that consumers would not value information that they deem “incredible”.

A need for a standard means to validate website content is apparent, and though quality certification for health-safety information is the best option, it would

involve a quagmire of paper work, the development of third party reviewers, and would require disclosure of research information—the certification process would be economically impractical.

The burden of presenting credible documents for the World Wide Web currently lies with—and will probably continue to fall upon—the site developer. To enhance the acceptance of health related information, it would be to the developer's advantage to drive traffic to sites in a respectable manner, and to build a credible site.

Though there is research on web credibility assessment, identification of key credibility factors remains to be established. Before the web documents can be regarded as reliable information sources, users and developers will require a consistent standard for evaluating and creating credible Internet information.

II. Literature Review

A. Resource Author Metadata

With the rapid growth of the Web, comes the growing need to effectively organize and classify documents. Problematically, the Internet's seemingly exponential growth is not being matched by the population growth of professional catalogers or indexers due the lack of funds. Ideally, we would like to automate the process of indexing electronic documents—automation would enable a greater number of documents to be indexed. However, automation is not the only possible solution to organization, alongside the development of automated cataloguing methods, there is increasing interest in metadata—data about data—as a "means of imposing pre-defined structure on Web content". A "minimal amount of information

ordering, such as that represented by certain metadata standards, may vastly improve the quality of an automatic index at low cost" [8]; especially since, at times, "some types of simple description may be indexed with little or no human intervention." [8]

Historically, metadata has been an effective tool in classifying and retrieving information; we not only use metadata to classify electronic documents, we use metadata to simplify our daily tasks. Imagine a book without metadata—without a table of contents or an index. Finding a particular topic within a book would be an extremely frustrating process as one would have to search the entire volume for a desired subject—this intricacy is eliminated with the use of metadata. Expounding on the example of a book, a library without a catalog describing the contents would further complicate the task of the user.

Unfortunately, the transparency afforded by a table of contents or the Dewey decimal system is not so easily achieved when retrieving documents and site components from the Internet. This is due, in part, to the ease of digital publication: anyone can build a presence on the internet, and this openness results in a flood of works—from shanties and stick drawings to architectural wonders and masterpieces on a massive variety of topics. Without a filter, or a standard set of guidelines for developing accurate descriptions for documents, the Internet will continue to expand unmanaged and transform from a forum of education, communication, and, most recently, self-expression to a massive production possessing the spite of Babel, a bad opera that lacks a proper script and appropriate stage direction.

Establishing a metadata standard is the first piece of a possible solution to document classification and retrieval on the Web. The "big task at hand" then is the

actual metadata creation: who will create the metadata, and how will they measure the "quality" of the "validity" of that metadata?

Professional metadata indexers have a good grasp, though their methodology is not quantified, of what makes "quality metadata." [11] Unfortunately, the profession does not lead to a proliferation of knowledge, the professional indexer's approach to metadata creation is subjective, and therefore somewhat mystical; it begs adequate scientific analysis.

Even if professional metadata indexers did follow a scientific approach, in its current state, the professional indexer's field is very limited and with scarce population and supply that enable those in the profession to demand "exorbitant" salaries. Without the inundation of new professionals, the cost of professional indexing is unlikely to decline sharply, and the widespread utilization of professional metadata indexers remains uncertain, at best, leaving smaller organizations without the means and justification to gain access to professionally generated metadata, and larger organizations hesitant to sacrifice the expenditures necessary to obtain the benefits of metadata indexing.

Economically, metadata creation would be much more effective if adeptly performed by resource authors, but because the method of creation is currently subjective, and lacks guidelines, most resource authors may be incapable of producing accurate, consistent, value-adding metadata.

Resource authors possess a familiarity and understanding of their works that no other can surpass [4]. Assuming that the author is motivated to gain respect or to properly serve the communities in which their works are utilized, authors already own

the “will” to accurately and appropriately describe their works so that they can communicate and share with the proper audience. They have a “way” to create the necessary traffic by using metadata. “How” they will achieve their goals remains hidden in the quality measurement of metadata, and current practices for metadata usage.

The “how” of micromanaging the Internet will prove to be a crossroad in establishing a large-scale organization scheme for Internet documents.

Despite the ability of authors’ to create acceptable metadata, a growing need for understanding and outlining what acceptable metadata is has yet to be discovered. Understanding metadata usage of resource authors may attempt to answer this problem. Before offering a template for metadata creation, we first have to examine how resource authors currently use metadata to define Internet documents.

In this study, we will also analyze the usage of metadata to note if it is currently used as an effective tool in document indexing for “alternative medicine websites.” The study will focus on a restricted set of criteria for metadata implementation involving term frequency and overlap among other factors.

B. HTML Meta Tags

Metadata exists in many shapes and formats; however, computers require conformity and a format that is standardized and highly structured before data recognition and processing capabilities are enabled. HTML, the most commonly used web format for displaying documents fails in conformity—it attempts to combine structure, style, and semantics—and makes it difficult for computers to

separate the human-readable from the machine-readable content. In terms of functionality, HTML has primarily served as a means of sharing documents among humans, not computers.

Luckily, the HTML standard supports a limited metadata resource definition. Meta information in the head tag is primarily used to communicate summary information about the page to indexers and robots, not users. For example, META tags can tell a search engine robot in which languages the document is inscribed or what authors to associate with the document. As an example, let us assume a web page author named Paul Jones wants to use META tags to describe his HTML web page entitled “Cool Penguins” written in English:

<META NAME=”language” CONTENT=”en-us”>

<META NAME=”author” CONTENT=”Paul Jones”>

<META NAME=”title” CONTENT=”Cool Penguins”>

<META NAME=”format” CONTENT=”html/text”>

The NAME attribute is used to name a property such as author or language. Interestingly, there are no restrictions on the values for this attribute. Although an optional scheme may define values, placing any text or metadata field as its value will not result in an invalid HTML document.

Despite the lack of control on META NAME attribute values, there are commonly used values used in combination with CONTENT values to assist in identifying the properties of a document as seen in the previous Paul Jones example. A common use for the META tag is to specify keywords “that a search engine may use to improve the quality of search results” [6]. Keywords are often delimited using

commas allowing for single and compound terms; unfortunately, the comma-delimited phrase rule is not instituted by the HTML standard, neither are there specifications from search engines on what they expect a keyword set to look like. Another common META tag is description; however, since the description META tag information is normally returned as part of the result-set [7], description META tags may not be evaluated by search engines.

Like the META tag, the TITLE tag is also contained in the head area of the page. Although it is a separate HTML tag, it is important to mention as it too is commonly incorporated into web pages as metadata.

C. Low Usage

Many authors do not incorporate META tags into their web pages for a number of possible reasons. Statistically speaking, the Lawrence study estimates META tag usage of server homepages at 34.2%, and of that percentage, only a sparse .3% uses a metadata standard scheme such as Dublin Core [9].

While many search engines and directories recognize META tags in determining placement, there are numerous others that ignore META tags altogether [11]. According to the Sullivan articles, only Inktomi and Teoma utilize keyword META tags – ironically, neither of these search engines have significant ratings when it comes to the total number of search hours used searching them. The more prominent search engines seem to completely ignore the tag altogether. Additionally, many of these engines implement spider programs that scan through web sites and index them according to a certain portion of the written content [12]. These web

search engines are primarily “being designed to search on ill-assorted collections of unstructured text”. [11]

The proprietary nature of search engines leaves room for uncertainty in the assessment of META tag employment. An author posting META tags to a web page knows little about what search engines look for or what the optimal set of tags look like. According to Lawrence, a great diversity in META tags was identified, “with 123 distinct tags, suggesting a lack of standardization in usage [9]”. With 123 distinct tags identified, the question of whether search engines utilize standards remains unknown and highly variable.

“Overall, few incentives have been shown for encouraging metadata creation, or the Internet already would be filled with resource descriptors. Ultimately, the only reasonable way to encourage widespread metadata implementation is to provide a strong potential for profit from use of the information” [16].

If correlated, metadata usage may be a useful factor in determining credibility. Indexers may be able to use metadata usage trends from the web sites, and based on credibility, they may be able to filter out or give lower rankings to sites partially based on metadata usage.

III. Methodology

A. Overview

In order to get an idea of how metadata is being used in the Alternative Medicine Internet community, a sample of web sites was gathered by inquiry from two search engines. In November 2002, fifteen sites were collected using two sets of

search phrases from two popular search engines, google.com and altavista.com; from these sample sites' home pages, two content-level pages were randomly selected and added to the sample. For example, we selected the NCCAM home page, http://nccam.nih.gov/, from the first search engine. Then, the http://nccam.nih.gov/health/hepatitisc/ and http://nccam.nih.gov/health/stjohnswort/ pages were randomly selected as the two content-level page requirement. See Appendix B for complete listing of web pages.

A rating scale for assessing the credibility and a set of measures for quantifying metadata usage of these web sites were then developed. From the sample, the websites' credibility and metadata usage were evaluated and tabulated, and then compared.

B. Establishing Measures for Credibility

We used an already existing credibility assessment scale as our foundation. Using this scale as a foundation, we added and subtracted factors where necessary. We looked at scales geared toward assessing credibility of web sites and selected the criteria developed by Web Feet—an organization that evaluates subject websites for library and school use—as our basis of assessing credibility of Alternative Medicine sites [18]. Acting as subject matter experts, librarians, educators, subject-area specialists, and editors review the collection of sites for appropriateness and pertinence using this established set of criteria or factors. Some of the factors include:

- Timeliness – site updated frequently, page lists date of most recent update

- Source – source of information is identified, primary sources used, contact information is displayed, expertise and reputation of source and host.

Existing literature identifies a number of factors that users claim add or subtract from the credibility of a web site, including: content, design and aesthetics, disclosure of authors, site sponsors, developers, currency of information, authority of source, ease of use and accessibility, and availability [10], and indicate that the most important source characteristics for medical information are: currency, accuracy, cognitive accessibility, credibility, physical accessibility, relevance, and confidentiality [2]. Most are criteria for credibility included in the Web Feet Criteria for Site Selection (see Appendix A for details). To ensure consistency among the measures, each factor was broken down into more granular measures or sub factors as identified by the Web Feet Criteria for Site Selection. Applicable factors that had a significant impact on the overall credibility score in the Fogg study [5] that were not included in the Web Feet Criteria were added to our measures. Some of the additional sub factors include:

- Information – links to a flagrantly incredible site
- Accessibility – paid access to some content

The factors developed by the Fogg study have broad implications: designers can use the factors as guidelines in developing credible websites and researchers can utilize them to fit their research needs. Although the study did not incorporate metadata usage as a credibility factor, it further established a foundation for this study in assessing credibility factors that affect Alternative Medicine web sites.

C. Establishing Measures for Metadata Usage

We used simple and objective, quantitative measures for identifying metadata usage.

1. Phrase count: a count of the number of phrases used delimited by a comma
2. Number of repetitions: number of times where any of the phrases repeated (exact match)
3. Home overlap: percentage of the keywords used in page A overlapped or were repeated in page B – a home to content-level contrast
4. Content-level overlap: same as 1st Overlap percentage except makes content-level to content-level contrast
5. Body word count: a count of the number of keywords found in the body of the page
6. Title tag content: did the tag contain information pertaining to content, title of the organization, source, or general subject terms?

In evaluating the influence of metadata usage on credibility, quantitative measures were needed to compare the two web site credibility categories, low versus high. With both sets of measures established, questions such as: 'Do highly credible sites use a greater number of phrases?' or 'Do lowly credible sites use a greater number of repetitions in keywords used?', could then be addressed.

D. Collecting the Sample

Although methods of assessment by search engines are proprietary and publicly unknown, it is likely that, because people tend to narrow a large number of

search results by considering only the first few pages, most search engines return a gradient of results that list the 'best' sites as the first string of returns, and then other relevant content afterwards. Due to the overwhelming number of bad quality web sites on the Internet, we assumed that the more popular search engines give higher rankings to the more "credible" websites, and collected our sample from the upper-portion of the search engine result sets—this assumption was not corroborated by this review. From each of the two result sets, every third site was reviewed to ensure that all of the following criteria were met:

1. Content presented as factual, or as consumer advice
2. Contained META keyword tags.
3. Were English
4. Did not return an HTTP error
5. Already collected from previous searches
6. Were not sponsor links

When the prospective sample did not meet the criteria, the next site was reviewed, and if suitable, collected.

E. Adding up the Scores

The granular measures for each factor were placed on an interval scale from 0 to 10, where 0 denotes the lowest score and 10 the highest.

Each site generated a total credibility score from the addition of all the measures. Due to the uneven number of measures, the total score for each factor was

the average score for each granular measure. This was essential for regression analysis.

Home and content-level pages were evaluated differently; in some cases, factor measures differed in semantics and presence. Therefore, content-level to home page level contrasts were invalid.

F. Analysis Procedures

All statistical analysis was conducted using SPSS v 10.0.7. Logistic regression analysis compared the total credibility scores to a superficial measurement of credibility to confirm the validity of the data. We wanted to ensure that the data measured up to the perceived or superficial measure of the site—i.e. the credibility scores from the rating scale should make sense.

Once the logistic regression model was validated, the next step was to evaluate the impact of each factor on the total credibility score among content-level and home pages separately. Stepwise linear regression analysis was also employed to illustrate the degree of impact of each factor[1], and indicate whether the impact was positive or negative, and determine the level of statistical significance. Because the total score was a linear combination of the factors alone, a perfect R squared value was expected.

ANOVA techniques compared the means of the metadata usage measurements to total credibility scores. Any significant differences among the means would signal a degree of impact of metadata usage on total credibility.

[1] SPSS stepwise regression enters the variables into the model based on the level of variance accounted for in order.

The Pearson product-moment correlation coefficient indicated how metadata usage variables were related to one another. Drawing relationships between these variables gauges how metadata is being used in Alternative Medicine web sites.

IV. Results

Stepwise regression analysis showed that each credibility factor impacted the total credibility score at varying levels. Due to the relatively large number of variables and small sample size, an adjusted R squared was deemed a more appropriate measure than R squared. As each factor was added to the model*, a significant change in adjusted R squared signaled a significant impact of a factor on the total credibility score. As expected, all factors for main and content-level pages were significant. Each factor, however, varied in level of impact. The amount of change in adjusted R squared is detailed Tables 1 and 2.

Table 1. Credibility Model Summary for Home Pages

Model	R Squared	Adjusted R Squared	R Squared Change	Sig. F Change
1	.732	.712	.732	.000
2	.884	.865	.152	.002
3	.960	.949	.076	.001
4	.990	.987	.031	.000
5	1.000	1.000	.010	.

Model 1: Source
Model 2: Source & Information
Model 3: Source, Information & Timeliness
Model 4: Source, Information, Timeliness, & Accessibility
Model 5: Source, Information, Timeliness, Accessibility, & Design

Table 2. Credibility Model Summary for Content-Level Pages				
Model	R Squared	Adjusted R Squared	R Squared Change	Sig. F Change
1	.479	.461	.479	.000
2	.772	.755	.292	.000
3	.955	.949	.183	.000
4	.975	.971	.021	.000
5	1.000	1.000	.025	.

Model 1: Timeliness
Model 2: Timeliness, Source
Model 3: Timeliness, Source, Information
Model 4: Timeliness, Source, Information, Design
Model 5: Timeliness, Source, Information, Design, Accessibility

ANOVA did not find a significant relationship between credibility and any of the metadata usage measures assuming a .05 level of significance.

In measuring relationships between metadata usage variables, for home pages, only one significant relationship was found. The number of keyword phrases and body word count had a high, positive relationship (r = .729, *p* = .002).

Content-level pages, on the other hand, showed a greater number of significant relationships among metadata usage variables. Phrase count demonstrated varying, positive relationships with three other variables: home overlap (r = .590, *p* = .001), content-level overlap (r = .449, *p* = .013), and with body word count (r = .782, *p* = .000). There were also significant relationships between the two overlap variables (r = .562, *p* = .001), and between home overlap and the body word count (r = .362, *p* = .049).

V. Discussion

A. Impact of Credibility Factors

The individual credibility factors have varying levels of impact on the total credibility score at the two page levels.

Table 3. Order of Precedence of Credibility Factors on Total Credibility Score		
Order	Home Pages	Content-Level Pages
1	Source	Timeliness
2	Information	Source
3	Timeliness	Information
4	Accessibility	Design
5	Design	Accessibility

The ordering of credibility factors in Table 3 shows a distinct difference of how the two page levels are evaluated. The factors vary slightly in terms of semantics for the two page levels; thus a true comparison of credibility alone between the two types of pages is not statistically valid.

For content-level pages, timeliness appears to take precedence over source; however, this may be an inaccurate judgment as timeliness is a highly variable factor given that its granular measures were 'binary' – either true or false. No intermediary scores were assigned; instead, the lowest and highest possible values, 0 and 10 respectively, were assigned. This added bias to the amount of variation explained by this predictor variable—an inherent weakness of the rating scale. Future study may look to address this issue.

For home pages, the adjusted R squared value makes the biggest change with the addition of the source factor. At both page levels, source contains a greater probability of predicting credibility in comparison to the information factor. We may

partially attribute this ordering to user difficulty, in this case the data collector, in judging the goodness of unfamiliar information [13] – in this case Alternative Medicine. Future research may look at testing this notion by assessing credibility in different subject domains.

B. Credibility and Metadata Usage

According to our results, credibility and home page overlap are somewhat inversely related; the higher a site's credibility, the less likely that it used the same META tags from the home page on its subsequent content-level pages. Initially, this seems a positive trait; ideally, authors should use different keyword phrases for the more specific topics of the content-level documents; however, given the non-significant relationship between credibility and content-level overlap, the trait becomes moot—many times, sites use one or more keyword tag sets for the home page and different or alternating sets of keyword tags for content-level pages. Using a constant tag or sets of tags for the content-level pages does not necessarily relate metadata to the content-level documents. In some cases, authors intend to mislead indexers by "loading" keyword tags to gain higher site rankings. Economic motivations tend to encourage the limited recycling of keyword tags, and thus may partially explain the use of alternating sets of keyword tags.

C. Metadata Usage in Alternative Medicine

At the home page level, phrase count and body word count are correlated. The number of phrases used in a META tag is balanced by the probability that those

phrases were contained in the body area of the document. Many web sites use their home pages as general "portals" or "gateways" to more specific information. As a result, a web site using a general set of Alternative Medicine keywords may have a good chance of using a portion of those words in the body area of an Alternative Medicine web site's home page. However, this is a strong assumption to make given the metadata usage variables used in the study. Other measures like number of words in the body of the page and level of keyword and document specificity in subject domain are other factors future research may wish to address in providing further explanation.

Table 4. Summary of Significant Correlations between Metadata Usage Variables

Page-Level	Relationship	r	Sig. p
Home	Phrase Count & Body Word Count	.729	.002
Content-Level	Phrase Count & Body Word Count	.782	.000
	Phrase Count & Overlap 1	.590	.001
	Phrase Count & Overlap 2	.449	.013
	Overlap 1 & Overlap 2	.562	.001
	Overlap 1 & Body Word Count	.362	.049

Likewise, phrase count and body word count are correlated at the content-level. Interestingly, phrase count was also correlated with both the percentage of overlap on the main page and percentage of overlap on the content-level page. The more keywords used at the content-level, the higher the likelihood a percentage of those keywords were repeated on its corresponding home page and content-level page. A high overlap signals a strong possibility in reuse of part or entire keyword phrases. Similarly, sites using a low number of keywords on a given content-level page tended to have low overlap with other pages.

Another significant relationship between two metadata usage measures: percentage of overlap with the home page and percentage of overlap with the other content-level page. If a site tends to repeat keywords on both content-level pages, it is also inclined to repeat a subset of keywords from the content-level page on its home page. This does not necessarily mean the same exact keywords are being reused on all the sites' pages.

The last correlation at the content-level is between overlap with the main page and body word count.

The two overlap variables at the content-level have a significant number of relationships among the metadata usage variables. High overlap, however, may not be an admirable trait for a site to have. From an indexer's standpoint, the goal of an efficient metadata system should be "to reduce fuzziness without unnecessarily reducing resolution or precision" [17]. Using the same keywords to describe multiple items weakens the realization of this goal as it definitely reduces resolution—"the ability to differentiate between two similar items" [17].

Until indexers develop more efficient ways of screening out patterns and signs of "abuse" and "misuse", many search engines may continue sacrifice META tags for other means of automatic indexing where the goal of reducing repeatability and increasing resolution and precision is less at risk.

VI. Conclusion

The heterogeneous nature of the web makes it difficult to "tame". As researchers, we must develop innovative ways to get the beast to behave in a

consistent and congenial manner [16]. Before we develop a set of guidelines for proper behavior, we must first understand what affects behavior, and current trends in interaction and transition. In order to assess the current patterns, some level of quality judgment should be made automatically. The problem, however, has been the inability for computers to create or even simulate human judgment. Until computers can effectively mirror this ability, a more feasible means of assessing these factors must be addressed. Human created metadata, on the other hand, the seemingly natural complement to automatically generated metadata, may offer a viable part of the solution to both assessment and consistency. Unfortunately, metadata usage on the web is largely unstandardized and inconsistent.

META tags, in comparison to other metadata formats, allow for metadata usage study at varying levels of standardization. Developing metadata usage measurements at such a low level of standardization provides a framework for evaluating measures of metadata usage at a very raw form.

Eventually, evaluating metadata usage on the web may give us a good idea of where future research initiatives in metadata schema developments, web site credibility rating methods, and automatic or assisted metadata generation techniques should head.

VII. Bibliography

[1] Amento, B., Terveen, L., & Hill, W. Does "Authority" Mean Quality? Predicting Expert Quality Ratings of Web Documents. (2000). *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* 296-303.

[2] Bunn, M. D. Consumer Perceptions of Medical Information Sources: An Application of Multidimensional Scaling. *Health Marketing Quarterly* 10 (1993): 83-104.

[3] Burkell, J. & Wathen, C. N. Believe It or Not: Factors Influencing Credibility on the Web. *Journal of the American Society for Information Science and Technology* 53.2 (2002): 134-144.

[4] Craven, Timothy. Changes in Metatag Descriptions over Time. First Monday, 6.6 (October 2001). Available at http://www.firstmonday.dk/issues/issue6_10/craven/index.html.

[5] Fogg, B, et. al. What Makes Web Sites Credible? A Report on a Large Quantitative Study. *CHI 2001, ACM:* 61-68.

[6] Global Structure of an HTML Document, The. *W3C*. Available at http://www.w3.org/TR/html4/struct/global.html#edef-META.

[7] Ianella, R. and A. Waugh. "Metadata: Enabling the Internet." 1997. Available at http://archive.dstc.edu.au/RDU/reports/CAUSE97/

[8] Lagoze, Carl. "Keeping Dublin Core Simple: Cross-Domain Discovery or Resource Description." D-Lib Magazine, 7.1 (Jan 2001). Available at http://www.dlib.org/dlib/january01/lagoze/01lagoze.html.

[9] Lawrence, S. and Giles, C. L. Information on the Web. *Nature*. 400 (1999): 107 – 109.

[10] Lim, Eng, Deering, & Maxfield. Criteria for evaluating the quality of health related sites on Internet. Available at http://atlas.ici.ro/ehto/medinf99/papers/criteria_for_evaluating_the_qual.htm.

[11] Milstead, Jessica and Susan Feldman. "Metadata: Cataloging by Any Other Name ..." *Online*, (Jan/Feb 1999): p24–31.

[12] Moura, A., Campos, M., Barreto, C. A Survey on Metadata for Describing and Retrieving Internet Resources. *World Wide Web I*. (1998): 221 – 240.

[13] Rieh, Soo Young. Judgement of Information Quality and Cognitive Authority in the Web. *Journal of the American Society for Information Science and Technology* 53.2 (2002): 145-161.

[14] Sullivan, Danny. “Jupiter Media Metrix Search Engine Ratings”. 12 Dec. 2002. Available at http://searchenginewatch.com/reports/mediametrix.html.

[15] Sullivan, Danny. Search Engine Features for Webmasters. Available at http://searchenginewatch.com/webmasters/features.html.

[16] Thomas, C. F. and Griffin, S. F. Who Will Create Metadata for the Internet? *First Monday* (1998). Available at http://www.firstmonday.dk/issues/issue3_12/thomas/index.html.

[17] Wason, T.D. and Wiley, D. Structured Metadata Spaces. *Journal of Internet Cataloging* 3.2/3 (2000): 263-277.

[18] Web Feet Criteria. Available at http://www.webfeetguides.com/criteria_WF.html.

VIII. Appendix A: Web Feet Criteria

Source

The source of information is identified.

Primary sources are used when possible.

The contact information for the source or site administrator is displayed.

The expertise and reputation of the source are considered. The source is preferably a qualified professional at a peer-reviewed site or a government (.gov), educational institution (.edu), or respected organization (.org).

The expertise and reputation of the site's host are considered. The host is preferably a government (.gov), educational institution (.edu), or respected organization (.org).

Information

The information is not easily available at other sources.

Reviewers (subject-area experts and researchers) make every effort to ensure that the information is free of errors.

The information and images are objective; balanced; and not politically, commercially, religiously, or otherwise biased.

The information is appropriate for all ages and has sufficient scope to cover the topic for the intended audience.

The information is readable and free of spelling and grammatical errors.

Sponsorship is clearly indicated, and advertising is minimal. When a site contains advertising, it should be neither intrusive nor presented in a way that may bias the user's understanding of the information.

The necessary disclaimers and privacy statements are posted.

Timeliness

Site is updated frequently, typically indicated by a recent "last updated" date.

Pages list the date of the most recent update or the dating of the information is made clear in an accessible area of the site.

Links

Links work, and they are relevant and appropriate.

At gateways sites, a large number of links are checked for inappropriate content (sexual references, profanity, violence, and other mature themes).

Chat Rooms and Message Boards

Chat rooms and message boards are reviewed for inappropriate exchanges or postings (sexual references, profanity, violence, and other mature themes).

Accessibility and Navigation

The site loads in a reasonably short time (less than 10 seconds).

The site is easy to access and navigate.

Navigation includes clear headings and intuitive icons, menus, and directional symbols that foster independent use.

Standard multimedia formats such as HTML are used.

Most information is accessible without special plug-ins such as Adobe Acrobat Reader.

Logical options are available for printing and downloading all or selected text or graphics.

Design

The site follows good graphic design principles.

Information for specific audiences, such as consumer information within a professional or commercial site, is easy to locate.

The site has a text size that is easy to read for the intended audience.

Product advertising is not intrusive and is clearly differentiated from original content on the site.

IX. Appendix B: Web Site Listing

bold – home page

non-bold – corresponding content-level page

http://nccam.nih.gov

http://nccam.nih.gov/health/hepatitisc/

http://nccam.nih.gov/health/stjohnswort/

http://www.healthy.net/asp/templates/center.asp?centerid=1

http://www.healthy.net/asp/templates/Article.asp?Id=1179

http://www.healthy.net/asp/templates/Article.asp?Id=1297

http://www.holistic-online.com

http://www.holistic-online.com/Remedies/Biot/biot_anthrax-nat-rem-home.htm

http://www.holistic-online.com/Remedies/Sleep/sleep_ins_breathing.htm

http://www.the-cma.org.uk

http://www.the-cma.org.uk/HTML/diabe2.htm

http://www.the-cma.org.uk/HTML/tcm1.htm

http://www.hcrc.org/sram

http://www.hcrc.org/contrib/basser/acup.html

http://www.hcrc.org/contrib/adair/fear.html

http://www.gems4friends.com

http://www.gems4friends.com/floweressence.html

http://www.gems4friends.com/oils.html

http://www.alternativedr.com

http://www.alternativedr.com/conditions/ConsHerbs/Lindench.html

http://www.alternativedr.com/conditions/ConsSupplements/VitaminCAscorbicAcidcs.html

http://www.quackwatch.org

http://www.quackwatch.org/01QuackeryRelatedTopics/quackvul.html

http://www.quackwatch.org/01QuackeryRelatedTopics/PhonyAds/bracelet.html

http://www.geocities.com/altmedd

http://www.geocities.com/altmedd/tcm_treatment.htm

http://www.geocities.com/altmedd/massage.htm

http://www.ivillagehealth.com

http://www.ivillagehealth.com/experts/fertility/qas/0,11816,166931_125563,00.html

http://www.ivillagehealth.com/experts/guests/articles/0,11299,166056_430056,00.html

http://www.alternativemedicinechannel.com

http://www.alternativemedicinechannel.com/coldsandflu

http://www.alternativemedicinechannel.com/ms

http://www.explorepub.com

http://www.explorepub.com/articles/washreport3.html

http://www.explorepub.com/articles/cardiactherapy1.html

http://www.alt-med-ed.com

http://www.alt-med-ed.com/Herbs/St_Johns_Wart.htm

http://www.alt-med-ed.com/Herbs/Astragalus.htm

http://drweil.com

http://www.drweil.com/app/cda/drw_cda.html-command=healthConditionDetail-articleType=Condition-pt=Condition-articleId=27

http://www.drweil.com/app/cda/drw_cda.html-command=TodayQA-pt=Question-questionId=68298

http://www.acsh.org

http://www.acsh.org/press/editorials/cryingboy080702.html

http://www.acsh.org/press/releases/anthrax100501.html