Are Electronic Editions Inherently Obsolete?

Phillip Berrie, Graham Barwell, Paul Eggert and Chris Tiffin

Australian Scholarly Editions Centre

Australian Defence Force Academy

p.berrie@adfa.edu.au

Abstract

This paper presents a challenge to the current print-based paradigm for creating electronic editions in terms of their long term viability. It discusses some of the problems inherent with the use of embedded markup and describes how these problems could cause the loss of such editions to the human record. This paper describes how the Authenticated Electronic Editions Project is using the techniques of Data Simplification, Standoff Markup and Just In Time Authentication in the creation of a new type of electronic text that does not suffer the same limitations.

Introduction

As a practising computing consultant. My day-to-day job is helping people get the most out of their personal computers. Having made this admission you might wonder why I have asked the question “Are Electronic Editions Inherently Obsolete?” You might think that it is in my best interest to promote all uses of computing technology to make sure that my bread continues to be buttered and that expressing concern over possible limitations of the technology threatens my livelihood. It probably does, but unlike a lot of people in the IT field, I am not driven solely by money.

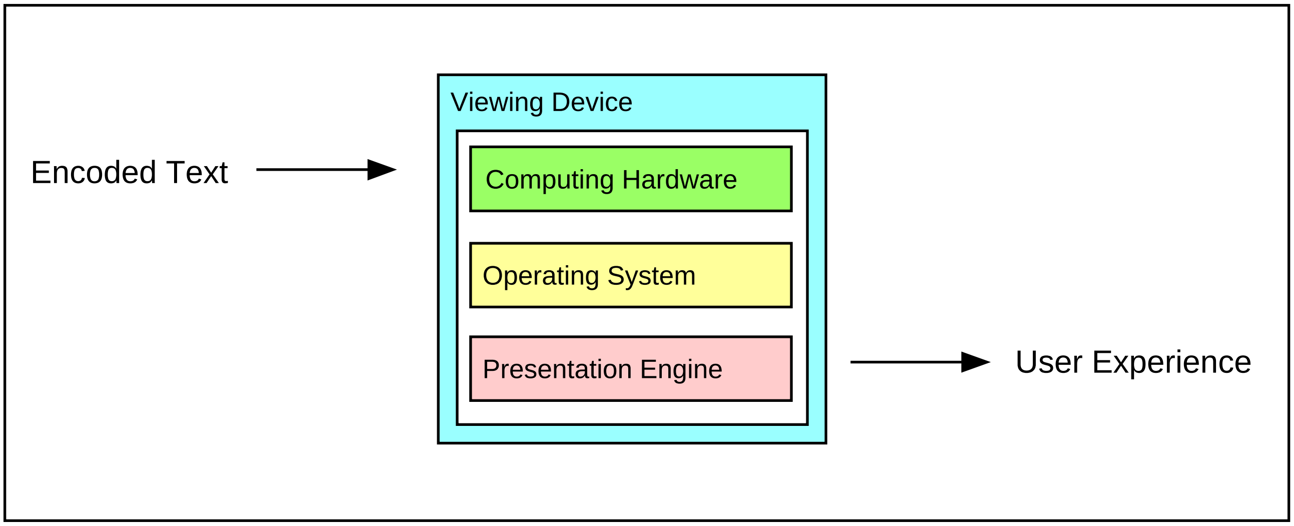
Having worked in IT at Universities for over twenty years I have seen many examples of the effect of the advancement of computing technology on the maintenance and integrity of stored data. One of the big advantages of books is their long-term archival property. The “Book of Kells” has lasted over a thousand years. How many people in the audience today could make use of data stored on 8” floppy disks? In fact, with the average life-time of a document on the World Wide Web being about seventy days[1] I feel that the growing fear amongst information scientists of this time becoming the electronic equivalent of the dark ages is all too real.

> **If you are involved in the development of electronic texts, you must think NOW about the long-term survival of your projects.**

Many factors can be responsible for the possible loss of digital information. This paper will concentrate on factors that can be inherent in the data itself, which leave it susceptible to loss through obsolescence.

Reasons for Obsolescence

The main cause of obsolescence with these technologies is that a "Viewing Device" is required to access the stored material. For the first time in history, information is stored in a form that is not directly accessible by the people that created it.



The diagram shows the main components of an electronic edition in use it will be used to show the parts of the whole experience, which are susceptible to obsolescence.

The term, "Presentation Engine", as shown in the diagram is a general term for any software that provides the "User Experience" of the "Encoded Text".

[1] Statistic attributed to Brewster Kahle [Kahle, 1998] of the Internet Archive in the paper "Is the Record of the 20th Century at Risk" by Diane Vogt-O'Connor [Vogt-O'Connor, 1999].

This could be anything from a simple text editor through to a fully TEI compliant SGML browser and includes research tools such as search engines.

The Obvious

Not to belabour the obvious, but it should be clear that the three sub-components that make up the viewing device are very susceptible to change and problems dealing with encoded texts because of obsolescence of some sub-component of the viewing device are very common. The following are some of the reasons why the sub-components of the Viewing Agent may become obsolescent.

- **Advances in hardware technology.**
    - **Keyboard entry over punched cards.**
    - **High quality output capability.**
    - **Better storage capacity.**
- **Advances in software technology.**
    - **Graphical User Interfaces.**
    - **Networking Protocols.**
- **The commercial imperative to sell product.**
    - **Competition drives the development of new technology.**
- **Changing standards.**
    - **ASCII, SGML, HTML, XML, UNICODE**

As can be seen from the examples, obsolescence need not be a bad thing. In fact, since the field of computing technology is so new, less than fifty years for most of our purposes, it is hardly surprising that the technologies involved are still evolving.

Clearly one of the problems is that we are working at the wrong time with tools that have not properly matured. This is an intractable problem.

The Not-So-Obvious

The not so obvious problems arise from the inter-dependency of the components in the diagram. The important thing to note is that every part of the model is heavily dependent on the others. These dependencies will now be explored in detail.

The Encoded Text

It should be emphasized here that we are talking about the encoded text as something separate from the hardware on which it is stored. Storage technologies have their own problems of obsolescence, but this discussion is about some of the not so obvious problems with the way information is being recorded that may be just as fatal to the long-term survival of an electronic text as any obsolete hardware would be.

There are two levels of encoding that should be considered here. As mentioned earlier this is an interesting time in computing technology. The simplest, but most profound level of encoding, that needs to be considered is the use of single or multi-byte encoding for recording the individual characters of the data. The commonly used ASCII character set records a possible 256 different characters in an eight-bit byte of computer memory. This was fine when the majority of computer users used the Roman character set. However, it has serious limitations when used for other languages. Multi-byte encoding schemes such as Unicode use two or more bytes to store character information allowing, in the case of two-byte encoding, up to 65,536 different characters and thereby providing support for the use of most non-Roman characters. Operating systems and software that use multi-byte encoding schemes are only now becoming commonly available, but with the globalisation of the IT industry, the advantages of using these encoding schemes are undeniable. It therefore becomes a distinct possibility that sometime in the future, software will no longer recognise single-byte encoding schemes, like ASCII.

Who will be responsible for ensuring that your ASCII encoded electronic texts will be accurately converted to the new encoding scheme? This can be done automatically, but the sheer volume of information stored on computers throughout the world make this task the responsibility of the owner of the material because the onus of scholarship requires that the new version of the file should be checked for accuracy of the conversion process.

The second level of encoding concerns markup languages. For an electronic edition developed under the prevailing paradigm, the encoded text is an amalgam of two different types of information. The first type of information is the transcription of the text, the actual words and punctuation, while the second form of information is any markup (i.e. codes embedded in the text) required to provide the appropriate user experience desired by the creator of the edition. Examples of this type of markup include, at the basic level, the recording of typographical emphasis such as italicisation but also includes more specialised markup for applications such as the analysis of dialogue or identification of proper nouns for purposes of annotations. The possibilities here are boundless.

Other people have written about the problems inherent with embedded markup [Raymond et. al. 1995], however I would like to concentrate on the problems that affect the long-term archival properties of electronic editions. There are two problems with the embedded markup paradigm.

Problem number one is that the syntax and data details of the embedded markup is dependent on the markup language and requirements of the presentation engine, which has already been pointed out as being an optional component of the viewing device. Change the presentation engine and the specifications for the markup in the encoded text may change[2]. The change may be for the better, adding richness and increased facility to the user experience. Nevertheless, who is going to do it and ensure the authenticity of the encoded text? We will all be well aware of the proofreading overheads involved in preparing scholarly texts. Should the creator of the edition be expected to do all the markup upgrades due to changing technologies to maintain the scholarly integrity of their original work? Moreover, who will take on this responsibility when the original creator can no longer fulfil their role as the guardian of their academic rigour?

The second problem with embedded markup is revealed by considering the last component of the model we have been using - the "User Experience".

The User Experience

The diagram might imply that the user experience is a passive component of the system. This is of course far from the truth as the user experience is the driving force behind the development of the electronic text.

What happens in the case when the required user experience of the edition is not supported by the rest of the components?

If the user is the creator of the edition, then the other components are modified until the desired experience is achieved. For example, if the creator decides to include more features into the edition then extra markup can be added to the encoded text to facilitate these requirements.

What happens when the requirements of a user other than the creator are not being provided by the edition? From a bibliographical point of view, the best result is that the creator makes the required changes to the edition to support the new user requirements. The intermediate case is that the user makes what use of the edition they can and does not follow up their real interest. The worst case is that the user creates their own edition.

[2] Newer browsers conform to version 4 of the HTML DTD. This version marks as obsolescent some markup that was legal under earlier versions. How long before browsers no longer know how to handle documents created with this older syntax.

It is impossible for the original creator to provide markup for all possible uses of an edition. It is not the complexity of the task or the problem of dealing with conflicting structures in SGML-based languages that makes this impossible. It is the impossibility of knowing all the future user requirements for the electronic edition. Therefore, in the long-term, with scholarship being what it is, the probability of new electronic editions being developed to cater for requirements that are not available in the existing electronic editions is high. The proliferation of variant states of a work is a well known problem. Surely we should be able to prevent this problem from reoccurring in the digital age.

The other side of this problem concerns the long-term survival of the individual electronic editions. Having multiple editions creates an environment where competition for available management and maintenance resources brings the forces of natural selection into play putting at risk those editions that are considered lacking, for whatever reason, no matter how meritorious the scholarship that went into their creation.

The Problem with Editions

Getting back to the original question:

> "Are Electronic Editions Inherently Obsolete?"

Of course, this is in fact a trick question. The problems we have been looking at are mainly due to the use of the word "Edition".

Problems with Publishing

The use of the word "Edition" when dealing with electronic texts is a carry-over from the printing paradigm. An edition is defined by the act of publication, at which time it becomes static and part of the written record and cannot be easily corrected or extended. In the case where the scholarly work of an edition is to be extended, this normally involves the publication of a new edition and any necessary corrections are normally incorporated into the new edition.

Today's electronic editions tend to exist in two formats: CD-ROM and Server-based. The first form is directly akin to a printed edition in that the original published edition cannot be changed after publication. It should be obvious that I have grave concerns about the longevity of this form of electronic text.

The Internet and the World Wide Web have allowed the development of server-based electronic editions where users can access the up-to-date scholarship of those responsible for making the electronic text available. In effect, this gets around the problems of publication as the work is always “in progress” and part of this ‘progress’ is the continual maintenance of the work so that it remains available. Archivists now believe that the best way of maintaining digital assets is to keep them in use so that public interest ensures that they are propagated onto new platforms as technology advances. However, embedded markup causes problems with this idea.

Subjective Data

In many cases, the accuracy of the transcription is based on the editor’s interpretation of the original and therefore it becomes a subjective process. Certainly, the science of paleography helps minimise the subjective component of the copyist’s work, but even in transcribing printed texts there are cases of ambiguous readings where the biases of the copyist will become incorporated into the transcription. Different copyists can therefore produce different transcription files. It will become the subject of discussion as to which more correctly represents the original work. This difference of interpretation is possibly an insurmountable aspect of the act of transcription[3]. This is brought to your attention as an example of how even an ‘accurate’ transcription may be unacceptable to some scholarly users.

The problems with subjective data are more subtle though. Under the current paradigm the markup is embedded into the electronic text. This markup is being used to record information about the text for use by the presentation engine to provide the user experience the editor considers best. This markup can include details of the structure of the original, which has to be specially encoded in the digital medium as it was represented in the original by typographical artefacts such as white space or line breaks. It can also include interpretive markup considered important by the editor for the appreciation of the text. For example, an editor who is interested in characterisation could markup all references to the different characters in a story. The important point is that the editor is making decisions as to what markup to include in the edition. Their selection is subjective, based on their own needs and opinions and is therefore possibly at odds with needs and opinions of other users of the edition.

[3] This problem highlights the necessity of allowing the user access to facsimile images of the original so they can make their own decisions.

As we have already discussed, embedding subjective markup into the electronic texts automatically creates the possibility that different texts of the same work may need to be created to cater for different views and needs of the work thereby putting at risk the less-favoured editions because of the overheads involved in maintaining digital resources.

How then can these problems be solved? Well perhaps they can't, but things can be made easier for the scholars of the future by changing the way we do things now. The printing paradigm is not a suitable methodology for dealing with electronic texts and we shall now examine some of the features of the Authenticated Electronic Editions Project, which have the potential for creating more robust and long-lived electronic texts.

Our Solutions

The following are the aspects of the Just In Time Markup (JITM©[4]) system that in our estimation promote the long-term survival of the electronic text we are working on. The most important aspect of the AEE Project is that it is a work in progress and new source material can be added without affecting either the authenticity of the transcriptions or at the expense of previous work.

Simplification of the Data

The strongest archival property of the Just In Time Markup (JITM©) system is the simplification of the data. By keeping, the transcription files as simple as possible we keep them very easy to maintain so that they will be easy to propagate onto different computing platforms in the future. This also means that they can be used by other people without them having to use any of the rest of our system thereby avoiding them having to re-transcribe the original and potentially creating variant electronic states of the work.

Another aspect of the system is that although there is an Academy Editions reading text of *His Natural Life*, which is the scholarly work of the editors of the project, we also plan to make available to users our source materials. These resources will include facsimile images of the original states as well as the authenticated transcription files of all the witness states used to create our reading text. This will allow users to make their own decisions and if

[4] The Just In Time Markup (JITM) system is Copyright 2001 by Berrie, Barwell, Eggert & Tiffin.

necessary extend our work through their own efforts. The nature of the AEE Project allows the inclusion of the works of others very readily without jeopardising existing work. New source material can be added and even reading texts established on different principles can be incorporated into the project without detracting from what is already there but using the same resources.

Standoff Markup

A natural consequence of these simplifications is the requirement to use "Stand-off Markup" to add features to the electronic text. With the standoff markup technique the detail of a document's markup is kept in a separate file. This markup is applied to the text as required to create a virtual document. By keeping the markup separate from the transcriptions it can be manipulated, emended and even ignored depending on the user's requirements.

Manipulation of the markup cannot affect the authenticity of the transcription files thereby reducing the overheads involved in working with the transcription in the short term and protecting the transcription files in the long term by making them as reusable as possible.

One of the most powerful features of standoff markup as used in the AEE project is that the mechanism used allows multiple users to create markup for the same piece of text simultaneously. This is inherently impossible with the embedded markup paradigm predominant today.

Just In Time Authentication

Just In Time Markup is a standoff markup system developed as part of the AEE Project. For those interested more information is available on the web at the following URL:

http://idun.itsc.adfa.edu.au/ASEC/

Briefly, the standoff markup is applied to the transcription of the text as required. The difference between JITM© and other stand-off markup techniques, few that these are, is that JITM© incorporates just in time authentication as part of the process to guarantee that the transcription is authentic.

Authentication of the transcription is verified in the process using a key/lock type approach. Each piece of markup is stored with an authentication key that must match the calculated value of the text it is modifying to be a valid insertion. All markup is effectively inserted simultaneously so that the calculated value of the transcription file is not changed by any previously embedded markup.

The virtual text so created is termed a “Perspective” of the work, it is an amalgam of the transcription files, and a user specified set of markup. Many different perspectives are possible and the number of different perspectives increases factorially[5] as new markup elements are added. Limitations with the SGML-based markup schemes prevent a perspective supporting conflicting markup, but the JITM© system does support conflicting markup using multiple windows.

The JITM© authentication mechanism promotes the archival potential of the work in the following manner. By keeping the authentication mechanism external to the transcription file (i.e. the authentication key is stored with the markup and is compared against a calculated value from the transcription file) we prevent the transcription files being put at risk because of the obsolescence of the authentication scheme. Authentication schemes are software algorithms (i.e. programs) and are just a subject to obsolescence as any other software so by abstracting the authentication mechanism away from the transcription files we prevent the transcriptions from becoming obsolete due to the obsolescence of the authentication technology. This is a potential danger with other authentication technologies such as digital signatures, which incorporate part of their mechanism into the files of the texts they are protecting.

The authentication mechanism also allows for multiple copies of the electronic resource to be available. Crosschecking of authentication between sites occurs automatically when markup created at one site is used at another site. If the authentication step fails for the markup at the new site then one of the two sites has a corrupt transcription file and this can be checked against the originals or an authenticated facsimile. In the case of a castastrophe where neither the original server nor authenticated facsimiles are still available, a consensus agreement between surviving sites could be arrived at to decide on the most likely correct reading.

[5] The number of possible combinations of features included in a perspective is based on the factorial of the number of features available. For example, if there are four features available, the number of possible perspectives is twenty-six (eg. 4 x 3 x 2 x 1 + 2). The twenty fifth perspective would be the case where all features were selected and the twenty-sixth perspective would be the case were no features were selected.

Conclusion

A New Type of Electronic Resource

The challenge of this paper was based around the weakness of the static nature of an edition. The idea of a continually evolving electronic work as described calls for a new term. The term I prefer is, “Study”, where the transcriptions and markup constitute a “study” of the work. This term gives the feeling of on-going scholarship, in keeping with an evolving scholarly effort.

Summary of paper

Hopefully this paper has shown some of the problems associated with the current paradigm of electronic texts. Hopefully the “not so obvious” points that have been raised, highlight some of the intractable problems with embedded markup and show how it make electronic editions susceptible to obsolescence because of its limited adaptability to changing user requirements.

The paper also reiterates some of the strengths of the JITM© technology, which we believe potentially, make the results of the AEE Project versatile enough for it to outlive it creators. Creating a “Study”, a work in progress, that any scholar can easily contribute to without interfering with the work of others will create an electronic resource more likely to find people who will maintain and propagate it onto the computing platforms of the future thereby ensuring its long-term survival.

Responsibility of Transcription

In conclusion, it is important to talk briefly about the responsibility of transcription. The nature of the digital medium is such that the transcription and checking process should only need to be done once for any particular work. To do this successfully the complications and compromises of markup should be removed from the equation. It is my contention that the greatest contribution of any electronic edition of an existing work is the transcription of the work onto the digital medium. Therefore, it is paramount that the accuracy of the transcription be as high as possible and that steps be taken NOW to ensure the continued authenticity and usability of the transcribed file.

Bibliography

Kahle, Brewster. “Setting the Stage: Summary of the Initial Discussion,” in *Times & Bits” Managing Digital Continuity*. Edited by Margaret MacLean and Ben H. Davis, Santa Monica, CA: The J. Paul Getty, Trust, 1998, p. 39.

Lynch C., “The Battle to Define the Future of the Book in the Digital World”, *First Monday*, Vol. 6, No. 6, June 2001. Also
URL: “http://firstmonday.org/issues/issue6_6/lynch/index.html”

Raymond, D. R. et. al. “Markup Reconsidered”, Presented at the First International Workshop on Principles of Document Processing, Washington DC, October 21-23, 1992.

Thompson, H. S. and McKelvie, D., “Hyperlink semantics for standoff markup of read-only documents” also
URL: “http://www.ltg.ed.ac.uk/sgmleu97.html”

Vogt-O'Connor, D., "Is the Record of the 20th Century at Risk?" in *CRM: Cultural Resource Management,* Vol. 22(2), 1999, pgs 21-24. Also
URL: “http://tps.cr.nps.gov/crm/archive/22-2/22-02-9.pdf”