AUTHENTICATED ELECTRONIC EDITIONS PROJECT

Graham Barwell, Chris Tiffin, Phillip Berrie, and Paul Eggert

The Authenticated Electronic Editions Project is developing a system for the secure storage of scholarly texts while at the same time allowing for multiple, simultaneous markup to enhance and enrich the usefulness of those texts for the whole range of presentational and analytical activities embraced by Humanities computing. Once an edited text has been established, often after considerable effort and expense, it is vital that subsequent copying and use of the text does not negate the meticulous and laborious processes of checking and proofreading that have gone into establishing that text. Users in ten years' time should be as confident that the electronic file they examine is as accurate a rendition of the established text as a printed copy would be. The process adopted to achieve this outcome, involving stand-off markup and checksum authentication, has been developed by Phillip Berrie as the Just In Time Markup (JITM) system. In order to maintain textual stability, and yet still allow the texts to be useful in the broadest possible way, this JITM (pronounced "**jit**-em") system follows three key principles in its structure and procedures: openness, separation, and security.

The JITM system takes advantage of the conventions that have been agreed to thus far in electronic publication. Text characters are represented by the ASCII (ISO 646) character set with built-in support for Unicode (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). JITM has adhered wherever possible to the Guidelines for Electronic Scholarly Editions drawn up by the Scholarly Editions Committee of the Modern Language Association of America.[1] In order to produce editions that are as robust and long-lived as possible the system is designed to conform to international standards and to be platform- and software-independent. The works reside on a server and are accessed using the usual browsers developed for the Internet. 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.

The JITM system separates core text (in "transcription files") from its ancillary aspects, the latter being held in stand-off files which can be applied by a user in multiple combinations. The transcription files consist of two elements: the succession of characters, numerals, punctuation and word spaces in the original, and SGML tags which demarcate hierarchical divisions and text elements to allow for accurate mapping and insertion of the tags held in the stand-off files. Since the representation of this core text is done in ASCII, all characters which do not occur in that character set are coded by entity references. Examples of text features which are held in stand-off files include font changes (italics, bold, larger font, simulated handwriting), and pagination in specific editions. There is no practical limit to the range of analytical tags that can be added to demarcate specific strings for linguistic, literary or other investigation. Since the tags are not part of the transcription file,

---

[1] Committee onScholarly Editions. "Guidelines for Editors of Scholarly Editions." Modern Language Association of America. 31 July 2002.
<http://jefferson.village.virginia.edu/~jmu2m/cse/CSEguidelines.htm>

they can be disregarded by users who have no interest in that aspect (or extension) of the text.

The JITM system is designed to maximize access to the text without allowing it to be changed or corrupted in the process. Since access to and usefulness of the text are enabled by markup, the crucial task is to ensure any markup applied does not alter the original. This is achieved by inserting the tagsets into a copy of the file. An automated process subsequently removes them along with a record of their placement, and compares the resultant stripped file with the original. Exact correspondence of the two verifies that there has been no corruption. A parallel process is employed when the file is subsequently used either for browsing or for analysis in that the mapped tagsets are applied only to a copy of the file to produce the display-formatted or analysis-ready version of the file.

Background

The project is an extension of the Academy Editions of Australian Literature, a series of carefully edited printed critical editions of important early examples of Australian literary culture. The text of each work is freshly edited with an editor's introduction, textual and explanatory notes, and in some editions a range of further aids. The Australian Academy of the Humanities sponsors the project which involves three Australian universities, Wollongong, Queensland, and the Australian Defence Force Academy, where the project is supported by the Australian Scholarly Editions Centre. Technical expertise is provided by Phillip Berrie from the Information Technology Services Centre at the Defence Force Academy. The project is designed to test the system on a prototype edition in two stages: the first stage covers electronic production of the complete text with page images of each significant early version of the novel (not possible in a single printed work), together with a full treatment of variant readings, plus the essays, notes and other matter from the printed edition. The second stage will see the incorporation of a wider range of related material revealing the reception history of the work. Because the potential benefits are best demonstrated in a work with a complex textual history, Marcus Clarke's novel of convict life, *His Natural Life*, was chosen for the prototype edition.

It is convenient to describe the project in terms of architecture, the JITM System and the prototype edition, collation, digitisation, and other features.

Architecture

The architecture adopted is flexible and robust, suitable for the two-stage development of *His Natural Life* and extendable to other editions. 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 (see Appendix) gives primacy to the edited text appearing in the print edition, in that specific ancillary material is coded to the edited text, while more general material will 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 is called a *perspective*. The markup may represent basic presentational features of the original, such as page and line divisions. In addition

the user can call up digitised images of the chapter in each variant state page by page. If users wish to undertake their own analytical work, they can download files of the edited text and as many of the variant states as they require, together with appropriate JITM tools to produce markup specific to their needs.

The Just in Time Markup (JITM) System and the Prototype Edition

Under the JITM System the transcription and markup files are always kept separate, being combined only as required on the user's screen.[2] A set of tools and algorithms allows users to input the transcription files, generate markup tags, extract those tags into a separate file and then authenticate the transcription files to establish that they have not been changed in any way. For 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 collated this file against the copy-text of the Academy edition (the Melbourne 1874 edition) in order to proof it.[3] We have provided a generic basic tag set in which tags are used to define the blocks of text. These are more flexible and allow a finer level of granularity than that provided by the <p> tag. We do not envisage the project team providing all the tags researchers might find useful, since the JITM system allows scholars to produce tag sets tailored to their specific needs.

The JITM toolset is still under development. Currently in prototype are the *JITM Preprocessor*, which translates the files used in collating for the book edition into the 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 transcription file. The toolset will be made platform independent in stage 2 of the project, prior to distribution. In that final form the toolset will also allow researchers to use SGML editors of their choice. On the website for the prototype edition, the part played by the JITM system is not overt when users connect. They are merely asked to select the perspective of the edition they wish to have generated for them, and 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 unresolved matter is the final home of the electronic 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. 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. We believe this is an important factor in the long-term continuance of the edition.

Collation

Collation for the book edition of the prototype was done with a Macintosh descendant developed at ADFA of the set of tools, Computer-Aided Scholarly Editing (CASE). Neither this nor the program, *Collate*, was completely suitable for the electronic edition, so we have developed a collation utility. Using this tool, users

---

[2]The project website (http://idun.itsc.adfa.edu/ASEC) gives full details of the system so only a brief summary is given here.

[3] All variants listed by the computer collation were checked against the Academy Edition's editor's emendations, both those explicitly listed and the silent categories declared.

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 trialed.

The development of this stand-alone application is indicative of an important aspect of the JITM system. Rather than being entirely server-based, it relies on client-based research tools; since these can be further developed in parallel by a number of workers, they can be used by researchers for their own work and thus reduce maintenance overheads on the server-side software, which is used primarily to generate perspectives for users. 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

Users can refer to digitised page images of the original versions of the novel to verify the accuracy of the transcription or collation, or to see how each page looked in its original format. These page images are 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.

These digitised images are generated from black and white microfilms, in some cases specially produced for this project. 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 two pages have been photographed on one frame when one page would have been preferable. Each page image is first digitised as a binary bitmap at 300 dpi then converted to a standard compressed format, JPEG, for delivery to users. Each page is available in two resolutions: a low resolution image for general use, with a higher resolution image available if required.

Other Work

We have also prepared a number of extra materials for the electronic edition, including the text of the *Australian Journal* version of the novel, which is much longer than the later book versions. Line-by-line collation of this version against other versions of the novel is not sensible, but an electronic mapping at approximately paragraph level has been prepared, enabling readers to find equivalent sections in the longer and shorter versions of the novel. The *Australian Journal* version, including its woodcut illustrations, has also been digitized.

Conclusion

Two clauses from the MLA's "Guidelines for Editors of Scholarly Editions" ask:

> How important is permanence or fixity? How can these be attained? Alternatively, is there a possible benefit to openness and fluidity (for example, the certainty that new material will come to light)?

These are the conflicting aims that the JITM system has been developed to overcome, although the "new material" that JITM caters for is that provided by the researchers and users of the texts as well as newly discovered additions to the corpus. Electronic

editions of the last decade have ranged from semi-automated programmes like the *Making of America* series which digitises hundreds of thousands of pages of single-version texts through scanning and OCR technology, to the meticulous reproduction and indexing of image-and-text in the Blake Archive or the Rossetti Archive. Techniques like on-the-fly conversion from TEI-SGML to HTML pioneered by John Price Wilkin at the University of Michigan have become widely used by E-text depositories to accommodate richly encoded SGML text to the current generation of web browsers. These manifestations of "openness and fluidity" among the scholarly and library communities have already produced unparalleled access to manipulable texts, enabling new types of analysis and reproduction. Yet the niggling question remains, "How important is permanence and fixity?" not just of the arrangement and apparatus edition, but also of the text that is being presented. How can that text be faithfully reproduced, manipulated, supplemented by multi-layers of markup and still retain the integrity accorded by its initial editing? The JITM system offers a practical method of maximizing both the stability and integrity of the text and its accessibility and usefulness to present and future scholars.

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