LIBRES: Library and Information Science Research
Electronic Journal ISSN 1058-6768
2001 Volume 11 Issue 2; September 31
Bi-annual LIBRES 11N2

DESCRIPTION Meta Tags in Public Home and Linked Pages

<http://publish.uwo.ca/~craven/index.htm>

Timothy C. Craven
Faculty of Information and Media Studies
Middlesex College
The University of Western Ontario
London, Ontario N6A 5B7 Canada
(519)-661-2111 ext. 88497. Fax: (519)-661-3506.
craven@uwo.ca

Abstract

Random samples of 1,872 web pages registered with Yahoo! and 1,638 pages reachable from Yahoo!-registered pages were analyzed for use of meta tags and specifically those containing descriptions. Seven hundred twenty-seven (38.8 percent) of the Yahoo!-registered pages and 442 (27.0 percent) of the other pages included descriptions in meta tags. Some of the descriptions greatly exceeded typical length guidelines of 150 or two hundred characters. A relatively small number (ten percent of the registered and seven percent of the other pages) duplicated exactly phrasing found in the visible text; most repeated some words and phrases. Contrary to documented advice to web-page writers, pages with less visible text were less likely to have descriptions. Keywords were more likely to appear nearer the beginning of a description than nearer the end. Noun phrases were more common than complete sentences, especially in non-registered pages.

Introduction

This article reports on work designed to extend a preliminary investigation (Craven, 2000a) of how people and organizations summarize their own web pages and, specifically, how and to what extent they make use of meta tags, especially those with the NAME attribute equal to DESCRIPTION.

The background to this investigation was research conducted over a number of years directed toward developing a prototype computerized abstractor's assistant (Craven, 1988, 1991, 1993, 1996, 1998). As a kind of writer's assistant, such a software package includes a simple word processor and other general writer's tools (Kozma, 1991) with their functioning adapted to fit the needs of abstractors and writers of other kinds of short summaries. In addition, the package integrates such tools as an automatic extractor, related specifically to the task of summarizing. In addition, Paice (1994) has provided a list of desirable features for such a package.

A hybrid system, in which some tasks are performed by human abstractors and others by software, appears to be an appropriate short-term goal since purely automatic abstracting methods (Paice, 1990, 1994; Endres-Niggemeyer, 1998, pp. 297-366; Pinto & Galvez, 1999) do not show immediate promise of totally superseding human effort.

At least two possible benefits are expected from the study of web-page authors' actual practice in summarizing their documents in meta tags. The first of these is in the design of software to assist authors in tag generation. This expectation is based in part on an assumption that author-created descriptions will reflect features that authors and other users consider desirable. The second anticipated benefit is in browser design: to know whether introducing a feature to display the description, as, for example, a title is commonly displayed in the caption bar, would benefit visitors to a substantial number of web sites (compare Beagle, 1999).

Other aspects of the content of web pages have been studied by various researchers. King (1998), for example, studied page layout of library home pages; Haas and Grams (2000) concentrated on characteristics of the anchors found in randomly selected pages; Almind and Ingwersen (1997) applied informetric measures; and Harter and Ford (2000) studied links to e-journals and their articles. Little investigation has been done of the meta tag with NAME='DESCRIPTION' (hereafter referred to as the DESCRIPTION tag). Turner and Brackbill (1998) do report the results of a small experiment that showed that the addition of a DESCRIPTION tag did not improve retrievability of web pages on Infoseek and Altavista.

Advice provided in both printed and web-based sources on the function, content, structure, and style of the DESCRIPTION tag has been reviewed elsewhere (Craven, submitted). What follows is a brief summary of the main recommendations garnered from that review.

1. The DESCRIPTION tag can be used for an abstract.
2. Tag contents should not be deceptive.
3. A description is particularly useful for documents with little text.
4. The description should be no more than two hundred characters (though some recommended maximum range from one hundred to 256). There is a more absolute technical upper bound of around one thousand characters.
5. The description should be concise.
6. The description should reflect the single page, a whole site, or both (advice varies).
7. A number of keywords should be included in the description.
8. The most important words should be near the beginning of the description.
9. The description should not be the same as the title.

Sample descriptions in the sources showed various patterns. Some contained what appear to be formulaic elements such as "Home page for"; others did not.

It has been noted that there is also a DESCRIPTION element in the Dublin Core (2000), defined as "an account of the content of the resource," with the further comment that it "may include but is not limited to: an abstract, table of contents, reference to a graphical representation of content or a free-text account of the content." As a meta tag, this element appears in standard form with the name DC.DESCRPTION.

Sample

Methodology

In order to estimate the total variety of use of the DESCRIPTION tag, it is desirable to obtain a broad, representative sample of publicly available web pages, especially of those pages employing meta tags, in particular the DESCRIPTION tag. A survey of forty-two search engines (in December 1999) (Craven, 2000) revealed no feature on any that permitted searching for specific meta tags. In terms of sampling web pages in general, Askjeeves and Webcrawler permitted peeking at sample queries, OpenDirectory selected random categories if an empty query was entered, and All-In-One's "What's New Too"

option showed the day's announcements of new pages.

Since a sufficient proportion of pages indexed by Yahoo! used meta tags, including the DESCRIPTION tag (twenty-six percent), it was feasible to proceed by sampling from these pages. At that time, an application program was created in Delphi 3 to access sample web pages and to log results. The NEWT ActiveX control included with Delphi 3 was employed to handle the hypertext transfer protocol (HTTP) and to display pages as they were downloaded.

In the preliminary study, only pages returned directly by Yahoo!'s random page service were used (level-one pages). Since such pages were presumably registered with Yahoo!, it may well be that they were also registered with other web search services. Although Yahoo! ignores meta tags, the pages' creators may have been particularly sensitized to the value of meta tags for some of these other search services. Thus, in order to investigate possible differences in non-registered pages, the present study added two other types of page: a page reachable by following a random link from a page returned by the random-page service (level-two page) and a page reachable by following a random link from a level-two page (level-three page). Requests for random pages were submitted to Yahoo! by a research assistant over a period of twenty-five days. The aim was to retrieve at least eight hundred pages at each level. Because the initial set of pages at level one was gathered in a mode that included page rendering and downloading of inline image files while the pages at levels two and three were retrieved with these options off, the assistant also collected an additional set of pages at level one with the mode matching that of levels two and three. This was performed about one month after the original level-one set.

Within each set, a test for duplication of uniform resource locators (URLs) was carried out for those pages containing DESCRIPTION tags. The number of links on each page was also recorded. Included in this count for purposes of this study were not only links defined by hypertext reference (HREF) values in anchor (A) and area tags but also links defined by source (SRC) values in frame tags.

Results

A natural multiplier effect meant that the proportion of successful downloads decreased as level increased. For example, to obtain the first two hundred level-one pages required only 274 requests, a success rate of seventy-three percent, but to get the first two hundred level-two pages required 448 requests, for a success rate of 44.6 percent. It had been established in the preliminary study that the most common reason for failure (nearly four out of five) was timing out, usually with a "not found" or "404" message displayed in the hypertext markup language (HTML) viewer. The number of successful downloads was 833 for the first set at level one, 821 at level two, 817 at level three, and 1039 for the second set at level one.

As shown in Figure 1, the proportion of pages containing meta tags held very close to sixty-seven percent at all three levels; the proportion in the preliminary study had been somewhat lower at fifty-seven percent. The proportion containing the DESCRIPTION tag

specifically was noticeably higher at level one (38.4 percent and 39.2 percent) than at level two (26.1 percent) and level three (27.9 percent).

Only one case of duplication of a URL was found among the pages with DESCRIPTION tags, at level 3.

Pages with descriptions typically contained a number of links, with means somewhat above the fifteen calculated in the preliminary study (19.7 at level one, 23.9 at level two, 22.7 at level three). The maximum number of links on a page was 326 (at level three). As in the preliminary study, few pages had no links (two and three at level one, one at level two, and one at level three).

Repeating the findings of the preliminary study, very few pages used Dublin Core elements (two and two at level one, one at level two, and one at level three). Three of these included a Dublin Core description, in one case without a DESCRIPTION tag, in one case with a DESCRIPTION tag with the same value, and in one case with a slight difference in the wording of the two tag values.

Discussion

The proportion of pages using meta tags was even more noticeably above that of the 24.4

percent reported by Qin & Wesley (1998) for pages in polymer chemistry than had been the case for the preliminary study. The proportion of pages using the DESCRIPTION tag at level one was noticeably greater than the proportion at levels two and three, which were close to the twenty-six percent found in the preliminary study and much above the figure of about twenty-one percent using meta tags with both the names KEYWORD and DESCRIPTION cited by Clark (2000). The higher rate of page tagging at level one would seem to support the hypothesis that developers are more likely to pay attention to tagging for pages that they will be submitting to search services.

Errors excepted, it appears that the pages returned by requests to the Yahoo! random-page generator are generally home pages. A rough confirmation of this conclusion can be obtained from statistics on the forms of the URLs. Only 178 and 243, or 21.4 percent and 23.4 percent, of the URLs for the level one pages contain the string ".htm" (which would include both the ".htm" and the ".html" extensions). Although some of the others will represent other file formats, most of them are simply references to the default pages delivered by servers for their root or other directories. By contrast, 611, or seventy-four percent, of the level-two pages and 621, or 76.1 percent, of the level-three pages contain ".htm" in their URLs.

The extremely low level of duplication of URLs confirms that the Yahoo! random-page service gives access to a large sample of web pages and suggests that page duplication as such would not be a significant concern in future studies involving this method of selection. On the other hand, there was some evidence of a minor concentration of results on certain sites, most noticeably that of the Toronto *National Post*, for which the service, at various levels, returned a total of thirty-three pages with the same description.

Comparison of Descriptions with Visible Text

Purpose

Two main questions were posed regarding the relationship of the description to what the user would actually see in the browser window. First, are pages with little visible text more likely to be given descriptions as is recommended by some of the sources? Second, to what extent do the descriptions merely repeat words or phrases that are visible on the pages?

Methodology

Visible text was defined as all text that was not part of a tag. This would typically include the page's caption title and most text normally displayed as text in the viewer. Visible textual material that would be excluded would include button captions in forms and any text loaded as part of a frame. Text appearing in graphic form would also be excluded as would any alternate text (ALT) values given in the image (IMG) tag.

A *description* was defined as the value given to a DESCRIPTION tag in a downloaded file.

Degree of match of each description to the corresponding visible text was calculated for

words and phrases. The measure used for words was the density of non case-sensitive matches of visible-text words within the description. The measure used for phrasing was density of non case-sensitive matches of visible-text two-word sequences within the description. A word was defined as any sequence of alphabetic characters delimited by other types of characters. In addition, the longest visible-text word sequence found in each description was logged; in the case of a tie, the tied sequences were all logged, separated by a delimiter, in a single record.

Results

Lengths of the visible text varied between 0 and 178,637 bytes. The median length increased somewhat with level (565.5 and 647 for level one, 764.5 for level two, and 896 for level three). Shorter visible texts were significantly less likely to be associated with descriptions, even leaving aside those pages with no visible text (with $p < 0.0001$ in t-tests applied to the logarithm of number of characters in visible text).

As shown in Figure 2, the descriptions were generally relatively short. The mean length was just over twenty words at all levels (21.9 and 24.2 at level one and 21.3 and 20.9 at the other two levels). But they could be fairly long, with the longest being 173 words (the preliminary study had found one at 294 words). The shortest that contained any words consisted of the single words "strong" and "test." Using a different measure, 181 and 238, or 56.6 percent and 58.5 percent, were no more than 150 characters in length at level one; 128, or 59.8 percent, at level two; 142, or 62.3 percent, at level three; and 243 and 310, or 75.9 percent and 76.2 percent, were no more than two hundred characters at level one; 165, or 77.1 percent, at level two; and 178, or 78.1 percent, at level three. The longest was 839 characters.

Fig. 2: Distribution of descriptions by length in words

The density of visible-text words in the descriptions ranged from zero percent to one hundred percent, with means of 87.1 percent and 91.3 percent at level one and 84.9 percent and 82.2 percent at the other two levels. As is clear from Figure 3, the distribution was heavily skewed toward the high end. This was especially true at level one, where 186 and 277, or 58.1 percent and 68.2 percent, of the descriptions had a visible-text word density greater than ninety percent, figures somewhat higher than the fifty-two percent observed in the preliminary study. The skew was also found, to a declining degree, at the other levels, where 105, or 49.1 percent, and ninety-three, or 40.8 percent, showed a density of greater than ninety percent. Visible-text word density was in fact one hundred percent for 117 and 168 descriptions, or 36.6 percent and 41.4 percent, at level one, sixty-eight descriptions, or 31.8 percent, at level two, and sixty-two descriptions, or 27.2 percent, at level three.

Fig. 3: Distribution of descriptions by density of visible-text words

In contrast, the distribution of density of visible text phrasing, as shown in Figure 4, appears to be bimodal, as noted also in the preliminary study, with one local maximum somewhere in the twenty to fifty percent range and another in the ninety to one hundred percent range. In a number of cases (thirty-six and forty at level one, thirteen at level two, and eighteen at level three), the density was one hundred percent, meaning that the entire description was word-for-word a sequence also found in the visible text. Many such descriptions were short, but the longest (at level three) was sixty-two words, exceeding the forty-six-word record set in the preliminary study.

Fig. 4: Distribution of descriptions by visible-text phrase density

This entire description of sixty-two words (<http://octopus-design.co.uk:80/octopus/home-nos.htm>) was also the longest word sequence in a description that exactly matched one in the visible text:

Octopus Design are a team of graphic design professionals utilising the latest computer technology to assist a wide range of skills effectively and creatively. If you are looking for a design company who are committed to supplying top quality graphic design solutions on time and on budget, why not view our portfolio here, e-mail us for more information or telephone. UK 0118 934 4209

Discussion

As in the preliminary study, the suggestion of one source that descriptions are especially important to pages with little visible text was not reflected in practice: pages with less visible text were actually less likely to contain the DESCRIPTION tag at all levels. This observation does not, of course, invalidate the advice.

The great majority of the descriptions conformed to the common maximum-length guideline of two hundred characters. A smaller majority conformed to the more restrictive guideline of 150 characters given by HotBot and others. The fact that more than twenty

percent of descriptions at all levels, as in the preliminary study, exceeded the two-hundred-character limit, in one case reaching more than four times that number of characters, may suggest that there is a need on the part of some authors for a way of including much longer descriptive information. The one description in the preliminary study that was almost nine times the recommended maximum length seems, however, to have been particularly anomalous, exceeding even the more technical limit of one thousand characters.

The bimodal appearance of the phrasing density distribution clearly suggests at least two approaches to authoring descriptions: on the one hand, production of an expression that is original but that substantially echoes wording found in the visible text and, on the other hand, exact copying of an entire expression from somewhere in the visible text. It is expected that future research will examine which parts of the visible text tend to be duplicated in the latter situation. For example, is it the first two hundred words as suggested by some of the existing guidelines?

Since most descriptions are not word-for-word repetitions of information provided in the visible text, a browser feature to display the description might be of value to some users. Such a feature would need to take into account that some descriptions are quite long. A single line, like the caption bar used for titles, would not be sufficient (indeed, even titles of web pages are sometimes too long to fit into the space available in the caption bar). The increasing complexity of HTML is making it more and more difficult for a user to access non-display text by the alternative expedient of viewing the page source.

Given the number of links on the typical page, it seems reasonable to assume that, in many cases, the descriptions are intended to apply to a site or collection of pages rather than to a single page. A related study, not yet reported on, used the variation of appending the visible text of linked pages to the visible text of the initial random page. As hypothesized, both word-match and phrase-match density within the descriptions increased substantially with the addition of the linked-page texts, suggesting that the descriptions are indeed intended to apply to multiple pages. Another test, to be carried out in the future, will compare the results of following only local links to see whether authors are more likely to use descriptions for sites rather than for multi-site groupings of pages.

Conciseness and Structure

Purpose

Questions to be addressed under the heading of conciseness and structure included the following. Are descriptions generally concise, as recommended, and how does their conciseness compare with that of scholarly abstracts? Do descriptions use complete sentences, as recommended for abstracts, or do they tend to consist of title-like noun phrases or other syntactic structures? Are home-page descriptions more likely to use complete sentences than descriptions for other pages? Is formulaic phrasing, as suggested by some of the samples provided by the sources, at all common?

Methodology

Simpson's I (Simpson, 1949), a measure of concentration or repetition of vocabulary equal to the probability that the words occurring at two different random locations in a text are the same, was computed for each description as a possible negative indicator of conciseness.

All descriptions were analyzed for general syntactic structure. For this purpose, each description was considered to be divided into segments by sentence-level punctuation marks: periods, exclamation marks, and question marks. Each segment was then categorized as a noun phrase or sequence of noun phrases (n), a verb phrase or sequence of verb phrases (v), an adjectival or adverbial phrase or sequence (m), a sentence in the indicative mood (s), a sentence in the imperative mood (c), or other (o). For example, the description "Marshall Media produces high-quality CD-ROMS for children and adults. Order online from our shop." would be coded sc.

To estimate inter-rater reliability, the research assistant was first asked to re-code sixty-five descriptions previously coded in the preliminary study. Consistency on this test was 90.8 percent on all codings and rose to 95.4 percent when all codings except n, nn, s, and ss were collapsed into a single "other" category. Reliability was thus deemed more than sufficient for the assistant to proceed with coding the main description sets independently.

Results

Values for Simpson's I were similar across the levels, with means of 0.0127 for both level one sets and 0.0108 and 0.0136 for the other two levels, and ranged from 0.0000 to 0.2273. The most common value was zero, generally in shorter descriptions, though the longest was thirty words: ""Desertcom - Just what you've come to expect!!! A 15-year history of tradition with Southern California businesses. Panasonic Digital Business System, PanaVOICE, Active Voice, Pacific Bell Network Services and more!" (<http://www.desertcom.com:80/telecom.htm>).

The most common words in the descriptions (at least one occurrence for every thirteen descriptions), apart from obvious stopwords, were as shown in Table 1.

Table 1: Common words in descriptions							
Level 1a		Level 1b		Level 2		Level 3	
DESIGN	27	BUSINESS	37	BUSINESS	23	DESIGN	20
INFORMATION	30	INFORMATION	33	ESTATE	19	INFORMATION	20
INTERNET	35	NATIONAL	51	INTERNET	31	INTERNET	22
NEWS	27	NEWS	43	NATIONAL	19	NATIONAL	27
ONLINE	25	SERVICES	34	NEWS	22	NEWS	22
SERVICE	25	WEB	41	ONLINE	17	ONLINE	22
SERVICES	30	WORLD	36	REAL	20	SERVICES	19
SITE	35			SERVICES	25	SITE	21

WEB	44			SITE	24	WEB	33
				WEB	33	WORLD	26
				WORLD	25		

Common syntactic patterns are shown in Table 2.

Table 2: Syntactic patterns								
	Level 1a		Level 1b		Level 2		Level 3	
n	113	36%	155	39%	91	44%	107	52%
nn	13	4%	27	7%	14	7%	5	2%
s	72	23%	73	18%	41	20%	35	17%
ss	30	10%	36	9%	14	7%	9	4%
other	87	28%	109	27%	48	23%	50	24%

In the "other" category, eighty-three descriptions contained imperative (c) segments, almost always at the end, in eighteen instances after a single indicative-mood (s) sentence. The largest number of segments in a description was eleven, where the segments all consisted of brief noun phrases separated by periods.

Discussion

The average values for Simpson's *I* for descriptions were similar to those in the preliminary study and only slightly lower than those observed in a previous study of abstracts produced with computer assistance (Craven, 2000b).

ONLINE, SERVICE, INFORMATION, INTERNET, and SERVICES had all been noted as among the most common words in the preliminary study. All the occurrences of ESTATE and all but one occurrence of REAL at level two were in the phrase REAL ESTATE, which was fairly concentrated in two descriptions but appeared in others as well. The word pair REAL+ESTATE was noted as the eighth most common in Wolfram's (1999) study of term co-occurrence in Excite queries. NUDE, PICS, and XXX were among words that were common in the queries analyzed by Wolfram but occurred only a couple of times each in the descriptions.

Noun phrases or sequences of noun phrases were much more common than one might expect in an abstract, especially as one progressed from level one to levels two and three. The increasing preference for noun phrases with level would be consistent with some tendency to apply abstract-like descriptions more to home pages or pages to be registered with search services and to apply subject-heading or title-like descriptions more to other, especially subordinate, pages on a web site.

Imperative-mood sentences would not be expected in abstracts, and they were relatively rare in the web-page descriptions in spite of the presumably promotional nature of many of

the sites and the use of imperatives in some of the descriptions provided as models.

Comparison with Keywords

Purpose

The main questions to be addressed in comparing keywords were the extent to which keywords were in fact found in descriptions and whether the advice to place keywords near the beginning was followed. The amount of repetition within keyword lists was also of interest.

Methodology

Wording and phrasing of descriptions were compared to the contents of any meta tag with a name attribute of KEYWORDS in a fashion similar to that applied for the visible text. In addition, the mean position of keyword matches within descriptions was calculated. Simpson's I was used to measure repetition within keyword lists.

Results

The density of keywords in the descriptions ranged from zero percent to one hundred percent but averaged close to the thirty-eight percent value observed in the preliminary study (36.9 percent and 36.7 percent at level one, 38.9 percent at level two, 39.9 percent at level three).

As in the preliminary study, density in the descriptions of two-word sequences from the keywords showed a local maximum in the zero-ten percent range at all levels. No keyword word pairs were found in just under one-third of the descriptions at all levels (ninety-three and 110 at level one, sixty-six at level two, and sixty-five at level three). Again, a small number consisted entirely of keyword word pairs (five at level one, four at level two, and four at level three); the longest of these was 147 words ([http://www.transdev.com/eudora="autourl">http://www.transdev.com:80](http://www.transdev.com/eudora=)).

The mean position of keyword matches was significantly more likely to be nearer the beginning of the description than the end at all levels (0.0000 and $p=0.0001$ at level one, $p=0.0015$ at level two, $p=0.0001$ at level three using a chi-squared test), with the average position being around forty-five on a 0-100 scale. As in the preliminary study, length of the keywords had a low positive correlation with length of description (0.2262 and 0.2146 at level one, 0.1962 at level two, 0.2934 at level three).

For keyword lists, Simpson's I was generally higher than for the descriptions, with means of 0.0301 and 0.0269 at level one and 0.0222 and 0.0260 at the other two levels, and ranged from 0.0000 to 0.5000. The highest value was observed in the clearly repetitive "sturgis, sturgis rally, STURGIS, STURGIS RALLY, Sturgis, Sturgis Rally" ([http://www.sturgiscamping.com/eudora="autourl">http://www.sturgiscamping.com:80](http://www.sturgiscamping.com/eudora=)).

Discussion

The results on average position of keyword matches represented an addition to the preliminary study where statistical significance had not been attained. They do supply some support for the hypothesis that description developers/writers are following the advice to put keywords near the beginnings of their page descriptions, but the tendency does not appear to be very strong.

The slightly higher but still modest values of Simpson's *I* for the keyword lists are not consistent with a view of developers as engaged in widespread "word stuffing" to increase retrievability of their pages. There are obviously some exceptions.

Conclusion

From this study, the following main findings have emerged:

1. Pages with little visible text are actually less likely to be given descriptions, contrary to what is recommended by some of the sources.
2. Descriptions vary greatly in their repetition of words and phrases that are visible on the pages: some repeat word for word; others repeat selectively.
3. Descriptions generally appear about as concise as scholarly abstracts.
4. Unlike abstracts, many descriptions tend to use noun phrases rather than complete sentences. Use of complete sentences appears to be slightly more characteristic of descriptions on home pages.
5. While some words are found fairly frequently in descriptions, there is little indication of widespread adoption of formulas.
6. Keywords are found in descriptions to various extents. They have a slight tendency to appear nearer the beginnings than the ends of descriptions, reflecting in a very weak fashion the advice to place them up front.
7. On average, keyword lists are not highly repetitive.

In future research, it would be useful to rate a sample of descriptions for quality, using either objective criteria or subjective human judgments or both. Even scholarly abstracts have sometimes been found to be of poor quality. Pitkin, Branagan, & Burmeister (1999), for example, demonstrated inconsistencies and other defects in published author abstracts. Inter-rater reliability might, however, be expected to be low. In a study in which subjects rated different abstracts of the same document on various criteria, agreement was sometimes very poor (Craven, 2000b).

A very simple kind of assistance for web-page developers is already provided in WordPerfect, namely, copying any abstract into the DESCRIPTION tag when exporting to HTML. Conceivably, the abstract or description might also be automatically copied to the corresponding Dublin Core tag. Dublin Core elements were, however, rarely encountered in this study and the DC.DESCRPTION meta tag appears to be redundant, especially if it merely duplicates the DESCRIPTION tag.

If more advanced tools are to be produced to assist in the adding of appropriate meta tags to HTML documents, it is likely that different tools will suit different types of users. That individuals use quite different approaches in writing abstracts has been noted in studies involving think-aloud protocols (Endres-Niggemeyer, Waumans, & Yamashita, 1991); similar observations are to be expected regarding the writing of other kinds of summary. For composing web-page descriptions specifically, results of the present study suggest that some authors might want a tool for copying text from elsewhere on the page while others might find automatically generated lists of key words or phrases to be helpful.

Acknowledgments

Research reported in this article was supported in part by individual operating grant A9228 of the Natural Sciences and Engineering Research Council of Canada.

The extensive assistance of research assistant Michael Dub in data gathering and categorization is also acknowledged.

References

- Almind, T.C., & Ingwersen, P. (1997). Informetric analyses on the World Wide Web: Methodological approaches to 'Webmetrics'. *Journal of Documentation*, 53 (4), 404-426.
- Beagle, D. (1999). Visualization of metadata. *Information Technology and Libraries*, 18 (4), 192-199.
- Clark, S. (2000). Back to basics: META tags / WebDeveloper.com. Retrieved January 20, 2000 from the World Wide Web: http://www.webdeveloper.com/html/html_metatags_part2.html.
- Craven, T.C. (1988). Text network display editing with special reference to the production of customized abstracts. *Canadian Journal of Information Science*, 13 (1/2), 59-68.
- Craven, T.C. (1991). Algorithms for graphic display of sentence dependency structures. *Information Processing and Management*, 27 (6), 603-613.
- Craven, T.C. (1993). A computer-aided abstracting tool kit. *Canadian Journal of Information Science*, 18 (2), 1993, 19-31.
- Craven, T.C. (1996). An experiment in the use of tools for computer-assisted abstracting. In Hardin, S., ed., *ASIS '96: Proceedings of the 59th ASIS Annual Meeting 1996 (Volume 33)*, Baltimore, Maryland, October 21-24, 1996. (pp. 203-208). Medford, New Jersey: Information Today.
- Craven, T.C. (1998). Human creation of abstracts with selected computer-assistance tools. *Information Research*, 3 (4), paper 47. On the World Wide Web: <http://www.shef.ac.uk/~is/publications/infres/paper47.html>.
- Craven, T.C. (2000a). Features of DESCRIPTION META tags in public home pages. *Journal of Information Science*, 26 (5), 303-311.
- Craven, T.C. (2000b). Abstracts produced using computer assistance. *Journal of the American Society for Information Science*, 51 (8), 245-256.
- Craven, T.C. (submitted). 'DESCRIPTION' META Tags in Locally Linked Web Pages. Submitted for publication.
- Dublin Core Metadata Initiative / documents / proposed recommendations / Dublin Core

Element Set, version 1.1.2000. Retrieved April 24, 2000 from the World Wide Web:
<http://purl.oclc.org/dc/documents/rec-dces-19990702.htm>.

- Endres-Niggemeyer, B. (1998). *Summarizing information*. Berlin: Springer.
- Endres-Niggemeyer, B., Waumans, W., & Yamashita, H. (1991). Modelling summary writing by introspection: A small-scale demonstrative study. *Text*, 11 (4), 523-552.
- Haas, S.W., & Grams, E.S. (2000). Readers, authors, and page structure: A discussion of four questions arising from a content analysis of Web pages. *Journal of the American Society for Information Science*, 51 (2), 181-192.
- Harter, S.P., & Ford, C.E. (2000). Web-based analyses of e-journal impact: Approaches, problems, and issues.
- King, D.L. (1998). Library home page design: A comparison of page layout for front-ends to ARL library Web sites. *College and Research Libraries*, 59 (5), 458-465.
- Kozma, R.B. (1991). The impact of computer-based tools and embedded prompts on writing processes and products of novice and advanced college writers. *Cognition and Instruction*, 8 (1), 1-27.
- Paice, C. (1990). Constructing literature abstracts by computer: Techniques and prospects. *Information Processing and Management*, 26 (1), 171-186.
- Paice, C.D. (1994). Automatic abstracting. In: Kent A., & Hall, C.M., eds., *Encyclopedia of Library and Information Science*, (Vol. 53 [supplement 16], pp. 16-27). New York: Dekker.
- Pinto, M., & Galvez, C. (1999). Paradigms for abstracting systems. *Journal of Information Science*, 25 (5) 365-380. *Journal of the American Society for Information Science*, 51 (13), 1159-1176.
- Pitkin, R.M., Branagan, M.A., & Burmeister, L.F. (1999). Accuracy of data in abstracts of published research articles. *JAMA*, 281 (12), 1110-1111.
- Qin, J., & Wesley, K. (1998). Web indexing with meta fields: A survey of Web objects in polymer chemistry. *Information Technology and Libraries*, 17 (3), 149-156.
- Simpson, E.H. (1949). Measurement of diversity. *Nature*, 163, 688.
- Turner, T.P., & Brackbill, L. (1998) Rising to the top: Evaluating the use of the HTML meta tag to improve retrieval of World Wide Web documents through Internet search engines. *Library Resources and Technical Services*, 42 (4), 258-271.
- Wolfram, D. (1999). Term co-occurrence in Internet queries: An analysis of the Excite data base. *Canadian Journal of Information and Library Science*, 24 (2/3), 12-33.

This document may be circulated freely
with the following statement included in its entirety:

Copyright 2001

This article was originally published in
LIBRES: Library and Information Science
Electronic Journal (ISSN 1058-6768) September 31, 2001
Volume 11 Issue 2.

For any commercial use, or publication
(including electronic journals), you must obtain
the permission of the author.

Timothy C. Craven
Faculty of Information and Media Studies
Middlesex College
The University of Western Ontario
London, Ontario N6A 5B7 Canada
(519)-661-2111 ext. 88497. Fax: (519)-661-3506.
Email: craven@uwo.ca

To subscribe to LIBRES send e-mail message to
listproc@info.curtin.edu.au
with the text:
subscribe libres [your first name] [your last name]

Return to [Libres 11n2 Contents](#)
Return to [Libres Home Page](#)