AUTHENTICATING ELECTRONIC EDITIONS

PHILL BERRIE, PAUL EGGERT, CHRIS TIFFIN, AND GRAHAM BARWELL

The scholarly edition's basic task is to present a reliable text.
—Guidelines for Editors of Scholarly Editions

A book is generally seen as a trustworthy carrier of text because, once printed, text cannot be changed without leaving obvious physical evidence. This stability is accompanied by a corresponding inflexibility. Apart from handwritten marginal annotation, there is little augmentation or manipulation available to the user of a printed text. Electronic texts are far more malleable. They can be modified with great ease and speed. This modification may be careful and deliberate (e.g., editing, adding markup for a new scholarly purpose), it may be whimsical or mendacious (e.g., forgery), or it may be accidental (e.g., mistakes made while editing, or minor mistranslations by a software system). The nature of the medium makes the potential effect of these modifications greater because the different versions of the text can be quickly duplicated and distributed, beyond recall by the editor. Does the electronic future, then, hold in store something akin to medieval scribal culture? If this lack of control is the risk, will scholars be willing to put several years of their lives into the painstaking creation of electronic editions of important historical documents or works of literature and philosophy?

How can textual reliability be maintained in the electronic environment? There is a major question here of authority and integrity; if not more acute than that in the print domain, it at least has different characteristics. Especially where it is crucial that a text be stable and long-lasting—for example, in legal statutes, cumulative records, or scholarly editions—a noninvasive method of authentication is required. Following a discussion of various problems associated with the markup (encoding) of electronic texts and the danger to ongoing textual reliability that markup poses, we describe a potential model.

THE SUBJECTIVITY OF MARKUP

Verbal texts being prepared in a scholarly manner for electronic delivery and manipulation need to be marked up for structure and the meaning-bearing aspects of presentation. In the electronic domain, the features of text that in the print domain have long been naturalized by readers demand explicit categorizing and interpretation. This task is not straightforward. The most trivial things can raise tricky questions. What, for instance, is the meaning of small capitals or italics in a nineteenth-century novel? Traditionally, italics are seen either as a form of emphasis (and therefore a substantive aspect of meaning) or as presentational (as in the name of a ship or painting). Neither function can be rendered in the ASCII character set. As they cannot responsibly be ignored, a decision about their function (and therefore their presentation by the software) must be made by the human editor. Under the current paradigm, the instruction is entered into the text file.

Similarly, electronic text editors are forced to decide whether line breaks are meaningful, whether a line of white space is a section break in a chapter or only a convenience dictated by the size of the printed page and the desire to avoid widows and orphans. Editors have to decide whether a wrong-font comma, a white space before a mark of punctuation, or a half-inked character is meaningful—should it be tagged or not? The instruction (recorded in markup) will be an editorial interpretation, made, probably, in the context of what is currently known about contemporaneous print workshop practice and convention. In making explicit what in the physical text was implicit, the editor is inevitably providing a subjective interpretation of the meaning-bearing aspects of text. A later editor, or the same editor returning with new information, may disagree with the earlier interpretation.

The arduous business of entering, proofreading, amending, and consequently reproofing a transcription containing the new interpretation (the print edition paradigm) can seemingly be avoided in the electronic medium; but in fact a new state of the text will have been created. Accidental corruption of the verbal text is very possible, so collation and careful checking of the new state against the old will be necessary. The same checking is needed if interpretation of other features of text is added—for example linguistic features, historical annotations, or cross-references. Even though markup is usually separated from text by paired demarcators, as its density increases, so does the practical difficulty in proofing the text accurately.

Consider the following scenario. No one expects any two scribal copies of the same work to be textually identical: scribes will almost certainly have changed or added things, large or small. This instability is not restricted to

pre-1455 or even the pre-1800 period, before the age of the steam-driven machine press. Optical collation in scholarly editing projects has proved again and again that no two copies of the same edition are precisely identical, even if printed in the industrial age. Printing involves change as well as wear and tear; inking varies, and paper has imperfections. While recent editorial theory has shown that the physical carrier can itself affect the meaning of text (e.g., McKenzie, *Bibliography*), the prospect of marking up text to record every physical variation in every known copy of a work would create a file of bewildering complexity whose reliability would be in serious doubt. No editor can foresee all the uses to which an electronic scholarly edition can be put or all the interpretative markup that will be required. The more the attempt to provide interpretative markup is pursued through increasingly heavy tagging, the more the reliability of the text is put at risk.

This situation shows the need for an automated authentication technique that separates verbal text from markup while retaining all the functionality of a computer-manipulable file. The proposal that we describe below involves such standoff markup. It also addresses another problem of markup that has often been observed. The current standard for the markup of humanities texts, that of the Text Encoding Initiative, requires an objective textual structuring to be declared on the assumption that if computers are to manipulate parts of text powerfully, then text needs to be seen as an ordered hierarchy of content objects with its various divisions and parts appropriately identified. The difficulty with this assumption is that texts are neither objective nor ideal things. They incorporate a stream of perhaps only lightly structured human decision making, of which traces have been left behind as part of the production process. Moreover, we as readers cannot help participating in the business of making meaning as we read and interpret what we see on the page. The advantage of our participation is that we, unlike computers and logic systems, can handle structural contradictions and overlaps with relative ease and safety. But if we then attempt to codify the texts for use with systems that cannot handle contradictions, the systems reveal their inadequacies. At present, only fudges—partly satisfactory work-arounds—are possible to deal with this problem.

AUTHENTICATION TECHNOLOGIES

Authentication technologies were developed by information scientists to provide a reliable basis for sending verifiable messages over networks. These technologies are based on the mathematical routines of cryptography but are

designed to work with clear-text messages. (The subtle forms of meaning-bearing presentation discussed above are not normally relevant here.) The goal of such technologies is not to obscure the information contained in the message but to verify that it was sent by the person claiming to have sent it and has not been altered in the course of transmission. Meeting these requirements has allowed the development of e-commerce with such services as Internet banking.

These services require a large amount of infrastructure to support them. Changes deemed necessary to the authentication protocols and procedures must be carried out quickly because of the potential risk of criminal exploitation of a weakness. While financial institutions have the money to pay for these high maintenance costs, such resources are not available to an academic community interested in authenticating its electronic editions.[1] Authenticated financial transactions over the Internet have a lifetime of minutes if not seconds, whereas full-scale electronic editions must have a life of decades if they are to justify the investment of an editor's time and energy. The chance of an electronic edition's becoming unusable because of the obsolescence of its authentication system rules out the use of proprietary and invasive solutions.[2]

Fortunately, authentication for electronic editions is not as exacting as that for e-commerce, where it is a requirement that the creator of the message be verifiable. In electronic editions, detection of textual corruption is the primary concern. An authentication system must protect the reliability of the encoded text, by indicating if and where a file has been corrupted, thus allowing it to be replaced from a trusted master copy. The best authentication method is bit-by-bit comparison of the working copy of the file against a locked master copy. Some electronic editions at present provide their master files on nonvolatile media (e.g., CD or DVD); working files are always generated afresh from the master files. Unfortunately, this solution is very weak for long-term storage, as the master files are bound to a particular storage technology. And the system does not allow for the possibility of revised or additional interpretative markup.

Most authentication methods involve the use of hashing algorithms.[3] In its simplest form, a hashing algorithm steps through the characters of a piece of text using a mathematical formula to calculate a hash value that is dependent on their sequence. The formula is such that the resulting hash value is highly representative of the text, because small changes in the text produce large changes in the calculated value. Authentication is achieved by comparing the stored hash value of the master copy with the calculated hash value

of a working file. If they are the same, it is extremely likely that the two files are identical. This technique prevents from going undetected corruptions of a file that are otherwise easy to overlook.

STANDOFF MARKUP AND AUTHENTICATION

> I want to discuss what I consider one of the worst mistakes of the current software world, embedded markup; which is, regrettably, the heart of such current standards as SGML and HTML.
>
> —Theodor Nelson, *Embedded Markup*

The problem of maintaining the authenticity of a text file across platforms is not a trivial one. In addition, it is desirable to prevent the proliferation of different versions of a text that would otherwise be brought about by (future) developments in or additions to markup, annotation, and cross-referencing. The use of standoff markup in an electronic text environment possessing strong authentication characteristics may allow these desiderata to be met.

To illustrate how such authentication might be achieved, let us take the case of a literary work extant in several typesettings. After the base transcription file of each typesetting was prepared, each such file would be a lexical transcription of the original but minimally marked up—since the editor's interpretative responsibilities could be fulfilled in standoff markup files. The verbal content of the base file would need to be contained in uniquely identifiable text elements. In prose, this containment could be done at the level of the paragraph; in verse, at the level of the line. The identifiers would need to be inserted in the text to act as markers, and the text proofed against the original. After proofing, the file's authenticity could be maintained by an authentication mechanism based on a simple hashing algorithm. Ideally, authentication would be done at the text-element level, so that a change to even one character would be immediately discernible when the hash value for the text element was checked. Such authentication would allow possible corruptions to be quarantined while leaving the rest of the text usable. Once the base transcription file had been prepared and proofed, markup (e.g., in SGML using a document type definition [DTD] conforming to the TEI guidelines)[4] would be inserted, its operation tested and then removed into a separate, standoff file. Standoff files would also store the hash value of the text element to which the markup could validly be applied. The result of this structuring is that the tags would carry a test of the authenticity of that portion of the text, and any attempt to reapply them to a corrupted version of the text element would result in a notified error.

A model developed along such lines would offer a number of advantages. First, by supporting the standard TEI-compliant SGML, it could be used in an SGML environment, giving access to all the available browsers and tools. But the base transcription file would not be dependent on SGML, and the separate markup files could be easily manipulated to comply with whatever markup schemes were required.[5] Second, this model would enable the text to be annotated or augmented with analytic markup, in parallel and continuously, while still retaining its integrity. Third, the levels of markup could be developed independently for different purposes and applied selectively to meet different user requirements. This independence would future-proof the edition against the obsolescence brought about by subjective markup, since any edition deemed unsuitable for a particular application is liable to spawn a competitor that will vie with the original text for maintenance resources.

To date, only one implementation of the proposed model has been developed for electronic editions: the JITM (just-in-time markup) system. It has utilities for inserting tags, subsequently removing them, and running the verification process. The embedding into the base file of the markup from the standoff files creates a virtual document—a perspective—that is inserted into a template conforming to the appropriate DTD.[6] Because any markup added to the base file is extracted into standoff markup files and the base file is authenticated "just in time" when a call is made to create the new perspective that incorporates the added markup, an automatic proofreading of the base file is in effect being continually carried out.[7] This procedure can significantly reduce the time it takes to create an electronic edition while maintaining the academic rigor required for such a project. The same authentication system continues to ensure the reliability of the edition after publication; and the same textual resources do not need to be newly transcribed or proofread afresh for each new editorial or other study.

There are further advantages. First, in the original creation of the base transcription file, proofing can, if desired, be simplified by separate checking of the markup on the one hand and the words and punctuation on the other. Second, different or conflicting structural markups can be applied to the same base file, because they are in different files and can be applied to the base file selectively. Finally, because the JITM system separates the transcriptions from the markup, the question of copyright is simplified. Since the markup is interpretative (as explained above, and more obviously with added explanatory and textual notes), a copyright in it can be clearly identified and defended. In all this, the base transcription file remains as simple as possible

(thereby greatly
cation mechanis
rather than a pro

Ensuring contin
for print edition
edition does not
the edition is pro
that gave it life.
ability, but they
term availability

The use of
potentially allow
enhanced and tl
future developm
without compro
labor that went

NOTES

1. The National
born-digital a
tication techn
cure files are c
2. These authent
where the file
calculated fro
owner. The u
to authenticat
private key gu
infrastructure
distribution o
numbers, at le
handled by so
3. The National
5 (MD5) hash
authenticity o
4. The standoff
embedded ma

(thereby greatly easing its portability into future systems) and the authentication mechanism remains noninvasive. JITM is, in other words, an open rather than a proprietary system.[8]

Ensuring continuing reliability is a bigger issue for electronic editions than for print editions. The creator's responsibility to the users of an electronic edition does not end with its publication; steps must be taken to ensure that the edition is protected against corruption by the very processes and medium that gave it life. Authentication technologies can provide the required reliability, but they must be applied in such a way that they protect the long-term availability and reliability of the edition against obsolescence.

The use of standoff markup and abstracted authentication techniques potentially allows editions to have their markup revised, reinterpreted, or enhanced and their protection mechanisms easily upgraded or replaced, as future developments require it. Such maintenance will be able to be done without compromising the base transcription files or wasting the editorial labor that went into establishing them.

Notes

1. The National Archives of Canada has decided to archive only clear text in its born-digital archives (Brodie). The extra costs involved in archiving the authentication technologies necessary to authenticate the original, cryptographically secure files are considered too great a burden.
2. These authentication solutions are largely based on the idea of the digital signature, where the file to be authenticated has attached to it a cryptographic signature calculated from the contents of the file and a unique private key registered to the owner. The user of the files uses a public key provided by the message originator to authenticate the file, and the correspondence between the public key and the private key guarantees that the file was sent by the owner of the private key. The infrastructure involved in this system involves the calculation, registration, and distribution of the authentication keys. Currently these key are unique prime numbers, at least one hundred digits in length. The distribution of these keys is handled by sophisticated servers that are expensive to maintain.
3. The National Library of Australia records in its online catalog a Message Digest 5 (MD5) hash value for its digital assets in the *PictureAustralia* service so that the authenticity of the files downloaded by users can be checked.
4. The standoff markup paradigm would readily support the use of other, normally embedded markup systems in parallel, if this support were a requirement.

5. When writing for this chapter began, the P4 version of the TEI DTD had just been released. Now, the technology has progressed such that XML is the requisite language and the P5 version of the TEI DTD is almost upon us. Trusting in the stability of embedded markup for long-lived e-texts is shortsighted at best.
6. While the base transcription file does not in itself adequately represent a historical state of the work being edited, the default perspective in JITM for new users is the one that records the physical presentation of the original. More experienced users, and scholars seeking to interpret the base file or turn it to new purposes, work with the base file.
7. Each tag markup instruction incorporates a hash value for the text element into which it is to be inserted. The comparison of this stored value against the value calculated for the text element provides the automatic proofreading of the JITM system.
8. The algorithms for the JITM system and the hashing algorithm it uses are to be made public in due course. For papers about the project, go to the Australian Scholarly Editions Centre Web page at www.unsw.adfa.edu.au/ASEC/JITM and see the JITM Web site itself at www.unsw.adfa.edu.au/JITM. Just In Time Markup is copyrighted 2005 by Graham Barwell, Phillip Berrie, Paul Eggert, and Chris Tiffin.

Jerome McGann is John Stewart Bryan University Professor at the University of Virginia. His development group, ARP (Applied Research in Patacriticism), recently released *Ivanhoe*, a collaborative online play space for generating and analyzing acts of critical reflection.

Katherine O'Brien O'Keeffe, Notre Dame Professor of English at the University of Notre Dame, has edited the C-text of the Anglo-Saxon Chronicle (2001) and is completing a study of the textual dimensions of the Anglo-Saxon subject. She was cochair of the Committee on Scholarly Editions.

D. C. Parker, professor at the University of Birmingham, is codirector with Peter Robinson of the Institute for Textual Scholarship and Electronic Editing, coeditor of the International Greek New Testament Project, and editor of *Texts and Studies*. He is currently working on editions of the Gospel of John in Greek and Latin.

Sebastian Rahtz, information manager in IT support for Oxford University, uses TEI XML extensively for creating Web pages. He is a member of the board of directors and of the Technical Council of the Text Encoding Initiative Consortium. He is a chief architect of the revised fifth edition of the TEI's literate programming language.

Peter Robinson is professor of English and textual scholarship in the Faculty of Humanities and director of the Centre for Technology in the Arts at De Montfort University. He is developer of the textual-editing program *Collate* and director of the *Canterbury Tales* Project. In 2000, he founded Scholarly Digital Editions, a new electronic publishing house specializing in high-quality electronic publications.

Bob Rosenberg oversaw the creation of the *Thomas A. Edison Papers* Web site, which currently has more than 180,000 document images, and laid the foundation for marking up the text edition's transcriptions. He is an independent scholar in the San Francisco area.

G. Thomas Tanselle, senior vice president of the John Simon Guggenheim Memorial Foundation and adjunct professor of English at Columbia University, is author of *Textual Criticism since Greg* (2005) and a coeditor of the Northwestern-Newberry edition of Melville.

Chris Tiffin teaches at the University of Queensland in Brisbane, Australia. He is the compiler of *Mrs. Campbell Praed: A Bibliography* and editor or coeditor of *South Pacific Stories*, *South Pacific Images*, and *De-scribing Empire*.

John Unsworth is dean and professor at the Graduate School of Library and Information Science, University of Illinois, Urbana. He has been director of