AUTHENTICATED ELECTRONIC EDITIONS PROJECT

Paul Eggert, Graham Barwell, Phill Berrie and Chris Tiffin

Background

Phill Berrie and I have been involved for several years with Chris Tiffin and Graham Barwell in a project that fills in a gaping hole. We take a long-term view about the survival prospects and the *ongoing* accuracy of scholarly electronic editions. Even when created using a standardised and widely accepted markup system, and even if not tied to proprietary software, electronic editions face an uncertain future. Electronic texts can be copied and modified effortlessly; the modification may be accidental, perverse, for the purpose of adjusting text or, more likely, adding markup for a new scholarly purpose. In addition, disaster planning is necessary if you're thinking a century into the future: keeping copies at many locations is sensible.

As the electronic edition is migrated to new platforms – *if* it is – or modified, who is to know which of the distributed, and now *different*, copies is the correct one? What we may be looking at is a duplication of something like medieval scribal culture. If that is the case, will scholars really want to put several years of their lives into the painstaking creation of electronic editions of important works of literature, philosophy and of historical documents?

Our electronic edition of Marcus Clarke's *His Natural Life* – which is our great, gut-wrenching convict novel from the 1870s – is a spin-off from the hardback critical edition that appeared late last year in the Academy Editions of Australian Literature series.

Using *His Natural Life* as our testbed, we have developed a simple, non-invasive mode of authentication that depends upon a system of stand-off markup: i.e. the markup is in files external to the text itself. First, I'm going to list the advantages of what we call Just In Time Markup (JITM). Then I'm going to discuss some theoretical-cum-practical problems we've encountered. Then Phill will talk, giving an on-screen presentation.

The advantages:

- Transcription files are authenticated in the act of using them for display or interpretative purposes. Phill will demonstrate this. The current MLA guidelines require authentication but don’t say how to do it.

- Different or conflicting structural markup can be applied to the same transcription file. The markups are in different files and can be applied to the transcription file selectively.
- Collaborative and simultaneous use of the same transcription file is possible, since this work is in effect producing stand-off markup files and not altering the text at all.
- Transcription files can be used for interpretative purposes other than those envisaged by the original editors but without compromising the integrity of the transcription file. This is done by the creation of new stand-off markup files.
- In the original creation of the transcription file, proofing can be simplified into notionally separable markup and any other interpretative markup on the one hand, *and* the words and punctuation on the other.

Theory hobbling practice

You won’t appreciate the nature of the breakthrough, however, unless I take you through some of the underlying theoretical and practical questions we worked through.

When you are heavily involved in marking up a text for structure and presentation the last thing you want to think about is the abstract issue of what text is or what a work is. That’s not the problem you’re trying to solve. You’re trying to make a text amenable to computer manipulation because there is going to be, you anticipate, a big pay-off in the end. The trouble is, the features of text that you had naturalised – had long since ceased to observe – in printed text now start to demand categorising; they ask awkward questions. As you begin to answer those questions you begin to realise that, in your markup, you are making explicit a theory about the way in which texts exist, about how they are meaningful.

The most trivial things can ask the hardest questions. What for instance is the meaning of small capitals in a 19th century novel? In WORD we have to type them as lower case and then apply the small caps command. Is there some wisdom in this? Does it somehow reflect the ghostly hand of, say, Sir Walter Greg applying Thomist abstractions, that they are accidentals rather than substantives: features of meaning rather than meaning itself? But if so, is this always the case with small caps? Can they be, say, semi-substantive in the way that punctuation can be – as Fredson Bowers, late in his life, argued about Shakespeare’s printed punctuation?

The normal approach would be to tag them with the <emph> tag, with the rendition value, small capitals:

<emph rend = "small caps"> ___ </emph>

But *are* small caps a form of emphasis? Or, do small caps have the same or a different meaning from full capitals, which we wouldn't tag? Or from lower case? We can't just ignore them since they are most definitely present. A typesetter working from his marked-up printer's copy or perhaps reflecting the conventions of the workshop deliberately inserted them. That fact is somehow a constituent of the text. Texts have histories, and we're not – we can't – deal with text in the abstract when we're marking-up.

In the event, we have decided to go with the HI> tag from the TEI DTD as it does not imply what the <emph> tag does:

<HI rend = "small caps"> ___</HI>

If you are marking up a text you have to decide, one way or the other, how you are going to mark up its features and apply that decision systematically. We've made a decision about small caps, but we could be wrong. Our markup could be, almost certainly will be, superseded.

This is an uncomfortable realisation. Even while you are doing it, it seems that as soon as you have made up your mind about a way to tag, the text turns refractory on you. You mark-up, say, italics with the <emph> tag and you decide that punctuation is just part of the text, not in need of special markup. Then you hit a ship's name enclosed in inverted commas. Fair enough: no problem here – until you find that the next version of the same novel that you're now working on has presented ships' names in italics. You've already decided that italics need special tagging. But here italics are being used not for emphasis (which scholarly editors might well interpret as meaningful and thus substantive) but are a presentational feature of meaning. So clearly it does have to be marked up, despite the fact that the tag signals a different status for the name than its counterpart had in the other version where markup was not appropriate.

The problem with trying to make text fit a logical structure and reflect categorical distinctions such as computers deal well with is that texts are not just objective things. They incorporate a stream of perhaps only lightly structured human decision-making of which traces have been left behind. Nor can we as readers help but participate in the business of making meaning as we read and interpret what we see on the page. The advantage of this is that we, unlike computers and logic systems, can handle contradictions with relative ease and safety. But when we then attempt to codify the texts for use with systems that cannot handle contradictions, the systems reveal their inadequacies.

Texts involve human participation, but they are recorded as documentary objects that go on existing independent of our reading of them: they exist as ink on paper or magnetic or electronic traces or as sound waves. As Lou Burnard has said in one of his oft-repeated papers about the TEI, markup forces you to the realisation that the basic textual data is very basic indeed. I'd put it this way: markup forces us to return to the basic distinction between document and text, a distinction that we have all but naturalised out of existence.

So we are forced to decide whether line-breaks are meaningful, whether a line of white space is a section break within a chapter or whether it was only a convenience dictated by the size of the page and the desire to avoid widows and orphans; we have to decide whether a wrong-font comma, a white space prior to a mark of punctuation or a half-inked character is meaningful – do we tag them or do we not? In medieval manuscripts, do we expand abbreviations on a whole-word basis because this is conventionally expected? But if we do, what information are we losing? The mark of expansion is physically present in the manuscript, yet expanding the word eliminates it.

No-one expects any two scribal copies of the same work to be textually identical since 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 (where different copies of the same edition are compared letter by letter to check for possible variation) proves again and again that no two copies of the same work are precisely identical, even if printed in the industrial age. Printing involves change, wear and tear; inking varies, so does paper, etc etc.

This line of thought starts to beg bibliographic and editorial-theory questions about the relationship between a text and a work. What, after all, is a text a text *of*? Perhaps none of this would make any discernible difference if it were not that the scholarly edition is a machine of consistency and the editor especially sensitised to textual variation. Scholarly editions grind exceedingly small; but electronic editions grind even smaller. The computer crunches ones and zeros, not abstractions such as work or text.

Theory into practice

What conclusions can we draw? What we think of as markup and what we think of as text are not necessarily distinct, as we saw with small caps and ships' names. But this doesn't mean we should abandon the attempt to employ the distinction for useful ends. It used to be common to say that a speech or an idea had been 'reduced' to writing. Similarly, to mark up is to reduce a text to a form that a computer can conveniently deal with. There is no wriggling out of this realisation that a marked-up text is an impoverished representation (based on a single interpretation) of what the text is saying and how it is saying it in terms of structure and presentation. There is always a reduction going on (the representation of that stream of human participation of which only traces are left); but it's not always easy to say, in any particular case, just exactly what the reduction is or how important our ignoring it is.

Of course, many people have serious reservations about what we are ignoring. The very business of marking up in valid SGML presupposes a DTD that enforces an assumption that text is an Ordered Hierarchical Content Object (OHCO). Disputes about whether texts are indeed OHCOs have been cropping up for the whole life of the TEI. If texts of the machine-press period are unstable, what hope have we got with manuscripts? The Wittgenstein project in Norway maintains that the OHCO assumption leads to a misrepresentation of Wittgenstein's manuscripts. For them, overlapping hierarchies are simply there and have to be represented, but the logic of the TEI is to forbid it. Yet, the recent medieval manuscripts description project that Peter Robinson is involved with is finding a way round an allied problem and is getting some following. The key seems to be in the necessity of making pragmatic compromises: reductions that will pay off. This doesn't mesh well with the high aims of the TEI project, but it may be more realistic. Scholarly editions are a high-minded compromise; electronic editions may be as well. So let us choose to be pragmatic: computer manipulation may be worth the price.

No matter how much markup I might want to pack into a transcription file to adequately express my understanding of what the text is saying, sure as eggs someone smarter or better informed than me will sooner or later point out what I overlooked or how better I could have arranged and structured my markup. Then I am going to have to go back into the file and start again, or make many changes. What is more, I can never predict all the uses to which the marked-up transcription might be put. But every time I change my mind about what should be in the markup and add or change the file accordingly I should ensure that I have not inadvertently corrupted it. We are, after all, talking about a scholarly edition here: it could be a biblical work, it could be a canonical work of literature; it could be an authoritative wording of a piece of legislation. We are talking about any text which has reached a point of stability and whose wording it is important to preserve precisely.

The fact that it is in electronic form does not absolve us in the least from achieving the same standard of reliability and accuracy that a print-based scholarly edition traditionally achieves. Strictly, therefore I should not just rely on search-and-replace mechanisms to make the now-desired changes. I *should* proof the revised version entirely, and certainly collate it against its prior version, checking the report of variants. This is a considerable added burden, especially given that I had already done it many times over in preparing the scholarly edition by collating its variant versions and checking the variants against the original hard copies. Furthermore, if the file is now replete with tags serving the various functions I have foreseen as needing to be there for different categories of users, then the job of proofing becomes very difficult indeed. The eye cannot deal easily with complicated tagging.

The pragmatic compromise we have adopted in our project is to step as lightly as possible on the still mysterious terrain of text by separating characters and punctuation from presentational and nearly all other markup. In the transcription file, we insert TEI <div> tags to define text elements (e.g. paragraphs), but that is all. These text elements have a unique id attribute to allow them to be manipulated easily.

- *Ours is an impoverished definition of text, granted. But the rest (including italics and other aspects of presentation, line breaks, page breaks and any other interpretative markup) is recorded in tagsets which can be applied to the file at will in order to create what we call a* perspective. *We'll show you this in a minute. In the act of creating the perspective, a checksum authenticates the text, ensuring that no alteration to it has happened since it was created.*

How far we should go with markup is another source of uncertainty. Scholarly editors have traditionally defined the stream of words and punctuation as the text, ruling out things like headers, pagination, line breaks for prose, paper stock and binding, leading and type size, thus relying on everyone's naturalised assumptions about the vagaries of textual embodiment in physical form, deeming that embodiment to be either meaningless or insufficiently important to be worth wasting time on or alternatively lamenting the inability of editorial processes to deal with it. Jerome McGann has more recently alerted us to the bibliographic codes which constitute meaning for the reader, or which affect the reading experience, and he has proved that the cases where physical form matters are by no means restricted to Blake.

Expanding this purview of the textual object to its logical endpoint where every copy is differently meaningful is in line with what I said before about the lack of complete identity of copies of the same edition of the same work. The trouble with this line of argument is that it incapacitates the editorial intervention that readers need. It puts such store on the importance of textual variation that it forbids the claim that a textual representation of a work can be any more reliable than any other. Life isn't long enough to sustain this degree of textual equalitarianism, I believe.

Thus we reach another justification for the pragmatic compromise I have discussed. Marking up for every variant feature in every extant copy would be possible in theory, but think of the intricacy of the resulting file! Nevertheless *if* you wanted this facility but without an impossibly heavily marked-up file, the JITM system would give it to you.

That's not the route we've taken with the testbed for our JITM system. We've chosen to prepare an e-edition – or really, as we call it, a study – of the great convict novel *His Natural Life* by Marcus Clarke. The critical edition, a result of many years work, appeared late last year. It documents variation between the 1874 first Melbourne edition prepared by Clarke, the three-volume first English edition published by Bentley in 1875, the first American edition, the second English edition, as well as a later serialisation in a Queensland newspaper. But even the critical edition could not contain the very first version that had appeared serially in Melbourne in 1870–72 and that is about one-third longer than Clarke's revision for first book-form of 1875. The e-edition can and does contain it. It has transcriptions and jpeg facsimiles. We are developing an automatic collation facility between the states of the short version and a look-up table that allows you to go between the material common to the long and the short versions.

We feel no need to apologise for our incorporation of an electronic equivalent of the reading text from the critical edition. Indeed its being made available as part of the textual archive or study (as we call it) keeps the editors honest since it can be collated against its own copy-text and all of the emendations noted. In the print edition on the other hand many categories of silent emendation (mainly of trivial corrections or normalisings) operate and only a few hundred emendations are actually listed in the apparatus.

The explanatory notes in our study are linked to this reading text via another tagset. Reworking the critical edition's explanatory notes for the study is turning out to provide another order of problem, as we try to conceptualise the needs of our reader. We believe that most users of the electronic study will want at least to start with this reading text and will be grateful for a point of reference. Distributing the notes into glossaries accessible from the transcription file of any state is a next step. More experienced users will be able to navigate from any of the versions to any other by means of the collation facility. Once again, this facility has not committed us to any special embedded markup in our transcription files. It uses the text element <div> tags for paragraphs, and beyond that counts the words.

- *Paul Eggert*