13. Citing, Identifying and Describing Networked Information Resources

As networked information resources are integrated into the body of information, both scholarly and popular, it will be necessary to extent traditional print-based methods of citation to accommodate reference to these new network-based resources. Here, the objective is to continue the functions served by citation in print literature: to permit a work to make reference to the contents of another work with sufficient specificity to permit the reader to obtain a copy of the cited work and locate the part of that work being referenced; to give the reader of the citation enough information to make some judgments about whether he or she is already familiar with the cited work, and to provide some information about the cited work such as date of publication, title and author which might help the reader to determine if it is worth obtaining a copy of this cited work. It is important to note that traditional print citations today serve both of these purposes; for example, citations consisting simply of document numbers assigned by some document registry are not typically used because while they would allow the reader of the citation to obtain a copy of the cited document, they don't tell the reader anything about the cited work to help in making a decision whether to obtain a copy of it. [83]

At the same time that the need to cite electronic information resources is being recognized, several other closely related requirements are emerging. These include the desire of libraries, bibliographers and other organizations and individuals that organize information to catalog the increasingly valuable and common electronic information resources; essentially, to extend the existing mechanisms of bibliographic description and control to facilitate access to these resources. The needs here are closely related to those of citations, but more extensive in that there is usually a requirement to include more information about how to obtain access to a given resource once identified, and also requirements to include subject access or other classification information.

It is interesting to note that both for citation and cataloging purposes a number of people have expressed a desire to have the citation or cataloging record include some information (such as document digests or signatures, as discussed earlier in this paper) that would permit the user to check that he or she had retrieved the same version of the electronic object that the creator of the citation or descriptive record had originally described (at least as an option: when one is talking about electronic documents this makes sense, but when one is making reference to a database that is continuously updated at the level of an information resource, rather than referring to the contents of a specific record in that database at a specific point in time, such version information does not make sense). Logically, this requirement makes little sense. Reasoning by analogy with the print world, if a citation specifies the second edition of a specific work, it is possible that the publisher might change the contents of the work and reprint it without updating the bibliographic information or date of publication, in effect creating two editions of the work that have different content but are not identified as distinct

[83] It is worth noting that in some areas of scholarship historically citation systems have been used *that* only address the identification of a work, or passages from it, without referring to specific editions. Examples include biblical scholarship and some types of literary criticism. Usually in these situations there is an implied canonical text, so it is not necessary to specify the specific edition of the intellectual content.

editions.[8] However, this does not happen often (at least for materials that are extensively cited and where very precise citation is important) in the print world and people don't generally worry about it much.[85] The emergence of this requirement for version verification in the electronic information world simply underscores the general perception that electronic information is more volatile and more easily changed, and that the contents of electronic objects cannot be trusted to retain their integrity over time without introducing special verification processes into the system of access and management of these resources. It is also worth recognizing that on a technical level this problem of version verification is largely unsolved as yet; while the digital signature and digest algorithms discussed earlier can readily ensure that a document is bit-for-bit the same as the one cited, citation typically is more concerned with intellectual content. As we move to an environment where software and protocols for retrieval of electronic documents (in the broad sense of multimedia objects) becomes more adaptive and mature, transfer of documents from one host to another may commonly invoke format translations and reformatting of various types automatically,[86] while such translations would presevee the intellectual content of the document (perhaps at varying levels of precision, depending on whether the transformations were lossless and invertable), the transformation would of course change the actual bits comprising the document and thus cause it to fail a version comparison test based on such bit-level algorithms.

[84] T. be clear: current library cataloging rules direct catalogers to explicitly differentiate works that are different even if the publisher has not done so.

[85] Indeed, if anything, the problems today with citation to printed material, as discussed earlier in this report include the difficulty that the creator of the citation often does not realize that the publisher is producing multiple editions targeted for different geographical regions or for different subsets of the readership (for example, trade magazines that include special advertising sections targeted at readers who work in specific industries) and hence doesn't create a sufficiently specific citation. From the publisher's point of view, there is often great economic incentive to keep repackaging and reissuing content with minimal changes as new editions or even new works; the notion of going to the trouble of producing an unadvertised and unlabeled new edition and quietly introducing it into the marketplace is relatively rare, at least for print; this practice does occur sometimes with electronic publications such as software, where minor corrections or improvements are sometimes shipped automatically without much publicity, although even there the publisher usually changes the version number. There are a few examples of audio materials where different versions have been shipped with the same cover and same publisher catalog number.

Also, in the print world, in cases where a citation is to a work where there is some question about the precise final form of the work, conventions have been developed such as indicating "unpublished draft" or "in press" to alert the user of the citation that there may be some problems. Of course, such citations are the exception rather than the rule.

[86] T. provide only a few examples of such translation, a document might be changed from one character set to another (ASCII to EBCDIC or UNICODE); fonts might be substituted, since fonts are copyrighted and the workstation receiving a document might not have the fonts used by the author, so it might be necessary to substitute similar fonts that are either in the public domain or that are licensed to the receiving workstation; an image or digital sound clip might be converted from one format to another, and the resolution or sampling rate might be altered; or more extensive format changes might occur, such as the rendering of a postscript document into a bitmapped image prior to transfer. The extent to which these transformations preserve the intellectual content of the work are highly dependent on the nature of the transformation and also the use to which the document will be put when it is transferred; for example, if it is only to be viewed, then a transformation from SGML markup to a bitmapped image makes no difference to the content in some sense, but if that same document is to be edited or analyzed by a postprocessing program, then there is a very large loss of information in the conversion from SGML to bitmapped representation.

A third set of requirements are more technical in nature but address some of the needs for both cataloging and citing networked information resources; while they solve neither problem they provide tools for developing solutions. In addition, a solution to these technical requirements is needed to enable the widespread development and deployment of a number of important networked information applications. These technical requirements are based on the need for standards so that one object on the network can contain a computer-interpretable “pointer” or link to another object on the network. This is needed for network-based hypertext systems such as the World Wide Web. It is needed so that document browsers can automatically follow references in a document when these references are to other network resources. It is needed so that bibliographic or abstracting and indexing records that describe electronic information resources can include information about where to find and how to access these resources. This last case is particularly important for a number of projects that are now underway where large bitmapped image databases of material are being created; because of the size of these databases it is desirable to store them at only at most a few sites on the network and to retrieve page images from them on demand; yet multiple databases of descriptive records, developed by multiple organizations, need to include links to these image databases. Further, in some cases, the descriptive records are being distributed under different license terms than the actual content; for example, some major publishers are exploring scenarios where they give away brief records analogous to tables of contents in printed journals, and then charge transactionally for retrieval of the actual articles.

The idea is that these pointers to networked information resources should be representable as an ASCII text string, permitting their inclusion in both electronic documents and in printed documents, as well as their easy transfer from machine to machine and from one application to another within a machine (for example, via cut-and-paste facilities now available in most graphical user interfaces, with the idea being that a user might view a document or an electronic mail message or a screen display from an online bibliographic database in one window, find a reference to a document that he or she desires to fetch, and simply highlight and drag the citation to another application, which would then fetch the object or open a connection to the service, using whatever access protocol is required).

These technical requirements are being addressed by a working group of the Internet Engineering Task Force. While the technical details of the IETF standards proposals are beyond the scope of this paper (and, indeed, some specifics of the standards are still under active debate within the IETF Working Group **as** of this writing) there seems to be some substantial consensus on the overall approach to be taken. It should also be recognized that there are some very substantial research problems in dealing with these technical requirements in full generality, and thus the IETF work should be regarded as a beginning and a framework that will undoubtedly undergo a great deal of extension and refinement in the coming years based on operational experience with the first generation standards, improved understanding of the theoretical issues and abstract modeling questions underlying the standard, and the continued development of protocols and applications for accessing networked information resources of various types.

Roughly, the IETF proposals call for the definition of a syntax for what they call a locator, which is an ASCII string that identifies an object or service that is hosted on a specific machine (typically specified by its domain name) on the network, the service (such as FTP, electronic mail, Z39.50 database query) that is used to obtain the object, and the parameters that are to be passed to that service to identify the specific object to be obtained (for example, in the case of FTP the fully-qualified filename). There are several problems with locators as a basis for citation, however. Machines on the network come and go over time, and files are migrated from one machine to another. Some commonly used files are duplicated on multiple machines; from the point of view of citation, one wants to refer to content and not instances of content, and thus should no more list machines containing copies of a file than one would list libraries holding copies of a book in a citation. An object may be accessible through multiple access methods (for example, FTP and database retrieval); indeed, the method of access may change over time and in response to improved technology, but the content being accessed remains unchanged. Further, one cannot tell whether two different locators actually refer to the identical content.

Thus, the IETF working group has proposed the definition of identifiers, which are strings assigned by identifying *authorities* to refer to content. An identifier for an object, then, is just a two-component object consisting of a specifier for the identifying authority (these would be assigned centrally, as a service to the Internet community, much like top-level domains or network numbers) and the identifier that the authority provided. These identifying authorities (and other organizations) may offer services that provide a mapping from an identifier to a series of locators, which could then be used to actually obtain access to a copy of the object. Some mapping services, particularly those operated by specific identifying authorities, might only resolve identifiers assigned by the operating identifying authority; others, perhaps operated by organizations such as libraries, might attempt to resolve identifiers issued by multiple identifying authorities into sets of locators. Locators would be viewed as relatively transient; at any time one could obtain a fresh set of locators corresponding to an identifier. Identifiers would be used in citations and other applications. It is important to note that the IETF model explicitly recognizes that deciding whether two instances of an object are "identical" is a subjective issue which is highly dependent on the objectives of a given identifying authority, and that there will be a multiplicity of such identifying authorities, which might include publishers, service organizations, libraries, or industry-wide standards implementations (such as the International Standard Book Number in the print world). The same content might be assigned identifiers by multiple identifying authorities; in some cases two objects might be viewed as identical by one identifying authority (meaning that the authority would return locators for both objects in response to its identifier) and yet viewed as distinct by another identifying authority. [87]

Rules for citations are typically set by editors of journals, or sometimes by professional societies (for groups of journals) or by style manuals (such as the *Chicago Manual of Sty/e).* While a number of journals (both print journals and electronic journals) have

[87] AS a specific case in point,one identifying authority might view a bitmapped image and a Postscript file of the same document as identical; another might view these as different objects. The issue of format variations and the extent to which these variations, as well as multiple versions of documents, should be recognized by and integrated into the locator and identifier scheme is still an active area of discussion.

already defined practices for citing electronic information resources it seems very likely that these practices will be altered over time to include identifiers which followed the IETF standard in order to facilitate both the identification and the retrieval of electronic objects. As these identifiers come into wide use, some of the other material that is currently specified in citations to electronic resources (such as the name of a machine holding a file available for anonymous FTP and the file name) might well be dropped. Some of the traditional citation data elements that help the reader to identify and evaluate the intellectual content of the cited work, such as author, title, and publication date, will almost certainly be retained. A few data elements used in some citation formats, such as the number of pages in a work, are problematic in an electronic environment; while it is clearly useful for the reader to have some sense of the size of a cited work, it is unclear how to most usefully measure this in an electronic environment that may contain multimedia works. The transformation of citation rules is likely to be a gradual process; it is important to note that, at least in practice, citation formats are really not national and international standards, but rather working rules that serve various communities, and there are a fairly large number of citation formats in common use.

Cataloging practices for networked information resources is an area that is currently under very active discussion. Several groups within the American Library Association (in particular, MARBI and CC: DA) are studying this issue and working on guidelines in association with groups that include the Library of Congress, OCLC, the Coalition for Networked Information, and the IETF. Some of the issues involved here are very complex, and not yet well understood; indeed, some of the questions involve very basic considerations about the purposes and objectives of cataloging. Taxonomies for classifying networked information resources are also needed, and still poorly understood. The current drafts [Library of Congress, 1991b; Library of Congress, 1993] from the American Library Association’s MARBI committee again recognizes the use of the IETF locator and identifier structure as an appropriate means of encoding some needed information, and foresees a conversion to these standards as they are established, while also supplying some provisional field definitions that can be used by catalogers who wish to experiment with cataloging network resources in the interim.

It is also important to recognize that cataloging is only a part of the broader question of how to provide information to help users to identify and select networked information resources. Cataloging is concerned primarily with description and organization of materials (for example, through assignment of subject headings within some classification structure and vocabulary, or through the development of name authority files that bring together works published by the same author under different names, or different variations of a single name); equally important information which would allow someone to obtain evaluative information about a resource or to compare one resource to another is outside the scope of cataloging. Such information is provided by book reviews, consumer information services, ratings services, critical bibliographies, awards given by various groups, sales figures and other tools. All of these services-and new ones, such a certification that software works properly in a given environment or is free from viruses, for example—will need to be evolved into the networked information environment but with some new and challenging additions. One key objective will be to preserve, and if possible to expand the diversity of evaluative sources that information seekers can consult if they wish; just as one promise of the networked information