ncse **ncse: Digitising Journalism**

Our editorial strategy has been structured by our encounters with the rich, but complex, material histories of our source objects. As Laurel explained, our decision to digitize them is simply the latest in a long line of material transformations these texts of undergone. They have been single numbers, cut up and bound in volumes, left to sit on library shelves, filmed as microfilm, and now, finally digitized. Our decision as editors of this material is which aspects of these transformations we acknowledge, and which we elide. For instance, an often overlooked consequence of the use of microfilm as a source for digital images is that we are working in a black and white environment that bears little relationship to the hard copy, whether in terms of colour images, different shades of 'black' letterpress, or the colour of the yellowing paper. This attention to material form is, of course, part of a much broader attention to the content of this material. In digitizing nineteenth-century periodicals we have had to ask difficult questions about what it is that makes these texts what they are. The process of digitization is more than just affecting a material change: we have to identify the source material, recognize its relevant structures (both in terms of form and content) and devise the means to implement them into digital form

1. What is ncse?

So, to begin, what is **ncse**?

- ncse:
    - It is a collaboration between BL, CCH, Olive software, but will produce a freely accessible online resource.
    - We are working with Olive Software and CCH to develop ways of identifying and capturing information beyond simply relying on OCR to produce an index.
    - Developing metadata structures to allow comparisons across the titles. These include indices of people, places, events, institutions, and publications, and data mining to provide subject indexes which we will link to thematic concepts.
    - It is not just a delivery system that serves an archive of content, but rather an edition which foregrounds the contextual and referential nature of C19th print culture.
- ncse contains:
    - Six periodicals:
        - *Monthly Repository* (1806-1838) inc Unitarian Chronicle
        - *NS* (1837-1852) inc portraits and Kew material. As far as we know this is the most complete NS in the world.
        - *LDR* (1850-1860)

    - *EWJ* (1858-1864). Sourced from Women's Library so that it includes some adverts.
    - *Tomahawk* (1867-1870). Sourced mainly from the BL St Pancras, but supplemented by two microfilm we found at Colindale that have wrappers.
    - *Publishers Circular* (1880-1890). Only 10 years of a much longer run. Includes large advertising sections, the *English Catalogue of Books*, and the supplements of illustrations that were issued with the Christmas numbers.

      This is a total of 98,565pp taken from across the C19th. It equates to something like just under a million distinct items.

  - Contextual material.

    - Headnotes explaining about the individual titles
    - Accounts of methodology, including editorial decisions, design of tools and interfaces, implementation of schema
    - But the process of mapping periodical form has also lead us to reflect upon the genre more widely. We will also contribute essays on aspects of form such as the importance of beginnings and endings.

- Periodical pages are rich in information. Simple text transcriptions are not sufficient to capture the diversity of material, and the complex relationships with which it interacts.

  - Text is, of course important: after all, nineteenth-century readers bought periodicals for what they say. However, what texts say is also a function of how they say it. Text can come in various fonts, is different sizes, and is located in a variety of places on the page.

  - Layout also plays a part in terms of length of articles, where they are on the page, and the order in which they appear

  - Therefore all periodicals rely upon the visual as an organizing category.

  - But pictures and images are important too. From fingerposts to full page cartoons. But also the important role images play in advertising material: from the Northern Star's illustrations to the more elaborate ones in the *PC*

  - And there are strange categories of content such as mastheads or library stamps.

  - All these, in combination, participate in denoting firstly

    - that these are periodicals, that they come out serially, and so identify their continuity over time.

        - That they are a part of a wider market. These aspects denote the individuality of a periodical, while referring to those to which it is similar in order to aid readers identify what it is they are buying.

    - Because of the genre's relationship with time, these relationships are dynamic. We must accommodate variation.(and use NS becoming the Leader [**slide**])

2. How it is organized

So, how do you capture the complexity of the periodical page, with all sorts of heterogeneous content, in a way that recognizes what stays the same (the generic categories) while not oversimplifying those that change (the bits that respond to the market, the wider cultural context, keep the journal fresh, keep readers interested)?

Well….

- The journals themselves are within an Olive application

    - This capitalizes on both their experience in digitizing complex source material

    - Allows us to build on their existing delivery platforms, which already specialize in zooming, printing, whatever, so that they are more responsive to our needs.

    - This application produces xml, and so can interact with other applications that we build.

- We're going to seamlessly house the Olive application within an interface designed at CCH and BBK. This interface will also:

    - Contain any supplementary material such as:

        - Additional visual material like the NS portraits. These were not issued with the *NS*, so it is appropriate to house them in a separate application. This will allow enhanced zoom functions
        - Contain the contextual essays etc, plus the documentation for the project
        - House any supplementary tools and databases such as indices etc that we are exploring with text mining.

- **Structure:** We're imposing a simple structural hierarchy on all the titles, despite their individual differences:

    - Structure of edition: edition>title>vol>number>dept>item [**slide**]

    - First 4 of these (up to number) are located in a hierarchy of folders

  - The last of these, department and item, are in a hierarchy that we've created. After segmentation, each number consists of a series of items. As we'll explain in a bit more detail later, we've provided instructions so that those items that mark where departments begin and end are identified so that the items that fall between them can be grouped under the appropriate label.

3. Segmentation

- **What is an Item?**

  - The most simple structural unit in **ncse** is an item. This roughly equates to a single textual component, like an article.

  - However, we also take it to include (**slide** with a page with a suitable array of items – maybe *NS*?) mastheads, department headers, articles (including advertisements), images, stamps etc…

  - It is impossible to define an item by content, instead we've used non-textual identifiers on the page. For instance, in most of the titles it is better to avoid defining items by the occurrence of headlines, and instead use the dividing lines that mark where they end

  - This is a little problematic as the lines do not appear when an item ends at the bottom of a column. As this is usually desirable for editors, this is quite common. However, we've found this usually coincides with the occurrence of a title at the top of the following column – because otherwise readers wouldn't know it was a new item – so it remains possible to recognize the boundary.

- **What is a Department?**

  - What we've described is a simple, single level hierarchy, yet nearly all of our titles posit another structural level, collecting types of content together. These sections recur in each number, and usually have regular titles.

  - In partnership with Olive Software, we've devised a way of reconstructing the division between items and departments.

  - The Olive viewer, called Viewpoint, which we are developing with Olive, has a Table of Components to the right of the page that is being viewed. This is taken from an existing application called Enterprise Publisher, that was designed to display books. The Table of Components was a way of recognizing the chapter divisions in a book: we wanted to adapt it so that it captured the departments

  - Because of the scale of the edition, it is beyond our resources to manually identify where departments begin and end or, alternatively, design rules for this to be done mechanically and then edit them accordingly.

- We have instead negotiated with Olive that all items appear in the Table of Components, but that they appear as 'untitled article' – the default label Olive use for those items. (**slide** – of whatever we have that shows this – maybe cheat with a page with no depts on it).

- We then provided simple rules that allowed the operators to recognize which items marked the beginnings of departments. These titles were not changes.

    - We've tried to use formal features rather than content to delimit departments. For instance, the correspondence in the *Leader* is called 'Open Council', while the editorial notes to contributors in the *MR* is called 'Correspondence'. The journals themselves denote this structural distinction through formal features, so it makes sense that we do too.

    - The most common structural feature marking a department is actually to do with where they end rather than begin: double lines!

- Using these rules, we've provided the means to identify those items that correspond to the beginning of departments. These are usually, but not always, some sort of heading. As department headers usually exploit some fancy font to distinguish them from plain old item headers, the entry in the ToC is gobbledegook.

- So, we are left with a ToC which has 'untitled article', or some other string of characters. By suppressing the 'untitled article's, we are left with a ToC of just departments

- We are currently deciding what to do next. Our options are to hand key each department title, or to use a small image from the page – a snippet.

    - Hand keying would be very time-consuming, but has the advantage that we can specify what each department is called. This is vital for departments that do not have titles, such as the regular section of advertisements – usually around ¾ of its total pages – that appeared in each number. We can opt to call it something like 'Advertising Department'.
    - The advantage of snippets is that it is relatively easy to import page images from the location of the text string, and would leave us with little or no hand correction. However, we are not sure how straight-forward it would be for users to recognize what a department is from a snippet. Often departments do not have titles, but are merely denoted by double lines: not a very useful image. Also, we are not sure

how elegant a series of images would be in the Table of Components. However, as images, they show users precisely the textual features that denote departments. The journal's structure here speaks for itself, with our interpretive function restricted to the design of the rules.

- Either option will provide us with markers that identify where a department begins. We can add metadata to these items, which we call 'department headers' to mark them out. We have also arranged for Olive that metadata added to department headers will be cascaded to all the items between one header and the next. This would make it possible to search all the 'Open Council's of the *Leader* for instance.

- However, this is inaccurate: department headers mark where departments start rather than end, and the widespread use of small, miscellaneous items to fill a space so a new department begins on a fresh page or column means that departments do not follow on from each other sequentially. We feel that the gain in recognizing departments is worth introducing this necessary inaccuracy.

- And, as Laurel said, this department>item structural model does not necessarily apply to all our titles:

  - EWJ [**slide**], and explain: 12 items in each number, some of which seem to be departments (Open Council, Passing Events, reviews, poetry). The journal mixes the levels, should we in the ToC?
  - THKs [**slide**], and explain: Tomahawk is a series of items. There are no recurring features other than the cartoon, the leading article, and the puzzle. Are these the only things that should appear? Or, as the cartoon is an insert, should we include all the items in the ToC?

  In a while you will be asked to give your opinions on this issue. It is one of the General Questions in our (blue?) handout that you will be asked to complete in the session before lunch.

- **Editorial issues**

In trying to edit 98,565 pages with various and complicated material histories into a reasonably consistent form that still captures their diversity, we have encountered a number of interesting editorial questions. Many of these revolve around the issues that Laurel has introduced: what is the final form of the edition that we are creating? Does it recreate the number, the volume, or something else entirely? We've also included spaces in the questionnaire of general questions about each of these, so feel free to let us know what you think during the day or on the questionnaire.

- *Publishers' Circular*. In the *PC* the front matter for the volume was issued in the advertising section of the first number in January. Our source, the bound volumes in the British Library at St Pancras, reveals inconsistent policies in binding this material:

  - Sometimes it appears at the front of a volume [**slide**]
  - sometimes it appears mixed up with the English Catalogue of Books, an annual supplement that operates as an index, also towards the front of the volume [**slide**]
  - and sometimes it has been left in the advertising pages of the first number of the subsequent volume. [**slide**]

So, what should we do? Moving them all to the front position of the volume would make it consistent, and represent the volume as a structural unit. However, it would mask the actual means of production of the journal, particularly its fortnightly periodicity by privileging this annual format. It would also mask the historical condition of the material object that we have digitized: we are transforming it into some sort of ‘ideal’ that we have imagined. We cannot, of course, move the frontispieces that have been bound at the front of volumes to the back, as they do not have page numbers, so we don’t know where they came from…

- *Tomahawk Almanacks*. A similar problem arises with the position of the *Tomahawk* almanack in the hard copy at St Pancras and the films at Colindale. Each December *Tomahawk* published an almanack: a lavishly illustrated supplement that contained a calendar for the year to come and some satirical surveys and projections. The almanack cost 3d, a penny more than the journal itself. The St Pancras run contains the first two almanacks, for 1868 and 1869, but they are bound in different locations within their respective volumes. *Tomahawk* marked the end of its volumes at the end of every six month period. The 1868 almanack is bound in at the end of June 1868 – so at the end of the first volume in 1868, but six months after it was issued; and the 1869 almanack is at the beginning of January 1869, a few weeks after it was issued. So what should we do?

  - This time 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.
  - The run of *Tomahawk* from St Pancras is already in volume format. The front and back matter already been moved from the respective portions of the individual numbers and the wrappers removed. Therefore the volume seems to be a sensible unit to posit as ideal text.
  - But, of course, the almanack’s do not have an allocated place within the volumes, and were not issued as part of the serial. They are supplementary.

- Such questions remind us of the other aspect of editing. While we seek to negotiate the means to acknowledge the important aspects of our source material, we do so knowing that we are presenting this material for a particular audience. Just like the contingent material form of our source material, our knowledge of how it was produced, and our understanding of what it contains, users too shape the final object that we produce.

4. Current work and where we're going

- At the moment we're in the middle of the processing schedule. Olive have already produced pdfs of most of the pages in ncse. Before this can be segmented each page needs to be checked and edited accordingly. While doing this, we are also designing segmentation policies. Processing is currently taking about three weeks per title, after this we check the segmented pages and produce a folder tree containing bibliographic information.

- Our next major job, which we are undertaking in close partnership with CCH, is the insertion of metadata. As you can see from the metadata schema in the newsletter, there is quite a lot of information, most of it bibliographic, that we need to add. We imagine this happening in three ways

    - Manual insertion: obviously this must be kept to a minimum. Because we can cascade entries down through the hierarchy, we have tried to restrict hand entries to the highest level of the schema, which is number.

    - Computer-generated metadata: with CCH we are exploring how we can use text mining to extract information from the ncse corpus. Text mining is a process described from computational statistics that compares the occurrences of words and phrases between selected corpuses. We are investigating how this technique can help us in two areas:

        - Firstly with the production of indices. Through an analysis of syntactical forms we can identify many of the proper nouns from the ncse texts. Through the generous cooperation of John North, we have been able to acquire the indices to the *Waterloo Directory*. By comparing our output of undifferentiated names with *Waterloo* lists, we can begin to build a database of people, places, etc… and, time permitting, start to link them together.
        - Secondly we can use text mining to identify generic and thematic content. Because certain words recur more often in certain genres of content, it might be possible to identify what type of article an item is by comparing its words against an average C19th corpus. For instance, genres like obituaries and correspondence are likely to have characteristic phrases that will allow us to find them. A similar technique might be possible to work out what an item is about. We can use text mining to find out which words, for instance, occur most

often in certain grammatical positions, giving a cluster of words that corresponds to what a text is about. Ideally we would like to link these clusters to some aspect of our own concept map – displayed on the wall behind you – that we produced in year one. It was designed to allow users to find articles that interest them by moving through a three stage hierarchy, much like a subject index. The bottom terms are quite specific, and we may find that in some cases the keywords that are extracted through text mining correspond closely to our terms.

- The third level of metadata will arise through our collaboration with NINES. As Jerry will explain later on, NINES is a suite of tools that permits users to search across associated projects, identify objects that interest them, and gather them in their own collections. An important component of this is the ability to mark-up objects and see the terms others have already used. This will allow users to identify objects in ways that are meaningful to them and the community as a whole.