2001

Authenticated electronic editions project: A progress report

Graham Barwell
University of Wollongong, gbarwell@uow.edu.au

Phillip Berrie

Paul Eggert
University of New South Wales

Chris Tiffin

Publication Details

Barwell, G, Berrie, P, Eggert, P & Tiffin, C, Authenticated electronic editions project: A progress report, Digital Resources for Research in the Humanities, 2001, University of Sydney. Original article can be found at: http://setis.library.usyd.edu.au/drrh2001/

AUTHENTICATED ELECTRONIC EDITIONS PROJECT: A PROGRESS REPORT

Graham Barwell, Phillip Berrie, Paul Eggert, Chris Tiffin

Paper presented at the conference, Digital Resources for Research in the Humanities, University of Sydney, 26-28 September 2001

Background

In 1991 the Academy of the Humanities established a series, the Academy Editions of Australian Literature, consisting of critical editions in book form of some of the major contributions to Australian literary culture of the nineteenth and early twentieth centuries. The works chosen for inclusion in the series do not currently exist in reliable editions, so the text of each work is freshly edited to be as accurate and reliable as possible. Each edition includes the editor's introduction, and textual and explanatory notes, while some editions offer background essays by other scholars, maps, chronologies and similar aids for readers. The project is modelled on similar projects in the United States and Canada.

Since the book project was designed from the start to make full use of computing resources, such as for collation, and the question of producing editions in electronic form was being widely canvassed at that time, it was soon recognised that an electronic version of the Academy Editions could take advantage of the possibilities offered by the new medium. The proposed electronic edition would be much more than a printed book in digital form. It was acknowledged that electronic editing of this kind was still a very new field and its potential benefits would best be demonstrated by its application to a work with a very complex textual history. The work chosen from those already being prepared for print publication was Marcus Clarke's novel of convict life, *His Natural Life.* This had first appeared in serial form in an Australian journal, but was then substantially reduced and altered by Clarke prior to publication in book form in Melbourne, before being serialised in this shorter form in another Australian journal, with subsequent book editions in both England and the United States. The Academy of the Humanities agreed to extend its support to this undertaking.

The project involves scholars at three universities, Queensland, Wollongong and the Australian Defence Force Academy, where it is supported by the Australian Scholarly Editions Centre. The work in preparing the electronic edition has been funded over the years by grants from a variety of sources, most recently from the Australian Research Council through its SPIRT scheme with the Academy of the Humanities and Apple Computer as industry partners, but the research and development work is still in progress. The project is designed in two stages: the first stage will see the production of the complete text and digitised page images of each significant early serial or book version of Clarke's novel, together with a full treatment of variant readings, plus the essays, notes and other matter from the print edition. The second stage will see the incorporation of a wider range of related material revealing the reception history of the work, such as illustrations, paintings, songs, film and television adaptations. A valuable feature of stage one is the inclusion of the full texts of all the major editions of the novel from Clarke's lifetime. This is simply not possible in a printed work and it has been usual for the longer first serial version of the work to be printed separately. Even in the print version of the Academy Edition the relationship between the shorter and longer versions of Clarke's novel can only be indicated by a look-up table showing equivalent chapters and a collation of a sample chapter illustrating, without fully documenting, the extent of the changes Clarke made in his revisions.

Technical expertise in the project is provided by Phillip Berrie from the Information Technology Services Centre at the Defence Force Academy, and he has devised a particularly valuable method of applying stand-off markup to the electronic transcription file of each print version of the novel to remove the possibility of introducing errors and to guarantee the integrity of the transcription file, no

matter how much we manipulate it. Phillip calls this the Just In Time Markup (JITM) System[1]. Implementation of this feature is an important part of stage one of the project and will be discussed in more detail below. Phill's paper briefly outlines the principles underlying the JITM system, which supports many SGML applications: for instance, TEI-compliant SGML and XML; HTML. The importance and innovative nature of the work has seen a shift in the project title, from Academy Electronic Editions Project to Authenticated Electronic Editions Project, as we recognise that the electronic *His Natural Life* implements principles and practices which are widely applicable in electronic publishing.

Because the project is complex and includes some novel features, it has required an extensive process of planning, research and development. The amount of time involved in producing electronic publications should not be underestimated, especially in situations like ours where we are working on a not-for-profit basis and fitting our work around other paid commitments. We want the project to be successful, so we have devoted considerable time to thinking through all the issues involved in making it successful. In the first instance, we have given top priority to ensuring the electronic edition maintains the same high standards of accuracy and reliability as the print edition. Since we have aimed at a similar market to the print edition, we have endeavoured to adhere where possible to the expectations of the appropriate professional body with regard to this material. In our case these are expressed in the Guidelines for Electronic Scholarly Editions drawn up by the Scholarly Editions Committee of the Modern Language Association of America.[2] Because we want the edition to be as robust and long-lived as possible we have designed it to conform to international standards and to be platform and software independent. Thus text characters are represented by the ISO 646 character set (with built-in support for ISO 10646), images are in JPEG, while the work is marked up in the dialect of the Standard Generalised Markup Language (SGML) recommended by the Text Encoding Initiative (TEI). The work resides on a server and will be accessed using the usual browsers developed for the Net. For the development stage of the project we have been using a special browser plug-in, Softquad's *Panorama,* to recognise the SGML markup. We anticipate moving to browsers with XML capability as they are developed.

At this point it is appropriate to report our achievements to date beginning with the architecture of our electronic edition.

Architecture

The challenge was to devise an architecture which is flexible and robust, suitable for the two-stage development of *His Natural Life* and extendable to editions of other texts. It is predicated on the user viewing and working with the edition via an ordinarily available web browser and connecting to a server containing the electronic files and processes for generating the on-screen representation of the text which the user requires.

The architecture we have agreed on (see Appendix) gives primacy to the edited text appearing in the print edition, in that ancillary material of a specific nature, for example, editorial annotations, would be coded to the edited text, while material of a more general nature, for example, maps, or an essay on Clarke and convictism, would be always available to the viewer of a JITM generated perspective via links from the webpage displayed onscreen. These links function like the contents page of a book.

Under the JITM system, users first specify what particular kind of activity they will undertake, then select the appropriate parts of the edition to be viewed and the markup which will be applied to it. The resulting on-screen representation of the electronic edition, generated by the JITM system and delivered by the server, is called a *perspective.* For example, a user wishing to examine the textual history of chapter 1 may select chapter 1 as it appears in the printed Academy Edition, together with

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

[2] Committee on Scholarly Editions. "Guidelines for Electronic Scholarly Editions." 1 Dec. 1997. Modern Language Association of America. 21 June 2000 <http://sunsite.berkeley.edu/MLA/guidelines.html>.

versions of that chapter as it appears in the variant states. Basic structural features of the original published versions may be important, such as page, paragraph and line divisions, so the appropriate tags are selected. The user can call up digitised images of the chapter in each variant state page by page, to confirm the reading in the transcription file. If users wish to undertake their own analytical work, we envisage the electronic edition offering them the opportunity to download files of the edited text from the print edition and as many of the variant states as they require, together with appropriate JITM tools to produce their markup.

The Just in Time Markup (JITM) System

Having considered at length the various ways in which markup could be generated and made available, we eventually decided on a process in which the markup and transcription file are always kept separate. They are combined only as required on the user's screen. In his paper Phill gives a brief account of the advantages of this system and the reasoning which lies behind its development. The project website (http://idun.itsc.adfa.edu/ASEC) gives full details of the system so there is no need to repeat them here.

The system has involved the development of a set of tools and algorithms to allow users to input the transcription files, generate markup tags, extract those tags into a separate file and then authenticate the transcription files to establish they have not been changed in any way. In the case of the variant states, the transcription files have been produced from the carefully proofread electronic files used to compare variant readings when the book edition was being prepared. The electronic file of the edited text from the printed edition is taken from the *PageMaker* file used to set-up the printed book. We have provided a generic basic tag set in which the , <div1> . . . <div9> tags are used to define the blocks of text. The tags are more flexible and allow a finer level of granularity than that provided by the <p> tag, especially for poetry. We do not envisage the project team providing all the tags researchers might find useful. The JITM system is designed to allow scholars to produce tag sets for themselves, tailored to their specific needs.

The JITM toolset is still under development and has been subject to extensive trials by members of the project team. The tools currently in prototype are the JITM Preprocessor, which translates the files used in collating the book edition into what are called "transcription files" in the JITM system, and the JITM Transcript Editor, which is used to create, insert and extract markup tags, and then to authenticate the transcript file. The toolset is platform-specific at the moment, but it will be made platform independent in stage 2 of the project, prior to distribution to the scholarly community. In that final form the toolset will also allow researchers to use SGML editors of their choice. In our prototype website for the edition, the part played by the JITM system is not made overt when users connect to the server. They are merely asked to select the perspective of the edition they wish to have generated for them. They do this by selecting appropriate buttons for transcription file and tag set, with the combination of the two then appearing on their screen.

One matter which is still not resolved is the final home of the electronic edition. Who is to be responsible for providing and maintaining the server, its files and the website for the edition? For a major work of national importance we believe that only a reputable organisation with a long-term commitment to the field is appropriate, such as a professional association, a university research centre or library, or a national cultural institution. It is worth noting that the architecture and authentication scheme supports multiple sites, so it is not necessary for one organisation to have sole responsibility for the hosting and maintenance of the editions or even a single edition, since files may be kept at separate locations. We believe this is an important factor in the long-term continuance of the edition.

Markup Experiments

Originally we began trialling markup with the specialist SGML-aware program *Author/Editor.* In order to display output from this program on the web using an ordinary browser and SGML-aware plug-in, Softquad's *Panorama,* we found we had to do some tweaking because *Author/Editor* avoids

calling to an ENT file and only calls to a DTD file in order to parse the file, whereas a file for display with *Panorama* has to contain a call to external entities or else it cannot resolve the TEI Lite DTD. In addition, *Author/Editor* embeds tag sets in the file being marked up and thus conflicting structures cannot be marked up in the one file.

With the development of the JITM system, we moved to the creation of a toolset which could be used to provide the mark-up we needed. Using these tools we experimented with marking up the editorial annotations using the <xref> tag. Here we found a problem in trying to display our material locally or online with a browser and *Panorama.* Through reference to an online TEI Lite tutorial on <xref> tagging, our research assistant discovered and reported that the plug-in requires a declaration within the TEI Lite DOCTYPE declaration in order to know what tags are eligible for linking. Unfortunately the small amount of documentation provided with the plug-in makes no mention of this need. If <?TAGLINK xref "TEI-P3"> is added to the DOCTYPE declaration, then *Panorama* can hyperlink the <xref> tag without problem. In addition we trialled the tag creation and extraction tool by tagging proper names in a single chapter using the <rs> element and the attributes **type** and **key.** Some of the simple tags for structure and pagination are automatically generated by the JITM toolset when the transcription files are generated from files used as the basis of collation for the book edition.

Collation for that edition was done with a descendant of the set of tools, Computer-Aided Scholarly Editing (CASE), developed by Peter Shillingsburg. This was ported from mainframe to desktop computer and further developed by Phillip Berrie. We were aware of Peter Robinson's program, *Collate,* which not only prepared collations and a master list of variants, but also automatically generated TEI-SGML tags linking the reading text to the master list of variants and thence to the location of the variant reading in the witness files. We experimented with this program to see if we could use it for the automatic generation of tags, but, like *Author/Editor,* we found it unsuitable for our purposes. In order to generate tags, it had to collate afresh using specially numbered paragraphs. But collation and the identification of variants had already been done for the book edition, and did not need to be repeated. In addition *Collate* does not generate tags to allow users to go back to main text from collation and witness files. It depends on the browser back button for that.

Rather than try to reproduce in digital form the kind of collation results included in the book edition, we have tried to make better use of the possibilities of the medium and are developing, through Phill's expertise, a collation utility, which will be available from the website of the electronic edition. Using this tool, users can collate transcription files produced with the JITM system, regardless of whether the files reside at the electronic edition website or locally on the user's machine. Users can thus collate existing states as well as any new state of the edition and authenticate that new state against the master files on the website. A prototype of this utility is currently being trialled.

The development of this stand-alone application is indicative of an important change in the conception of the JITM system. Whereas it had been initially conceived as server-based, current thinking emphasises the virtue of client-based research tools, since they can be further developed in parallel by a number of workers, they can be used by researchers for their own work and they reduce maintenance overheads on the server-side software, which in this new paradigm is used primarily to generate perspectives for users. In addition, reducing the complexity of server-side tasks makes it easier to maintain the site and reduces the problems in migrating the data to new servers in the future.

Digitisation of facsimile page images

It has always been part of our project that users should be able to refer to digitised page images of the original versions of the novel, primarily to verify the accuracy of the transcription or collation, but in addition to see how each page looked to nineteenth-century readers. In the case of the work's first appearance in print, serialised in the pages of the 1870-72 issues of the *Australian Journal,* the digitised facsimiles are further useful in that they include the illustrations produced for the journal's readers but not subsequently reproduced in the book editions which followed. These illustrations will be linked to the electronic files of the variant states, so that users can go directly from the transcription of the page to the image of the page.

We have decided to produce our digitised images from black and white microfilms. In the case of the serial publications of the novel, microfilms of the relevant serials already existed, but we had to have microfilms specially prepared of the book editions. The State Library of New South Wales generously provided their copies of the relevant Australian and English editions and had them microfilmed for us. A microfilm copy of the rather rare first US edition has proved less easy to obtain. Using existing microfilms, we are reliant on the original photographers for the layout of the page on each frame of film. This is not usually a problem, but occasionally there are times where two pages are photographed on one frame when one page would have been preferable. While specialist microfilm producers can produce digitised images of the work being photographed, it has proved easier for us to do all the digitising at the one location. We have used a Canon Microfilm Reader/Scanner connected to an ordinary desktop PC running appropriate scanning software to acquire the digitised page images.

Each page image is first digitised as a binary bitmap at 300 dpi and is burnt on a CD in this form for preservation. Each bitmap image is large and needs to be compressed in order to bring down to a size which can be served over the Web and not take too long to download. We had chosen a standard compressed format, JPEG, for the images we would deliver to users. Compression however does result in the loss of some information, so it was resolved that we would prepare each page in two resolutions: a low resolution image for first use, with a higher resolution image available if further detail was necessary. After some experimentation, we settled on satisfactory image sizes for web delivery and usability, though it has proved difficult to reduce images of individual broadsheet pages from the *Australian Journal* to a small size without the type become illegible. With each page scanned first as a bitmap, it is possible to use an image manipulation program, like Adobe *Photoshop,* to batch process the conversion of the images to 8 bit greyscale and the saving of them in two resolutions as compressed JPEGs.

Other Work

Since the electronic edition is designed to provide more than the printed edition, we have had to prepare a number of extra materials in addition to the digitised page images. A major undertaking has been to have the text of the *Australian Journal* version of the novel prepared in electronic format. Since this first version of the novel is much longer than the later book versions, it had not been converted into electronic format at the time of the production of the printed Academy Edition, since it is impossible to usefully collate two so dramatically different versions of the same work. The text of the long version was converted into electronic format by a professional transcription bureau, using double-entry keying, then very carefully proofread to the same high levels of accuracy achieved for the other versions of the novel. Since collation is not sensible, an electronic version of the look-up table found in the Academy Edition has been prepared, so that readers can find equivalent chapters in the longer and shorter versions of the novel.

Conclusion

This project has been in existence since the early 1990s and in the course of that time we have learned a good deal about what we think an electronic edition should be and how best it can remain useful for a time commensurate with the effort which goes into making one. Our choice of a textually complex work has undoubtedly slowed our progress, but we have produced a more robust and flexible design as a result. Our desire to rely as much as possible on non-proprietorial software and delivery mechanisms has meant that the choice of software, for display purposes, for instance, has been less wide and less ideal than we might have liked, but we believe the work will be more useable over a longer period of time. We hope that our commitment to the highest standards will make this kind of electronic edition something which will have the same standing among researchers and students as scholarly editions on paper have long enjoyed. Time will tell if we have reached our objective.

Revised Academy Electronic Editions Architecture for *His Natural Life*

NB: The final version of an AEE will have more than one interface, e.g., for reading or for analytical work. This model represents the architecture as seen by the user who simply wants to read. Options to download tools or transcriptions might be part of the interface for analytical work