COMMISSION ON PRESERVATION
& ACCESS

Intellectual Preservation: Electronic Preservation of the Third Kind

By Peter S. Graham, Associate University Librarian
for Technical and Networked Information Services
Rutgers University
March 1994

The advent of electronic information introduces new preservation requirements. In contrast with print materials, where to preserve the artifact is to preserve the information contained in it, electronic information is easily transferred from one medium to another with no loss. [1]

Medium preservation and technology preservation

Medium preservation has been addressed by some librarians and computing experts in discussions of environmental and handling concerns for tapes, magnetic disks, optical disks, and the like.[2] The preservation of the medium on which the bits and bytes of electronic information are recorded is an important concern. But such solutions will inevitably be short-term, and will not in themselves be the means of preserving information over long periods of time. Michael Lesk, in a report for the Commission, has urged that the greatest attention should instead be directed to the obsolescence of technologies rather than simply of the media.[3]

Lesk describes the rapid changes in the means of recording, in the storage formats and in the software that allows electronic information to be of use. Urging what might be called *technology preservation* he asserts that for electronic information, "preservation means copying, not physical preservation." That is, the preservation of electronic information into the indefinite future requires its being "refreshed" from old to new technologies as they become available and as the old technologies cease being supported by vendors and the user community.

Intellectual preservation

There remains a third preservation requirement, *intellectual preservation*, which addresses the integrity and authenticity of the information as originally recorded. Preservation of the media and of the software technologies will serve only part of the need if the information content has been corrupted from its original form, whether by accident or design. The need for intellectual preservation arises because the great asset of digital information is also its great liability: the ease with which an identical copy can be quickly and flawlessly made is paralleled by the ease with which a change may undetectably be made.

Professor Barry Neavill of the University of Alabama library school has written of the "malleability" of electronic information, that is, the ease with which it can be transformed and manipulated.[4] Patricia Wilson Berger, 1989 President of the American Library Association, noted:

> [I]n a democracy, information access requires an information base secure from intrusion, distortion, and destruction; one protected from both physical and technological deterioration.[5]

Clifford Lynch of the University of California President's Office has noted:

> It is very easy to replace an electronic dataset with an updated copy, and...the replacement can have wide-reaching effects. The processes of authorship...produce different versions which in an electronic environment can easily go into broad circulation; if each draft is not carefully labeled and dated it is difficult to tell which draft one is looking at, or whether one has the "final" version of a work.[6]

Professor D.F. McKenzie, in his Centenary Lecture of The Bibliographical Society (London), wrote in urging a new direction for the Society, that

> It's the *durability* of those textual forms [books] that ultimately secures the continuing future of our past; it's the *evanescence* of the new ones that poses the most critical problem for bibliography and any further history dependent upon its scholarship.... As the late Northrop Frye said, "Society, like the individual, becomes senile in proportion as it loses its continuous memory," and [electronic] texts are now part of that memory, significant products of our civilization.... [There is] a new urgency with the arrival of computer-generated texts. The demands made...by the evolution of texts in such forms, the speed with which versions are displaced one by another, and the question of their authority, are no less compelling than those we accept for printed books.[7]

The problem

The problem may be put in the form of several questions that confront the user of any electronic document (whether it is text, hypertext, audio, graphic, numeric or multimedia information):

- How can I be sure that what I am viewing is what I want to see?
- How do I know that the document I have found is the same one that you used and made reference to in your footnote?
- How can I be sure that the document I now use has not been changed since the last time I used it?
- To put it most generally: How can a reader be sure that the document being used is the one intended?

We properly take for granted the fixity of text in the print world. The printed journal article I examine because of your footnote is beyond question the same text that you read.[8] Therefore we have confidence that our discussion is based upon a common foundation. With electronic texts we no longer have that confidence.

Taxonomy of changes

There are three possibilities for change in electronic texts that confront us with the need for intellectual preservation techniques:

- accidental change
- intended change that is well-meant
- intended change that is not well-meant, that is, fraud

Note that backup is not the issue or the solution. In question is how we know what we have (or don't have).

Accidental change

A document can sometimes be damaged accidentally, perhaps by data loss during transfer or through inadvertent mistakes in manipulation; for example, data may be corrupted in being sent over a network or between disks and memory on a computer. This no longer happens often, but it is possible. More frequent is the loss of sections of a document, or a whole version of a document, due to accidents in updating.

Intended change -- well-meant

There are at least two possibilities for well-intended change:

New versions and drafts are familiar to us from dealing with authorial texts, for example, or from working with successive book editions, legislative bills, or revisions of working papers. It is desirable to keep track bibliographically of the distinction between one version and another. We are accustomed to visual cues to tell us when a version is different; in addition to explicit numbering we observe the page format, the typography, the producer's name, the binding, the paper itself. These cues are not available or dependable for distinguishing electronic versions.

Structural updates, changes that are inherent in the document, also cause changes in information content. A dynamic data base by its nature is frequently updated: *Books in Print,* for example, or architectural drawings, or elements of the human genome project, or a university directory. How do we identify a given snapshot and authenticate it as representing a certain time?

Intended change -- fraud

The third kind of change that can occur is intentional change for fraudulent reasons. The change might be of one's own work, to cover one's tracks or change evidence for a variety of reasons, or it might be damage to the work of another.

In an electronic future the opportunities for a Stalinist revision of history will be multiplied. An unscrupulous researcher could change experimental data without a trace. A financial dealer might wish to cover tracks to hide improper business, or a political figure might wish to hide or modify inconvenient earlier views. Imagine if you will that the only evidence of Reagan's Iran-Contra scandal was in an electronic format, or that the only record of Bill Clinton's draft correspondence was in e-mail. Consider the political benefit that might derive if each of the parties could modify their own past correspondence without detection. Then consider the case if each of them could modify the other's correspondence without detection. Society, as well as the parties involved, needs a defense against both such cases.

A potential solution

The need is to fix, or authenticate, a document so that a user can be sure of the unaltered text when it is needed.[9] Such a technique must be easy to use so that it does not impede creation or access. The technique must provide generality, flexibility and openness where possible, as well as document security where desired. It must be available at low cost and - most of all - be functional over long periods of time on the human scale. A solution will have to be based on simple software rather than on hardware, which rapidly becomes obsolete. This would seem to be a problem, for software, like documents themselves, can easily be tampered with and modified.

One example of how the problem can be solved has been developed by a small group of researchers. They have named their proposal *digital time-stamping* (DTS).[10] It calls upon the cryptographic technique of one-way hashing and uses the concept of the "widely-witnessed event." DTS is a means of authenticating not only a particular document, but its existence at a specific time. The technique is analogous to rubber-stamping incoming papers with the date and time they are received. In electronic form, its use is proposed to be by a document's creator (or other responsible intermediate party) to set up the necessary conditions for later authentication by an eventual user.[11]

The researchers were initially prompted to develop DTS by charges of intellectual fraud made against a biologist. They became interested in how to demonstrate that there had been no tampering with electronic evidence. In addition, they were aware that the technique could be useful as a means for determining priority of thought (e.g., in patents). The technique they developed makes use of cryptographic theory but does not require the encryption of documents.

Hashing

Any document may be viewed by a computer as a collection of numbers. A hash function is an algorithm which converts any collection of numbers into a single, distinct number (perhaps of a score or a hundred digits) which has no meaning in itself but which will uniquely represent the set of numbers from which it was derived. A one-way cryptographic hash of a document may be created using mathematically complex, but computationally speedy, techniques. The process ensures the uniqueness of the hash and also its non-reproducibility; that is, it is not humanly or computationally feasible to create another document which would result in the same hash. Therefore it is not possible to change the given document and still to preserve its original hash. The hashing technique is called "one-way" because the original document cannot be recreated if one has only the document's hash.

Note that by using this technique the document itself may be kept private if its creator wishes. However it need not be and in many cases would not be. For librarianship and scholarship generally, the public accessibility of documents without human intervention is a necessity and the one-way hash allows both a document and its hash to be public without fear of change. Note also that the algorithm (software) for creating the hash may also be public; its mechanism need not be private, for knowledge of it will not affect the uniqueness or the one-way nature of the created hash.

The widely witnessed event

For many purposes in librarianship the document and its secure hash will be all that is necessary to assure authentication; one can imagine bibliographic citation formats, for example, which would include a form of the hash as a means of identifying a specific version of a particular work and allowing its authentication.

But one can also imagine situations some years, decades, or centuries from now in which it will be desirable to be assured as to when the document first existed. In patent and contract law, which DTS will also serve, this is a daily necessity. In scientific research the need is clear, as it is if one considers stylistic analysis of an author's growth using electronic manuscripts as evidence.

The "widely witnessed event" is a concept that draws on the difficulty of tampering with a fact that is known to many outside the circle of interested parties. State lotteries prevent both collusion and the appearance of collusion by publicizing drawing of the winning numbers, often on television. Everyone sees the numbers drawn as they are drawn so that it is not possible for officials of the lottery to arrange the winner in advance. Similarly, sunshine laws are intended to make it impossible for members of legal bodies to collude in agreements at the meetings which the public and press attend; open contract bidding serves a similar purpose.

DTS draws on the principle of the widely witnessed event by openly intertwining the hash of a given document with the hashes of other documents submitted unpredictably by unknown other parties. The combined hashes for each document (known as "certificates") depend upon a visible chain of actions of other similar parties such that tampering cannot occur without being immediately evident to an observer.

Digital time-stamping in operation

In practice digital time-stamping requires the existence of time-stamping server software. Client software on a networked computer is also required: to create the hash of a document, to communicate with the server and (at a later time) to perform authentication.[12] None of the software need be computationally complex, large or time-consuming.

The user at a client workstation, perhaps a PC, creates the hash of a document (this can be very quickly done at the click of a button) and then sends the hash over the network to a time-stamping server, which combines the hash with a hash previously received (see Figure 1). The resulting number is called the "certificate" for the present hash, and is sent back to the user's workstation. This certificate becomes part of the authentication means for the original document whether used in the next half hour or the next half century. Note that the certificate is inextricably intertwined with those previously created for hashes received in unpredictable order from unknown (and unpredictable and uninfluenceable) users. The time-stamping server might easily be constructed to serve a region as large as the United States.

The time-stamping server creates a *root* certificate which is widely published at regular intervals. As a demonstration, the technique's authors for several years have published a root certificate once a week in the personals column of *The New York Times*. Such a widely witnessed event, available for centuries on microfilm or other means, is a tamper- proof tool of authentication. In real-world practice, the intervals would be much shorter, say perhaps one minute.

In theory, the user wishing to authenticate a document would have to recompute the chain of certificates for every document hash received from it on to the widely witnessed, published, root. This is not practical, and the researchers have solved this problem in three ways: by "publishing" the root more often, by creating a tree structure of certificates that logarithmically (and drastically) reduces the number of computations required for any given document, and by supplying all the other necessary certificates as part of the document certificate (see Figure 2).

It makes sense that all documents be time-stamped, even the seemingly most trivial. It is often not evident until well after the fact that the authenticity of a document and its timing are important; e.g., a telephone log of someone who later becomes a Supreme Court Justice, or laboratory notes of a researcher who later realizes that a result may lead to an important patent. For some documents, third parties will have an interest in their authentication (as in the former case, above); for others, the document's creator will have the interest (as in the latter case). The unpredictability of need, combined with the ease of time-stamping, should encourage techniques that would make authentication routine (and in some environments required) with little effort.

One could therefore conceive of a high-volume time-stamping server providing perhaps a million certifications per minute. But using the logarithmic tree structure described above, the number of certificates necessary to validate to the one-minute root is at most only about 20. On average, therefore, the number of certificates necessary is only ten. All 20 (or less) of these numbers can be returned promptly to our user within the minute (and the average user will wait only half a minute, probably continuing with his or her work while certificate management goes on in the background). Even if a root certificate is provided as often as once a minute, a ten-year list of them could be recorded and published on a single CD-ROM. In practice, servers providing roots and associated times could easily be made available, and this could become an important function for research, government and corporate libraries.

A later user of the document then has available the necessary information for its authentication: the time of its creation (which leads to the root certificate on the public server, which is tamper proof because widely known), the document's own certificate, and the certificates of the other documents necessary to compute to the root. This information may be saved with the document or in other locations (e.g., as part of a citation to the document).

In another month, or in 50 years, when another user wishes to be sure that he has the actual document to which his research trail led him, he uses the information provided along with the document to locate the proper root. He calculates the hash of the document and further uses the additional certificates until he reaches the root. Using the client software, this takes only a moment. If the calculations match the root certificate, he has confidence that he has the desired document.

Real-world implementation

DTS is being presented as a solution of value to a number of information communities, for example banking, law, pharmaceutical companies, and government. Its proposers have been intrigued by unique library requirements including long functional life on the human scale. But if DTS were to be used in research librarianship, several practical matters would have to be worked out.[13] These include:

- means and forms of bibliographic citations using DTS;
- means of associating certificates with documents;
- long-term accessibility of roots;
- the utility and practicality of time-stamping servers or repositories dedicated to library needs;
- financial implications, and effect on desirability, if DTS is marketed as a proprietary service.

Digital time-stamping may provide for many of the needs of the library and archival communities for long-term authentication of electronic information. If this approach turns out not to be suitable, however, it is likely that one relying on similar techniques will be found. In any case, it is important that libraries identify some solution that allows scholars, students, readers, publishers and information users to have confidence that their electronic resources are authentic.

Endnotes

1. For a concise summary of the implications of the "sharp distinction between the carrier and the intellectual knowledge it contains," see Patricia Battin, "From Preservation to Access -- Paradigm for the Future," *Annual Report July 1, 1992-June 30, 1993* (Washington, DC: Commission on Preservation and Access, 1993), p. 1-4.

2. See especially Lesk (below), but also Janice Mohlhenrich, ed., *Preservation of Electronic Formats: Electronic Formats for Preservation* (Fort Atkinson, Wis.: Highsmith, 1993), the proceedings of the 1992 WISPPR preservation conference. In it, Karen L. Hanus provides an extensive "Annotated Bibliography on Electronic Preservation" (p. 121-136). See also "Implications of Electronic Formats for Preservation Administrators," *Newsletter Insert: Newsletter, Commission on Preservation and Access* No. 62 (Nov.-Dec. 1993), p. 1-2.

3. Lesk, Michael, *Preservation of New Technology: A Report of the Technology Assessment Advisory Committee to the Commission on Preservation and Access* (Washington, DC: CPA, 1992; available from the Commission for $5.00, prepayment required).

4. Gordon B. Neavill, "Electronic Publishing, Libraries, and the Survival of Information," *Library*

Resources & Technical Services 28: 76-89 (Jan. 1984), p. 78.

5. As quoted in "Patricia Wilson Berger Inaugurated as ALA President," *Library Hotline* Vol. 18, No. 27 (July 10, 1989) p. 1.

6. Clifford Lynch, *Accessibility and Integrity of Networked Information Collections* (Office of Technology Assessment, Congress of the United States, July 5, 1993; 107 pp.), p. 68.

7. D.F. McKenzie, '*What's Past is Prologue,'* (n.p.: Hearthstone Publications, 1993), The Bibliographical Society (London) Centenary Lecture, 14 July 1992; pp. 21-22, 27.

8. This is of course true only for modern printing. In the hand-press period, one must take account of changes made during printing; see e.g., Charlton Hinman, *Printing and Proof-Reading of the First Folio of Shakespeare* (Oxford, 1963). For more substantial changes see Martin Boghardt, "Partial Duplicate Setting", *The Library (Transactions of the Bibliographical Society)* Sixth Ser., Vol. XV, No. 4, Dec. 1993, pp. 306-331.

9. The archive community speaks of the importance of provenance in establishing that a piece of information is in fact a record. Electronic information by itself can have no demonstrable provenance; the authentication solution hereinafter described may be able to provide the equivalent.

10. Haber, Stuart, and W. Scott Stornetta, "How to Time-stamp a Digital Document," *Journal of Cryptology* (1991) 3: 99-111; also, under the same title, as DIMACS Technical Report 90-80 ([Morristown,] New Jersey: December, 1990). See also D. Bayer, S. Haber and W.S. Stornetta, "Improving the Efficiency and Reliability of Digital Time-stamping," *Sequences II: Methods in Communication, Security, and Computer Science,* ed. R. M. Capocelli et al (New York: Springer-Verlag, 1993), p. 329-334. A useful brief account is in Barry Cipra, "Electronic Time-Stamping: The Notary Public Goes Digital", *Science* Vol. 261 (July 9, 1993), p. 162-163 (I have used Cipra's diagram as the basis for my own).

11. This is consonant with what Battin notes as in the future for librarians: "For analog information, we must develop triage strategies for the past; for digital, prospective triage strategies at the point of acquisition or creation"; Battin, p. 3-4.

12. Client/server software assumes a planned, cooperative relationship between two computers. The server typically provides a generalized source of information or a generalized service to a wide clientele, while the client provides computing intelligence physically close to the user and tailored to the user's specific machine and needs.

13. The Research Libraries Group has determined to embark on a pilot project to develop a repository of electronic research collections. Identifying authentication requirements and solutions is seen as one task of such a project, and the Haber/Stornetta technique is under consideration.

Return to CLIR Home Page >>

In library and archival science, digital preservation is a formal endeavor to ensure that digital information of continuing value remains accessible and usable. It involves planning, resource allocation, and application of preservation methods and technologies, and it combines policies, strategies and actions to ensure access to reformatted and "born-digital" content, regardless of the challenges of media failure and technological change. The goal of digital preservation is the accurate rendering of