

# *External Descriptions of Web Pages: Their Features and Their Relationships to Web Page Elements*

TIMOTHY C. CRAVEN

Faculty of Information and Media Studies, Middlesex College,  
The University of Western Ontario, London, Ontario, Canada

---

Fifteen sets of external descriptions of Web pages were examined for common phrases, general syntactic structure and content. For the seven largest sets, the value of meta tag descriptions and keywords, the first 200 characters of the body and text marked with common HTML tags as extracts helpful for writing external descriptions was estimated by applying two measures: density of external description words and density of two-word external description phrases. Syntactic

patterns were found to vary between sets, with larger sets tending to be more internally consistent. Generally, titles showed the highest match densities (means between 50.6% and 69.4% for words and between 30.1% and 61.3% for phrases); match densities were also generally high for meta tag descriptions and for the first 200 words of the body, and low for text tagged A, with mixed results for keywords and for text tagged B, CENTER, or FONT.

---

## *Introduction*

The research reported in this paper originates in the author's ongoing research aimed at developing a computerized abstractor's assistant (Craven, 1988, 1991, 1993, 1996, 1998). In addition to a simple word processor and other general writer's tools, the assistant integrates tools, such as an automatic extractor, related specifically to the task of summarizing. Apart from the author's own work, Paice (1994) has given 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 (Endres-Niggemeyer 1998; Paice 1990, 1994; Pinto & Galvez 1999) do not show immediate promise of totally superseding human effort.

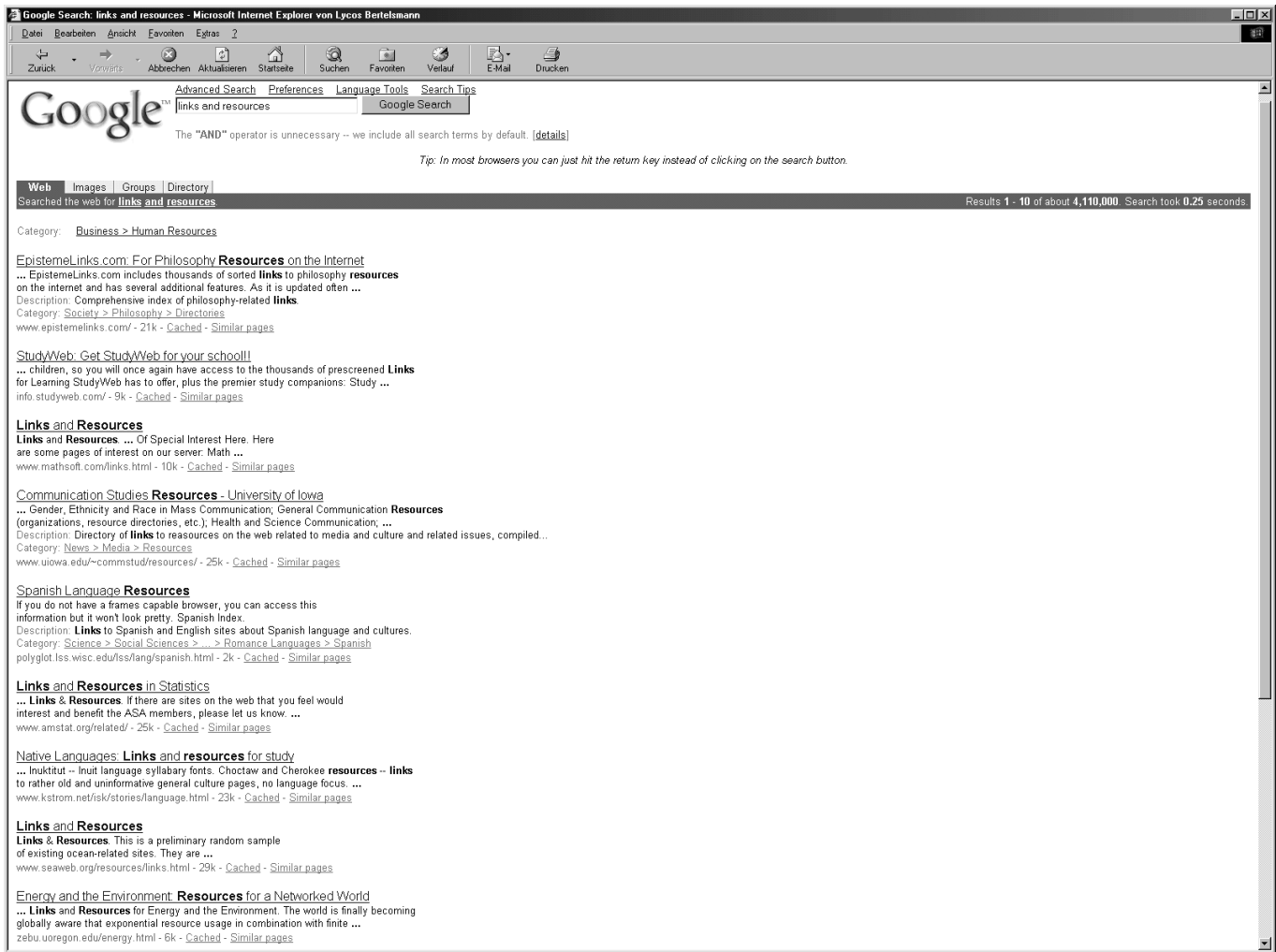
With a view specifically to applying the computerized assistant to summarizing Web pages, this development work led to an investigation (Craven 2000) into how people and organizations in fact summarize their own Web pages, specifi-

cally in meta tags on the pages themselves. An underlying assumption of that investigation was that author-created descriptions reflect features that authors and other users may consider desirable. In addition to providing input specifically to the development of the computerized assistant, results of the investigation might also have wider implications; for example, for advice to be given to Web page creators on the kinds of information to supply in meta tags or to search engine designers on the kind of information to be expected.

Other aspects of the content of Web pages have been studied by various researchers; for example, page layout of home pages (King 1998); characteristics of anchors (Haas and Grams 2000); informetric measures (Almind & Ingwersen 1997); links to e-journals and their articles (Harter and Ford 2000). Little investigation had been done into descriptions in meta tags. Turner and Brackbill (1998) had, however, reported results of a small experiment that showed that addition of a description did not improve retrievability of Web pages on Infoseek and AltaVista; similar results have been reported more recently for the two

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. E-mail: craven@uwo.ca

Figure 1: Google search of „links and resources“ used to identify the set of Web site descriptions used in this study



search engines and five others by Henshaw and Valauskas (2001).

Another article by the author (Craven 2001) has reviewed advice given in both printed and Web-based sources on the function, content, structure, and style of meta tag descriptions.

A question raised in the course of these studies related to the specifics of repetition of wording from the text of the page body. Were there particular parts of the body that were more likely to be used than others? If so, suitable cues might be employed to extract such parts automatically and so to assist in the writing of descriptions for those pages. Another question was the extent to which passages so extracted might be better or worse candidates for extraction than the page titles or any keywords included in meta tags on the page.

Since the aim was to develop generalized tools and the subject matter and language of Web pages various greatly, it was decided to look for cues in

the HTML tags, which tend to occur over and over again independently of the specific topic or level of the page.

In the study of meta tag descriptions (Craven submitted), titles and keywords generally showed the highest densities of description words; parts of the body showed densities not much different from the body as a whole: somewhat higher for the first 200 characters and for text tagged with CENTER and FONT; somewhat lower for text tagged with A; not significantly different for TABLE and DIV.

The present study looks at external descriptions; that is, descriptions that do not form parts of the pages themselves and that are written by someone other than the page author. This is not the first time that external descriptions on the Web have been examined. Wheatley and Armstrong (1997) compared abstracts from Internet subject trees, subject gateways, and three

online databases for readability, length, contents, and style. Amitay (2001) developed a tool called SnipIt to extract descriptive passages with URLs from Web pages and another tool called InCommonSense to select from among these the “best” descriptive passage for each URL.

The present study, however, focuses specifically on descriptions in sets, each set typically compiled by the same author, in order to discover possible patterns or rules that may be followed in description construction; moreover, it looks specifically at common phrasing, syntactic patterns, and the degree to which description wording corresponds to different parts of the pages described.

### *Selection of description sets*

It was decided not to take the sets of descriptions from major Web search services, where the tendency is to display miscellanies of descriptions from meta tags, descriptions supplied by people registering sites, automatic extracts based on rules that vary over time, and descriptions generated by in-house staff. Instead, more specialized sets each related to some specific subject area were sought. For this purpose, random sampling of Web sites, examining each to see whether it contained a substantial set of descriptions of pages on other sites, was not deemed feasible. The method in fact adopted was to search Google for the phrase “links and resources” and examine the URLs returned in sequence, looking for sites that satisfied the following requirements: at least 30 descriptions linked to Web pages on other sites, with each description consisting of at least five words, clearly set out in a separate paragraph or similar structure, and having a valid and appropriate link to an external page (see Figure 1).

For each set of descriptions so identified, the following was recorded for each of up to 300 descriptions: the URL pointed to and the full wording of the description.

For convenient reference, each set of descriptions was assigned a one-word name. In each case, the word chosen was the most unusually frequent word (using a chi-square-like measure) that was not an obvious stopword or otherwise non-distinctive word. The following is a list of the sets, giving their names, URLs, and site titles and a typical description from each:

1. MAPS. Mercator’s World.  
<http://www.mercatormag.com/links.html>.  
“Paths to History – Archaeological Map of Beirut <http://www.lebanon.com/construction/beirut/pathstohistory.htm> provides photos, maps, and information about Beirut and its history. “
2. SECURITY. Encryption and Security-related Resources.  
<http://www.cs.auckland.ac.nz/~pgut001/links.html>.  
“Chris Vidler’s Cryptography Page Links to FTP archives, bibliographies and e-journals, disk and file-system encryption, laws and regulations, network security, newsgroups and mailing lists, protocols and standards, software, and vulnerabilities. “
3. HISTORY. SocialStudies.org Internet Resources & Links.  
<http://www.socialstudies.org/links>.  
“UCLA Center for East Asian Studies As a U.S. Department of Education designated National Resource Center, CEAS works to support teachers in their efforts to bring Asia to their students. Each year we conduct an intensive two week summer seminar for K-12 teachers and we have recently initiated school-site training programs which integrate content and new media training. Our website offers a comprehensive guide to curriculum materials, a index of educational films, and an annotated list of recommended web resources. Additional features include ‘Today in Asian History’ and our ‘Statistical Index.’ CEAS educational resources: <http://www.isop.ucla.edu/eas/resource.htm>”
4. AUDIO. Audio Related WWW Links.  
<http://www.aes.org/resources/www-links/>.  
“Audio and Three Dimensional Sound Links is a page with links to 3D and Spatial Sound related resources on the Web. Reference, Research, Educational, Tools, People, Commercial Systems and Products. “
5. HR. Society for HR Management.  
<http://www.shrm.org/hrlinks/allLinks.asp>.  
“American Immigration Center AIC offers employment visa information, kits and H-1b status material. “
6. CONSUMER. AAFCS: Family Resources and Links.  
<http://www.aafcs.org/resources/>.  
“Cooperative State Research, Education, and Extension Service (CSREES) [www.ree.usda.gov](http://www.ree.usda.gov) works with several higher education institutions to advance a global system of research, extension and higher education in the food and agricultural sciences and related environmental and human sciences to benefit people, communities, and the Nation. This site also links to several of CSREES’s land-grant university partners. Children, Youth and Families Education and Research Network (CYFERNet) [www.cyfernet.org](http://www.cyfernet.org) offers pages of resources for families and professionals who work with families and children. “
7. DIGITAL. Digital Imaging and Media Technology Initiative Linklist.  
<http://images.library.uiuc.edu/resources/links.htm>.  
“Digital Imaging Tutorial provided by Cornell University – (<http://www.library.cornell.edu/preservation/tutorial/contents.html>) A tutorial offering basic in-

- formation on digitization and digitizing cultural heritage materials.”
8. HIV. HIV InSite Links.  
<http://hivinsite.ucsf.edu/InSite.jsp?page=Links>.  
 “GMHC’s Living with HIV or AIDS Features medical care and treatment, nutrition, mental health and counselling services, support groups, financial concerns, legal concerns, alcohol and drug use, insurance, and emergency services.”
  9. BIOLOGY. Science: resources, information, links, courses.  
<http://mindquest.net/>.  
 “Cell Biology Tutorials. Nice introductory cell biology tutorials.”
  10. ETHICS. American Society for Bioethics and Humanities Links.  
<http://www.asbh.org/links/>.  
 “Kennedy Institute of Ethics Syllabus Exchange Catalog <http://www.georgetown.edu/research/nrcbl/syllabus/> A large number of syllabi are available from the National Reference Center for Bioethics Literature at the Kennedy Institute.”
  11. HOME. Homeschool Central.  
<http://www.homeschoolcentral.com/>.  
 “F.A.I.T.H. – Family Association for Instruction and Teaching at Home - F.A.I.T.H. is an active support group for homeschoolers in the western Arkansas/eastern Oklahoma (Fort Smith, Arkansas) area. Visit our site for lots of links and great information for homeschoolers everywhere!”
  12. DISABILITIES. Disability Links.  
<http://www.irs.org/disability.htm>.  
 “Professional Fit Clothing – Clothing apparel and accessories for people with disabilities.”
  13. PEER. Peer Resources – Links from the WEB to Peer Resources.  
<http://www.peer.ca/Links.html>.  
 “S.O.S., Inc. – Students for Other Students – A non-profit organization dedicated to the funding, development and operation of peer tutoring programs in public primary and secondary schools. They offer a number of resources to schools seeking to operate student-to-student tutoring programs, including sample documents, regulations and protocols. Their online bookstore features copies of studies that demonstrate the benefits of peer tutoring, books for sale, and bibliographic information. They also provide links to similar programs that involve peer tutoring and related ideas.”
  14. STORYTELLING. Storytelling on the Internet.  
<http://www.storynet.org/resources/links.htm>.  
 “<http://www.funfelt.com> Fun Felt for Kids – The Story Teller Imaginatively printed felt boards, felt books, felt masks and puppets for storytelling with children. Fairy tales, nursery rhymes and classic stories come to life with the washable, durable felt visual aids from The Story Teller, Inc. Free catalog and felt sample on request.”
  15. RIVER. Watershed Education Resources.  
<http://www.igc.org/green/resources.html>.  
 “Connecticut River – Connecticut River Education Initiative CREI consists of a unique consortium of non-profit educational institutions from four New England states working together to establish a broad spectrum of educational resources inspired by the study of the Connecticut River watershed.”

Mean words per description were as follows: MAPS, 26.3; SECURITY, 14.8; HISTORY, 30.1; AUDIO, 17.9; HR, 15.1; CONSUMER, 48.3; DIGITAL, 38.6; HIV, 20.0; BIOLOGY, 9.2; ETHICS, 44.8; HOME, 26.9; DISABILITIES, 30.6; PEER, 36.7; STORYTELLING, 39.1; RIVER, 29.0. These average lengths correspond to the lower half of average lengths in the study of Wheatley and Armstrong (1997), with the shortest length similar to that for Yahoo! in that study, the longest length similar to that for Magellan and less than half that for the Lycos Top 5%.

### Common phrases in descriptions

Table 1 shows the most common phrases of two or more words in all fifteen sets of external descriptions taken together. Several broad categories can be identified, though the correct category to which to assign a given phrase may not always be immediately clear. The first category may be referred to as *stop phrases*, word sequences that are generally common in English and whose occurrence is of relatively little significance; in this category, we may place AND TO, WITH THE, AS A, AT THE, AND A, BY THE, AND THE, TO THE, FOR THE, IN THE and OF THE.

The second category may be called *general description formulas*, word sequences that are likely to be more common in descriptions of Web sites or pages than in other types of texts, though they will not necessarily be used in all, or even most, sets of such descriptions. These may categorize the form of information found ([AND] LINKS [TO], RESOURCE[S] FOR, COLLECTION OF, [OF/AND] INFORMATION [AND], GUIDE TO and [A] VARIETY [OF]); mark descriptions of subject matter or intended audience (ABOUT THE, INFORMATION [FOR/ABOUT/ON], and DEDICATED TO and ON THE), of the author or publisher (FROM THE) or of the site or page in

general (WEB SITE, SITE IS, HOME PAGE, and IS [AN/THE]); indicate aims (TO PROVIDE and IS TO), geographical or other scope (U S, THE WORLD, THE WEB and THE INTERNET); form part of the identification of the author or publisher (UNIVERSITY OF, CENTER FOR and DEPARTMENT OF); connect parallel parts of the description or allude to information which could not be included in the description (AND MORE, AND OTHER and AS WELL [AS]); or simply be common parts of the URLs, which are quite often included in the visible text of the descriptive paragraphs (HTTP WWW).

The third broad category may be termed *description-set-specific formulas*; these are word sequences that are especially common in particular sets of descriptions, either because of subject matter or because of individual stylistic preferences or policies. In this category we may place HISTORY OF (14 occurrences in MAPS), EDUCATION AND, HOME EDUCATORS (34 occurrences in HOME), RESEARCH AND (10 occurrences in DISABILITIES), FOR PEOPLE (16 occurrences in each of HIV and DISABILITIES), SOCIAL STUDIES (36 occurrences in HISTORY), SUPPORT GROUP (34 occurrences in HOME), HIV AIDS (35 occurrences in HIV), (IS A) MANUFACTURER (OF) (40-41 occurrences in AUDIO), NON PROFIT, TO HELP, (PEOPLE) WITH (DISABILITIES) (16 occurrences of the 3-word sequence in DISABILITIES) and THE NATIONAL (13 occurrences in HISTORY).

Sequences that fall into this last category may not appear in Table 1 because they are relatively infrequent in the descriptions as a whole. Examples in the stylistic subcategory are rare, but include THIS SITE HOSTS (10 occurrences in MAPS); instances in the subject-related subcategory are more common and include AUDIO (AND/EQUIPMENT) (22 occurrences each in AUDIO), DIGITAL AUDIO (22 occurrences in AUDIO), INC IS (27 occurrences in AUDIO), LEARNING DISABILITIES (22 occurrences in DISABILITIES), and LESSON PLANS (22 occurrences in HISTORY). Some sequences cited in the second broad category above are near the borderline of the third category; for example, HOME PAGE (20 occurrences in SECURITY, 11 in HISTORY) and IS A (151 occurrences in AUDIO)

One word sequence that was difficult to place was ON LINE.

Table 1: Most common phrases in external descriptions

385	HTTP WWW	46	UNIVERSITY OF
363	OF THE	46	WITH DISABILITIES
288	IS A	41	A MANUFACTURER
189	IN THE	41	INFORMATION FOR
145	FOR THE	41	IS A MANUFACTURER
127	INFORMATION ON	41	VARIETY OF
124	TO THE	40	A MANUFACTURER OF
123	LINKS TO	40	ASSOCIATION OF
122	ON THE	40	IS A MANUFACTURER OF
113	FROM THE	40	NON PROFIT
112	THIS SITE	40	TO HELP
106	AND THE	39	GUIDE TO
87	THE WORLD	39	HIV AIDS
80	AND OTHER	38	AT THE
80	IS THE	38	SUPPORT GROUP
73	BY THE	37	DEPARTMENT OF
70	AND MORE	37	SOCIAL STUDIES
68	DEDICATED TO	36	FOR PEOPLE
66	HOME PAGE	36	ON LINE
65	MANUFACTURER OF	35	AS A
63	IS TO	35	RESEARCH AND
63	THE NATIONAL	35	THE WEB
62	PEOPLE WITH	34	A VARIETY
61	IS AN	34	A VARIETY OF
60	AS WELL	34	HOME EDUCATORS
60	U S	34	TO PROVIDE
56	INFORMATION ABOUT	33	ABOUT THE
54	AND INFORMATION	33	EDUCATION AND
54	AS WELL AS	33	SITE IS
54	WELL AS	33	WITH THE
53	CENTER FOR	32	AND TO
53	INFORMATION AND	31	COLLECTION OF
53	RESOURCES FOR	31	OF INFORMATION
52	WEB SITE	30	AND LINKS
47	AND A	30	HISTORY OF
46	THE INTERNET	30	RESOURCE FOR

### Quoting

Use of quotation marks to set off exact repetition of phrasing from source texts is discouraged in the writing of scholarly abstracts; it is very rarely found in meta tag descriptions because of potential conflicts with HTML syntax. Generally, the external descriptions in the present study showed relatively little tendency to employ quotation marks, as indicated in Table 2, which gives the proportion of descriptions in each set containing at least one double-quote mark.

It is only the shorter sets (of fewer than 100 descriptions) that have double-quotes in more than

Table 2: Use of double quotes

Set	Descriptions containing ""	Percentage of descriptions
MAPS	11	3.9
SECURITY	3	1.0
HISTORY	19	7.3
AUDIO	9	3.0
HR	7	2.5
CONSUMER	7	20.0
DIGITAL	5	8.5
HIV	10	3.4
BIOLOGY	1	1.7
ETHICS	10	15.6
HOME	8	3.9
DISABILITIES	14	4.7
PEER	2	3.7
STORYTELLING	10	12.0
RIVER	1	2.2

about 7.5% of descriptions. At least some of this may be due to higher random variation in smaller sample sizes (compare Egghe 2001), though it might be interesting to pursue this point with more powerful analysis techniques after gathering data on a much larger number of description sets.

### *Syntactic structure of descriptions*

For the purpose of analysing general syntactic patterns, each description was considered to be divided into sentence-like segments. Segment divisions could be marked either by sentence-level punctuation marks or by HTML-coded line breaks. The sentence-level punctuation marks (periods, exclamation marks, and question marks) had been applied in the previous studies of meta tag descriptions; the HTML-coded line breaks, which do not appear in meta tag descriptions, were added in the present study. Each segment was 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 training and evaluation of inter-rater consistency, 181 meta tag descriptions gathered in one of the earlier studies were employed. After studying the first part of these and their syntactic coding, the research assistant coded the second part on his own. For the latter 80 descriptions,

there was 81% agreement between the assistant's coding and that of the researcher. When syntactic categories were collapsed to **n**, **nn**, **s**, **ss**, and **other**, agreement increased to 90%. The assistant then proceeded to code the external descriptions gathered for the present study.

Results are summarized in Table 3. The specific syntactic categories **n**, **nn**, **s**, and **ss** are included to allow comparison with the earlier studies; other categories specified were found at least 15 times in at least one of the sets.

It is evident that different sets show quite different general syntactic patterns. Most sets strongly favour one or two patterns: **nn** is the pattern of the overwhelming majority of descriptions in SECURITY, as is **s** in AUDIO; HR and HIV both also give majorities to **nn**; STORYTELLING gives a majority to **nns**; in MAPS, CONSUMER, and ETHICS, a single pattern (**nnn**, **s**, or **nns**) accounts for close to half; in HISTORY, **ns** combined with **nn** forms a majority.

The tendency to concentrate on particular patterns, as measured with Simpson's *l* (Simpson 1949), showed a significant rank correlation (0.584869;  $p < 0.05$ ) with number of descriptions in the set. Since Simpson's *l* is not biased by sample size (though its variation tends to be larger with smaller samples), the observed correlation must be due to other factors. It may be that compilers of longer lists are more likely to fall into or otherwise establish particular patterns of description than compilers of shorter lists. On the other hand, concentration varied considerably, even among lists of similar length; longer lists were by no means always more consistent in their syntactic structures.

The earlier studies of meta tag descriptions had showed a predominance of **n**, followed at some distance by **s**, **ss**, and **nn**, in that order. The dominance of somewhat more complex structures in the present study no doubt reflects in part the greater freedom of authors to expand when writing descriptive paragraphs to appear in list of links than when writing meta tags where the common advice is to limit length severely. Additional content elements appear often to be included as a result. The effect of the sampling criterion that excluded sets of descriptions of less than 5 words likely also had some influence on the results, though this is assumed to be relatively minor.

Table 3: Syntactic structures

	Maps	Security	History	Audio	Hr	Consumer	Digital	HIV	Biology	Ethics	Home	Disabilities	Peer	Storytelling	River
n	1 (0%)	3 (1%)	11 (4%)	12 (4%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	15 (25%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
nm	0 (0%)	25 (8%)	24 (9%)	0 (0%)	32 (11%)	0 (0%)	0 (0%)	28 (9%)	6 (10%)	0 (0%)	16 (8%)	23 (8%)	2 (4%)	0 (0%)	1 (2%)
nn	1 (0%)	253 (84%)	62 (24%)	3 (1%)	167 (59%)	0 (0%)	1 (1%)	158 (53%)	12 (20%)	0 (0%)	66 (32%)	57 (19%)	20 (37%)	0 (0%)	10 (22%)
nnm	26 (9%)	0 (0%)	0 (0%)	0 (0%)	9 (3%)	0 (0%)	2 (3%)	19 (6%)	2 (3%)	0 (0%)	3 (1%)	1 (0%)	1 (2%)	0 (0%)	3 (7%)
nnn	114 (41%)	0 (0%)	4 (2%)	0 (0%)	0 (0%)	0 (0%)	12 (20%)	22 (7%)	0 (0%)	9 (14%)	16 (8%)	2 (1%)	2 (4%)	7 (8%)	5 (11%)
nns	59 (21%)	0 (0%)	0 (0%)	0 (0%)	1 (0%)	0 (0%)	8 (14%)	0 (0%)	0 (0%)	29 (45%)	9 (4%)	6 (2%)	1 (2%)	47 (57%)	8 (18%)
nnss	11 (4%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1 (1%)	0 (0%)	0 (0%)	16 (25%)	3 (1%)	0 (0%)	1 (2%)	9 (11%)	0 (0%)
nnv	25 (9%)	1 (0%)	0 (0%)	0 (0%)	5 (2%)	0 (0%)	1 (1%)	11 (4%)	0 (0%)	1 (2%)	4 (2%)	1 (0%)	0 (0%)	0 (0%)	0 (0%)
ns	0 (0%)	5 (2%)	86 (33%)	5 (2%)	14 (5%)	1 (3%)	0 (0%)	7 (2%)	0 (0%)	0 (0%)	27 (13%)	52 (17%)	4 (7%)	0 (0%)	8 (18%)
nss	0 (0%)	2 (1%)	13 (5%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	14 (7%)	15 (5%)	2 (4%)	0 (0%)	3 (7%)
nv	0 (0%)	4 (1%)	3 (1%)	0 (0%)	44 (15%)	0 (0%)	0 (0%)	17 (6%)	1 (2%)	0 (0%)	8 (4%)	81 (27%)	3 (6%)	0 (0%)	4 (9%)
nvs	0 (0%)	0 (0%)	1 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	20 (7%)	0 (0%)	0 (0%)	0 (0%)
s	0 (0%)	0 (0%)	15 (6%)	234 (79%)	0 (0%)	16 (46%)	0 (0%)	1 (0%)	0 (0%)	0 (0%)	1 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
ss	0 (0%)	0 (0%)	1 (0%)	15 (5%)	0 (0%)	5 (14%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
other	44 (16%)	7 (2%)	37 (14%)	29 (10%)	13 (5%)	13 (37%)	34 (58%)	35 (12%)	24 (40%)	9 (14%)	39 (19%)	42 (14%)	18 (33%)	20 (24%)	3 (7%)
total	281	300	257	298	285	35	59	298	60	64	206	300	54	83	45

The frequency of structures beginning with **n** can largely be accounted for by the use of titles and/or URLs as introductory elements in the descriptions, much as citations are included before the bodies of scholarly abstracts.

### Content types

The following inductively derived categories were applied for analysing the types of content found in the descriptions: **a** = personal author name; **e** = evaluative wording; **m** = miscellaneous (e.g., instructions); **o** = organization or company name; **p** = purpose of organization or products or services of company; **s** = subject scope of page or site; **t** = title-like phrase (more than just organization or company name); **u** = URL; and **w** = who are the intended users. The order of inclusion of different content categories was disregarded. For example, the following description (URL: <http://www.sonic.net/yronwode/sacredland.html>) would receive the categorization **stu**:

The Sacred Landscape

<http://www.sonic.net/yronwode/sacredland.html>

About archaeoastronomy, sacred geometry, symbolic landscaping and more.

Where an element in a description fitted more than one content type, all applicable codes were assigned.

Figure 2 shows the relative frequency of the nine content categories in the 15 sets.

Evidently, subjects and title-like phrases (**s** and **t**) predominate in almost all sets (with the exceptions being AUDIO and CONSUMER); URLs (**u**) are very common in a minority of sets and almost entirely absent in the rest; purpose or product information (**p**) varies considerably in frequency, but in none of the sets does it approach universal inclusion; other categories – personal authors, evaluation, miscellaneous, and intended users (**a**, **e**, **m**, **w**) – are in a minority in all sets, with the exception of intended users (**w**) in CONSUMER and DISABILITY.

Figure 2: Relative frequency of content categories

	a	e	m	o	p	s	t	u	w
MAPS	3%	11%	11%	23%	14%	84%	81%	100%	18%
SECURITY	5%	9%	1%	21%	4%	96%	97%	0%	1%
HISTORY	4%	23%	6%	36%	15%	88%	81%	4%	35%
AUDIO	2%	29%	2%	77%	74%	33%	7%	2%	7%
HR	2%	6%	1%	45%	11%	89%	78%	7%	15%
CONSUMER	0%	20%	29%	69%	54%	60%	0%	100%	66%
DIGITAL	29%	14%	9%	63%	15%	83%	83%	92%	5%
HIV	3%	10%	5%	49%	11%	91%	83%	4%	21%
BIOLOGY	2%	30%	37%	17%	2%	100%	95%	2%	12%
ETHICS	13%	13%	0%	69%	38%	66%	61%	100%	16%
HOME	2%	8%	5%	51%	42%	61%	53%	0%	23%
DISABILITIES	3%	13%	3%	52%	45%	58%	55%	0%	76%
PEER	6%	20%	6%	43%	39%	72%	61%	0%	37%
STORYTELLING	10%	8%	6%	4%	12%	87%	77%	100%	22%
RIVER	0%	2%	2%	33%	24%	78%	93%	0%	9%

Content type pattern concentration showed a non-significant rank correlation with number of items (0.373363), weaker than that shown by concentration of syntactic patterns. The highest concentration ( $l=0.449766$ ) was found in SECURITY, where 195 out of 300 items were coded as **st** (subject scope with title-like phrase); the second highest ( $l=0.336364$ ) was found in a rather small set, RIVER, where 25 out of 45 items were again coded as **st**. The same pattern **st** was also the most common in HISTORY, HR, HIV, BIOLOGY, PEER, and second most common in HOME and DISABILITIES; it was never found in MAPS or STORYTELLING, in both of which, however, the related **stu** was by far the most common, or in ETHICS, where, however, **stu** was the second most common, or in CONSUMER or DIGITAL, and only 7 times in AUDIO, where **eop** and **op** predominated.

#### *Density of description words and phrases*

The sets consisting of more than 250 external descriptions (MAPS, SECURITY, HISTORY, AUDIO, HR, HIV and DISABILITIES) were used for statistical comparison with textual features of the pages described.

The degree of match between the external description and any meta tag descriptions on the pages themselves was measured by the density of external description words and two-word phrases. These densities were significantly more often higher in the meta tag descriptions than in the bodies of the pages for MAPS, HISTORY, AUDIO, HIV and DISABILITIES for both words and phrases and for words only for SECURITY and HR. Mean word densities were 40.3% for MAPS, 28.1% for SECURITY, 44.8% for HISTORY, 38.5% for AUDIO, 29.4% for HR, 39.2% for HIV and 40.9% for DISABILITIES; mean phrase densities were 16.7% for MAPS, 13.7% for SECURITY, 22.7% for HISTORY, 16.8% for AUDIO, 11.1% for HR, 16.0% for HIV and 18.1% for DISABILITIES.

In the earlier study of meta tag descriptions (Craven submitted), titles had consistently performed better in density of description words and phrases in comparison with the body of the visible text. In the present study of external descriptions, the same finding applied with high statistical significance ( $p=0.000000$ ) to all seven sets examined. The average densities were also considerably higher than those in the earlier study: in that study, word densities averaged around 40% and phrase densities around 20%; in the present

study, means for description word densities in titles were 61.9% for MAPS, 69.4% for SECURITY, 58.0% for HISTORY, 47.3% for AUDIO, 50.7% for HR, 60.4% for HIV, and 51.6% for DISABILITIES, and means for description phrase densities in titles were 43.0% for MAPS, 61.3% for SECURITY, 44.1% for HISTORY, 30.1% for AUDIO, 34.5% for HR, 42.2% for HIV, and 38.4% for DISABILITIES. It may also be noted that external description word and phrase densities were generally much higher in titles than even in the pages' meta tag descriptions.

The external descriptions in these sets all appear frequently (but not always) to include the page titles word for word. Even if the intention is in fact to repeat the page title exactly, at least two factors may contribute to discrepancies: (1) insufficiently frequent updating of the data in the list in comparison with the frequency of update of the pages referenced; (2) derivation of titles in the list from other portions of the pages than those marked with the TITLE tag, such as from text marked with the H1 tag.

In the earlier study, meta tag keywords, when compared to the page bodies, had also shown significantly greater densities of meta tag description words in all four sets analysed, and of phrases in all but one. In the present study, keywords showed significantly higher densities of external description words than page bodies in SECURITY ( $p=0.0022326$ ), AUDIO ( $p=0.0050784$ ), HR ( $p=0.0184334$ ) and HIV ( $p=0.0001304$ ), but not in MAPS, HISTORY or DISABILITIES. Moreover, keywords showed significantly lower densities of external description phrases than page bodies in both MAPS and HISTORY.

Other portions of the page that had shown at least some significant indication of higher description word or phrase density in the earlier study were the first 200 characters of the body and text marked with the CENTER and FONT tags. In the present study, all seven sets showed significantly higher description word and phrase densities for the first 200 characters ( $p < 0.001$ , except  $p < 0.05$  for phrases for DISABILITIES). Densities were substantially lower, however, than those found in titles: word density means were 27.1% for MAPS, 20.3% for SECURITY, 28.0% for HISTORY, 20.3% for AUDIO, 20.0% for HR, 26.7% for HIV and 29.1% for DISABILITIES; phrase density means were 9.8% for MAPS, 7.9% for

SECURITY, 10.2% for HISTORY, 7.4% for AUDIO, 8.8% for HR, 11.8% for HIV and 13.8% for DISABILITIES. Portions marked with the CENTER tag were not significantly likely to show either higher word density or higher phrase density than page bodies as a whole in any of the seven sets, except for words in SECURITY. The FONT tag did show a significant difference for words in MAPS and for phrases in AUDIO, HR, HIV and DISABILITIES, but not for either in SECURITY or HISTORY.

In the meta tag description study, text tagged with A had throughout performed significantly worse at description word and phrase density than the body text as a whole. The A tag thus appeared to be a potential negative cue, though a weak one because the differences in densities were in fact generally small. This same significant difference was observed in the present study for both words and phrases in all seven sets except SECURITY, where it applied to phrases only. A similar significantly worse performance for text tagged B was observed for phrases in MAPS, SECURITY and HIV, but not for either words or phrases in HISTORY, AUDIO, HR or DISABILITIES.

#### *Clumping of description words in page texts*

A measure of "clumpiness" had previously been developed (Craven submitted) to determine whether any particular parts of the bodies of Web pages were more likely to contain words from the meta tag descriptions and thus whether further investigation of cue-based passage extraction might be helpful. For the present study, this measure was applied to occurrences of external description words in the bodies of the referenced pages. If significantly more pages showed positive values for clumpiness than showed negative values, then it could be concluded that, in at least some of the pages, words important to the description did tend to concentrate in particular parts of the page body.

In fact, although there had been a significant preponderance of positive clumpiness for internal description words in one set examined in the previous study, no significant preponderances were discovered for external description words in MAPS, SECURITY, HISTORY, AUDIO, HR or HIV. Indeed, for MAPS and SECURITY, there was

a (statistically non-significant) preponderance on the negative side. A significant preponderance on the positive side, however, was discovered for DISABILITIES.

The negative results for general clumpiness seem slightly at odds with the significant word and phrase density results for the first 200 characters. One possible explanation is the presence of other factors that generally tend to promote negative clumping. For example, many common words, such as "the", which tend to occur in many descriptions, naturally almost never occur in immediate proximity in English text. On the other hand, a tendency for stopwords to repel one another does not appear to hold strongly over longer distances. For instance, a test of the clumpiness of the single stopword *and* (the most common word in the descriptions in this study) in a typical text yielded a value very close to zero (-0.018). Moreover, when more different words are to be matched, any tendency toward repulsion seems to be even less. For instance, a test applied to the nine stopwords *on, is, in, a, for, to, of, the,* and *and* (the nine most common words in the descriptions in this study) in the same text yielded a low positive value (0.023).

### Conclusion

Results of this study suggest that people composing longer lists of Web page descriptions tend to adopt some degree of formalism, in terms both of specific phrasing and of syntactic structures. Tools to assist in the application of formulas might be of use; for example, a form dividing the description into different content elements, with reminders as to what sort of syntactic structure fits into each slot, and possibly with the automatic supplying of certain standard phrases. Such tools might be especially helpful for large-scale ongoing projects; Thomas and Griffin (1998), for example, think that the future of meta data may lie with commercial indexing services, which might be driven by the profit motive to produce products of higher quality than other sources.

Formulas need not necessarily be supplied only in computerized assistance tools. Useful formulations might also be suggested by printed or on-line guides or cataloguing codes, where they might be presented either explicitly or as parts of more extended examples of suitable practice.

As general recommendations for automatic extraction to assist in description writing, the results might be taken as suggesting the following: extract the title, which is almost always available and of a reasonable length; extract also any meta tag description that does not simply duplicate title wording; if the title and description are short, or at the user's request, extract also the first 200 characters of the body; if using other extraction methods, such as word frequency, as well, give a small negative weight to text with the A tag. This advice is similar, but not identical, to advice inferred from the earlier study of meta tag descriptions; the major differences lie in the addition of the meta tag description as an extractable element and the downgrading of meta tag keywords.

Again, these findings could be applied also outside the immediate context of automatic assistance tools. Guides or textbooks aimed at those who will be summarizing Web pages could draw especial attention to the page title and meta tag description, and, failing these, the first 200 characters of the body, as sources to be examined.

The present study used external page descriptions as the standard of comparison for estimating importance of parts of web pages for summarization. It is not clear that such descriptions are particularly good, though they appear more consistent in content and form than meta tag descriptions. Previous investigations by the author have clearly shown that the latter are certainly not particularly consistent in content or form. Inconsistencies and other defects have also been demonstrated in published author abstracts (Pitkin, Branagan, & Burmeister 1999).

One possible future research approach might be to employ other evidence to estimate the perceived quality of sets of page descriptions and then investigate possible associations between perceived quality and objective description features. Potential indicators of perceived quality might include number of links to a page, evaluative language associated with those links, and appearance on best-site lists of search services.

Greater quality and consistency in descriptions might be expected to be one possible benefit of adoption of computerized assistance tools. Availability of more complete published guidelines might have a similar effect, as might greater centralization and professionalization of description

construction, in the same or similar context to library cataloguing or periodical indexing.

Although some writers may find formulation tools and automatic extracts useful, it should not be assumed that all description authors would wish to employ them. Instead, it is likely that different tools will suit different types of users. It is expected that individuals will use quite different approaches in writing descriptions, just as they have been reported to do in writing abstracts of scholarly articles (Endres-Niggemeyer, Waumans, & Yamashita 1991). Outside of the automatic assistance context, this observation would also be an argument in favour of guidelines, advice, and examples rather than rigid rules.

At fifteen and seven, the numbers of sets examined for content and syntax on the one hand and word and phrase matches in parts of the described pages on the other is quite small. Research based on a larger sample of sets may therefore be worth undertaking.

The word and phrase-matching portion of the study focused only on exact duplication. Further research might look at the effect of applying stemming algorithms or thesaurus lookup. Some preliminary experimentation with the latter has, however, proved somewhat unpromising, on account of the great variety and specificity of the vocabulary of Web pages combined with the relative brevity of the descriptions.

### Acknowledgements

Research reported in this article was supported in part by the University of Western Ontario Office of Research Services with funds provided by the Natural Sciences and Engineering Research Council of Canada. The extensive assistance of research assistant Emmett Macfarlane in data gathering is also acknowledged.

### References

- Almind, T.C. and Ingwersen, P. 1997. Informetric analyses on the World Wide Web: methodological approaches to 'Webmetrics'. *Journal of Documentation* 53 (4): 404–426.
- Amitay, E. 2001. What lays in the layout. URL: <http://www.ics.mq.edu.au/~einat/thesis/> [viewed February 20, 2002].
- 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): 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 59<sup>th</sup> ASIS Annual Meeting 1996 (Volume 33)*, Baltimore, Maryland, October 21–24, 1996. Medford, New Jersey: Information Today: 203–208.
- Craven, T.C. 1998. Human creation of abstracts with selected computer-assistance tools. *Information Research* 3 (4): paper 47. URL: <http://www.shef.ac.uk/~is/publications/infres/paper47.html> [viewed February 20, 2002].
- Craven, T.C. 2000. Features of DESCRIPTION META tags in public home pages. *Journal of Information Science* 26 (5): 303–311.
- Craven, T.C. 2001. 'DESCRIPTION' META tags in locally linked Web pages. *Aslib Proceedings* 53 (6), 203–216.
- Craven, T.C. Submitted. HTML tags as extraction cues for Web page description construction.
- Endres-Niggemeyer, B. 1998. *Summarizing information*. Berlin: Springer.
- Endres-Niggemeyer, B., Waumans, W., and Yamashita, H. 1991. Modelling summary writing by introspection: a small-scale demonstrative study. *Text* 11 (4): 523–552.
- Egghe, L. 2001. A noninformetric analysis of the relationship between citation age and journal productivity. *Journal of the American Society for Information Science and Technology* 52 (5): 371–377.
- 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. *Journal of the American Society for Information Science* 51 (13): 1159–1176.
- Henshaw, R.; Valauskas, E.J. 2001. Metadata as a catalyst: experiments with metadata and search engines in the Internet journal, First Monday. *Libri* 51 (2): 86–101.
- 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.
- 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., and Hall, C.M., eds. *Encyclopedia of library and information science*, volume 53 (supplement 16). New York: Dekker: 16–27.
- Pitkin, R.M.; Branagan, M.A.; Burmeister, L.F. 1999. Accuracy of data in abstracts of published research articles. *JAMA*. 281 (12): 1110–1111.
- Simpson, E.H. 1949. Measurement of diversity. *Nature* 163: 688.
- Thomas, C.F.; Griffin, L.S. 1998. Who will create the metadata for the Internet? *First Monday* 3 (12). URL: [http://firstmonday.org/issues/issue3\\_12/thomas/](http://firstmonday.org/issues/issue3_12/thomas/) [viewed February 20, 2002].
- 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.
- Wheatley, A.; Armstrong, C.J. 1997. Metadata, recall, and abstracts: can abstracts ever be reliable indicators of document value? *Aslib proceedings* 49 (8): 206–213.

*Editorial history:*

*final version received 29 January 2002;*  
*accepted 12 February 2002.*