# *Imagining the Digital Archive: The InterPARES Project*

**Bonnie Mak**
School of Library, Archival & Information Studies
University of British Columbia

*Presented at the 119th Annual Meeting of the American Historical Association*
*Seattle, Washington*
*January 8, 2005*

In the June 2003 edition of the *American Historical Review*, Roy Rosenzweig called to attention the impact of digital technology on archival research in a forum entitled, "Can We Save the Present for the Future?".[1] He observed that although more and more records are being generated on the computer, there are very few guidelines for the preservation of these electronic documents. Many record-creating bodies have no overarching system in place for archiving their files, and may choose to save nothing or everything of their work.

Rosenzweig argues that historians have a professional obligation to help address this lack of foresight; these records will be the foundations for historical scholarship in the future. By re-establishing ties with archivists, he says, historians can come to a better understanding of the challenges facing those who are trying to preserve cultural heritage in a digital environment. With this understanding, they may be able to contribute productively to the conversations that are already taking place in archival circles.

And so I come to you, not as a voice from the other side; not as an archivist, but perhaps as a liaison. I am trained as a medieval historian, but now work as a part of an international team that conducts research on digital records and archives.

Conceived and directed by archival scientists, InterPARES is an inter-disciplinary project that examines how electronic records are currently being generated. We assess what elements of these records must be preserved, and compare our findings with what can be preserved, given technological and financial constraints. One of the goals of InterPARES is to draw up policies and guidelines to assist record-creating and record-keeping bodies better prepare their digital records for long-term preservation. In order to help, as Rosenzweig says, "re-establish ties with archivists," in this paper I will outline a few of the concerns that the archival community has regarding the preservation of digital records, and the ways that it is seeking to resolve these issues.

In almost all sectors of life, people are now working on-line, from government to business to scientific research to the performing arts. The documents that chart and record these daily activities, once created on paper, are now generated and maintained on the computer. Records that we are accustomed to having in hard copy no longer take that form.

[1] "AHR Forum Essay: Can We Save the Present for the Future?" *American Historical Review* 108.3 (June 2003), p. 734; Roy Rosenzweig, "Scarcity of Abundance: Preserving the Past in a Digital Era," *American Historical Review* 108.3 (June 2003), pp. 735–762, <http://www.historycooperative.org/journals/ahr/108.3/rosenzweig.html> (11 January 2005).

These new records are not just born in a digital environment; many of them exist only in that form. These documents are created, modified, sent, received, read, maintained, and stored on the computer. They can range widely in type, and may include e-mails, tax forms, MRI and CAT scans, and artistic works. What distinguishes these entities from their predecessors on paper is not merely that they are digital. Rather, it is the added functionalities afforded by computer technology that make these records importantly different. They are frequently interactive and dynamic, changing in response to different inputs from different users; for example, a single record may draw information from multiple sources at the behest of a user.

One of the obvious hurdles for preserving electronic records is the technological. Certainly the records can be saved, but the concern for archivists is whether the records will remain legible and accessible as time passes and technology changes. It is probably clear to everyone by now that in order to keep their own electronic documents legible and accessible, they need to transfer them every time a change in software occurs — from WordPerfect to Word, perhaps, or from Word 97 to Word 2000. Migration, that is, the transferring of documents to a new technology, has become the most common method of preservation employed by archivists. However, it is by no means guaranteed that our files can be migrated through successive upgrades in hardware and software without experiencing some kind of change. It is therefore unlikely that digital records will remain exactly as they were when set aside, or that they will retain all of their characteristics.

Given that certain changes will have to take place to keep digital records legible and accessible, the challenge then is to ensure that these changes are kept to a minimum. Moreover, the records should be guarded with care during this process of migration and afterwards, so that no unnecessary changes are introduced that would ultimately undermine their validity. In this way, the job of the archivist is not only to preserve records and keep them legible and accessible in the long term, but also to maintain their authenticity, or trustworthiness.

Authenticity refers to the trustworthiness of a record as a record. This should be distinguished from reliability, which, in archival terms, refers to the trustworthiness of the record with regard to its content — whether the record can be considered an accurate statement of historical fact.

Archivists are interested in authenticity because they are called upon by courts of law to testify that a record is the same as it was when it was originally set aside, and that the record has been duly protected from tampering by an unbroken chain of custody. As you see, the concern is about the record's status as a record, rather than having anything to do with its content. A recent example that springs to mind is the furor over the memos allegedly written by a squad commander in the National Guard, which documented a young George W. Bush's lacklustre performance. One of the major problems with these records, as CBS discovered, was that their authenticity could not be verified — that is, no one could vouch for where the records had come from or where they had been in the intervening years — the chain of custody had been broken.[2] Whether the information in the records is accurate or not becomes irrelevant if no one can guarantee the ownership and custody of the records themselves.

[2] Among others, Jacques Steinberg and Bill Carter, "CBS Dismisses 4 Over Broadcast on Bush Service," *The New York Times* (11 January 2005), A6; and Maureen Balleza and Kate Zernike, "Memos on Bush are Fake but Accurate, Typist Says," *The New York Times* (15 September 2004), A5.

Authenticity or trustworthiness of records is based on two complementary factors: identity and integrity. The identity of a record assures us that the record is the same as it was when originally written. For traditional records on paper or parchment, this usually refers to sameness in its material, its form, and its content. Integrity more specifically refers to whether the record was corrupted after it was set aside. Digital records pose a special challenge because authenticity must now be guaranteed through successive technological advancements — namely, even in the face of necessary mutations in the records, archivists must be able to show that these documents are the "same" as they were when they were set aside, and have not been tampered with subsequently.

While archivists in recent years have focussed increasingly on the authenticity of records, historians have rarely concerned themselves with this aspect of their primary sources. They recognize that documentary evidence is, at some level, firstly an expression of subjective experience when it was created, secondly subject to preferential methods of preservation, and thirdly open to interpretation by the modern eye. Given these variables, it is of lesser interest to historians to know whether a record was preserved in a careful, authentic manner. They are more interested in preservation — in the sense that there is documentary evidence and that it is legible and accessible.

However, authenticity may soon become an issue for historians because, in the electronic environment, it is tied more closely to keeping records legible and accessible. As we have said, in order to preserve electronic records in a useful way, we will have to modify them to some extent. These necessary changes have ramifications for both the accessibility of the record and authenticity of the record. The dilemma is this: the record must be modified to be able to be read and accessed (goes to preservation); but the record must be modified in such a way that it still retains its identity and integrity (goes to authenticity). What can we reasonably call "sameness" under these circumstances? Data that is recorded on the computer can be re-constituted in different ways, none of which may be the way it had originally been instantiated. The identity and the integrity of the record in the electronic environment can therefore not be found in its re-constitution, or in other words, in how it looks or in how it is presented.

To help uncover what might be called the identity of electronic records, InterPARES is using the very traditional methodology of diplomatic analysis. As you may remember, this branch of study that focuses on the critical reading of documentary evidence began in the seventeenth century, as a result of a dispute between Daniel van Papenbroeck and Jean Mabillon.[3] Van Papenbroeck threw into question the credibility of the founding charters of the great Benedictine abbey of St. Denis. In order to defend the rights and privileges of the abbey and his monastic order, Jean Mabillon conducted a study of around 200 records, and proved that the charters of St. Denis were in fact genuine by comparing their medium, ink, language, form, and seals. With the publication of Mabillon's lengthy proof in 1681, the field of diplomatics was born.

Today, diplomatics is still mainly used as a tool for historians to help analyze written records from the Middle Ages. Diplomatics identifies the formal elements of documentary evidence. These elements encompass both the external characteristics of a given record, such

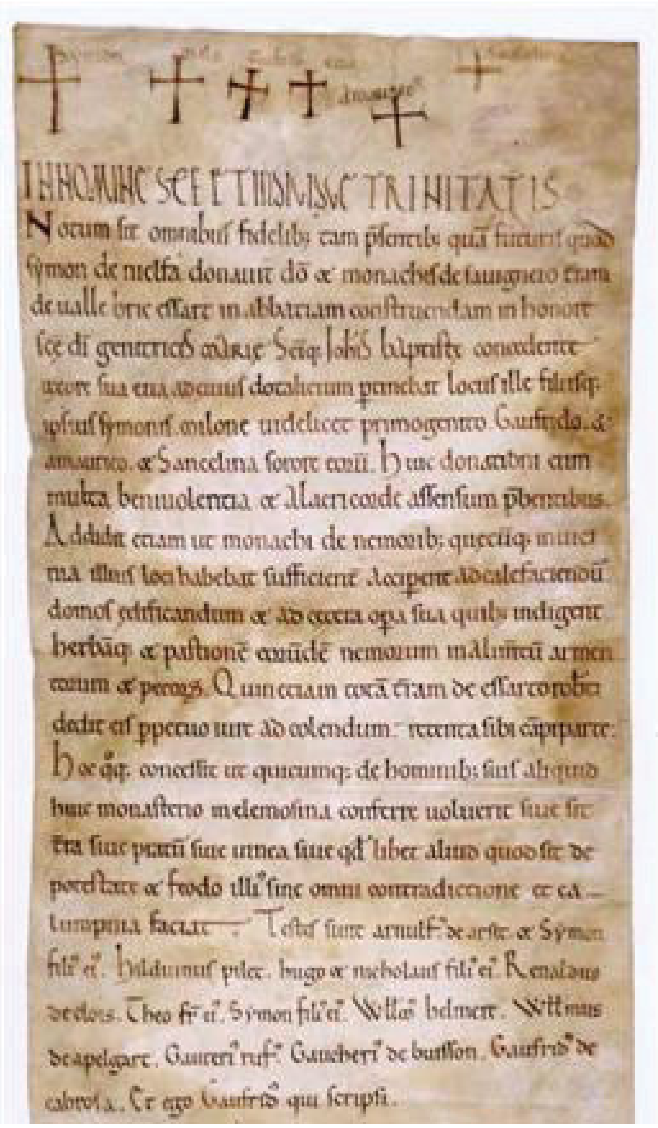
[3] Daniel van Papenbroeck, "Propylaeum antiquarium circa veri ac falsi discrimen in vetustis membranis," *Acta Sanctorum, Aprilis* II (Antwerp, 1675), pp. 1–52; and the response, Jean Mabillon, *De re diplomatica*, 2 vols. (1681, rpt. in Naples: Vincento Ursino, 1789). In general, see, Oliver Guyotjeannin, Jacques Pycke, and Benoît-Michel Tock, *Diplomatique médiévale* (Turnhout, Belgium: Brepols, 1993); and Georges Tessier, "Diplomatique," *L'histoire et ses méthodes*, Charles Samaran, ed. (Paris: Gallimard, 1961), pp. 633–676.

as its material construction, its shape, its format, and the like, and the internal characteristics, such as the sections of its text, and the type of act that it describes. Using this method of analysis, historians can quickly sort through records and readily identify what elements of the documentary evidence are important for their particular research.

For instance, in a record from 1118, in which Simon de Neauphle donates land to the monks of the abbey of Savigny to build a new abbey, a diplomatic analysis can very quickly identify all the important elements for us (Fig. 1).

One of the most obvious extrinsic or external characteristics of this charter is its medium — namely, its physical means of conveyance — , which is parchment. As for the *mise-en-page*, the text is written parallel to the shorter side of the parchment. This layout, called *charta recta*, was characteristic of private contracts of the time. The script, also an extrinsic element, is a Carolingian minuscule, with Rustic Capitals used for embellishment. Finally, the language of the text is Latin.

The first internal or intrinsic characteristic of the charter that is visible is the invocation. These few words are used to place a record under the patronage of a saint or God. In this case, the words *in nomine sancte et individue trinitatis* are clearly written in Rustic Capitals; the invocation calls upon the holy Trinity. The *notum sit*, or the announcement, is followed by the addressees of the charter. Here, the charter addresses *omnibus fidelibus tam praesentibus quam futuris* — all those faithful, present and future. The author's name is next; it is Simon de Neauphle who has the competency to issue the charter.



IN NOMINE SCE ET INDIVIDUE TRINITATIS

Notum sit omnibus fidelibus tam presentibus quam futuris quod
Symon de nealfa donavit deo et monachis de saviniaco totam
de valle brie essart in abbatiam construendam in honore
sce dei genitricis marie et sci iohannis baptiste concedente
uxore sua eva ad cuius dotalicium pertinebat locus ille filiisque
ipsius symonis milone videlicet primogenito Gaufrido et
amaurico et Sancelina sorore eorum. Huic donationi cum
multa benivolentia et alacri corde assensum prebentibus.
Addidit etiam ut monachi de nemoribus que eiusque in interi-
ma illius loci habebat sufficienter accipiant ad calefaciendum
domos edificandum et ad cetera opera sua quibus indigent
herbagiumque et pasturam eorundem nemorum in alimentum armen-
torum et pecorum. Qui etiam totam etiam de essarto roberti
dedit eis precio iure ad colendum retenta sibi campiparte.
Hoc quoque concessit ut quicumque de hominibus suis aliquid
huic monasterio in elemosina conferre voluerit sive sit
terra sive pratum sive vinea sive quodlibet aliud quod sit de
potestate et feodo illius sine omni contradictione et ca-
lumpnia faciat. Testes sunt arnulfus de arde et Symon
filius eius. Hilduinus piler. hugo et nicholaus filii eius. Renaldus
de clois. Theo frater eius. Symon filius eius. Willelmus belmeit. Willelmus
de apelgart. Gauterius rufus. Gaubertus de buisson. Gaufridus de
cabrosia. Et ego Gaufridus qui scripsi.

Fig. 1. Donation of Simon de Neauphle. 1118.

The disposition, or act, can easily be identified by the verbs — in this case, *donavit* and *addidit*. Simon is donating land to the monks, the addressees of the act, and he adds that they should also receive the right to use the neighbouring forest and pasture. The corroboration, the *testes sunt*, is the announcement of the signs of validation by the author and witnesses. Finally, in the attestation we find the names of the author and witnesses. The names and signs of Simon and his family can be seen running across the top of the charter, while the names of the witnesses are listed at the bottom.

Applying the same method of analysis to new records that have been created on the computer shows us that although electronic records may appear vastly different from their predecessors at first blush, we should remember that all documentary evidence has been (and continues to be) generated by the same processes, and for the same purposes: we are charting our own, human activities. As a result, all records, regardless of their medium, use similar kinds of formulae and similar kinds of guarantees.

For electronic records, there appear to be at least six areas in which diplomatic forms still apply. Without these six elements, the record cannot be recognized from an archival point of view. Consequently, these elements need to be preserved in order to have a digital record that can be considered complete with respect to identity and integrity.

1) form (in the sense that the message can be rendered with the same documentary form that it had when it was first set aside; for instance a memo remains in memo form; a letter remains in letter form);
2) content, which especially involves:
    - the author, addressee, and writer; and
    - the action in which the record participates (disposition);
3) explicit links to other records through a classification code or other unique identifier; and,
4) administrative context.

You will notice that these aspects do not include extrinsic characteristics, such as the material form, of the electronic record. In modern diplomatics, the medium, or the means of conveyance of the record, is no longer taken by archivists to be meaningful.

But the recent studies in book history by the likes of Anthony Grafton, Robert Darnton, and D.F. McKenzie have shown us that it is not simply the content held within a text that can be of use to researchers.[4] Indeed, evidence offered by the material and formal elements of written sources can be as important for research as the textual content that they support. For instance, the scent of correspondence from the Middle Ages has been used to establish which areas were experiencing an outbreak of the Plague; letters were routinely dipped in vinegar in hopes of halting the spread of the disease. Or scrap material used in the binding of certain manuscripts has been used to establish patterns of dissemination of French polyphonic music in England.[5] Given that we do not know how electronic records will be used — whether they will be examined for content, for form, for functionality, or for something else as yet unimaginable — in my view we should try at least to consider all of their characteristics when preserving them.

[4] Lisa Jardine and Anthony Grafton, "'Studied for Action': How Gabriel Harvey Read his Livy," *Past and Present* 129 (November 1990): 30–78. Robert Darnton, "What is the History of Books?" *Daedalus* 111 (1982), pp. 65–83. Reprinted in, David Finkelstein and Alistair McCleery, eds., *The Book History Reader* (New York: Routledge, 2002), pp. 9–26; and Robert Darnton, *The Kiss of Lamourette. Reflections in Cultural History* (New York: Norton, 1990), pp. 107–135. D.F. McKenzie, *Making Meaning: "Printers of the Mind" and other Essays*, Peter McDonald and Michael Suarez, eds. (Amherst, Mass.: University of Massachusetts Press, 2002); and D.F. McKenzie, *Bibliography and the Sociology of the Texts* (London: British Library, 1986).

[5] Among others, Edward H. Roesner, "Who 'Made' the Magnus liber?" *Early Music History* 20 (2001), pp. 227–266; Olga Malyshko, "Three Newly Discovered Fragments at Worcester Cathedral: Another 'Magnus liber organi' Flyleaf," *Scriptorium* 52.1 (1998), pp. 66–82; and Rebecca Baltzer, "Notre Dame Manuscripts and their Owners: Lost and Found," *The Journal of Musicology* 5.3 (Summer 1987), pp. 380–399.

However, part of the difficulty with preserving the original extrinsic characteristics of electronic records is that the process of migration obliterates them. Once you transfer the document that you created in WordPerfect 5.1 over to Microsoft Word 2003, there is no record of its original state. One of the proposed solutions to this problem is to create a metadata file that would accompany the electronic record. This metadata file would describe the original characteristics of the record, both its intrinsic and extrinsic elements. So, a file would be attached to your document that stated that you had originally composed it in WordPerfect 5.1, when you had composed it, and when it was transferred to Word 2003, etc. This method of preserving the extrinsic elements may mean that much of our "non-textual" evidence that has been of such interest to historians of the book will only be able to be found in a description that accompanies the electronic record.

One extrinsic and non-textual element that has already presented itself as a challenge to preservation is the digital signature.[6] The digital signature is a code which is attached to a document by the signatory, sealing the file from alteration.

Historians are probably more familiar with the predecessor to the digital signature, the seal. Many medieval records are accompanied by seals that served to validate or confirm signatures and to provide solemnity to the documents. Seals became the main method of validating records by the eleventh century. A ball of wax, lead, silver, or gold was affixed to the paper or parchment and impressed with matrices, front and back. The seal could be placed on the charter itself, by cutting a hole in the material and placing the wax in the hole; or could be attached on a loop of parchment, silk or leather. It verified the origin of the record, and attested to its identity, integrity, and indisputability. As a physical attachment, the seal is not an intrinsic element of the record, but rather an external and "non-textual" element.

Similarly, the digital signature verifies the origin of the electronic record, and attests to its identity, integrity, and indisputability. In current practice, when the digital signature is affixed to an electronic document, that document is "sealed," or locked in that particular form. In order to open the document, the recipient needs to check the signature by using an algorithm that matches the digital code with the signatory's key. Like the seal, the digital signature is considered an extrinsic characteristic of the record, and contains a wealth of information that may be of great interest to researchers in the future.

However, there are a number of problems that archivists have encountered while attempting to preserve digital signatures and the records to which they are attached. Firstly, there is the issue of whether the appropriate software will remain for reading and verifying the signatures. That is, it is difficult to deduce the signing key or algorithm even a few years after the digital signature has been placed on a document. Without the signatory's key or the algorithm, the signature cannot be read.

Secondly, if we choose to maintain these elements, and freeze the digital signature, the algorithm, and signing key to keep the signature legible, the record to which the signature is attached must also be frozen. In this case, while the digital signature may be readable, the record will not be.

Finally, if the record is migrated to newer software in order to maintain its legibility and accessibility, the signature will be rendered unreadable. Although the record can be read

[6] See, for instance, Jean-François Blanchette, "The Digital Signature: To Preserve or Not to Preserve," *Imaging Science & Technology Archiving Conference*, 20–23 April 2003 (San Antonio, TX); Heather MacNeil, "Providing Grounds for Trust: Developing Conceptual Requirements for the Long-Term Preservation of Authentic Electronic Records," *Archivaria* 50 (2000), p. 61–63.

in its new format, the signature cannot. The legibility and verifiability of the signature depends on the record remaining precisely the same, particularly with respect to its bit-content.

From this example, the challenges of preserving electronic records become clearer. To keep non-textual evidence accessible in the long term may paradoxically mean the loss of textual evidence, and vice versa. As the chief users of archives, it is imperative for historians to become more involved in discussions about the preservation of our digital heritage. Archivists are not generally trained in research methodologies, and are therefore unaware of the variety of ways in which records can be read. Consequently, they have little idea of what historians deem valuable in records, and what historians would wish to be preserved.

Conversations between historians and archivists will result in more informed choices in the preservation of public memory. Archiving can be understood as a method of self-representation, and we should now begin to think about what traces of ourselves we wish to leave. How do we wish to shape our own legacy? I leave you, then, to imagine the architecture of your own digital archive.