Citations in Hypermedia: Implementation Issues

Peter Jörgensen

Internet sources are increasingly used in scholarly work at all levels, yet it is often difficult to collect the information needed to cite these sources properly. The author proposes a method by which bibliographic information embedded in electronic sources could be automatically extracted when needed and discusses existing standards that could be utilized to accomplish this and impediments to implementation.

Electronic sources of information, also known as *e-sources*, have become an important repository of knowledge. E-sources have the unique quality of being accessible to anyone who might want to cite or quote their content in a form that allows direct transfer from the source to the derivative through the copy-and-paste functionality of information technology (IT). While this transfer of content has been made easy, it is still a difficult and manual process to capture and properly utilize the bibliographic information describing e-sources. In many cases, notably Web pages, the bibliographic information is already embedded in the source, yet it is not easily accessible.

It has been more than ten years since issues regarding the need for standards for citing hypermedia works were first raised.[1] During that time an explosive growth and evolution of hypermedia mainly in the form of Web sites, but also in CD-ROM, DVD, and other formats has occurred. The Internet in particular is increasingly used by scholars at all levels in support of their research and writing.[2] IT has been firmly embraced by scholars in every field, at the least for manuscript preparation and in many cases for the entire research and publication cycle. Yet scholars haven't begun to take advantage of the potential mechanisms that IT can afford for accurately identifying sources and linking these sources to our citations of them.

Citation standards

Citation standards fall into two categories: formatting of citations in derivative works and inclusion of bibliographic identifying information in original works.

Standards for formatting citations for e-sources have been established by a number of authorities. These authorities include the Modern Language Association (MLA), the American Psychological Association, *The Chicago Manual of Style*, the National Library of Medicine, and others.[3] These standards for e-sources differ from one another just as their print counterparts do (see appendix A). However, the actual metadata (i.e. title, author, year published, etc.) are, of course, the same for any given item. Only the presentation of this information varies. Because computer-aided formatting of citations and references is handled now by a variety of bibliographic database managers this aspect of citation reference will not be discussed further.

Very little, however, has been done on the part of e-publishers to make proper citing of works straightforward. The few exceptions are mostly in the form of written instructions at the beginning or end of a work, or on a separate page, stating the preferred wording and format of a citation to that work. Some e-publishers provide a formatted (such as in MLA format) citation for copying which is helpful but also renders them less useful for those who need to use other formats. The American Library Association (ALA) provides a button labeled "cite this page" on the pages of their Web site.[4] It produces a page with the ALA-formatted citation ready for copying and subsequently pasting into a references list, again, less useful for those not adhering to the ALA's preferred style. Until recently PsychInfo provided a button at the beginning of each paragraph of full-text articles available through their Web site. This button displayed the formatted citation in a separate window ready for copying. This feature has, however, been removed from the Web site. The ACM Portal provides links on citation pages for displaying bibliographic metadata in a format suitable for importing into two popular bibliographic database management systems (EndNote and BibTex).[5] The resulting output must then be saved as a text file and imported into the database.

These are just a few of the solutions that have been implemented by a variety of organizations. The lack of standardization in and, in some cases, the ephemeral nature of providing bibliographic information to the reader will limit the use of these aids. The remainder of this paper will focus on a method that can be applied across a range of digital media types and operating systems and could form the basis for a much needed standard.

Enabling technologies

The digital nature of e-sources should make capturing the information necessary to properly cite them especially easy. Viewing e-source material requires the use of a digital information system of some sort, often a Web browser or other viewer running on a computer.[6] Most, if not all, of these programs now include the capability to

Peter Jörgensen (pjorgensen@ci.fsu.edu) is Assistant Professor in the College of Information at Florida State University, Tallahassee.

select, copy, and paste portions of material on screen into a system-wide data structure commonly known as a clipboard or pasteboard. Taking notes is thus facilitated and direct quotations are no longer subject to typographical errors introduced by the quoter.[7] The author previously suggested that this copying facility could be harnessed to automatically gather bibliographic information which could then be attached directly to a quote or linked to through a bibliographic database.[8] In order to accomplish this goal, the necessary bibliographic information must be embedded in the e-source in a standardized, or at least parsable, form. Because it has been shown that resource authors can provide good metadata in Dublin Core (DC) format, it is reasonable to propose that authoring systems should make it easier to include properly formatted meta information in e-sources.[9] In addition, the software that is used to access e-sources must have the capability to harvest the bibliographic information when required and make it available to the user. These requirements will be addressed in the following sections, which will use Web document technology as the frame of reference, keeping in mind that very similar techniques can be applied to other digital publication forms, such as DVD.

Proposed solutions

Embedded bibliographic information

Metadata tags are now quite common, although their use is primarily for embedding information regarding the authoring software and instructions to Web "robots" and caching servers.[10] The addition of bibliographic metadata is no more difficult and, in fact, is already being done by some authors and publishers using three largely compatible standards.[11] These standards, the DC, the Metadata Object Description Schema (MODS), and the Metadata Encoding and Transmission Standard (METS) are discussed briefly in the following sections.

DC

The DC metadata standard was initially developed by an ad hoc group of scholars invited to the OCLC/NCSA Metadata Workshop held in Dublin, Ohio, in March 1995. Since then the DC Metadata Initiative (DCMI) has been established to provide formal leadership in developing and maintaining the standard.[12] The standard contains eighteen basic elements (table 1) and has been incorporated into ISO Standard 15836-2003, NISO Standard Z39.85-2001 and CEN Workshop Agreement CWA 13874, the Open Archives Initiative, and implemented by more than two hundred data providers.[13] Tags of primary interest to the scholar wishing to make proper attributions are dc:title, dc:creator, dc:date, dc:source, dc:publisher, and dc:identifier. The DC is missing elements that would be useful for proper citing, such as editor (which is included but not differentiated in the dc:creator element) and conference-related information. These items could be included in one of the catchall elements, such as identifier (or better yet, its refinement bibliographicCitation).

Table 1. DC elements

audience	format	rights
contributor	identifier	rightsHolder
coverage	language	source
creator	provenance	subject
date	publisher	title
description	relation	type

MODS

The U.S. Library of Congress has developed MODS as an alternative to DC. In their words, MODS "is intended to be able to carry selected data from existing MARC 21 records as well as to enable the creation of original resource description records."[14] This schema utilizes XML and is richer than DC but not as rich as MARC.[15] Its richness has made it the format of choice for intermediate records in several conversion processes.[16]

METS

METS was developed as part of the Digital Library Foundation initiative as an open standard for digital object description.[17] METS is the most finely grained metadata standard of the three considered here. It is intended to be used strictly for creating documents which describe digital resources, although it does have a provision (the <FContent> element) for including the actual content of the object the METS document describes, rather than the usual file specifications only.[18]

Of these three standards, DC is both designed and most widely implemented as an embedded source of metadata in e-sources. Therefore the remainder of this paper will focus on DC meta tags, recognizing that the principles being presented can just as well be applied to other embedded metadata standards.

Of course, the inclusion of metadata elements in e-sources is a requirement for the utilization of this information by citing authors. These tags could be added by authors themselves, as supported by Greenberg et al., or it might become the added value that publishers provide as the scholarly press becomes increasingly digital and Internet-based.[19] At any rate, there are now several (largely experimental) solutions to the problem of automating the creation and addition of well-formed meta tags. One model generates meta tags based on an analysis of the content of any Web page whose URL is supplied by the user.[20] After retrieving and analyzing the Web page, a form is typically presented which shows the values that the program has assigned to the DC elements. This form is similar to those presented initially by other sites which then use the input to the form to generate meta tags.[21] It seems, however, to this author that the logical place for this functionality may lie within the authoring environment itself. This paper will not treat this issue further except to stress that facilitating the creation or generation of valid bibliographic meta tags is as important to their widespread adoption as creating the tools to extract them.

The final step in the process of automating the capture of bibliographic metadata is the harvesting of this information by someone wishing to properly cite the source. The next section will present a model for this process and a proof of concept in the form of an AppleScript program.

Automatic citation production

Enabling display software to gather the bibliographic information when needed could be implemented in the following nonexclusive ways.

- Display software could automatically gather the information whenever material is copied and append it to or embed it in the copied material on the clipboard.[22]
- A separate "copy citation" command could be included in Web browsers that would copy just the bibliographic information to the clipboard. The information would then be available to the system for insertion into the user's reference management database (with appropriate information embedded in the record linking it to the copied material) or inserted as a formatted citation into a document.
- The bibliographic information could be automatically added to the user's bibliographic database when a document is viewed, possibly adding a date and time stamp to help the user manage this information.

Certainly this list does not represent all of the possibilities. Any one or a combination of these methods could easily be incorporated directly into Web documents by the use of such embedded software as Java or JavaScript. Alternatively, these methods could be provided by the browser program or as a system-wide service. A service which extracts metadata is available from DC-dot, but rather than create a formatted reference or an importable representation, it displays the results in an editable table.[23] This is certainly useful for metadata editors, and serves as a proof of concept for but does not implement the functionality that this paper proposes.

A model for bibliographic metadata extraction

Extraction and utilization of bibliographic metadata requires the following four separate steps.

1. The e-source must be parsed and the metadata extracted
2. The metadata must be normalized
3. The metadata must be converted into a form that can either be imported into a bibliographic database, pasted as a formatted citation into a word processor document, or otherwise made use of
4. The formatted metadata must finally be inserted into the desired database, document, or other destination

Parsing the e-source document

Parsing the bibliographic metadata in a document containing DC meta tags is a trivial matter due to the nature of HTML. Each meta tag must begin with the characters "<meta" making it easy for a program to locate the beginning of each tag. All meta tags are in the <HEAD> element of the HTML file, so the parser can stop processing the file when it reaches the </HEAD> or <BODY> tags. Checking for both of these provides some protection against poorly formed HTML files that might be missing one or the other.

Normalizing the metadata

The second step, normalizing the metadata, could include correcting the sequence of authors' names, checking authority files, enforcing capitalization standards, or even applying controlled vocabulary term substitution. This aspect of the process is not considered further here.

Formatting the metadata

Before the metadata can be inserted into a database or a document (directly as a reference), it must be formatted according to the requirements of the application and the user. In the case presented here, the metadata is saved into a tagged text file suitable for importation into an Endnote bibliographic database. It could also be stored on the system's clipboard data structure as formatted text.

Inserting the metadata

Finally, the formatted metadata could be inserted into a document by pasting. It is also possible, given sufficient information about the programming interface of the application, to insert the data into a newly created record in a database. An increasing number of programs provide methods whereby other processes can interact with them in this way.

Proof-of-concept implementation

An application has been developed in AppleScript which implements the process as previously outlined. The application interacts directly with the Safari Web browser to get the HTML source of the current page, although it could easily be modified to work with other browsers. It parses the <HEAD> element of the page, extracting any DC meta tags that it encounters. The application next writes the corresponding data to a file formatted for importing by EndNote. Finally, it sends a system message to EndNote instructing it to import the file. As of version 8, Endnote does not provide AppleEvents support. Therefore, an AppleScript program cannot "tell" EndNote to insert data directly into its database.[24] AppleScript is, however, capable of mimicking such user actions as menu choices and keystrokes. This facility has been used with limited success to paste text directly into EndNote fields. This is less than satisfactory because of the lack of inter-process communication to coordinate the timing of the simulated user actions. The proof-of-concept implementation, therefore, creates a text file that EndNote then imports. Because the importation of the file is a separate step, it could be postponed until a number of e-source documents had been processed and their data appended to a single import file.

Conclusion

Bibliographic metadata, especially in the form of DC meta tags, is increasingly embedded in e-documents. A process by which this metadata can be extracted by readers for their use has been described. A proof-of-concept implementation has been produced and is available under the GNU license from the author. It is hoped that the existence of this tool, and similar ones which will surely follow, will accelerate the inclusion of bibliographic metadata in e-documents.

References and notes

1. Corinne Jörgensen and Peter Jörgensen. "Citations in Hypermedia: Maintaining Critical Links." *College & Research Libraries* 52, no. 6 (1991): 528–36. Although the term "hypermedia" has been dropped by and large and replaced with "World Wide Web" and "Internet," it and "e-sources" is used in this paper to signify not only the ubiquitous Web but other current forms of digital information delivery, such as CD and DVD and those which will undoubtedly be developed in the future.

2. Martin Rees, "Not Worth the Paper," *New Scientist* 176, no. 27 (Nov. 23, 2002): 27.

3. Joseph Gibaldi, *MLA Handbook for Writers of Research Papers*, 6th ed. (New York: Modern Language Association, 2003); American Psychological Association, "General Form for Electronic References," American Psychological Association Web site, 2001. Accessed Feb. 18, 2004, www.apastyle.org/elecgeneral.html; The University of Chicago Press, *The Chicago Manual of Style*, 15th ed. (Chicago and London: Univ. of Chicago Pr., 2003); Karen Patrias, "National Library of Medicine Recommended Formats for Bibliographic Citation Supplement: Internet Formats," National Library of Medicine Web site, 2001. Accessed July 22, 2004, www.nlm.nih.gov/pubs/formats/internet.pdf; Andrew Harnack and Eugene Kleppinger, Citation Styles, Online! A Reference Guide to Using Internet Sources Web site, Bedford/St. Martin's, 2003. Accessed Feb. 18, 2004, www.bedfordstmartins.com/online/citex.html.

4. American Library Association, "Promotional Products," American Library Association Web site, 2003. Accessed May 21, 2004, www.ala.org/Template.cfm?Section=promotional.

5. Association for Computing Machinery, "Rapid Serial Visual Presentation: A Space-Time Trade-off in Information Presentation," ACM Digital Library, ACM Portal, 2005. Accessed Feb. 4, 2005, http://portal.acm.org/citation.cfm?id=345309&jmp=cit&coll=GUIDE&dl=ACM&CFID=37999606&CFTOKEN=2560252#.

6. It may not seem so, but even a DVD player is more closely related to a computer than to a VCR. Computer here means a digital electronic device for presenting information.

7. Of course some may be tempted, by virtue of the ease with which it can be done, to pass off others' work that they simply copied as their own. The solutions that will be offered here will not change this unfortunate fact, but may, in fact, reduce it by making proper attribution easier.

8. Jörgensen and Jörgensen, "Citations in Hypermedia: Maintaining Critical Links."

9. Jane Greenberg, Maria Cristina Pattuelli, Bijan Parsia, and W. Davenport Robertson, "Author-Generated Dublin Core Metadata for Web Resources: A Baseline Study in an Organization," *Journal of Digital Information* 2, no. 2 (2001). Accessed Feb. 2, 2005, http://jodi.tamu.edu/Articles/v02/i02/Greenberg/.

10. A quick tally of 632 scholarly articles retrieved from .edu Web sites using Google Scholar (http://scholar.google.com/) reveals that 83 percent contain at least one meta tag. However, very few of these meta tags were bibliographic in nature, and the vast majority (431) were robot directives.

11. Data on the number of Web pages that contain standardized bibliographic (DC) meta tags is being collected and will be published in a forthcoming paper by the author.

12. DCMI, "History of the Dublin Core Metadata Initiative," Dublin Core Metadata Initiative Web site, 2004. Accessed Feb. 12, 2004, http://dublincore.org/about/history/; DCMI, "The Open Metadata Registry," Dublin Core Metadata Initiative Web site, 2004. Accessed Mar. 2, 2005, http://dublincore.org/dcregistry/navigateServlet.

13. Carl Lagoze, "Open Archives Initiative FAQ," Open Archives Web site, 2002. Accessed Dec. 8, 2004, www.openarchives.org/documents/FAQ.html; "Dublin Core Metadata Element Set, Version 1.1: Reference Description. 2," Dublin Core Metadata Initiative Web site, 2003. Accessed Aug. 20, 2004, http://dublincore.org/documents/dces/; Carl Lagoze, Herbert Van de Sompel, Michael Nelson, and Simeon Warner, "Open Archives Initiative Protocol for Metadata Harvesting v.2.0," Open Archives Initiative Web site, 2003. Accessed Aug. 20, 2004, www.openarchives.org/OAI/openarchivesprotocol.htm#MetadataNamespaces.

14. Library of Congress, "MODS," Library of Congress Web site, 2004. Accessed Feb. 8, 2005, www.loc.gov/standards/mods/.

15. Roxanne Missingham, "Reengineering a National Resource Discovery Service: MODS Down Under," *D-Lib Magazine* 10, no. 9

(2004). Accessed Dec. 7, 2004, www.dlib.org/dlib/september04/missingham/09missingham.html

16. Missingham, "Reengineering a National Resource Discovery Service"; Chris Putnam, "Bibutils," The Scripps Research Institute Web site, 2004. Accessed Feb. 2, 2005, www.scripps.edu/~cdputnam/software/bibutils/; Marie-Louise Ayres, "Music Australia: Building on National Infrastructure," paper presented at the Twelfth Biennial VALA Conference and Exhibition, Melbourne, Feb. 3–6, 2004.

17. Library of Congress Help Desk, METS: An Overview and Tutorial, Library of Congress Web site, 2004. Accessed Feb. 2, 2005, www.loc.gov/standards/mets/METSOverview.v2.html.

18. D. T. Hawkins, "Metadata Practices on the Cutting Edge," *Information Today* (United States) 21, no. 7 (2004): 28.

19. Greenberg, Pattuelli, Parsia, and Robertson, "Author-Generated Dublin Core Metadata for Web Resources."

20. Two examples of systems that generate meta tags based on an analysis of Web pages whose URLs are provided by the user are www.ukoln.ac.uk/metadata/dcdot/ and www.kb.nl/cgi-bin/donor-mg.pl.

21. The following URLs are examples of Web-based systems that provide a form or template into which a user can enter values from which meta tags will be generated: www.lub.lu.se/cgi-bin/nmdc.pl; www.mathematik.uni-osnabrueck.de/cgi-bin/MMM3.1.cgi; http://physnet.uni-oldenburg.de/services/mmm/; http://metabrowser.spirit.net.au/prodClient.htm; and http://www.mkdoc.org/.

22. Most users probably do not realize that information copied onto the clipboard is often represented simultaneously in a variety of ways in the computer's memory. This facilitates data interchange between programs of different types. For instance, formatted text copied from a word processing document is stored on the clipboard in the native format of the word processing application and as plain, unformatted text (and possibly in other forms). This allows the user to paste this text into another word processor document, retaining the formatting, or into an e-mail message or database field, which will not generally display the formatted text. This transparent (to the end user) manipulation of the data on the system clipboard could be extended to include metadata about the material which has been copied.

23. Andy Powell, "UKOLN: DC-Dot Dublin Core Metadata Editor." UKOLN Web site, University of Bath, 2001. Accessed Aug. 2, 2005, www.ukoln.ac.uk/metadata/dcdot/**.**

24. "Tell" is an AppleScript keyword, as in *tell application EndNote to set field "author" to "Peter Jörgensen."*

Appendix A. Electronic source citations in different formats

These formatted citation examples were produced by EndNote 8 from a record entry using the styles shown.

APA

Cailliau, R. (1995, Feb 16, 2001). *A little history of the world wide web.* Retrieved August 28, 2001, from http://www.w3.org/History.html.

Bioscience

Cailliau, R. 1995. A little history of the world wide web. W3C.

CBE Style

1. Cailliau R. 1995 August 28. A little history of the world wide web. 1.24. W3C <http://www.w3.org/History.html>. Accessed 2001 August 28.

The Chicago Manual of Style

Cailliau, Robert. 1995. A Little History of the World Wide Web. In, 1.24, W3C, http://www.w3.org/History.html. (accessed August 28, 2001).

IEEE

[1] R. Cailliau, "A little history of the world wide web," vol. 2001, 1.24 ed: W3C, 1995.

Library Quarterly

1. W3C. "A little history of the world wide web." http://www.w3.org/History.html August 28.

MLA

Cailliau, Robert. "A Little History of the World Wide Web." 1995. Web Page. 1.24 (Feb 16, 2001): W3C. August 28 2001. <http://www.w3.org/History.html >.

NLM

Cailliau R. A little history of the world wide web. [Web Page] 1995 [cited 2001 August 28].

Turabian (bibliography)

Cailliau, Robert. *A Little History of the World Wide Web* [Web Page]. W3C, Feb 16, 2001 1995, accessed August 28 2001; Available from http://www.w3.org/History.html.