Information Research, Vol. 7 No. 3, April 2002,

What is the title of a Web page? A study of Webography practice

Timothy C. Craven
Faculty of Information and Media Studies
University of Western Ontario
London, Ontario N6A 5B7
Canada

Abstract

Few style guides recommend a specific source for citing the title of a Web page that is not a duplicate of a printed format. Sixteen Web bibliographies were analyzed for uses of two different recommended sources: (1) the tagged title; (2) the title as it would appear to be from viewing the beginning of the page in the browser (apparent title). In all sixteen, the proportion of tagged titles was much less than that of apparent titles, and only rarely did the bibliography title match the tagged title and not the apparent title. Convenience of copying may partly explain the preference for the apparent title. Contrary to expectation, correlation between proportion of valid links in a bibliography and proportion of accurately reproduced apparent titles was slightly negative.

Introduction

It is by now well known that the growth of the World Wide Web since its inception in 1993 has created the need for new forms of citation in bibliographies, lists of references, and similar texts. The problems are not simply ones of citation format, but also ones of citation content. That is, it is not simply a question of what data elements are to be included, in what order, and with what punctuation, font, and capitalization, but also of how the values of those data elements are to be derived.

This paper is concerned with what may be considered a key data element in most bibliographic references: the title of the item cited. Specifically, it addresses in a preliminary way the question of what people making references to Web pages think the titles of those pages are.

Although there are a number of style guides that cover citation of Web pages to some degree, few give any indication from where the title of a cited item is to be taken, unless the item exactly duplicates a printed publication, in which case the title of the printed form may be used.

By contrast, Land (2001) is very specific about using the title marked with the TITLE tag, with the option of adding the contents of a heading (marked, for example, with H1) as a subtitle, if substantially different.

Ivey's (1997) observation that "a properly constructed HTML document must have a title" likewise suggests that the TITLE element should be the preferred source. A similar observation is found in Fletcher and Greenhill (1997). Ivey (1997) also notes that "it may be appropriate to include the title of the larger publication" as well, if this is known.

The opposing view, that preference should be given to the title as it appears in the main display, is found in Estivill and Urbano (1997), who recommend that the TITLE element be used as the title only if a main display title is absent (they also allow the TITLE element to be included in a note if it differs from the main display title and if it is

considered significant).

In contrast to the usual vagueness about the source of the title for Web pages, two commonly used style guides provide much more specific advice regarding e-mail messages and postings to news and discussion groups. For this type of document, the *MLA Handbook* (Gibaldi, 1999, p. 199) states that the title should be taken from the *subject* line; the APA manual (American Psychological Association, 2001) says, differing slightly, that the subject line should be used in the citation *instead of* the title.

A minority of authors of Web bibliographies may be familiar at least with the existence of the *Anglo-American Cataloguing Rules* (Joint Steering Committee for Revision of AACR, 1998), but the advice given there seems not particularly helpful for deciding about Web page titles. The general AACR rule for the information source for titles is the following: "Take information recorded in this area from the chief source of information for the material to which the item being described belongs" (p. 17). As applied to computer files, the chief source of information is the *title screen* (p. 222), defined (p. 624) as "a display of data that includes the title proper and usually, though not necessarily, the statement of responsibility and the data relating to publication". Strictly, the definitions are circular: the title is defined as text found on the title screen, and the title screen is a display that contains the title. The reference to the optional presence of information on responsibility and publication, however, does suggest that preference should be given to the more complete display in deciding what is to be viewed as the title screen. Preference for the more complete information source is certainly stipulated in cases where nothing can be identified as a title screen:

> If there is no title screen, take the information from other formally presented internal evidence (e.g., main menus, program statements, first display of information, the header to the file including "Subject:" lines, information at the end of the file. In case of variation in fullness of information found in these sources, prefer the source with the most complete information. (p. 222)

Advice given by Olson (2001) on the OCLC Web site as to what constitutes the title of a Web page for cataloging purposes is still rather vague. "The title of an Internet resource is taken from the chief source of information.... The chief source of information for computer files available by remote access is the title screen or similar display from the terminal or a printout of that information. If there is no special display, information may be taken from the home page, web page, or file itself: 'readme file,' 'about' screen, TEI (Text Encoding Initiative) header, HTML tagging, documentation file, internal menus, labels, subject line, program statements, etc." Failing these sources, "the cataloger may use a title from any published description of, or citation to, the file" or the file name "if there is no other title given", or, failing these, "must supply a title." The idea of using the file name if no title is in evidence is also mentioned by Rudolph (2001).

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 & Grams, 2000); informetric measures (Almind & Ingwersen, 1997); links to e-journals and their articles (Harter and Ford, 2000); effects of meta-tags on retrievability (Turner and Brackbill, 1998; Henshaw & 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.

Related more specifically to the study of how others describe or identify Web pages is the work of Amitay (2001), who 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.

Methodology

A sample of bibliographies was identified by searching on Web search engines (chiefly Google and Webcrawler, with one bibliography being found with Yahoo!) on the query "bibliography web" (or "bibliography of web bibliographies" or "bibliography online web"). To be included in this study, a bibliography had to satisfy the following criteria.

- It was in HTML format (not Word, RTF, etc.).
- Most items in the bibliography appeared to be in some standard bibliographic format, with each listed in a separate paragraph, list item, or the like. Bibliographic elements, such as title, author(s), and date were clearly

distinguishable.
- When duplicates were eliminated, at least 30 items in the bibliography had valid links to HTML versions of the items to which they refer.
- It was not a bibliographic search engine interface, though it could be spread over several Web pages.

For each bibliography chosen, the following data were recorded.

- URL.
- Bibliographic data (title, author, date, publisher, etc.) to the extent these could be determined.
- Total number of items.
- Number of items with links.
- Number of links attempted in order to reach 300 valid links (or all links in the bibliography, whichever was less).
- Number of valid links followed..

For each valid link in the bibliography to an HTML page (up to the maximum of 300 per bibliography), the following were recorded.

1. The title in the bibliography;
2. The subtitle in the bibliography (if any).
3. The title tagged with the TITLE tag in the item linked to.
4. The title as it would appear to be from viewing the beginning of the page in the browser (it might be centered, in larger type, in a distinctive font or color, immediately before the author's name, or otherwise identifiable).
5. The subtitle as it would appear to be from viewing the beginning of the page in the browser.
6. Match categories (as described below) for the tagged title.
7. Match categories for the apparent title.

The following codes were used to categorize the degree of matching between a title in the page itself and the title as given in the bibliography:

- **e** - an exact match of the bibliography title formed all or part of the title to be matched (ignoring capitalization, punctuation, and initial "the" and "a");
- **p** - the title to be matched contained a rewording or abbreviated version the bibliography title;
- **n** - there was no match with the bibliography title in the title to be matched;
- **a** - the title to be matched adds an author name or initials;
- **w** - the title to be matched adds a Web site name;
- **x**- the title to be matched adds other information, such as the name of another larger containing unit (series, journal, etc.) or body (publisher, etc.)

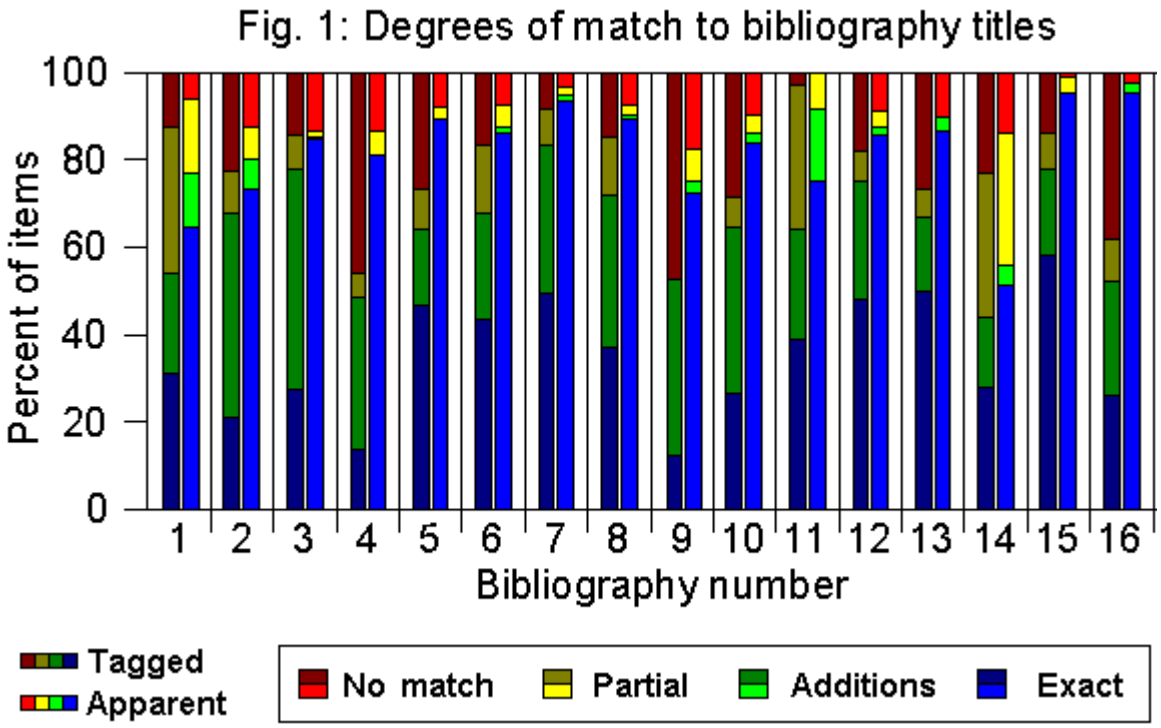
The codes **e**, **p**, and **n** were mutually exclusive; each of them could stand alone or be combined with any combination of **a**, **w**, and **x**.

Results

Details of the 16 Web bibliographies examined are given in the Appendix. The proportion of bibliography items with links ranged from about 12% (for bibliography 8) to 69% (for bibliography 13). In only one case (bibliography 3) were any of the links in the bibliography disregarded because the upper limit had been reach. The proportion of attempted links that proved valid at the time of examination ranged from 19% (for bibliography 12) to 96% (for bibliography 1).

To simplify the analysis of results for Figure 1, the match categories have been collapsed into the following:

- **exact**: code **e** alone, with no codes for additional information;
- **additions**: code **e** with any combination of codes for additional information;
- **partial**: any coding involving code **p**;
- **no match**: any other codings.



For each bibliography, the percentages are shown first for categories of match to the tagged title and then for categories of match to the apparent title.

The proportion of exact matches for tagged titles varies from 13% (for bibliography 9) to 58% (for bibliography 15); that for apparent titles varies from 51% (for bibliography 14) to 95% (for bibliographies 15 and 16). In all cases, the proportion for tagged titles is much less than that for apparent titles, with the smallest difference being 23% (for bibliography 14).

The gap is narrowed somewhat if exact matches with additions are included (with a minimum of about 11%, for bibliography 7), but the apparent titles remain ahead in all cases.

As can be seen from Figure 1, matches with additions are much more unusual for apparent titles than for tagged titles; this is no doubt true in large part to the judgment exercised by the research assistant in recording apparent titles.

Only rarely did the bibliography title match the tagged title and not the apparent title: there were two clear examples in bibliography 1, three in bibliography 2 (in one of which the apparent subtitle did, however, match the bibliography title), three in bibliography 3 (in one of which the apparent title combined with the apparent subtitle did match), one in bibliography 4, one in bibliography 5, one in bibliography 6 (with a match to the apparent subtitle), one in bibliography 7 (with a match to the apparent subtitle), none in bibliography 8, one in bibliography 9 (with a match on the combination of apparent title and subtitle), four in bibliography 10 (with one match to the apparent subtitle and one to the combined apparent title and subtitle), one in bibliography 11, none in bibliography 12, none in bibliography 13, three in bibliography 14, none in bibliography 15, and none in bibliography 16.

Discussion

One might expect that one factor affecting the accuracy with which bibliography entries reflect the titles on the original pages might be the currency of the information in the bibliography: recent changes to titles would not be reflected if the entries had not been checked recently for continued accuracy. Since currency would also be indicated by the proportion of valid links, one might expect some positive correlation between proportion of valid links and the proportion of accurate titles. This seems, however, not to be the case: the correlation between proportion of valid links and the proportion of apparent titles that were exact matches to the bibliography titles or matches with additional information was in fact slightly negative (-0.156225).

The convenience factor should probably not be ignored as a possible explanation for at least some part of the preference for the apparent title. Popular Web browsers, including Netscape Navigator, Internet Explorer, and

Opera, all make it fairly easy to copy text from the main display. To copy a tagged title, however, is more involved. Typically, one must call up a view of the source code, find the tagged title within that source, and only then copy the title. To the beginner, it may even not be clear (as in the case of some versions of Navigator) how text can be copied from the source. Addition of features to browsers to allow easier capture of obscured or hidden page elements, such as tagged titles and meta-tags, might have a substantial effect on future preferences.

Given the continuation of present browser models, compilers of Web bibliographies are probably best advised to use apparent page titles rather than tagged titles, on the grounds both of their own convenience and of consistency with more common usage.

This advice is, of course, based on a sample of only sixteen bibliographies. Although there is currently no reason to believe that substantially different results would be found with a larger sample, especially given the universality of the preference shown by all sixteen, further investigation might possibly be of value. For example, the study might be extended to the more common type of listing in which only very brief citations are given for most pages, typically a title with a link to the corresponding URL.

Some possibility of bias existed in the present study, since the research assistant who extracted the apparent title and subtitle from a Web page had already seen the bibliographic entry. A more rigorous methodology might have involved two assistants, one collecting the bibliographic items and links and the other subsequently examining the pages referenced. Having another research assistant revisit the pages at a later date might in any case be an interesting followup.

In the area of followup, future work might address the relative stability of tagged and apparent titles. If tagged titles turned out to be substantially more constant over time, that might be an argument in favour of employing them in citations, in spite of their other disadvantages in comparison with apparent titles.

Given the vagueness of the standard cataloguing rules with regard to the chief source of information for the title of a computer files, a study of actual cataloguing practice might be of interest. Do cataloguers of Web pages in fact tend to take titles from the main display window? In the few cases where information in the main display window is less complete than in the tagged title, is the tagged title preferred, or is some other information source employed?

Acknowledgments

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. & Ingwersen, P. (1997) "Informetric analyses on the World Wide Web: methodological approaches to 'Webmetrics'." *Journal of Documentation* **53** (4), 404-426.
- American Psychological Association (2001) *Publication Manual of the American Psychological Association*. 5th ed. Washington: American Psychological Association.
- Amitay, E. (2001) *What lays in the layout*. http://www.ics.mq.edu.au/~einat/thesis/.
- Craven, T.C. (2001) "'DESCRIPTION' META tags in locally linked Web pages." *Aslib Proceedings* **53** (6), 203-216.
- Gibaldi, J. (1999) *MLA handbook for writers of research papers*. 5th ed. New York: Modern Language Association.
- Estivill, A.; Urbano, B. (1997) "Cómo citar recursos electrónicos." http://www.ub.es/biblio/citae-e.htm.
- Fletcher, G. & Greenhill, A. (1997) *Academic referencing of Internet-based resources*. http://www.spaceless.com/WWWVL/refs.html.
- 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.
- Ivey, K.C. (1997). *Citing Internet sources*. EEI Press. http://www.eeicommunications.com/eye/utw/96aug.html.
- Joint Steering Committee for Revision of AACR (1998) *Anglo-American cataloguing rules*. 2nd ed., 1998 revision. Ottawa; London; Chicago: Canadian Library Association; Library Association Publishing; American Library Association.
- Land, B. (2001) *Web Extension to American Psychological Association Style*. http://www.beadsland.com/NN/ARC/1996/beadsland/ROOT/weapas/html/index/.
- 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.
- Olson, N.B. (2001) *Cataloging Internet resources*. OCLC Online Computer Library Center. http://www.oclc.org/oclc/man/9256cat/toc.htm.
- Rudolph, J. (2001) *Style manuals*. http://www.lib.memphis.edu/instr/style.htm.
- 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.

Appendix: List of Bibliographies

Bibliography 1 (Google Directory search "bibliography web")
http://www.lib.vt.edu/research/libinst/evalbiblio.html
Bibliography on Evaluating Internet Resources
Virginia Tech University Libraries
Last updated: Feb. 26/01
total items 118
items with links 58
links attempted 58
valid links followed 48

Bibliography 2 (Google Directory search "bibliography web")
http://www.archiveimpact.com/bibliography/index.html
Archive Impact Bibliography
total items 367
items with links 130
links attempted 130
valid links followed 71

Bibliography 3 (Google Directory search "bibliography web")
http://info.lib.uh.edu/sepb/toc.htm
Bailey, Charles W., Jr.
Scholarly Electronic Publishing Bibliography.
Houston: University of Houston Libraries
1996-2001.
total items 1301
items with links 354
links attempted 329
valid links followed 308

Bibliography 4 (Yahoo! Search "bibliography web")
http://arts.ucsc.edu/gdead/agdl/powers.html
Richard Powers: A Bibliography
By David G Dodd
copyright 1997-2001
Last Revised July 30/2001
total items 291
items with links 61
links attempted 61

valid links followed 37

Bibliography 5 (Google web search "bibliography of web bibliographies")
http://www-sor.inria.fr/projects/relais/biblio/
Web caching bibliography
Guillaume, Pierre
total items 344
items with links 108
links attempted 108
valid links followed 75

Bibliography 6 (Google web search "bibliography of web bibliographies")
http://www.oasis-open.org/cover/biblio.html
SGML/XML Bibliography
By: Robin Cover
Last modified: April 19, 2001
total items 2188
items with links 320
links attempted 320
valid links followed 221

Bibliography 7 (Webcrawler search "bibliography web", following list from "Contemporary Philosophy of Mind": http://www.u.arizona.edu/~chalmers/biblio.html)
http://www.cogs.susx.ac.uk/users/ezequiel/alife-page/alife.html
A Life Bibliography
Compiled by Ezequiel A. Di Paolo (c) 2000
total items 684
items with links 226
links attempted 226
valid links followed 94

Bibliography 8 (Webcrawler search "bibliography web")
http://library.usask.ca/~dworacze/BIBLIO.HTM
Electronic Sources of Information: A Bibliography
Marian Dworaczek
Last Revised: June 15/2001
total items 1362
items with links 157
links attempted 157
valid links followed 121

Bibliography 9 (Webcrawler search "bibliography web")
http://library.usask.ca/~dworacze/CENS.HTM
Censorship on the Internet: A Bibliography
Marian Dworaczek
Revised: Oct 25/00
total items 101
items with links 56
links attemptsed 56
valid links followed 40

Bibliography 10 (Webcrawler search "bibliography web")
http://www.gseis.ucla.edu/iclp/bib.html
Cyberspace Law Bibliography
by The UCLA Online Institute for Cyberspace Law and Policy
19 Aug. 2001
total items 509
items with links 212

links attempted 212
valid links followed 174

Bibliography 11 (Webcrawler search "bibliography web")
http://www.kcl.ac.uk/humanities/cch/bib/
Bibliography of Humanities Computing
Willard McCarty
5/96
total items 417
items with links 90
links attempted 90
valid links followed 36

Bibliography 12 (Webcrawler search "bibliography web")
http://www.counterpane.com/biblio/all-by-author.html
Crypto Bibliography
Copyright Counterpane Internet Security, Inc., 2001
total items 1498
items with links 292
links attempted 292
valid links followed 56

Bibliography 13 (Webcrawler search "bibliography web")
http://www.markus-enzenberger.de/compgo_biblio/compgo_biblio.html
Computer Go Bibliography
Markus Enzenberger
July 21/01
total items 144
items with links 100
links attempted 100
valid links followed 30

Bibliography 14 (Webcrawler search "bibliography web")
http://www.ala.org/alcts/publications/netresources/bib_main.html
Standardized Handling of Digital Resources Bibliography
Preston, Ahronheim et al.
American Library Association, 2000
total items 170
items with links 79
links attempted 79
valid links followed 43

Bibliography 15 (Webcrawler search "bibliography web")
http://www.cs.wpi.edu/~webbib
Webbib Online Bibliography - WWW as a distributed system
Note: items retrieved from:"Core set of bibliography references" section
Craig E. Wills
total items 805
items with links 274
links attempted 274
valid links followed 122

Bibliography 16 (Google Search: "bibliography online web")
http://www.enolagaia.com/Bib.html
BIBLIOGRAPHY - Autopoiesis and Enaction
Dr. Randall Whitaker
total items 525
items with links 70

links attempted 70
valid links followed 42

How to cite this paper:

Craven, T. (2002) "What is the title of a Web page? A study of Webography practice" ***Information Research***, **7** (3) Available at: http://InformationR.net/ir/7-3/paper130.html

 Updated: 20th March 2002

Check for citations, using Google Scholar

Contents **Home**