‘All change’:

Digitising Journalism: ncse Year 2

Some of you were with us exactly a year ago, and like us, when you came away you understood that ncse was a project planning to digitise six 19C journals of diverse types that span the century as first announced, with the addition of multiple editions of two weekly titles, the *Northern Star* and the *Leader*, issued in order to reach readers distant from the place of publication on the same date as those within easy reach. This increased the total pagination to nearly 100K. One result of this decision to include rather than ignore the multiple editions in the BL collections, taken in the first year, was the concept of a ‘core’ of 30K pages (the original figure we had envisaged) on which full processing would take place, by which we meant multi-level segmentation of the text, reflecting our attempts to theorise the structures and organization of our six titles and more generally the genre of 19C serials (our cluster is strategically not *either* periodicals *or* newspapers, but both: it consists of a weekly broadsheet newspaper; a general weekly covering politics and arts; a satiric illustrated weekly; a fortnightly trade paper; and two monthly magazines). While the full resource would be digitally available to users (although not searchable because unsegmented), the ‘core’ would benefit from segmentation; the full weight of our deliberations, and more elaborate processing. We duly presented to the gathering a year ago maps of date distribution over the century for the proposed core, percentages of each title that would be included, and the categories and criteria by which we selected it. Much interesting thought had gone into it, and as some of you perhaps remember, we were ready to roll.

CHANGE

However, soon after we met last February, after a number of iterations and demos between the research team and the software developers, it became clear that the level of realisation of multi-level segmentation attainable was to neither team’s satisfaction, and the experiment was halted, so that we could take stock

and re-consider our options. An alternative emerged, which was for the *entire resource* of 100k to be segmented –thus making all 100K pages searchable and amenable to metadata – but only at a single level. This would enable future work on the entire resource by subsequent researchers and offer users 70K more segmented and searchable pages than the core model allowed. It was a difficult decision for us; we were attached to the possibilities of multi-level segmentation, but the combination of greater potential for future work on the data, evenness of application across the resource, and time and money constraints on the finessing of multi-level segmentation to a point of acceptable levels of accuracy led us to opt for the single-segmentation alternative. We went for segmentation at the lowest level – the item or 'article', as that appeared to be the basic structural unit of the serial. However, we have tried to maximise the possibilities of single level segmentation to provide a modest functionality at two levels: while users will be able to search, read and print *items*, as well as browse full-page facsimiles, we have tried to indicate the structure of numbers (and titles at different periods of their runs) through a table of components window on the left of the screen that shows 'Departments' only (with the article level items suppressed).

———————————

However, you cannot see the inverted commas around the words 'articles' and 'departments' in my script, but believe me they are there. That is because definitions of articles are not the same in all titles nor do they follow the same 'rules' of identification across titles (separation by a rule, or initial small caps, or headings for example), nor are all six titles even organised into Departments. Although Departments or Sections might be thought a structural principal of all titles, they appear only in some of our 6 titles, and then with variable frequency. Even when they did seem to exist (it isn't always clear); as with articles the rules to mark them differ. Moreover, departments don't all have discrete headings (it could be the date or the masthead for the 'leader' or simply column and page position), let alone italic type. As in any edition, consistency has been a value, yet our commitment to the concept of the 'department' and the 'article' has meant

that, as we moved from title to title, we have had to settle for different definitions of these terms – sometimes even constructing departments to appear in the table of components, to give the user some idea of the organization of each issue, and the title over time. Where differences across titles that are widely separated by period, and infrequently overlap in time, are perhaps to be expected (although not *these* differences we murmur weakly), there is also considerable variability within individual titles, even when the run is as short as a decade. Departments disappear, and sometimes reappear, format changes as even the titles of the serials do, so our hard-won rules about 'rules' may be undermined even within a title.

It is interesting to contrast this attempt to find common structures across the cluster of titles while remaining primarily committed to the relatively inchoate variability of the press and the collection modes in which it has come down to us (its rich diversity we maintain) with the clearly interpretative categories we devised to select the core – beginning and ends of runs; changes of editor, coverage of a common topic, multiple editions, and incidence of visual material which allowed us to provide a higher proportion of *Tomahawk* than its short existence permitted. While these categories of selection sought to foster comparability across the cluster and to pre-determine some frameworks of interpretation, the contents of the issues selected as the 'core' otherwise retained their integrity, with the minimum unit in the Core being the 'Volume', accompanied by an attempt to favour contiguous volumes. So, although the 'core' *was* selective and interpretative, these frameworks determined the selection of volumes only, *not* their content which was miscellaneous and variable, and ungoverned by the editors. The interpretative frameworks were there to be drawn on by users, or not. In so far as the project has been from the start primarily interested in theorising the 19C serial and its 'translation' into digital form, it might be thought that the abandonment of the multi-segmented and selected core has reduced the possibility of achieving that end. No doubt there have been losses, but the concept of the Core, and its adoption and abandonment will be there for media history scholars in our foregrounding of the

process of producing this digital edition. By the time ncse goes live, the digital map of 19C serials will have changed beyond recognition from when it began. In addition to the BL Pilot, ILEJ, the Germ embedded in the Rossetti archive, the Modernist Journals Project at Brown, and the American periodicals in The Making of America on the Library of Congress site that were about at the start, at least three new and numerically large projects to digitise complete runs of 19C serials from BL (newspapers), Thomson Gale and ProQuest will mean that scholars far more readily can do cross-serial studies on topics suggested by the core, such as changes of editor, beginnings and ends of runs, and multiple editions. We are still experimenting with text mining to see whether some of the top layer elements of the concept map might be marked in the resource through automatic processing. We plan to draw on the indexes of the *Waterloo Directory* as authority lists to assist in the process. .

Although we expended considerable effort in Year 1 in theorising the potential *contents* of the data through the construction of a concept map with three different if interrelated levels, it has been our work on the structure of serials that has occupied us most to date, both in terms of segmentation, and the selection of generic metadata. As I have implied above, the necessity to create rules by which segmentation can be marked up in the text has made us very conscious of the ways our six periodicals have been constructed, by a combination of their editors, compositors, and printers, not only what kind of contents and what it says, but also how (if) it is organised, and what issues were operative in the course of 'layout', such as for example some recognisable, reiterated pattern to each issue for the ease of readers on the one hand, and on the other the editorial signalling of the import and category of each item. First, the presence or absence of contents which may constitute structural elements of serials, such as correspondence from readers, obits, and leaders is established; but predictably, the format, position, visibility and naming of leaders (for eg one version of leader in Tomahawk is identified as the War Whoop, untypical to be sure) are so various that some mark up will have to be done by hand.

IMAGES

Images in media history, history of the book, and bibliography tend to be separately studied, with notable exceptions such as that of Peter Sinemma's study of the ILN which discusses images in their serial context. For ncse, format does raise one of the problems of our text-based ocr technology; although graphics and images are usually identifiable through generic metadata (they are segmented), they are not searchable, as ocr does not 'read' or include them. So all mark up of images for search purposes in ncse will be done by hand, as the standard software used by art historians for this purpose is too complex a model for a mixed media resource like the ncse texts, which include mastheads, various standard printers' blocks (eg, finger posts), cover graphics, and advertising cuts, including the graphic arrangement of type, which ocr normally garbles or cannot read. The primary tasks for ncse with respect to images is to find means to make them an identifiable and visible part of all the titles, even those which are not 'illustrated'; to try and distinguish between different types of visual copy in these titles; to establish visual material as an important and integral thread and component of the 19C press; and to make it searchable at a simple level of subject, at the least. It is unlikely that engravers and artists will be marked up, even when signed, although the origins of blocks and the identities of artists and engravers known to work for particular titles can figure in the 'headnotes' we shall be including for each of the six serials. We also plan to situate our serials in the history of 19C illustrated journalism by including a timeline/chronology of the illustrated press, and a dedicated headnote about images and serials: our titles may thus be mapped against a representation of the appetite for illustration that spanned the period from the 1830s onwards.

Another structural element arising in our research and implicated in 'segmentation policy' has been multiple editions, about which several of us have given papers, and which we talked about a year ago. In addition, we have had productive exchanges about, for example, how we treat originally unbound or 'additional' materials -- paratextual items such as front and back matter (Tables of Contents and Indexes for example, and title pages; and what might be termed

Supplements such as those given by the *Northern Star* to subscribers only, prints which were distributed by newsagents in different geographical locations on different dates, as they were supplied by the publisher, so that attaching them to a single issue or indeed edition in a temporal sequence is problematic and perhaps inappropriate; or a more traditional supplement to the *Monthly Repository*, the *Unitarian Chronicle*, which had subsequently been bound separately by BL.

The debate about front and back matter makes some of the policy issues clear: First of all, front and back matter pertain to a specific format of the resource, that of the bound volume, which is the dominant format in which later readers, including ncse, have had access to these texts. We need to theorise the volume format as distinct from the issue which, in a different, collected form, comprises the volumes. We also need to remember that in addition to the annual or semi-annual volumes, weekly publications were often bound and sold monthly. So, the forms in which we read 19C serials are mediated and not transparent or without meanings; but let's agree on the Volume, which is the format in which most of the BL titles are preserved, as is the case in most libraries. In Volumes there is always additional matter, which may include some of all of the following: timely and informative annual review Prefaces, title pages, frontispieces, Table of Contents, Indexes, and not least the binding, some of which is customised, by the library, the original publisher, or individual owners. But volumes may also differ from the single issues they 'contain' in their *exclusion* of original materials in the issue, notably the adverts, and the issue's covers or wrappers, as well as any loosely bound in materials. If most of the additions emanate from the publisher of the serial, the exclusions have variable agents – the publisher again, who produced bound volumes at regular intervals to sell to libraries; the library or individual which bound their single issues acquired serially into volumes themselves, and the dealer, who stripped out valuable illustrations to sell as separate items. While motivation may have varied, the effect of the removal of the most ephemeral aspect of the serials supported the transformation of single

issues into a more enduring *book*, to take its physical, material place on shelves of libraries – institutional or private – with other books. The question of where among the issues and volumes to include the title pages, Indices, T of C, etc in the electronic edition involves theories of editing. Do we represent the volume and move the paratextual materials from the issues with which they were distributed, and arguably a part, back into the volume where we distribute them as the Publisher intended? When they are 'illogically' mis-bound by the Library's binders, do we adopt that phase of the mediation of these materials, and reflect that sequence, or do we make good their error and erase the binding process? To which moment of publication do we 'return' it? The issue? The invisible monthly? The volume? In most cases, then, we are not reading issues but a 2nd and perhaps 3rd redaction of the serial, already 'transformed' and significantly 'later' than the dates on individual copies. Which one(s) do we represent? And alongside of these historically contingent editorial questions – all of which involve a hypothetical fictional text -- is that of the reader of the electronic edition, and how our choices clarify or confuse their experience of the resource.