ncse Newsletter July 2007

Welcome to the fifth **nineteenth-century serials edition (ncse)** newsletter. Since our previous edition, in December 2006, the project team has made progress in the following areas:

Activities and Publicity:

- Presented papers / workshops at four events, including '**ncse**: Digitizing Journalism', held at the Centre for Computing in the Humanities at King's College London in February .
- Maintained the **ncse** website.
- Produced an **ncse** poster.
- Contributed a chapter for a forthcoming book entitled *Text Editing, Print, and the Digital World* edited by Marilyn Deegan and Kathryn Sutherland and published by Ashgate.

'Not Quite a Saint', ***Tomahawk*****, 6, 18 June 1870, pp. 240-1.**

Research:

- Made further editorial policy decisions regarding the organization of material within the edition.
- Finalized all six segmentation policies.
- Obtained a further two portraits of Chartists for the *Northern Star*.
- Obtained 56 full-colour images of cartoons from the hard copy of *Tomahawk* at the British Library.
- Continued to process the ncse periodicals, including editing pdfs prior to processing and then evaluating the outcome of the segmentation.
- Continued to investigate the application of text mining techniques for metadata and indexing.
- Worked with Olive Software to design the functionality and look and feel of Viewpoint, the application through which users will access the periodicals.
- Undertook user testing as part of '**ncse**: Digitizing Journalism.'

* * *

Activities and Publicity

The Project Team have been busy presenting papers about **ncse** at various events in Britain and the US. Many of our papers and presentation can be found on our website [http://www.ncse.kcl.ac.uk/activities/conferences.html].

- Laurel Brake, 'Journalism and Modernism: Culture Wars or Intimate Relations?', Modernist Magazines Conference, July 12-14th 2007, De Montfort University, 12-14 July 2007.

- Jim Mussell and Suzanne Paylor 'From Life on the Shelves to Digital Shelf-Life: ncse and the digitization of periodicals' at the British Library, 28 June 2007.

- Jim Mussell, poster session, British Printed Images to 1700, Friday 13 July 2007.

- Jim Mussell and Suzanne Paylor hosted a workshop at 'Victorian Studies: Pasts and Futures', a one day conference to celebrate the 40th anniversary of the Centre for Victorian Studies at the University of Leicester, 31 March 2007.

- The Project Team presented a panel at the Society for Textual Scholarship biennial conference held at New York University, 14-17 March 2007. The panel was entitled "Editing Journalism: the Past in the Present" and consisted of three presentations: Laurel Brake presented a paper entitled 'Periodical Problems: Clusters, Runs, and Editions'; Suzanne and Jim gave a paper called 'From Life on the Shelves to Digital Shelf Life: Representing Journalism in the Digital Domain'; and Mark Turner and John Stokes presented a paper from their work on Wilde's collected journalism for the Oxford English Texts *Complete Works of Oscar Wilde* entitled 'Editing Journalism: the Case of Oscar Wilde'.

- On the 24 February 2007 **ncse** hosted '**ncse**: Digitizing Journalism' at the Centre for Computing in the Humanities at King's College London. This well-attended event featured contributions from Isobel Armstrong, Laurel Brake, Hilary Fraser, Jerome McGann, Ed King, Jim Mussell, Suzanne Paylor, Joanne Shattock and Harold Short. There was also an opportunity for participants to explore a demo of ncse that featured some volumes of the *Leader* in Olive's Active Paper interface.

- Jim Mussell and Suzanne Paylor presented a paper entitled 'A Picture or a Thousand Words: the use of images in the nineteenth-century periodical press and how they are reproduced today' at the Open University's seminar

series ‘Publishing Periodicals: Seminars in Book History and Bibliography’ on the 29 January 2007 at the Institute of English Studies in the School of Advanced Study at the University of London. The paper was an extended version of the one presented at ‘The Verbal and the Visual’ in June.

Forthcoming presentations include:

- Jim Mussell and Suzanne Paylor, ‘**ncse** and the Production of Victorian Text in the Digital Age’, British Association of Victorian Studies, University of Salford, 30 August – 1 September 2007.

* * *

‘Sowing Tares!’, *Tomahawk*, 3, 11 July 1868, pp. 14–15.

Research

Interesting Page of the (half) Year!

As part of our agreement with the British Library, the majority of the material in **ncse** has been sourced from new and existing microfilm. This is common practice for many digital projects as not only is it quicker to scan from microfilm than from hard copy but many institutions see the production of microfilm as an important archival outcome. However, sourcing material from microfilm has a significant effect on the final digital product. Microfilms are usually in black and white, and so do not capture either the colour of the paper or the ink of the source material. As the tonal contrast on microfilm is usually set quite high in order to get good definition letters on a white page background, the range of greys produced from the colours on the hard copy are often lost. There can be a further reduction in the quality of the grey tones if a bitonal palette (i.e. black and white) is used when it is digitized rather than greyscale. This is particularly problematic for nineteenth-century engravings, where combinations of black and white lines are used to render quite subtle greys.

'The People's Guide!', *Tomahawk*, 1, 19 October 1867, pp. 242-43. Digital image taken from hard copy at Colindale.

In **ncse**, we have found this particularly troublesome with *Tomahawk*. This weekly magazine published a large cartoon, often over two pages. As you can see from the example above, these cartoons are engravings on ink washes. This means that they lose their colour when in black and white and, because of the high tonal contrast, often much of their subtlety. As this is a well-recognized problem, the microfilm operators will often alter the settings in an attempt to produce a better image. This means that duplicate images are present in the microfilm: for instance the examples below from the **ncse** film show the operators filming at the usual settings for text, and then altering the contrast to provide a much lighter image that can bring out the dark engraved figure. As the microfilm operators cannot see the results of each shot, they have to use their experience in order to estimate the correct settings for a particular image.

'The People's Guide!', *Tomahawk*, 1, 19 October 1867, pp. 242-43. Digital images taken from microfilm.

This means that our run of Tomahawk is bibliographically complex. The journal ran from 11 May 1867 to 27 August 1870. The British Library holds a run at St Pancras from 11 May 1867 to 25 June 1870, and two incomplete and overlapping runs at the Newspaper Library at Colindale from 9 January 1869 to 30 July 1870 and 15 January 1870 to 16 July 1870. Where no film existed, new microfilm was created and then the whole lot was scanned into tiff images and then bound into pdfs. We then edited this material and amalgamated the runs, checking the images for quality as we went. Where we judged that a new digital image from the hard copy was needed, we substituted it for the existing image sourced from microfilm. This means that our run of *Tomahawk* consists of three sets of microfilm and a set of digital images all sourced from at least three different paper runs in the BL.

The idea of the 'original' is largely illusory in journalism as not only are texts often derived from other sources, but the material with which we deal is usually compiled from other objects. Our run of *Tomahawk* exemplifies this: not only is it a composite of three different paper runs, but the contents of each run is slightly different. In editing the pdfs and amalgamating the runs we have, to some extent, hidden this history in representing a single unbroken run of numbers in digital form. However, the material content of this composite run resists the imposition of our ideal editorial copytext. For instance, the presence of advertising wrappers for some

numbers marks them as being from a Colindale run, and the presence of some full-colour pages reveals that they are sourced directly from hard copy rather than from microfilm. Just as the bibliographic history of the run persists despite our reordering of the material, so too does the material history persist despite the fact it is now all in the same material form, i.e. the digital. Although there is little historical interest in microfilm as an object, it is still an important stage in the history of some of the digital objects in our edition and so it is appropriate that it is duly acknowledged. It is just such hauntings of digital editions that gesture towards the complicated processes that underpin their creation.

Launch date

ncse are pleased to announce that the project will be officially launched on the 13 May 2008 at the British Library. The event will consist of a series of papers during the day, with a formal launch in the evening with a keynote address by Alan Rusbridger, editor of the *Guardian* newspaper. Further details of the launch will be posted on our website as the day approaches.

ncse: Digitizing Journalism

On the 24 February 2007, we hosted the second of our annual symposia, this time at the Centre for Computing in the Humanities at King's College London. The day began with remarks from Harold Short and presentations about the project from Laurel Brake, Jim Mussell and Suzanne Paylor. After a break for coffee Jerome McGann presented the latest developments from NINES, updating the group about the successful application of Collex both within the NINES environment and beyond. Lunch provided an opportunity for informal discussion over excellent organic fare sourced from local farmer's markets provided by Jo Foster. The afternoon featured a panel session and two plenary presentations. On the panel were Isobel Armstrong, Joanne Shattock and Harold Short, Jim Mussell and Suzanne Paylor (deputizing for Gerhard Brey). Isobel discussed her own encounter with the digital text of the *Monthly Repository*, and compared this experience with the earlier one that informed her reading of the title in her *Victorian Poetry*. Harold, Suzanne and Jim's presentation complemented Isobel's as it discussed a further form of information retrieval from large textual corpuses, text mining. Joanne's presentation related to Thomson Gale's *19th Century Periodicals*, a large resource of nineteenth-century periodicals due to be launched later this year. The two plenaries were from Ed King and Hilary Fraser. Ed, Newspaper Librarian for the British Library at Colindale, surveyed various online resources for historical and contemporary newspapers. This talk offered an interesting contrast with Hilary Fraser's, which considered how the periodical embodies and represents time as both artefact and archive. The day also featured an opportunity to explore a portion of the *Leader* through Olive's Active Paper Archive. Although this is not the

Taken from ''*Publishers' Circular*, 50, 1 October 1887, p. 1996.

application we are going to use in the edition, it was useful to gauge participant's opinions as to its functionality.

Viewpoint

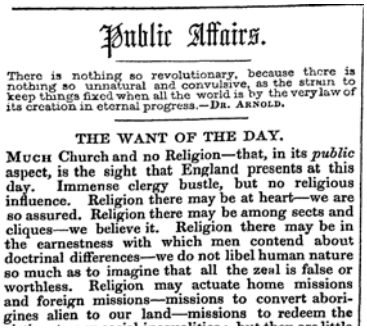
The team have continued to work with Olive Software on their new application, Viewpoint. Although it will not be ready until August 2007, we have seen some mock-ups to enable us to specify the functionality that we require. As mentioned in the last newsletter, we are using Viewpoint to combine the functionality of two existing products, Active Paper Archive, which is designed to handle newspapers, and File Cabinet, which is designed to store and display electronic books.

As the development of Viewpoint has been in parallel with the project, we have been able to contribute to its design and ensure that its functionality corresponds with the demands of presenting visually-rich, complex pages from nineteenth-century serials. Viewpoint will allow users to browse the edition, undertake complex searches across all its contents, view pages or the individual components that make them up. The Olive application will be situated within a larger resource designed by CCH that will contain the various contextual material produced by the project team. We are currently designing the interface between these two components, especially with regards to the indices that we are constructing through text mining. These will provide users with alternative ways to access the material such as by subject, and provide digests of the people, places, events, and publications that are mentioned within its pages.

Processing

As reported in our last newspaper, the **ncse** project team are in the midst of transforming simple pdfs into segmented numbers. This process creates facsimile pages and OCR transcripts, but encodes the structural hierarchies in xml documents. Although each title is segmented so as to conform to the ncse architecture:

Edition > title > volume > number > department > item



Public Affairs.

There is nothing so revolutionary, because there is nothing so unnatural and convulsive, as the strain to keep things fixed when all the world is by the very law of its creation in eternal progress.—Dr. Arnold.

THE WANT OF THE DAY.

Much Church and no Religion—that, in its *public* aspect, is the sight that England presents at this day. Immense clergy bustle, but no religious influence. Religion there may be at heart—we are so assured. Religion there may be among sects and cliques—we believe it. Religion there may be in the earnestness with which men contend about doctrinal differences—we do not libel human nature so much as to imagine that all the zeal is false or worthless. Religion may actuate home missions and foreign missions—missions to convert aborigines alien to our land—missions to redeem the

'Public Affairs', ***Leader*****, 1, 13 April 1850, p. 58.**

The way this is implemented in each title varies. For instance, in the *Leader* there are well-differentiated hierarchies within each number. In the example below, 'Public Affairs' announces a department, and is followed by a motto and the first article, 'The Want of the Day.' As **ncse** treats all components on the page as items, this example actually contains three items – the department header, the motto, and the article. The process at Olive identifies these as items and lists each in the 'Table of Components' that will appear on the left of the screen. However, these items actually represent a hierarchy as 'Public Affairs' marks a department that contains the motto and the articles that follow. By instructing the system to recognize gothic type, we can distinguish between those items which are department headers and which are not. By labelling items that do not have headings in gothic type 'untitled article' (the default heading where one does not appear), it is

possible to strip out all the items and leave a 'Table of Components' that contains only those items that correspond to department headings. As the actual contents of the 'Table of Components' will be a small image from the page, users will be able to see the typographical features through which serials signal their structure.

SEPTEMBER.

Table of contents, *English Woman's Journal*, 1, 1858, unpaginated.

I.—ON THE ADOPTION OF PROFESSIONAL LIFE BY WOMEN.

WE do not propose to consider in these pages the theory of woman's mission. It is a vexed question which will not be settled by words, nay, which words have rather a tendency to embitter, and we do not imagine that any reluctant mind was ever argued into a belief that it was good for a woman to leave her own fireside. Two only means of conviction can be employed with success, the presentation of facts

Anonymous, 'I. – On the Adoption of Professional Life by Women', *English Woman's Journal*, 1, September 1858,

However, two of our titles – the *English Woman's Journal* and *Tomahawk* – do not really employ departments as organizational categories. As you can see from this table of contents, the *English Woman's Journal* simply consists of a series of articles. Some of these, like 'Open Council' are like departments in that they reappear in every number and contain a series of articles. However, they are represented typographically as structurally equivalent to the other essay-type articles in the number. As the page from *Tomahawk* shows (below), it consists of a miscellany of stand-alone articles. The cartoon does function a bit like a department as it appears in the same location of every number, but it is the only component that does. In the terms of our hierarchy, we have to consider each article as a separate department. This allows us to represent the department structure where it appears, and consider those items where it does not as being departments that contain only one item. For instance, in the example shown left, 'I. – On the Adoption of Professional Life by Women' is the item that we identify as a department header (and so appears in the 'Table of Components'), but there is only one item within the department, i.e. the text that follows ('We do not propose…').

214 THE TOMAHAWK. [MAY 30, 1868.

object of its devotion as of its own. We are prepared for the howl of furious execration from those who are the sole objects of their own patriotism, and whose desire it is to see the Crown reduced to the utmost insignificance and such a precedent established as may pave the way for its entire abolition. These may extravagantly laud the woman at the expense of the Sovereign. We think Victoria can appreciate their devotion for what it is worth; for our own part we are sure that the course which we advise is the happiest way of escape from a threatening cloud of unhappiness which grows larger every day.

Released from the ties of ceremonial duties, relieved of a source of continual disappointment and vexation, and purged at once from all suspicions, however ungenerous, our beloved Queen will be able to enjoy an honourable retirement, cheered by the undimmed affection of her subjects—a peace, let us trust, undisturbed by any private or public trouble. She will be able to revel in the congenial solitudes of Osborne or Balmoral without any reproach, and to devote her leisure time to any pursuits which her inclination may select, and, encouraged by her previous success, to give to her country a history of her life as Queen as well as wife, which may add one of the most valuable treasures to the store-house of History.

CUM GRANO.

THE set of ruffians, cut-throats, bankrupts, and savages who have assumed the *rôle de gloire* of the "Republic of Mexico" have not of late forced themselves much upon public notice. However, they are once more apparently beginning to show signs of life of a certain sort, for a week or two ago they were knocking at the door of the Foreign Office in the hopes of getting an acknowledgment; and only the other day they managed to monopolise a whole trans-Atlantic telegram to themselves. The despatch in question informed Europe that the Congress of the Republic of Mexico had abolished the punishment of death. Notwithstanding the fact that this is the same body that recently declared the "execution" of Maximilian "illegal," we are obliged to regard its judgments from a very unamiable point of view. A set of drunken coal-heavers passing a resolution condemnatory of beer-drinking and swearing would have greater claims upon our serious attention. In short, the announcement that "capital punishment" is extinct in Mexico reads like a good joke, and if we are forced to take it seriously, we can only do so on the supposition that a wise legislature has come to the conclusion that if the law is to claim the life of every murderer in Mexico, there will very soon be an end of the republic altogether. This sounds illiberal, but it is nothing of the kind. As some portion of Italy, and nearly the whole of Greece, may be regarded as the sink of Europe, so Mexico, in an aggravated degree, discharges the same dirty function on the American Continent. It is, in a word, the worst place for a respectable man to set up house on the face of the globe, civilised or the contrary. Such news from such a place suggests a good deal in the shape of analogy, and those who have been good enough to take it in confidently and cheerfully, had better prepare themselves for one or more of the following announcements with the least possible delay:

Archdeacon Sinclair has written something in the manly, terse, and convincing style of Swift and Johnson, every copy of which has been bought up by the working classes.

Turkish 3 per Cents. are at par.

Mr. Disraeli has refused to eat his own words, and has joined the Conservative party.

Prussia has ensured the peace of Europe by disbanding another 10,000 men. As a further guarantee for his pacific intentions, Count Bismarck has had the military chemists severely reprimanded, their specimens of solidified nitro-glycerine taken away from them, and all their apparatus put into the fire.

Park lane was safely traversed three times yesterday. A brewer's dray, four-wheeled cab, and watering cart, successively crossed from Piccadilly to the Marble arch without breaking their horses' legs.

A controversial meeting, for the glory of God and the spread of true Christianity, has been held in the north of England, at which only fifteen people received gun-shot wounds, twenty-nine serious injuries, and thirty-five were carried to the hospital. As the Riot Act was read only five times, the disturbance lasted only three days, the destruction of property was limited to only 215 Irish houses, and lastly, as only 200 special constables were sworn in, and 300 military summoned from a distance, this may be regarded as one of the most orderly and edifying things of the kind that have occurred in the locality for some time past.

A London statue has just been put up which the best critics have pronounced to be not "boisterously comic," but merely "quietly funny." It is like somebody.

The Court will not leave town till the beginning of August.

A debate has occurred in the House of Commons, the course of which was interrupted by no personal attack, oath, scuffle, stand-up fight, song, bet, charge of dishonesty, or unconstitutional manœuvre. The matter under discussion was merely the Herring (Newfoundland) Fisheries Bill, but the occurrence is unprecedented.

The Ritual Commissioners have not all gone mad.

The hero and heroine of the new novel in *Once a Week* will not pass their honeymoon in Messrs. Charles Reade and Dion Boucicault's undiscovered island in the South Pacific.

Thanks to the precautions taken by the Government, there will not be any serious rioting in the country in the course of the present year.

A PEOPLE'S PARADISE.

MR. LABOUCHERE, one of the Members for Middlesex, has at length raised the question in the House of Commons if cabs are to be admitted within the gates of Hyde Park. The Honourable Member, in urging his case, pointed out, amidst some fifty other excellent arguments, that London was the only city in the world where the chief promenade was exclusively devoted to the service of the upper ten thousand; and that while in Paris, all classes of society have free access to the Bois de Boulogne in any kind of *voiture* their fancy or resources may suggest, in London a large body of the tax-paying British people are, by the exclusion of their national vehicle—the four-wheeler—debarred from the enjoyment of healthful recreation in a public Park, for the maintenance of which they are called upon to provide. With so much force has Mr. Labouchere argued his cause, that already the Commissioners of Woods and Forests have felt themselves bound to take the matter into their serious consideration; and with a view to rendering Hyde Park in future an agreeable resort for all classes of Her Majesty's subjects, the following regulations have been framed, which will come into action as soon as an additional thousand constables have been added to the Metropolitan Police Force to superintend their observance.

Regulations to be observed in the admission of vehicles into Hyde Park, and rules for rendering the Park a convenient resort for all classes of the public.

1.—Cabs, either with or without occupants, may enter the Park at all the gates. Cab stands will be formed at each extremity of Rotten Row, but cabs may ply for hire in the principal thoroughfares at the option of their drivers.

2.—The Kensington omnibuses will henceforward enter at Queen's gate, and the Brompton line will enter at Albert gate, all omnibuses leaving the Park at Hyde Park Corner.

3.—Heavy waggons conveying coal, stone, or merchandise, which can only proceed at a slow pace, may pass through the Park, but must take their place in a rank which will be formed in all the drives near the railings. Light vehicles, butchers' carts, &c., will only be allowed within the gates provided they travel at a speed of not less than twelve miles an hour.

Page from *Tomahawk*, *Tomahawk*, 2, 30 May 1868, p. 214.

For both the *English Woman's Journal* and *Tomahawk* the 'Table of Components' therefore gives much fuller information about a number's contents than in the other titles: even though it still gives a list of departments, many of these departments are actually synonymous with articles they contain.

Text mining and the Waterloo Directory

There has been considerable progress on the text mining component of **ncse**. Because it contains well over 100,000 pages, there is far more material within the edition than could be marked up by hand. At the Centre for Computing in the Humanities at King's College London we are attempting to do two things: to identify names (possibly including those of institutions), places, events and publications; and to produce a form of subject index based upon our concept map. As our texts are derived from uncorrected OCR they contain quite a number of illegible or nonsensical words. Some titles are better than others: so far the *English Woman's Journal* has produced a text of which 90% of the words are recognizable as genuine, but we expect this percentage to be much lower on those titles with denser print.

We have begun to use GATE (http://gate.ac.uk/) to produce lists of names from the **ncse** corpus. These lists can be generated from rules, or by comparing words against authority lists. Thanks to the generous assistance of John North, we have been able to use the indices from the *Waterloo Directory of Victorian Periodicals* (http://www.victorianperiodicals.com/series2/default.asp) as a source of authority lists, allowing us to differentiate between certain names and further demarcate the data. The *Waterloo Directory* is the most exhaustive reference work on nineteenth-century serials in the world, currently providing information on over 50,000 periodical titles. As such, its indices are invaluable: not only do they contain over 48,000 different personal names, but these names are connected through their relationship to the press.

Poster

In preparation for the launch we have overhauled all of our existing publicity material. With the assistance of Damien Doherty (http://www.damiendoherty.com/) we have designed a large poster for display at conferences, and a new set of A4 posters and A5 leaflets for distribution. Electronic copies will be available from our website, but if you want paper copies then contact either Suzanne Paylor or Jim Mussell.

* * *

Future work

Within the next (and final) six months the **ncse** team will undertake the following:

- **Complete the processing**

Processing has been underway in earnest since last October. This has been a difficult process as the material requires a high level of editorial supervision at every level. As a result, we have had to revise production estimates and the last title, the *Monthly Repository*, will go into production by the end of July.

Once the titles have been processed they need to be checked in order to gauge the success of the segmentation. This is a time-consuming process, but we have recruited a team of postgraduates to assist us in this important task.

- **Further explore text mining**

The preliminary results achieved through text mining are promising. As mentioned above, we hope to use text mining techniques to create indices of names, places, publications and events to help people browse the edition. We are also going to try and use rules-based approaches to identify types of content and what they are about. By searching for certain text strings or clusters of words we hope to find certain categories of text, such as letters or poems. Using similar techniques to analyze clustering of certain terms, we hope to be able to extract keywords from the text in order to cross reference them against our concept map. This work is experimental at the moment, and will deliver interesting methodological results as well as something that we can use to assist in populating the edition with metadata.

- **Implement metadata**

Much of the metadata within **ncse** is derived from the processing. For instance, the date and page number of every item is recorded in the xml. However, there are some metadata categories that have to be entered by hand such as volume numbers or the specific title of a serial at a particular moment of its run. As detailed in the last newsletter, we can cascade metadata down through the hierarchy, making additions to the metadata for large numbers of items by altering a field at a higher level. There is also metadata that we hope to input via automated means: for instance, information that is in the edition, but not identified by the Olive application. The text mining work is partly an exercise in developing practical sources of metadata from the edition: what we are currently doing is working out a strategy that will allow us to allocate resources to implement any metadata that it produces, alongside that which we already have to hand. As we are in our final months, we are working to ensure as much of our schema will be implemented as possible.

- **Continue to develop both the Olive application and the resource as a whole**

As the processing draws to an end, we will increase the amount of work undertaken in London to develop the edition as a whole. The development of the Olive application is well underway, and on track to be released for trial by **ncse** in August. We are beginning to think seriously about the design of the

resource within which the application is the central part, working out search strategies, and the overall layout of the other contents.

- **User testing**

An important part of this work will be based upon feedback from our intended users. We are hosting a user testing session on the 7 September 2007 at 12pm at CCH. Lunch will be provided from 12-2, and participants will be free to work through the resource over some of Jo Foster's excellent organic breads and cheeses. There will be a plenary at 2pm in which participants can ask questions of the **ncse** team.

- **Complete contextual essays and user documentation**

As we enter the final phase of the project, our attention is again focused on the other materials that we need to produce prior to the launch in May. These include discursive headnotes for each of the periodicals, the text that will be on the site (in both the Olive application and the resource as a whole), and user documentation that will provide both technical details and a methodological account of the project.

* * *

If you would like any further information, or wish to contact the project team, please visit our website:

www.ncse.kcl.ac.uk

Jim Mussell and Suzanne Paylor

Nineteenth-Century Serials Edition (**ncse**)
Faculty of Continuing Education
Birkbeck College
26 Russell Square
London
WC1B 5DQ

Project Director: Professor Laurel Brake