ALCTS
Association for Library Collections & Technical Services

Committee on Cataloging: Description & Access

Task Force on Metadata and the Cataloging Rules

Final Report (*continued*)

Table of Contents

Appendix:
Cataloging Problems with Web Sites

By Ann Sandberg-Fox

[The following report and recommendation was submitted by a consultant to the Task Force. Although the Task Force did not accept this recommendation, this report makes some interesting observations about the use in *AACR2* cataloging of metadata and other information found in HTML source code.]

Following is a report on cataloging problems involving web sites. Web sites constitute a growing category of Internet resources that are currently being cataloged. These sites are characterized by their extensive information and hyperlinks to other related resources, which have made them valuable information resources for library users. Web sites are made up of "pages" that contain text and graphics (and sometimes sound and video). Web pages to date are predominantly HTML-based, that is, they have been created using the Hypertext Markup Language or HTML. The terms "web page" and "HTML page" are often used interchangeably. However, for the purpose of the cataloging discussion presented, they are hereby distinguished: A web page (including home or main page) refers to visible, eye-readable, or so-called "surface" data that are displayed to the library user when accessing a web site, or URL; an HTML page refers to the coded commands called "tags" with "underlying" data that are used by a web browser to retrieve and display, in addition to assembling and arranging, the web page. To see the coded commands for any web page, select View Document Source when using Netscape Navigator or View Source when using Microsoft Internet Explorer.

The cataloging problems associated with web sites discussed here concern the title and statement of responsibility area (MARC 245 $a and c).

Title Proper

Very early on, catalogers reported difficulties in determining the title proper. Electronic discussions, such as those that took place on the INTERCAT list (Sept. 1995), centered on the following question: What do catalogers use as the title of a web site — the information presented as the title on the web page, or the information shown, sometimes in a title bar or box, at the very top of the Netscape or Microsoft Internet Explorer window, as illustrated in Figure 1. (The titles in both cases may be the same; however, they most frequently are not.)

In the case of the title shown at the top of the window in Figure 1, its source is the text that has been recorded between the <TITLE> tags in the HTML page. There can also be additional title information coded as ALT (alternative) in the HTML page, which is used by older web browsers that aren't able to display graphics. An alternative title can, in turn, differ from the title

given in the <TITLE> tags. Also seen in the HTML page is the text encoded in the <CENTER> tag which displays as the title on the web page in Figure 1. It is this text that the cataloger transcribed as the title proper in field 245 $a (OCLC record 36996298).

There was no consensus among the catalogers who posed the question on INTERCAT as to which title source to prefer. The argument for using the HTML title in field 245 $a was that it was assigned by the creator or producer of the file. The counter argument for using the web page title was that it was presented as the title to the user; it was also in conformance with *AACR2* rule 9.0B1 (including its recent amendment) to use formally presented eye-readable internal evidence as the chief source. Examples of formally presented eye-readable internal sources listed in the rule that can be used in the absence of a title screen are: "main menus, program statements, first display of information, the header to the file including 'Subject' lines, and information at the end of the file."

A recent informal sampling of the source of title notes in computer file bibliographic records in the INTERCAT database indicate the cataloger's preference for giving the web page title in field 245 $a with the HTML title being treated as a variant title that is recorded in field 246. This seeming resolution of the problem masks a complicating factor that was noted in the INTERCAT discussion mentioned, namely, the treatment of HTML titles that are found to contain fuller or more complete information than the titles displayed on the web pages. An example is illustrated in Figure 2. Following amended rule 9.0B1, the cataloger is directed to prefer the source with the most complete information when there is variation in fullness of information found in the sources cited above. However, the HTML page containing "hidden," non-visible, coded data technically does not qualify for consideration as a preference (although the data could qualify as being formally presented). In the bibliographic record for this example, the cataloger conformed to 9.0B1 and transcribed the web page title in field 245 $a and the fuller HTML title in field 246 (OCLC record 36861284). In some few cases, the sampling of the source of title notes showed opposite treatment, with the fuller HTML title being transcribed in field 245 $a and the web page title being given in field 246.

Statements of Responsibility

In addition to problems with the title proper, there are also problems in recording statements of responsibility for web sites. These problems have been identified recently as catalogers have begun to examine HTML pages more closely. The problems reported concern textual differences between statements of responsibility displayed on web pages and the encoded statements given in the HTML pages. The example of the Orange County, California, web site illustrates this situation in Figure 3. On the web page, there appears the following information: "Joshua Wallingford, County Webmaster." The HTML page (2nd line) identifies "Joshua Wallingford" as having "created and coordinated" the Orange County web site. This text, however, appears in an HTML "Comments" tag; it does not display on the web page and can be seen only by viewing the HTML page. The HTML page also contains a <META> tag which names the "author" as "Joshua Wallingford, County Webmaster/Internet Coordinator." Finally, the HTML page (about two screens down) contains a <CENTER> tag giving update and responsibility information; it is the text contained within this tag which is displayed on the web page shown in Figure 3. In the bibliographic record for this web site, the cataloger chose not to record this statement (OCLC record 36782187). This is not atypical. Individuals named in conjunction with these and other responsibilities, such as web curators, web owners, web administrators, and web maintainers, frequently change, which has led catalogers to question their usefulness in notes and as potential access points to the resource.

If the cataloger had viewed and examined the HTML page, it is most likely that she/he would have added the statement of responsibility contained in the "Comments" tag somewhere in the bibliographical record, perhaps in a 500 note, with an access point for Wallingford.

In any case, the problems described above beg the question of whether HTML pages qualify as a possible chief source of information in the cataloging of Internet resources. So far, as noted, the examples of alternate chief sources listed in 9.0B1 are limited to those that are eye-readable or displayed to the user.

This is much the case with the *ISBD(ER)* (K.G. Sauer, 1997) in which stipulation 0.5.1 lists the examples given in 9.0B1 and, in addition, includes “home page,” “TEI header,” and “other identifying information prominently displayed.” In the case of OCLC’s *Cataloging Internet Resources* (OCLC, 1997), the sources of information noted are more expansive and include “HTML tagging” in addition to “home page,” “web page,” and examples of the “file itself.”

What to Do?

It is not sufficient to merely recommend that HTML pages/tagging or “metadata” be added to the list of examples in rule 9.0B1. First, the examples listed in 9.0B1 are designated for use expressly as substitutes in place of an absent title screen. This is not the situation with web sites where the title screen (web/home/main page) is always in evidence, having been assembled and displayed from the encoded commands given in the HTML source page. To account for this situation, the text of 9.0B1 requires rewriting, more along the lines of stipulation 0.5.1 in the *ISBD(ER)*, which begins with the comprehensive statement: “Sources internal to the electronic resource shall be preferred to all other sources. Such information must be formally presented (e.g. in the title screen…).”

Second, the terms “HTML” and “metadata” are problematic with the former more than likely to be replaced by the new markup language XML (Extensible Markup Language) that has been designed for use over the Web, and the latter being too generic to be of help to the cataloger. There are after all many forms of metadata. Cataloging data, for example, qualifies as a form of metadata as does the Dublin Core.

Rather than getting bogged down in definitions or proposed wording at this late stage of work, it is suggested here that the TEI-TG recommend there be a future re-write of 9.0B1 that takes into account the problems noted. Of particular importance is the need to consider the use of any internal source with formally presented information as a chief source (this would be in line with 0.5.1 in the *ISBD(ER)*, and would depart from the present text in 9.0B1 which recognizes only the title screen as the chief source and everything else as an alternate). Also of importance is the

need to consider two possible categories of formally presented information to be used as appropriate in the cataloging of Internet resources: 1) that which is visible, eye- readable, or “surface" information displayed on the screen (e.g., a web page); and 2) that which is non-visible, hidden, or “underlying,” coded information that is displayed with a web browser. (These categories could also be tweaked to apply to non-Internet resources.)

Since there is a CC:DA Task Force on the *ISBD(ER)* which is looking into harmonizing the rules in Chapter 9 with the stipulations in the *ISBD(ER)*, any recommendation the TEI-TG proposes regarding 9.0B1 should be referred to this task force.