[**slide**: title]

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

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 historical artefacts, and instead posits the new edition as a digital surrogate with a direct relationship to a source of some kind. However, 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 afternoon we want to present our answers to some these questions. Drawing on the Nineteenth-Century Serials Edition – **ncse** for short – we will explore the editorial decisions that have to be made in presenting an object from the past in a different form in the present.

*　　　*　　　*

Journalism, as a print form, presents a distinct set of 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 creates its own set of problems, and the structure of ncse is, in some ways, an attempt to reconcile the often exclusive positions that result. 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 more closely 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:

[**slide of ncse deliverables**]

[**slide of ncse research objectives**]

ncse offers a model of textual scholarship that enables the republication of periodicals and journalism in digital critical editions.

1. Scale

The nineteenth-century periodical archive is characterized by its large scale and deteriorating, fragmentary condition. John North, the compiler of the most comprehensive list of nineteenth-century periodicals to date, estimates there were around 125,000 individual titles published in England between 1800-1900 alone. Making selections from this considerable collection can cause bibliographic headaches. As a cluster, the six titles in **ncse**, although all historically important in their own way, gesture towards the diversity of forms that serials could take in the period. The current crop of large-scale digitization projects from Proquest, Thomson-Gale and, of course, the British Library's *British Newspapers: 1800-1900* 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 contents of these projects. Rather than predicate themselves on exclusive contents then, competition will instead be based upon project scope, and the ease with which each project can be incorporated within wider searches.

Once selections from the periodical archive 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: the six titles in ncse contain 98,565 pages in total, but over half of these are either in the *Leader*, for the most part a 24 page weekly and the *Monthly Repository*, a monthly with between 60 and 70 pages. With around 24,000 pages each, the different time spans for each title demonstrate the difference in bulk that periodicity makes: whereas the *Leader* took 11 years to publish its 24,000 pages, the *Monthly Repository* took 33.

The size of these runs means that the republication of periodical texts is only possible as digests or anthologies. Even the shortest of our runs, the *Tomahawk* – a satirical weekly alternative to *Punch* that only ran for just over two years – still consists of some 3000 pages. To further complicate things, the existing runs of periodicals often do not fall into neat sequences of numbers. 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. [**slide: *MR* family tree**] The *Northern Star*, for instance, published up to 9 editions a week and as this slide of the life of the *Monthly Repository* shows, journals often have complex lives, developing offshoots that become publications in their own right, and changing their titles over a run. 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. However, as the **ncse** material has constantly reminded us, this linearity is resisted by the textual remains of the journals: 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.

2. Depth / Breadth

[**slide: front page *NS***].

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.

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. [**slide: page of the *Leader* in File Cabinet**] 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 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 are predicated upon a similarity to their source objects, then these structural categories ought to be reproduced along with the 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 changes 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***] 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.

3. Material Remains

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. 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 the various material forms that it has taken up to the present. There are single numbers, organized in a sequence that usually corresponds to time, but these often appear 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 itself, perhaps the most stable unit, 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 British Library are important partners in **ncse**, and the majority of material has been sourced from here. However, we have supplemented the BL holdings with material from a range of other institutions and private collections. This means that it is impossible to maintain the fiction that the history of these objects somehow stopped when they were published in the nineteenth century. Even if one wanted to offer the runs of periodicals in a way that disguised their later history, this is indelibly inscribed into their material form. For instance, **ncse** has largely been filmed from microfilm. Although there is little historical interest in this intermediary stage, it is nonetheless present in the digital resource and so raises the question of how we acknowledge. The microfilm stage can be problematic: for instance, *Tomahawk* contains large cartoons which are printed on ink washes; because microfilm is in black and white, and is often produced at a high tonal contrast in order to make the text stand out from the page, we lose not only the colour of the ink but also the subtle tones that are necessary for the images's dramatic effect.

[**slide: Tomahawk**]

The persistence and reproduction of these material traces across different material forms reminds us that the idea of a single 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. Perhaps rather than look to historically contingent criteria for editorial principles, we should look to the intended user of the digital resource for guidance. However, just as editorial principles derived from hypothetical ideal texts can be problematic, so too can those derived from an equally hypothetical user.

Conclusion: ncse, bibliographical control and editorial choices

These three aspects of journalism – scale, depth and breadth, and the material condition of the archive – demand editors provide bibliographic control at a number of structural levels, while accommodating the needs of intended users. 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. **[slide: *Penny Illustrated Paper*]** It is also the logic behind useful projects such as the British Library's own digitization of the *Penny Illustrated Paper*. Although wonderfully easy to use, and interoperable with the rest of the material in Collect Britain, the digitization has not retained the generic aspects of the periodical. For instance, page numbers are attributed which do not correspond to those printed, and the user interface (necessarily) presents articles as standalone units from a larger corpus, rather than as part of a structured text in their own right.

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 serials serials 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 them to take into account this audience, even if they are not the intended users. The expectation that a digital edition should address a very broad audience can mitigate against close scholarly care of its contents, making interest in formal features like these mastheads, for instance, 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.

[**slide: of edition structure**]

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. These deviations largely arise from the differences between the form of the material form of paper serials and the digital version that we are creating from them. In order to accommodate the diversity of periodical form, we have imposed a structural hierarchy onto the contents of **ncse**.

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.