12-2018

Romantic Poetry, Technical Breakthrough and the Changing Editorial Role

Paul Eggert
peggert@luc.edu

Desmond A. Schmidt
Queensland University of Technology, desmond.allan.schmidt@gmail.com

Recommended Citation

ROMANTIC POETRY, TECHNICAL BREAKTHROUGH AND THE CHANGING EDITORIAL ROLE

THE CHARLES HARPUR CRITICAL ARCHIVE

By Paul Eggert (Chicago) and Desmond Schmidt (Brisbane)

Abstract:

This report provides (1) a short history of the Charles Harpur Critical Archive (CHCA), which has been in preparation since 2009. Harpur was a predominantly newspaper poet in colonial New South Wales. Writing from the 1830s to the 1860s, he was unable to publish in book form because of the undeveloped state of the local literary publishing scene. Approximately 2700 versions of his 700 poems in newspaper and in manuscript form have been recovered. (2) A summary of the technical approach, a new one for special-purpose digital archives, is provided. The principal innovation is the use of a Multi-Version Document (MVD) file format. Because it is not dependent on conventional XML encoding, overlap is no longer a problem and automatic collation of versions, and of layers of revision in individual manuscripts, has become possible. Synchronous scrolling of facsimile image and transcription has also been achieved, lessening the need for detailed encoding of document elements and physical features. (3) The report then reflects on the theoretical implications of the concepts and methods used for the CHCA and on the changing role of the editor.

This is an account of the Charles Harpur Critical Archive. It provides a brief history of the project and its background, a survey of the novel technical approach taken, and finally a discussion of the theoretical and other implications of the underlying concepts and methods. A separate publication deals in more detail with the technical aspects of the project and its rationale.[1]

History of the Charles Harpur Critical Archive

The Charles Harpur Critical Archive (CHCA) has been in intensive preparation since 2013 although it actually began in 2009 at a time when the size and nature of the challenge that lay ahead were far from clear. Funding from the Australian Research Council paid, initially, for data input and research assistance. The development of the technical side only came later: two false starts followed by the one now settled upon. The

---

[1] Desmond Schmidt and Paul Eggert, 'The Charles Harpur Critical Archive: A History and Technical Report', forthcoming. Unavoidably there is overlap between the two reports but their differing emphases have warranted separate treatments.

development site is at <charles-harpur.org>.[2] Formal publication is planned for mid 2018, the 150th anniversary of Harpur's death. This archival expression of the project will lay the groundwork for more properly editorial and other interpretative endeavours, firstly by Paul Eggert as editor and as soon as practicable by collaborators.

Romanticists and others reading this report are unlikely even to have heard of Harpur, who was born in 1813. Yet we believe that assessment of his works, once they are available in a properly contextualised form, will lead many to form the view that he was a major talent who, for book-historical and other reasons, has been almost entirely lost sight of. Living and writing in colonial New South Wales from the early 1830s until his death in 1868 Harpur cut a considerable figure locally: his verse made at least 900 appearances in the colonial and intercolonial press. But publishing opportunities for him other than in newspapers were almost non-existent.

He maintained clippings files and retained some manuscript copies from the start, but he yearned for a magisterial volume of his collected verse to appear in London. He prepared the way for it, late in his life, by copying, sequencing and then, in successive manuscript books, revising and re-sequencing his poems. We have recovered approximately 2700 versions of his 700 poems.[3] In addition many have extensive authorial notes in which Harpur sought to key his newspaper poetry into the moving agendas of the day. The complexities of this evidence, together with the disarranged state of his manuscripts in the Mitchell Library in Sydney, defeated or compromised a series of editorial projects from the 1940s in Australia that aimed to capture the full range of his poetic achievement in book form.[4]

---

[2] Paul Eggert is project leader and editor of the CHCA (2009–). Desmond Schmidt is the technical lead and programmer (2013–), and Meredith Sherlock served as digital archivist 2009–2016. Other principal contributors include Elizabeth Webby (annotations), Chris Vening (biographical entries) and Michael Falk (qualitative assessments of poems). The Australian Research Council funded the project 2009–2014, and Loyola University Chicago has provided funding since 2015.

[3] The cut-off date for the project is 1900. The 2700 items include the newspaper and manuscript forms, and a heavily bowdlerised collection in book form arranged by his widow fifteen years after his death. Its texts were the source of future anthologies and most Harpur collections until the 1980s.

[4] See further Paul Eggert, 'Charles Harpur: The Editorial Nightmare', *Journal of the Association for the Study of Australian Literature*, 16.2 (2016), at <https://openjournals.library.sydney.edu.au/index.php/JASAL/article/view/11011>, accessed 2 November 2017.

The technical approach in summary

Problems of encoding and the Multi-Version Document format

Skilled inputters – MA students in the School of Cultural Records at Jadavpur University in Kolkata – were employed to transcribe the manuscript books using simplified XML-TEI codes.[5] Their job was to code, quickly, what they saw in front of them, in other words to transcribe documents as a series of scribal acts, to ignore problems of work-and-version differentiation, and to leave the encoding of problematic deletions and additions for later resolution. The resulting files were carefully checked, in part-time work over several years, by Meredith Sherlock along with other duties, notably the transcription of the newspaper cuttings. She also broke out the document transcriptions into work-and-version designations and maintained bibliographic control, basing her designations on an existing finding list but correcting and adapting it as new scrutiny of the evidence dictated.[6]

However, there was no satisfactory off-the-shelf solution available for the project as a whole. Various experiments were undertaken, first using an existing platform called Heurist, which had recently been adapted to deal with XML-TEI input, and subsequently using the newly developed archival, tools, storage and workflow system called AustESE.[7]

Then in 2013 Desmond Schmidt, who had been working along with AustESE to provide a collation tool, proposed a bold simplification. Schmidt's proposal was in part a response to the complexities of inline TEI markup, which, despite utopian claims made for it in its early days, makes interoperability achievable only within but not across projects.[8] It was also an acknowledgement of the fact that interest on the part of developers and tool-vendors in XML, the language in which TEI encoding is expressed, has been rapidly

---

[5] I.e., eXtensible Markup Language – Text Encoding Initiative: the current standard for encoding transcriptions of literary and other documents.

[6] Elizabeth Holt and Elizabeth Perkins, *The Poems of Charles Harpur in Manuscript in the Mitchell Library and in Publication in the Nineteenth Century: An Analytical Finding List* (Canberra: Australian Scholarly Editions Centre, 2002).

[7] Intended for archives and scholarly editions, only remnants of AustESE on the web remain. It was hacked in 2015, effectively destroyed, for there has been no funding available to reinstate it. It was a complicated system that addressed a wide variety of use-cases. It used Drupal as a content-management system (CMS) and could deal easily with XML-TEI, although it was in fact agnostic as to the language used. The CHCA now uses a CMS of its own.

[8] Discussed in Eggert, 'The Archival Impulse and the Editorial Impulse', forthcoming in *Variants*, 14 (2018). For Schmidt's papers on the CHCA tool set (which he calls Ecdosis and which will be released for more general use in 2018), and for the thinking behind it, see <http://ecdosis.net/papers>, accessed 1 November 2017.

diminishing, leaving the digital textual-studies research community, with its dependence on XML-TEI, potentially vulnerable.[9] We concluded that future support for domain-specific vocabularies such as TEI can no longer be assumed.

XML requires an ordered hierarchy of content objects, a tree structure. However, as every transcriber soon learns, literary texts are neither orderly nor hierarchical. Since there are usually multiple ways of encoding the same textual phenomenon, encoding is unavoidably subjective. Within any one project the element of subjectivity can be constrained with stringent editorial control but across projects it cannot. The result is that transcribers and tool developers will not necessarily share the same understanding of the deployment and therefore of the meaning of TEI codes. Elements frequently overlap, and workarounds have to be resorted to. Each workaround is an artificiality that makes things still harder for tool developers. Deletions and additions, for example, are normally encoded in TEI as if they are formats of a single linear text when they are actually content: variant readings that can only be read in parallel.[10]

The Harpur Archive has pioneered a simpler route that is less resistant to current technical trends but will nevertheless support what scholarly editing minimally requires. Instead of assembling an ever more complex TEI file as a single source to serve all future visualisations and analysis the CHCA has followed the devolved approach of text and external stand-off markup that Phill Berrie proposed in the late 1990s and developed for a few electronic editions in Canberra, Australia, in the early 2000s.[11] The introduction of the MultiVersion Document Format (MVD, described below) has been a radical step

---

[9] See Andy Patrizio, 'XML Is Toast: Long Live JSON', *CIO*, 9 June 2016: <http://www.cio.com/article/3082084/web-development/xml-is-toast-long-live-json.html> and his graph at <http://images.techhive.com/images/article/2016/06/stackoverflowqs-100665502-orig.png>: both accessed 1 November 2017.

[10] Available tools for textual collation such as CollateX encounter problems caused by the range of TEI codes employed by any one project to capture the subtleties proper to the particular author or set of documents. Such collation tools, if they are to be generalised, need to be able to anticipate every possible encoding of deletion and addition. Given the very many different ways and different positions by and in which the inscribing hand can make marks of alteration on a manuscript page the extreme difficulty of writing such a tool is all too obvious. For instance, an analysis of the manuscript books transcribed for the CHCA showed that 14 different ways of encoding first layer (i.e. initial) deletions and additions had been used, all of them valid TEI; 31 ways of encoding layer 2; and 37 of layer 3. Thereafter the numbers decline because of Harpur's habits of rarely going beyond level 4 (20 ways). Thereafter the figure is 6, 1, 1 and so on until the eleventh level (which is the most he ever employed), also 1.

[11] See, e.g., The Jerilderie Letter edition at http://asecentre.org/JITM/index.html, accessed 2 November 2017. Click on *Change Settings* then on *Create Perspective*. Just In Time Markup is applied as the perspective is created. Various options are available. The project ceased in 2004 before it had had the benefit of a professional page design; it looks antiquated now.

forward. Berrie's XML markup, which assumed a hierarchical document structure, is now reconceptualised as Schmidt's stand-off properties, which do not. But the goal of simplification is continuous from Berrie's work.

The CHCA transcriptions were already encoded in XML–TEI. Schmidt developed a Splitter tool to process them in batches. It divided the texts of manuscripts into their layers of revision. Then each layer was split into plain text and its properties. These two sets of data for layers and versions are then merged to create two MVDs – one for plain text and one for markup – for the poem as a whole: that is, for the work.[12] The MVD storing the markup as a series of range attributes points to the plain-text MVD. Standoff properties do not require TEI anchors to be recorded in the linear text.

The overlapping hierarchies problem that has dogged XML from the start[13] is eliminated by this combination of the MVD format and stand-off markup. Together they permit the free overlap of textual variants and of their properties. The use of the layering technique for recording deletions and additions (there are around 10000 such revision sites on Harpur's manuscripts) takes care of the other principal cause of overlap.

The present account makes the progress of the CHCA sound smoother than it actually was. We knew especially that the sequence of alterations at any one point would not have been accurately captured in TEI in a form suitable for generating the MVDs automatically. This is because the XML-TEI codes that we had originally used are typically deployed to capture the appearance on the page of revisions rather than their logical sequence. Schmidt had found ways of mitigating the problem by searching for XML-TEI patterns of recording deletions and additions. But the results needed a close check since the layer numbers are crucial to the MVD's functioning. As part of this checking process page images were prepared to permit synchronised scrolling alongside the transcriptions using a tool, developed in 2016, called Twinview. The check consumed most of 2017.

The experience confirmed us in the belief that if we were to start again from scratch we would not use TEI at all. Instead we would use something like the Markdown Editor recently devised for the project. Equipped with synchronous scrolling, this Editor was

---

[12] The MVD is a file format that is best understood as a variant graph that merges the separate transcriptions of each version of a work into the one file. The text held in common among the versions is stored once only. Where readings deviate one version has its reading recorded on the top loop and the other's on the bottom loop. Each loop re-enters the main text when the variance of its version ends.

[13] See Elena Pierazzo on the long-recognised problem of overlapping hierarchies for TEI encoding: 'the reality is that the problem has no real solution': *Digital Scholarly Editing: Theories, Models and Methods* (Farnham, Surrey: Ashgate, 2015), p. 120.

used during 2016–2017 to transcribe the complete extant correspondence (about 180 letters) of Harpur and his circle to 1900.

A minimal markup language (MML) is used in the Editor. Surprisingly few codes are needed when XML requirements are ignored and one works from the assumption that a facsimile image of the transcribed page will be visible to the user alongside the transcription. MML codes are used only for editing. For viewing, the text is converted into HTML, which is the standard language of the Web.[14] For storage it is split into plain text and standoff properties, as already described.

Once the transcription files for any one work have been processed to create the MVDs, various visualisations become possible. One is the Compare tool, which places the texts of any two versions or layers side by side with their variant readings highlighted. Another is the Table of variant readings. In this tabular format each version is stacked one above the other. The text of each transcription scrolls across the window under the user's control. A new text-segment is imposed as each variant appears. The position of each documentary text in the stack can be altered as the editor tests out possible transmission scenarios. The editor's nominated base text can be changed at will. The Table thus provides much of the requisite evidence for establishing the line or lines of textual transmission and for editorial decision-making about emendations.

This collation display is still in development. If fully successful, the Table should render preparation of the traditional textual apparatus unnecessary since every variant is given in context, more informatively and intuitively for the user than in any lemmatised apparatus. Finally, a specially written Timeline tool brings together, as an alternative point of access, biographical events, letters, Harpur's successive addresses, along with composition and publication events.[15]

The editorial role

First-time readers have to start somewhere. They need help, and even experienced readers need something considered to bounce off. For the CHCA the editorial aim (as opposed to the archival one) involves the establishment of reading texts of the successive versions of any one work. This is not the same thing as the archival capturing of the texts of the relevant documents. A version is a text that is postulated from the

---

[14] See further, Desmond Schmidt, 'Formatting Poetry for the Web', <http://digitalvariants.blogspot.com.au/2016/10/displaying-poetry-on-web.html>, accessed 2 November 2017.

[15] Timeline JS was used at first. It was replaced by one of Schmidt's devising because of its restriction on file sizes, which did not meet the CHCA's growing requirements, and because the tool is no longer supported by its creators.

documentary evidence and then editorially established. The same manuscript page may bear evidence of two or more versions, distinct in time;[16] or that page may not be evidence of a new version at all if it is an identical or nearly identical copy of one whose text is already witnessed. An editor intervenes between the documentary evidence and a readership. Editions are reader-oriented, whereas the archivist's transcription is necessarily document-facing.[17] Organising the workflow and the storage to reflect this reality, rather than loading up a single TEI file with the range of archival and editorial interpretations has made practical sense.[18]

Eggert's properly editorial role on the CHCA begins when the archival evidence is published. It will be a matter of deciding when to call a moment of pause in the development of the text of any one work and then to capture that text, emended as necessary, as a version. The decision will rest in part on whether it offers a good place from which to consider the evidence of textual development in the other relevant documents, looking backwards and forwards from the chosen point, both chronologically and textually. This will not be the German editorial definition of version considered as a semiotic system whereby a single variant creates a new one.[19] Rather the editorial decision will acknowledge the requirements of an audience needing an editor to have made sense of a complex set of evidence while not losing touch with the archival documents that the edition rests on.

---

[16] E.g., 'Trust in God' appears on the inside front cover of A87-1 (Mitchell Library, State Library of New South Wales). Its base layer is a second version of what appears on p. 109 of the same manuscript book. That base layer appeared in *Empire* on 20 June 1853; the second layer with some further development (perhaps via a lost subsequent draft) was printed in *Australian Home Companion* on 5 November 1859. For a theoretical defence of versional editing, see Eggert, 'Versional Editing of a Romantic Poet', *Tipofilologia*, 10 (2017), 11–24.

[17] For the relationship of the archival to the editorial see Eggert, 'Archival Impulse' (cited above) and 'The Reader-Oriented Scholarly Edition', *Digital Scholarship in the Humanities*, 31.4 (2016), 797–810.

[18] The reading text of a version, as subsequently established by the human editor, will be processed by the MVD as merely one more layer. If the reading text is to represent the whole work it will be treated as one more version.

[19] See Hans Zeller, 'Record and Interpretation: Analysis and Documentation as Goal and Method of Editing', in *Contemporary German Editorial Theory*, ed. Hans Walter Gabler, George Bornstein and Gillian Borland Pierce (Ann Arbor: University of Michigan Press, 1995), pp. 17–58; originally in German in *Texte und Varianten: Probleme ihrer Edition und Interpretation*, ed. Gunter Martens and Hans Zeller (Munich: C. H. Beck, 1971).

Expanding the commentary function

With the archival effort already discharged as described above, the commentary function may become a more far-reaching expression of the editorial role than it typically is in print. The editor will want, minimally for any one work, to explicate the relationships of the documents as well as their texts to one another and to characterise the kinds of revision and other change witnessed within each of the documents.[20]

The commentary, in the case of Harpur, will extend outwards to embrace comparative evidence found in other of his works, in his biography and in publishing contexts if they seem to have had determining effects. Finding an interpretative voice that is more free-wheeling and adventurous than what scholarly editors have previously felt appropriate in print form will be desirable. With the archival expression of the project joined at the hip to the editorial expression, interpretative claims will be able to be linked more precisely and historically to the texts, including their revisions, than was feasible in print form. The reader will be able to inspect the evidence for the editor's arguments directly. Interpreting the motivation for, and aesthetic effects of, the patterns of textual variation will be an attractive undertaking for editors and commentators in this more reader-oriented editorial environment.

Ushering the historical Harpur into our contemporary literary-critical space as a figure well worth encountering or revisiting in new textual forms and contexts will involve the taking of some risks as the editorial function expands. Although such commentary will date, it will be readily replaceable, without disturbing the archive, by future scholars as they usher Harpur's poetry into their new world. This is ultimately why the Harpur project is called a 'critical' archive. Text-critical certainly, but also further afield than that. Some mistakes will doubtless be made as the editorial discourse adopts new accents and takes new directions. Yet experimentation is necessary if the most is to be made of the new, unaccustomed logic of the digital medium.

When the archival expression of the project is formally published in mid 2018 it may involve adopting a university press as publisher. Although the website will be free and open access, spin-off print publications may later eventuate, for which a publisher will be necessary. But there are other, better reasons to involve a publisher: promotion of the project, peer-reviewing, imprimatur and ongoing responsibility for the critical archive into the future when the present collaborators are no longer involved.

---

[20] Technically, commentary on the text of whatever nature – textual, historical or literary critical – will point to the ensemble of versions and layers in the work's plain-text MVD. Consequently, such annotations do not need a TEI element to act as a target; and if an annotation is applied to text shared by several versions or layers then it automatically applies to all of them. The writing of the annotation tool for the CHCA is still in its early stages but this outcome is a consequence of the MVD file format

The intention is to publish the editorial expression of the project in tranches as editions and commentary are prepared. It is hoped that steady progress will attract new collaborators to contribute commentary of one kind or another to the project, with each receiving formal recognition for so doing and with a promise of preservation of their contributions.

Paul Eggert

Loyola University Chicago

pauleggert7@gmail.com

Desmond Schmidt

Queensland University of Technology

desmond.allan.schmidt@gmail.com

3764 words including notes