From Life on the Shelves to Digital Shelf-Life: Representing Journalism as an Historical Artefact in the Digital Domain

1. Introduction: From Life on the Shelves to Digital Shelf-Life

It is easy, when looking and using a digital edition, to forget the various stages through which material passes before it can be served up on screen. There is a kind of amnesia that elides the necessary transformations and editorial interventions that produce digital versions of an historical artefacts, and instead posits the new edition as a digital surrogate with a direct relationship to a source of some kind. However, as Laurel as argued, such easy translations between media rarely occur; rather, such correspondences between digital edition and source are often reconstructed, are effects produced rather than representative of the actual processes involved.

The question, of course, is whether this matters. Is it necessary to signal the passage of material, to acknowledge its complexities, when often the purpose of digitization is to overcome these difficulties? If a project is designed to bring diverse sources together (or take users to them), should the geographical or institutional distribution of these objects be recognized at the digital level? Equally, is it important to gesture towards the editorial tidying that goes on, whether by empowering users to shape their own editions, or by providing open resources where the editor's version can be one of many? This morning we want to present our answers to some the questions that Laurel has raised, and in doing so explore further the decisions that have to be made in presenting an object from the past in a different form in the present.

2. ncse and the Politics of Periodical Form

Journalism, as Laurel, Mark and John have already stressed, presents a distinct set of generic challenges for digitization. There are three main aspects to this: the size of the periodical archive, the complexity of the information that it contains, and the state of its material remains. Each of these presents a distinct set of editorial problems, and the structure of ncse is, in some ways, an attempt to reconcile exclusive positions. Whereas the problem of scale demands an archive-type model to organize the data and provide ready access to it, the complexity of this data, and the various different forms in which it appears, requires close editorial attention, which is linked to the model of the edition. Whereas **ncse**, to an extent, will function as an archive, directing users to articles in which they may be interested, it is much more than this: **ncse** offers a model of textual scholarship that enables the republication of periodicals and journalism in digital critical editions.

Scale

Laurel has already mentioned the large number of periodical titles published in the period, and the corresponding bibliographical headaches that arise in making selections from it. The current crop of large-scale digitization projects from Proquest, Thomson-Gale, and the British Library attempt to tackle this problem by incorporating as many titles as possible: indeed, as this material is out of copyright, there will inevitably be overlaps between the projects so that competition between them is predicated upon the breadth of their contents, rather than individual titles.

Once selections are made, there are additional problems caused by the size of the runs that each title represents. Serial texts do not usually have prescribed end points (and, indeed, some nineteenth-century titles are still being published today) so each title potentially represents hundreds of thousands of pages. Of course, the nineteenth-century press was also correspondingly competitive, and so only a small proportion of these titles survived beyond a few volumes. This is reflected in ncse: as Laurel mentioned, the six titles in ncse contain 98,565 pages in total, but over half of these are either in the *Leader*, a weekly which ran for 10 years, and the *Monthly Repository*, a monthly which ran for 32.

The republication of periodical texts is near impossible in paper. For instance, even a new edition of *Tomahawk*, a satirical weekly alternative to *Punch* that only ran for just over two years, would still require some 3000 pages. To further complicate things, the existing runs of periodical often do not fall into neat sequences of numbers. As Laurel has explained, 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 are individual runs quite diverse in appearance, but the archive is full of supplements, multiple editions, odd numbers, inserts, and other hard-to-place matter. Much of this material was excised when it was forced into a series of book-like volumes by whoever bound it. However, as the **ncse** material has constantly reminded us, this linearity is resisted by the textual remains of the journals: as we'll go on to show, the persistence of multiple editions and supplements in the 'wrong' places in the bound hard copy remind us of the need for a more flexible delivery system than a sequence of pages.

Depth / Breadth

This leads to the second of the difficulties in digitizing journalism: the range and depth of the information contained within it. 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 **[slide: front page *NS*, with Fergus in the corner]**.

The fact that pages of periodicals carry structured information should alert us to the presence of wider organizational structures in this material. On a simple level, where an item appears on a page, or as with the case of O'Connor's letter, which page it appears upon, affects its meaning. In addition, periodicals could employ a well-recognized system of major and sub-headings to group items in 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 the text that they contain. Any attempt to capture periodical form then, requires the editorial eye of a human operator.

There is a need to recognize the ways in which publications structure their contents, and this applies to all levels of the text. Just as where a number appears in a run is important – perhaps in terms of its relation to wider historical events or its proximity to Christmas – so too is the position of the individual item on the page and within the number. In addition to their historical importance, these structural categories should also provide the basic structural units for digital editions of periodicals. If digital editions necessitate similarity to their source objects, then these structural categories ought to be reproduced along with text and page images. However, the increase in complexity that such tiered information represents – you are not only dealing with words and pictures, but these are gathered in items, in departments, in numbers, in volumes etc… – is an unwelcome addition to the amount of content that needs to be handled and structured. This is further complicated when we recall that, as serial texts, periodicals embody movement within their form as well as their contents. These informational categories, although conceived to organize and signpost changing content, are also subject to change and modification. For instance, as this slide shows the form of a journal can change radically over its life. **[slide: Northern Star becoming Leader: describe history]**   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. Although they do provide structures that can be abstracted and used to organize digital editions, it is important that we incorporate them in ways that provide for their variation.

Form

The malleability of the periodical also complicates the material form of its remains. As we just mentioned, the bound volumes of periodicals often contain components that gesture to previous material incarnations. The marginalization of journalism within broader literary culture despite its centrality to print culture more generally means that often its historical descent is complex. As Laurel suggested, the material that we digitize is not single numbers as they were issued from the press, but rather a motley collection of bound volumes, microfilms, odd issues, and occasional supplements. What we digitize then, bears the traces of its various material forms. There are single numbers, organized in a sequence that usually corresponds to time, but these appear often with portions of their text moved to other parts of the volume, or excised entirely. In addition there are supplements, lacunae when things are absent, and repetitions introduced, perhaps, by the issuing of multiple editions.

The bound volume has resided on a library shelf for over 100 years, introducing elements of wear and tear as well as subjecting it to institutions' various conservation policies. The ncse titles are predominantly from the British Library, but due to various historical accidents – whether this be bomb damage from the second world war or misplaced volumes – some portions (even the occasional run) have been sourced from elsewhere. This means that the fiction that the history of these objects somehow stopped when they were published in the nineteenth century is impossible to support. Even if one wanted to offer the runs of periodicals in a way that disguised their later history, this is indelibly inscribed into their form.

The same is true of their subsequent history. Aside from digitization, the only other way to reproduce runs of periodicals is on microfilm. Because of the nature of the institutional relationships that underlie ncse, our source – as Laurel mentioned – is not the hard copy itself, but microfilm produced from it. Although there is little historical interest in this intermediary stage, it is nonetheless present in the digital resource. **[slide of Tomahawk muck rake and variants]** For instance, *Tomahawk* contains large cartoons which are printed on ink washes. Because microfilm is in black and white, we lose the colour of the ink and, often, the tones that are necessary for the images's dramatic effect.

What the persistence and reproduction of these material traces across different material forms reminds us is that the idea of a unitary source for journalism is largely illusory. Microfilms of bound volumes are haunted by prior material forms, whether these are the bound volumes of paper from which they are filmed, the individual numbers that constitute them, or the gaps that often signalled excised material such as advertisements. Editing journalism then cannot be predicated on the idea of the original, as this simply did not exist. Equally, because – like ncse itself – periodicals and newspapers are the products of many people, and often these people change over the course of a run, it is difficult to draw upon any of these to provide editorial principles. Laurel suggested that such historically contingent editorial questions be considered alongside that of the intended user of the digital resource. However, just as editorial principles derived from hypothetical ideal texts can be problematic, so too can those derived from an equally hypothetical user.

Bibliographical Control and Editorial Choices

These three aspects of journalism demand editors provide bibliographic control at a number of structural levels, while accommodating the needs of intended users. Of course, the same is true of critical editions in paper; however, as we have argued, paper provides a limited model for the republication of periodicals and newspapers. The predominant electronic model for this material is that of the archive: a relatively open collection of discrete units housed within a database structure and accessed though an easy to use front end. This logic informs well-known projects such as JSTOR and Project Muse, which conceive of themselves as providing access to articles, rather than the journals of which they are a part. It is also the logic behind useful projects such as *The Times Digital Archive* which, although it permits browsing by issue and enables users to keep individual articles in the context of their page, presents itself in the first instance as a portal through which to access an undifferentiated archive of text. Although the allocation of classifiers such as 'News' or 'Classified Advertisements' demonstrates a concern for the type of content an article is, the imposition of 21st century labels elides the historicity of both its contents and the way in which they are arranged.

The archive model tends to privilege content over form, subsuming structural differences in order to offer the appearance of unmediated access to information. However, this ignores the fact that not only must archives adopt editorial principles through which to organize their contents but, because they republish documents every time they are accessed, these principles also apply when presenting contents to users. As so much of what makes periodicals and newspapers periodicals and newspapers is in the formal features that differentiate them from other print forms, an over-reliance on OCR transcripts that abstract text and metadata categories adopted from different genres risks misrepresenting the contents of archives, even while simultaneously making them much more accessible. In concentrating on the archival properties of digital editions, much of what makes content interesting can be lost.

The question of accessibility is important: of course editorial decisions must take into account those who are going to use the edition but, because digital editions have a much larger potential audience than paper editions, there is a tendency to expect digital editions to take into account this audience, even if it is not primarily intended for them. The expectation that a digital edition address a very broad audience can mitigate against close scholarly care of its contents, making interest in formal features like, for instance, mastheads, seem specialist and antiquarian. If editors privilege this broader audience, then the archive model predominates.

Paradoxically, this is the model that is least suited to the financial resources of academic projects, at least in the UK. The use of public money usually necessitates free access to the public, even if the resource itself is of interest only to a minority of them. However, funding is limited in timescale, usually to three years, and tends to be awarded for projects that have a definite deliverable that can be launched at the end of the funding period. Funding patterns seem caught between the two models: the idea of a fixed period of research leading to a finished output is borrowed from the world of the academic monograph, while the accessibility and capacity of such projects favours archives that can be maintained and updated indefinitely.

ncse, with its three year lifespan and closed cluster of texts, is very much an edition. While recognizing the importance of the archive as a repository for textual content, **ncse** starts from the assumption that the identity of a title as a periodical or a newspaper is inseparable from the articles that it contains. As such, we acknowledge an editorial responsibility to periodical form, and undertake to account for differences in our rendering of it. As we will go on to show, these deviations largely arise from the differences between the form of the material forms of the periodical and the digital form into which we are translating it. In order to accommodate the diversity of periodical form, we have imposed a structural hierarchy onto the contents:

edition>title>vol>number>dept>item

For which the first four categories are organized in a folder tree, and the last two, department and item, are distinguished at the same level through metadata.

The following examples demonstrate the understandable differences between our means of reproducing the periodicals in ncse (in an edition, in a cluster, and all at once within this structure) and that of their original means of production (which is serially, incorporating variation, and with added bits such as supplements). For instance, the persistence of the two manifestations of the *Publishers' Circular* in its bound volumes has required us to adopt one or the other in our edition. The *Publishers' Circular* published numbers fortnightly, but also provided the means through which these individual numbers could be bound into volumes at the end of the year. The first of the January numbers contained the titlepage to the volume within its advertising pages. [**slide: showing advertisements and then vol titlepage from 16 Jan 1882 pp. 74-77**] Notice that although the volume titlepage is not paginated, it is accounted for in the sequence, allowing us to assume that this is where it was originally issued. Now, although this number resembles how the journal was issued, we've filmed from bound volumes from the British Library and sometimes these titlepages have been moved, and sometimes – as in this case – they have not. When the titlepages have been moved, they have sometimes been moved to unexpected places: for instance within a supplement to the *Publishers' Circular* called the *English Catalogue of Books*. Our options are:

- Leave it so it reflects the state of the hard copy. This would encompass both the histories of how the journal was issued and also how it was bound, even if 'wrongly'.
- However, this maintains an inconsistency that might be confusing for users. So, we could recreate the individual numbers by moving those titlepages that have been moved back into the advertising sections of the appropriate numbers. This recreation of the number might gesture towards a nominal 'original', but it is a form that has been superseded even within the nineteenth-century history of the material. Also, the adoption of the number as a run would in a partial edition: often there are integral portions such as advertising wrappers missing from the numbers of the *Publishers' Circular* bound up in volumes.
- The remaining option is to make them into tidier volumes, which is closer to the hard copy than restoring numbers but still necessitates an editorial intervention.

We have opted for the third, nominating the bound volume as the hypothetical copy text as it is closest to the last paper manifestation of the journal while also representing a logical organizational structure for twenty first century readers.

There is a similar problem in the two weeklies *Tomahawk* and the *Northern Star*. Both of these titles published supplements: *Tomahawk* published an almanack each year; and the *Northern Star* published a series of portraits of leading Chartists intermittently over its run. Although both of these titles survives in bound form, only *Tomahawk* published the various textual apparatus such as indices and front matter that signal that its editor and publisher intended it to be so preserved. The *Tomahawk* almanacks were issued separately from the weekly numbers, and cost a penny more. Three were published over *Tomahawk's* life, and two appear in the BL run. However, the problem is that they are in different places: appear in different places: the 1868 almanack is bound in at the end of the volume for January to June 1868; but the 1869 almanack is bound at the beginning of the January to June 1869 volume. Once more there is an inconsistency but, in this case, moving them seems quite unproblematic. Once again it obscures the unique historical condition of the hard copy but, in the case of this title, we have already done this by conflating together two runs, one from the British Library at St Pancras, and one from its newspaper outhouse at Colindale. As the St Pancras run provides the bulk of the material, and this is in irremediable volume format, it seems sensible to retain this unit for the digital edition.

For some time we pondered over where in the volumes to put the almanacks. This, in itself, is revealing: although the almanacks are incorporated into the material for of the bound volume, they are not accounted for in the volume's textual apparatus. We were so seduced by the logic of the volume, that we forgot that we could leave them outside of it. This, in fact, was our editorial decision regarding the *Northern*

Star portraits. Like the almanacks, these were issued separately from the weekly numbers of the periodical but, unlike the almanacks, they do not exist in any of the bound volumes. The portraits were discussed and advertised at some length in the letterpress of the *Northern Star,* but it is difficult to establish exactly when they were published. When we remember that there were possibly up to nine editions of the *Northern Star,* it becomes even more difficult to ascertain a place in the sequence of numbers in which to locate them. In this case, we quickly came to the conclusion that they should exist outside of the series of volumes and, when we considered the generic similarities of these supplements rather than the generic differences of the two periodicals, it became apparent that the Tomahawk almanacks, likewise, should exist separately from the volumes.

The final examples that we want to share with you relate to segmentation. We are delivering content through an application designed by Olive software called Viewpoint. Each number of a periodical is encoded as a series of items and, working with Olive, we have devised rules that enable those items that correspond with department headers, to be identified and marked up. This allows us to recreate the department > item hierarchy within the level of the number, and display it through the user interface. A useful aspect of this is that it provides data that we can extract and put into a table of components, allowing users to see the structure of each number at a glance, and move directly to the portions in which they are interested. Although the nomination of the item – which can correspond with the article, but not always – as the base unit resembles the archival model, our treatment of it in terms of structure and our delivery of it in terms of the application, keep items very much in context. However, as Laurel mentioned, the department > item distinction does not exist in the same form across all the titles. For instance, neither *Tomahawk* nor the *English Woman's Journal* posit two layers of hierarchy in which those things in the top layer repeat with every number. Tomahawk does have recurring features – there is always a cartoon, a leading article, and a puzzle – but, where these have titles, they are formally undifferentiated from other types of content. Equally the *English Woman's Journal* has recurring features like 'Open Council' (its correspondence column), 'Passing Events' (its survey of the news), and 'Notices of Books', its reviews column. However, as you can see in the slide, these are not differentiated formally from those items that are unique to that number. For both titles, these recurring items do not structurally organize content within them – i.e. they do not group together the same type of content – or if they do, they do so rarely, so that they seem to serve a different function than departments in the other four titles. Our decision here is whether to identify these recurring features as departments, and so make them available for navigation, or to identify them with the other items, and bring the lot into the table of components. Our decision once more took into account the needs of users: by bringing all the items into the table of components we do elide the presence of recurring features; however, this is something that the formal features of the journal does anyway. As our table of components is likely to consist of snippets from the page image, these formal similarities are foregrounded in our display of the journals's structure.

Conclusion

The distinction between archives and editions is a useful model but, we suggest, one that is of decreasing value in the digital age. The archive model, drawn from library science etc…, ignores the additional role that digital projects play as publications. The whole language of portals, gateways and links that characterize digital archives serves to elide the work that goes into gathering material, standardizing its data structures, and presenting it to users. However, there is a politics to all archives in both the selection of material and the way in which it is accessed. At ncse we have attempted to reconcile our fidelity to the source material with the awareness that this material represents a historical process as much as a discrete set of objects on the shelf. We have attempted to inscribe the characteristic forms of this process into our own publication, reproducing structural units, the order of pages, the appearance of text etc within our digital edition. It would be naïve, of course, to create an exact digital facsimile of this material and, indeed, as a cluster it already represents a different configuration of it. Accordingly, we have developed metadata schema and concept maps that will allow access to the edition as cluster, allowing it to be traversed in novel ways. However, like the various other instances where translations of material form from one into another requires an editorial change, we have been guided in the creation of these supplementary features under the influence of a posited user. Holding ourselves to account in this way demands that we make our users conscious of our interventions: as an edition rather than a facsimile, ncse makes explicit its workings, recognizing that digital literacy is a necessary component of understanding the past through digital means.