ncse and the Production of Victorian Text in the Digital Age

Critical editions of nineteenth-century periodicals and newspapers are rare despite the well-made case for their importance in nineteenth-century culture. In this paper, I want to explore why this is. Is it because the expense and difficulties of working with paper limit the reproduction of these often lengthy publications? Or are there more subtle discursive barriers that prevent journalism being edited on its own terms? One of the aims of the Nineteenth-Century Serials Edition (ncse for short) is to investigate whether it is possible to publish a digital edition of this material. Freed from the material constraints of paper, we have attempted to edit six nineteenth-century serials as a cluster that emphasizes their properties as serial texts. However, not only have we struggled to define these properties, but we quickly encountered a host of practical difficulties applying rigorous scholarly care to a publication of 100,000 pages.

Our current model of the critical edition is resolutely book-like: an editor, usually professionally qualified, selects a body of material – which may or may not already be a book – from the archive, edits it as a whole with some metatextual material, and then publishes it as a discrete unit. The editor here functions as a member of the cultural elite, making selections from often inaccessible archives in order to create a new textual object, privileged over any number of others that remain on dusty shelves. Although much editorial work is devoted to explaining contemporary references within the work, seeking to make its context understandable to readers today, this runs counter to the actual practice of editing, which involves abstracting the text from the historical objects that bear the marks of both its cultural life, whether this be production, circulation, or survival. Editing, necessarily, is a top-down form of cultural production that recognizes some historical objects as more valuable than others, and so worthy of being published as new books.

Nineteeenth-century journalism resists scholarly editing largely because it resists being turned into a book. Periodicals and newspapers are too long, too ephemeral, have too many authors, have too many textual manifestations, cover too many subjects, are too variable, and are too reliant on visual information. None of these pose a problem when working with digital media, so it ought to be possible to republish editions of periodicals, for the first time, without committing the violent textual acts that turn them into books. However, a large degree of editorial control is still required to handle this often complex material and, although size is not a problem for republication or functionality, it can prevent close scholarly attention across the project. As such, editors of digital editions, especially those of editions of journalism, tend to operate more as archivists, organizing the material and making it available to scholars with any editorial interventions well-hidden. This morning I want to present some of the solutions devised by ncse to allow us to produce a scholarly edition of nineteenth-century serials in digital form. These solutions, I suggest, allow us to retain a more traditional editorial role even while embracing some of the more 'open' aspects of digital publication.

* * *

The Nineteenth-Century Serials Edition is an AHRC-funded collaboration between Birkbeck College, King's College London, the British Library and Olive Software.

In May 2008 we will publish free and online an edition of six nineteenth-century journals: the *Monthly Repository*, the *Northern Star*, the *Leader*, the *English Woman's Journal*, *Tomahawk*, and the *Publishers Circular*. Although each is important in its own right, we believe that by publishing them as a cluster we foreground the interconnected nature of nineteenth-century print culture while also exploring its remarkable formal diversity. As our primary interest is in the journals themselves, we consider ncse to be not only a useful resource but also an exploration into the creation of scholarly digital editions of serial texts.

As projects such as the Internet Library of Early Journals, the Times Digital Archive, the British Library's Penny Illustrated Paper, or even JSTOR demonstrate, it is possible to publish editions of periodicals that avoid the textual violence necessary if they were to be republished in book-like form in paper. However, this seems to come at the price of losing the editorial rigour and metatextual apparatuses that constitute a critical edition: indeed, due to the increasing recognition of the cultural importance of nineteenth-century serials, their amenable copyright status and their demonstrable commercial potential (I'm thinking of *The Times* here), there are large projects underway from Thomson Gale, Proquest and the British Library that follow in this archival, database model. Simply by making huge swathes of nineteenth-century print available online and searchable these projects are likely to have a far-reaching and lasting impact on nineteenth-century studies. However, all three adopt a relatively unsophisticated editorial stance towards their material, preferring to elide their properties as serial texts and instead present them as archives of discrete articles. This has the result of making the contents visible and accessible – at least to those who have access to libraries that subscribe to them –

but at the cost of ignoring the journals themselves: rather than produce critical editions of nineteenth-century periodicals, they instead provide archives of their contents.

This is important as a model that separates articles from their contexts severely misrepresents the seriality of journals and newspapers. Although serials were sold as individual numbers, what survives on the library shelf is usually a volume bound up of these numbers, perhaps some front matter and an index. Of course, serials were not just sold and distributed as individual numbers, and many library holdings were actually bought as volumes from publishers, but what the form of the volume tends to impose is an over-determined linearity. The seriality of the periodical permitted publishers and editors to employ a range of publishing strategies to respond to the contingencies of the market. This means that not only do individual titles change in appearance over their runs but, often despite the well-meaning attention of overzealous archivists, the archive still contains odd supplements, multiple editions, odd numbers, inserts, and other hard-to-place matter. For instance, of our titles in ncse, the *Leader* published two editions and the *Northern Star* perhaps up to 9. Although much of this material was excised when it was forced into a series of book-like volumes by whoever bound it, or separated out by the archival strategies of whatever institution housed it, the traces that remain, remind us that the bound volume is simply one stage of a much broader history. A model of digital publication that mimics the form of the bound volume at least potentially recognizes this; one that separates bound volumes not into numbers but individual articles elides this history entirely.

The need to attract readers, whether as purchasers or subscribers, and then reattract them with each number, means that the identity of periodicals is not so much what they have to say as how they say it. The reliance on OCR technology, in which a textual transcript is generated from the page image which can function as a searchable index, privileges abstract text over its visual components. As this front page of the *Northern Star* shows, layout, typography, and images all play significant parts in ensuring what is written is more than a property of the words that are used.

This is further complicated when we recall that, as serial texts, periodicals embody changes within their form as well as their contents. In other words, these informational categories, although conceived to organize and signpost changing content, are also subject to change and modification themselves. For instance, as this slide shows the form of a journal can change radically over its life:

[**slide: Northern Star becoming Leader**]

The editor of the *Northern Star*, George Julian Harney, reimagined what the Northern Star was when he altered its format.   In the last number before its reduction in size he wrote that, and I quote, ‘as it is designed to make the paper of more than passing interest, its more compact form will with many be an additional inducement to preserve each consecutive number for binding in half-yearly volumes.’ What Harney is doing here is severing the title’s link with the news, making it into a periodical and not a newspaper. If the articles within the *Northern Star* in its later years are presented in a way that abstracts them from their formal context, then such a radical transformation in the genre of the title is lost. It is only by being aware of the wider meanings that these formal structures carry that we can understand what it means when they change. As such meanings are rarely made explicit within the text, it becomes doubly important that we retain the formal aspects of printed objects when we create digital resources from them.

The fact that pages of periodicals carry structured information should alert us to the presence of wider organizational structures in this material. For instance, many periodicals employ a system of subdivisions, dividing their contents into separate departments. An editorial policy that separates each number of a periodical into a series of items – something quite common in digital editions of more journal-type publications – will not register the difference between, for instance, a letter published in a correspondence column and a letter like O'Connor's that has a very different function. While these structures do exist – and they are present in some form in all of our titles – it is important to recognize that they are signalled through visual means: we can recognize these hierarchies by the way in which they present the headings, not through what the text says that they contain. Any attempt to capture the structure of periodicals, then, requires the editorial eye of a human operator.

The necessity for editorial intervention at the structural level causes problems with open-ended texts such as serials. Even in ncse, a relatively small project of only six periodicals, we are dealing with just under 100,000 pages. When we remember that each page is part of a number that is itself part of a volume, and contains departments and items, the number of units at each level, not to mention their hierarchical relationship, introduces considerable complexity. Yet if digital editions are predicated upon a similarity to their source objects, then these structural categories ought to be reproduced along with the text and page images. At ncse we have worked closely with the Centre for Computing in the Humanities at King's College London and Olive Software to develop tools that allow us to automate as much of this editorial work as possible. This work has been focused in two main areas: recognition of different textual units within the cluster; and marking up of contents with metadata.

For the first of these, the recognition of different textual units, we worked with Olive software to see if we could program software to identify visual clues on the page and then allocate textual units the appropriate place in the hierarchy. [**slide of *Leader* in old pilot]** This slide shows the most successful of these: as you can see, it has recognized four levels within this number, which is itself located within a volume and a specific title. First we have "Portfolio," which is the name of the department; then *The Apprenticeship of Life* (the title of a serial novel within the department), the next level, "First Episode", tells us which instalment of the novel we are reading; and then lastly comes the chapter itself "Chapter One. The Young Sceptic." This hierarchy is captured in this table of components on the left, and 'First Episode' is highlighted on the page on the right.

Although these results were impressive, they took a great deal of work from everyone for quite limited functionality. For instance, as this process works from the bottom-up – i.e. it segments what is on the page rather than working from abstractions – it is difficult to map the levels identified within the system across all six titles at all periods in their runs. Also, the level of detail here was felt to be of limited use – do we really need the "Episode" level displayed in this hierarchy? Instead, we refined the system so that it worked to much simpler model: **[slide of**

ncse hierarchy] As you can see, we have only allocated two levels within each number, departments and items, on the basis that departments are relatively easy to identify through visual clues and stay relatively consistent across runs, whereas different levels of item are much more difficult to map consistently. With some hand correction, we have managed to segment all six titles to item level, with items being located within the appropriate departments.

We have also been working with Olive on a suitable application to display the items. This slide [**slide of *Leader* with 4 levels**] shows an Olive application called File Cabinet that is a publishing tool designed mainly for electronic books. Although this is great at organizing tiered information in an archive – it locates individual books in a collection, and then allows you to drill down to chapters and then pages – it is not very good at dealing with newspapers, which often have lots of different articles on a page. Consequently, we have also worked with another application called Active Paper Archive [**slide Active Paper Archive**]: this is designed specifically for historical newspapers and is the application that drives the British Library's *Penny Illustrated Newspaper*. As such, this application is very good at displaying newspaper articles, but does not represent information in a structural hierarchy like File Cabinet. Some of our titles are more book-like than others: the *Northern Star* represents itself as a weekly newspaper with lots of often miscellaneous articles on large pages; the *English Woman's Journal,* however, consists of long articles that resemble chapters and are printed on smaller pages in numbers designed to be bound in volumes. Although periodicals are not simply mixtures of books and newspapers, but rather a genre that contains features that gesture to both, the needs of our material have driven Olive to develop a new application, Viewpoint, that can handle a wider range of print forms.

This is what the segmentation process looks like through Viewpoint [**slide Viewpoint**]. As you can see, the application is working on the structured information defined by the segmentation process. In this way users can search directly for articles, or they can browse through the edition: in either case, items are always located within their immediate textual contexts, whether this is at across editions, or at page, department, number, volume, or title level.

The segmentation displays the formal of the journal, the other area in which we have been working concerns the contents. At over 100,000 pages worth of articles, there is too much of ncse to read. However, we do have machine-readable transcripts of the entire corpus and, using an open-source computational linguistics program, have been able to extract certain categories of information [**slides**]. This work is quite experimental and we expect these techniques to work better on some types of information better than others. However, we're confident that the end result will be a browsable index of the contents of ncse, and the automated population of some of our metadata fields.

* * *

Part of the challenge of editing historical journalism is the difficulty in abstracting copytext from the material objects in which it is to be found. As such, we have worked to develop various tools that allow us to reproduce both the form and the content of ncse in our digital publication. It is in the production and implementation of these tools that much editorial work occurs, whether this is in the design of the profiles that allow the system to identify items according to textual marks, the applications that present items to users, or the algorithms and authority lists that allow us to identify names and ascribe them to historical actors. In addition to this, of course, is the editorial policy that informs the cluster as a whole. ncse is very much an edition, not an archive, and even though it has archival features, these are carefully designed to make arguments about the contents. By foregrounding the contextual nature of serials, whether in terms of form or content, within the edition, we are making the size of the archive work towards our editorial ends: rather than abstract content from the archive in order to turn it into a new, discrete object; we are keeping our objects in a contextual environment. Of course, there is still an element of selection involved – why choose the six journals we have? – but by choosing six titles, rather than one, adopting common standards and protocols, and affiliating with hubs such as Jerome McGann's NINES environment, ncse can play a part in a much wider digital landscape.

Advocates of digital culture often use a rhetoric of democracy, whether this is in terms of broadening access or the user-generated content that characterizes web 2.0 applications such as Flickr, Wikipedia, or Facebook. It is important, however, to acknowledge the editorial work that goes into the creation of the large digital archives that are beginning to emerge. Although they represent themselves as portals that allow users to exhaustively search their large corpuses, providing ready access to the texts themselves, there is a significant editorial role in selecting material, organizing it within the archive, and encoding it for functionality. This need for a digital literacy on the part of users, of course, is the often overlooked aspect of digital culture. People's expertise in using these tools varies, and many of them are not available to all free and online. At ncse we acknowledge this as best we can, making the edition available, keeping the interface simple, and reflecting upon and making explicit our editorial interventions.

The dominant model of the critical edition, even in the digital domain, remains that of the book: i.e a standalone resource modelled on a self-contained cultural object. Funding schemes are not geared to open-ended projects, career progression demands well-defined publications, and the need for definite pre-defined deliverables also impacts on the creation of novel tools and solutions. The archives produced by the private sector, although extremely valuable, do not ask serious questions of the material they digitize and work to conceal the editorial hand that has created them. It is, I suggest, the role of academics to apply their editorial skills more widely: just as the digital medium allows us to engage more critically with non-book print forms – and indeed non-print forms in general – so we need editors willing to design resources that respond to the demands of their source material. As we have repeatedly learnt with ncse, it is when you transform one cultural artefact into another that you fully realize its complexity.