Doing and Making: History as Digital Practice

Mussell, James

[This was published as James Mussell, 'Doing and Making: History as Digital Practice', *History in the Digital Age*, edited by Toni Weller (London: Routledge, 2013), pp. 79-94. Please cite the published version.]

Doing and Making: History as Digital Practice

> Like all media revolutions, the first wave of the digital revolution looked backward as it moved forward. Just as early codices mirrored oratorical practices, print initially mirrored the practices of high medieval manuscript culture, and film mirrored the techniques of theater, the digital first wave replicated the world of scholarly communications that print gradually codified over the course of five centuries: a world where textuality was primary and visuality and sound were secondary (and subordinated to text), even as it vastly accelerated the search and retrieval of documents, enhanced access, and altered mental habits. Now it must shape a future in which the medium-specific features of digital technologies become its core and in which print is absorbed into new hybrid modes of communication.[1]

In 'A Digital Humanities Manifesto 2.0', the authors argue that the digital revolution has entered a second phase, in which digital objects (and environments, tools and technologies) are considered on their own terms, rather than as derivatives or surrogates for those from the nondigital world. In the first phase, the manifesto suggests, the digital

[1] Todd Presner and Jeffrey Schnapp, 'A Digital Humanities Manifesto 2.0' (2009). Online. Available HTTP: <http://manifesto.humanities.ucla.edu/2009/05/29/the-digital-humanities-manifesto-20/> (accessed 20 August 2011).

revolution reproduced versions of print forms and the disciplinary apparatus that sustained them and gave them meaning. The second phase, alive to the specificities of different media (and the technologies upon which they depend), decenters print and so reconfigures the conditions under which the disciplines produce and codify knowledge. In this chapter, I examine how historians might engage with and benefit from this next phase of the digital revolution. Historians of all kinds already practice digital scholarship, whether this is composing papers using word processors, communicating via email and *Twitter*, or using digital resources to locate and access documents of various kinds. Historians are also actively building resources, perhaps collaborating with or identifying themselves as digital humanists. Digital resources, tools and technologies have become integral instruments through which we interrogate and understand the past. As these instruments continue to change, so too does the practice of history.

The manifesto is a necessarily provocative document that heralds the digital humanities as the set of methodologies necessary to reimagine scholarship in the digital age. One of its authors, Todd Presner, calls the rejuvenated digital humanities, liberated from their previously servile position, the 'digital humanities 2.0.'[2] While wary of such predictions, I think there is a case for imagining a corresponding 'digital history 2.0.' The manifesto posits a world in which print 'is no longer the exclusive or the normative medium in which knowledge is produced and / or disseminated' and 'digital tools, techniques, and media have altered the production and dissemination of knowledge in the arts, human and

[2] Todd Presner, 'Digital Humanities 2.0: A Report on Knowledge', *Emerging Disciplines*, edited by Melissa Bailar, Houston: Rice University Press, 2010. Online. Available HTTP: <http://cnx.org/content/m34246/latest/> (20 August 2011).

social sciences.'[3] While the digital humanities, described by the manifesto as 'an array of convergent practices', are undoubtedly well-placed to interrogate and participate in such a world, the manifesto, in its bid for disciplinary space, overlooks the extent to which other disciplines have a stake in the digital and can offer frameworks within which its significance and meaning can be understood.[4] This revolution is, after all, an historical event, its momentum sustained by a set of contingent circumstances that are open to analysis. It is the transformation of our heritage accomplished through media transformation, a process subject to scrutiny in a number of humanities disciplines. If the digital humanities are to have an influence upon the more established disciplines in the humanities, it can only be through collaboration and this means that influence will work in both directions. The transformation from digital humanities 1.0 to 2.0 was initiated by broader shifts in digital culture but it is sustained by continued interactions with the older disciplines of the humanities. Rather than a messianic digital humanities 2.0 rejuvenating these disciplines for a new era, I see digital humanities 2.0 as one of the many disciplinary transformations that will result when existing expertise is brought to bear on new objects and methods.

In what follows I assume that digital resources of various kinds are already integral to historical practice but argue that if the discipline is to take full advantages of the digital revolution, then it must engage more closely with the digital properties that give these resources their character. If the first phase of the digital revolution focused on the computer's capacity for simulation, then the next phase will be the result of once more

[3] Presner and Schnapp, 'A Digital Humanities Manifesto 2.0', unpaginated.
[4] Presner and Schnapp, 'A Digital Humanities Manifesto 2.0', unpaginated.

making it strange. The chapter is arranged in two parts. In the first I describe what I understand as digital history 1.0 and set out how this would differ from digital history 2.0. Taking my own field, nineteenth-century media history, as an example, I describe this transition as one predicated on a shift from documents to data. In the second part, I turn to the resources that will effect this transition arguing that they are not just tools, but legitimate objects of historical enquiry in themselves. One of the questions raised by the shift from documents to data concerns the status of the archive. As long as the archive is considered distinct from historical practice, a static set of documents against which history refines itself, then digital resources can only ever be instruments that provide access. However, if we recognize that in transforming the archive and rendering it processable it becomes something different, then these resources become constitutive parts of the archive and so subject to analysis in their own right. At stake in the shift from digital history 1.0 to 2.0 is the recognition that the traditional techniques of historical scholarship remain relevant, but that they are also necessarily transformed by the radical reconstitution of the archive. The name 'digital history 2.0' marks a break from digital history 1.0, but it remains history nonetheless.

From documents to data

The shift to history 2.0 requires a change in focus from document to data. It is the ability of the digital to sufficiently represent other media while bestowing upon them a particular form of materiality that has ensured digital resources are widely used in the humanities. However, the facility with which digital media can simulate nondigital forms means that

it is easy to overlook the extent this depends upon digital properties. As N. Katherine Hayles has argued in a discussion of digital editions of printed texts, this amazing capacity for simulation is only possible because the computer is 'completely unlike print in its architecture and functioning'.[5] The more successful the reproduction, the easier it is to be seduced by the simulation and so treat the simulated media as if it was the nondigital material. The same might also be said for born-digital objects. As many applications exploit a repertoire of learned behaviour online, resources tend to correspond to recognized genres in order not to bewilder their users. Just as those resources that translate nondigital material attempt to reproduce their corresponding – and so familiar – patterns of use into the digital environment, so resources using born-digital material appropriate and exploit conventions from the digital environment and beyond in order to create a recognizable interface. The shift to history 2.0 depends, to an extent, on using digital resources against the grain of their interfaces in order to access the data they contain. It is a shift that depends upon defamiliarization, on recognizing what is distinct about digital media and technologies and then exploiting this digital difference for scholarly ends.

My field, nineteenth-century media history, is very much entrenched in digital history 1.0. The industrialized presses of the nineteenth century produced an enormous amount of material, in a wide range of forms, issued in single editions, in parts, and in reprints, for a variety of audiences. Large amounts of this material have survived in the archive, but in a fragmentary condition. Despite some herculean bibliographic work (the

[5] N. Katherine Hayles, 'Translating Media: Why We Should Rethink Textuality', *The Yale Journal of Criticism*, 16, 2003: 264.

Wellesley Index, the various series of the *Waterloo Directory*, the *Dictionary of Nineteenth-Century Journalism*),[6] the archive remains difficult to work with. Many publications survive in runs too long to read; they are almost always incomplete, perhaps because not all issues survive, but also because they were routinely transformed on accession; and all periodicals and newspapers demand a high degree of contextual knowledge from the researcher, often requiring them to work between and across disciplines with a diverse, and nearly always, unsigned text. As a result, research into the press has necessarily been patchy, focusing on particular publications or people and relying upon existing disciplinary structures (certain authors, types of text, key events etc), at the expense of a rigorous analysis of the mechanisms of journalism or publishing more broadly.[7]

Digital resources of nineteenth-century newspapers and periodicals have existed online since the publication of resources such as Cornell University's and the University of

---

[6] *Wellesley Index to Victorian Periodicals, 1824-1900*, edited by Walter E. Houghton, 5 vols, Toronto: University of Toronto Press, 1966-1979; *Waterloo Directory of English Newspapers and Periodicals, 1800-1900*, edited by John S. North, first series, Waterloo: North Waterloo Academic Press, 1994. CD ROM; *Waterloo Directory of English Newspapers and Periodicals: 1800-1900*, edited by John North, second series, Waterloo: North Waterloo Academic Press, 2003. Online. Available HTTP: <http://www.victorianperiodicals.com/> (accessed 20 August 2011); *The Dictionary of Nineteenth-Century Journalism*, edited by Laurel Brake and Marysa Demoor, Gent and London: Academia Press and the British Library, 2009.

[7] One of the strengths of the field is its methodological reflexivity and these bibliographic challenges are well documented. See, for instance, Michael Wolff, 'Charting the Golden Stream: Thoughts on a Directory of Victorian Periodicals', *Victorian Periodicals Review*, 13, 1971: 23-8; Scott Bennett, 'The Bibliographic Control of Victorian Periodicals', in *Victorian Periodicals: A Guide to Research*, edited by J. Don Vann and Rosemary T. VanArsdel, New York: Modern Language Association, 1978, pp. 21-51; Joanne Shattock and Michael Wolff, 'Introduction', in *The Victorian Periodical Press: Samplings and Soundings*, edited by Joanne Shattock and Michael Wolff, Leicester: Leicester University Press, 1982, pp. xiii-xix; Laurel Brake, Aled Jones and Lionel Madden, 'Introduction: Defining the Field', in *Investigating Victorian Journalism*, edited by Laurel Brake, Aled Jones and Lionel Madden, Basingstoke: Macmillan, 1990, pp. xi-xiv; Laurel Brake, 'Tacking: Nineteenth-Century Print Culture and its Readers', *Romanticism and Victorianism on the Net*, 55, 1-44. Online. Available HTTP: <http://www.erudit.org/revue/ravon/2009/v/n55/039555ar.html> (accessed 20 August 2011).

Michigan's *Making of America* (1995), ProQuest's *Periodicals Contents Index* (1997), the *Internet Library of Early Journals* (ILEJ, 1999) and Heritage Microfilm's *newspaperARCHIVE* (1999). Many of these early resources were ambitious in scope, but they have been surpassed by more recent resources such as ProQuest's *ProQuest Historical Newspapers* (2001-) and *British Periodicals* (2007), the British Library's *British Newspapers 1800-1900* (published by Gale Cengage (2007) and also called *19th Century British Library Newspapers*), and Gale Cengage's *19th Century UK Periodicals* (2007). More recently, the British Library and Brightsolid have published the *British Newspaper Archive* (2011-), providing access (at a cost) to four million pages, with more to come up to a projected forty million. Over the last twenty years we have gained access to hundreds of publications from anywhere with a web browser (and, in most cases listed above, appropriate access rights).

These digital resources have transformed the field by altering the conditions of access to its primary materials. Built around searchable keyword indices, nearly always developed from uncorrected transcripts produced from optical character recognition (OCR) technologies, these resources subject the archive to a degree of bibliographic control. Not only do they render individual publications searchable but, by using text as the basis for the index, create the conditions for a cross-searchable database that can open up the press as a whole. There were printed reference resources that provided a degree of access across the archive, and many individual publications produced indices for volumes of periodicals as they were produced, but there was nothing like the coverage provided by a single one of the new digital resources. The ability to cross search not only makes it easy

to locate content, but also allows it to be traced across publications, exposing many of the connections occluded when reading one publication at a time. Although users might be drawn towards certain events or historical figures, the search engine's algorithms will return hits that it predicts are relevant, often returning familiar figures in unfamiliar contexts or providing articles that supply an unexpected perspective on an event. Search can disrupt existing hierarchies, complementing major figures, events or themes with a host of others who might otherwise be overlooked. As the search engine does not discriminate between types of content (unless instructed by the user), it also has the potential to displace or supplement existing canons of periodicals or newspapers, reminding users, for instance, of the diversity of periodical publication, or the importance of the provincial press. As portions of page images are returned (to compensate for the errors in the transcript), the resources reproduce the bibliographic codes on the page, providing access to the typeface, layout and any other visual features that constitute the printed text. By making its verbal content processable, these resources manage the scale and complexity of the print archive. They are designed to allow researchers to locate and recover articles in a form that reproduces the appearance of the printed page. The principle gains are efficiency and access: tasks that would have been prohibitively labour-intensive are routine; and the archive itself is now available to more people, in many more locations.

Digitization is a radical transformation of material form and so takes place in an economy of loss and gain. However, the rhetoric of surrogacy that underpins many of these resources masks the extent to which they differ from the printed material in the archives.

There is much that can be learned about the press through the process of digitization, but for users, who are often positioned as passive consumers of content, there is little that could not be learned from looking at the appropriate hard copy. In fact, as is frequently noted, there are aspects of printed media such as weight, texture, smell and (to an extent) size that can only be appreciated by considering the material in the archives. In exchange for the non- or partial representation of various material aspects of printed objects, the user gains considerable increases in searchability and accessibility. These gains are predicated on the material properties of the digital objects produced during digitization, but the user is not in a position to evaluate these processes or their efficiency. For instance, as is well-recognized, OCR, despite the claims of its vendors, rarely achieves the success rates in recognizing text from historical newspapers, usually due to the condition of the surviving hard copy.[8] This means that the searchable index only provides a partial representation of the text printed upon the page and reproduced on the scanned facsimile images. Yet very few resources allow users to see the OCR-generated transcript upon which their search queries are executed. Equally, even though many resources compete on scope, the corpus of publications that make up their contents is skewed towards certain types of publications in certain types of institution. Again, despite this, commercial vendors seldom provide information about the derivation of the content in resources and rarely provide contents lists so that users can ascertain what, exactly, they are searching. In both cases this data is present but is witheld from users so that they can get on with searching for and reading articles, one by one. These resources

[8] Simon Tanner, Trevor Munoz and Pich Hemy Ros, 'Measuring Mass Text Digitization Quality and Usefulness: Lessons Learned from Assessing the OCR Accuracy of the British Library's 19th Century Online Newspaper Archive', D-Lib Magazine, 15, 2009: unpaginated. Online. Available HTTP: <http://www.dlib.org/dlib/july09/munoz/07munoz.html> (accessed 20 August 2011).

are designed to mimic (a version of) the print object upon which they are based and so, often deliberately, do not allow their users to take full advantage of the data that they contain.

Despite the long-standing interest in this material (because it is free of troublesome rights issues), the core methodology and functionality of the resulting digital resources have remained remarkably consistent for almost twenty years. Nearly every single resource offering access to nineteenth-century British newspapers and periodicals privileges search over browse, redefining the serials on the library shelves as a database of discrete articles. This means that users are expected to be looking for something and that this can be mapped to the occurrence of words in a transcript and described in a search query. Serendipity still applies, as the scale of the archive means that the user, with little sense of the archive as a whole and restricted to browsing lists of metadata to determine which hits to read, will probably be surprised at what has been returned. Yet restricting the use of the underlying data to an index, which can only be queried in ways delimited by the interface in order to return articles, limits the interpretive potential of this data.

Print has a foundational relationship with repetition, and so printed objects lend themselves to computational analysis. Corpus linguistics has been connected with humanities computing since the work of Father Busa in the 1940s, but its techniques and methods have been marginalized in media history, particularly that carried out in English departments.[9] Wedded to close reading yet aware of the abundance of material in the

[9] For the history of humanities computing see Susan Hockey, 'A History of Humanities Computing', in *A Companion to Digital Humanities*, ed by Susan Schreibman, Ray Siemens and John Unsworth (Oxford:

archive, research has proceeded via detailed analyses of isolated case studies, whose significance is evaluated against an extrapolated print culture. This is not to suggest that quantitative analyses have had no effect on media history, or that digital resources and tools have not played a part in the analysis.[10] Rather, there has been an institutionalized preference for the exceptional – what makes a particular text or publication important or different – over the repetitive and generic. Like the historiographical preference for the study of great men or the emphasis on particular social classes, this methodological bias has ensured that some printed objects are preserved over others, and of these, a select few deemed worthy of analysis within the academy. Yet without the tools and methodologies to interrogate the repetitive (and so the generic and the abundant), analysis is restricted to generalizations based on the exceptional without really establishing the grounds for exceptionality within the culture of the period.

Such an orientation makes it difficult to understand the patterns that characterized both the production of journalism and the textuality of print culture. Corpus linguistics is usually concerned with naturally-occurring language, but the marked-up verbal content that underpins large digital resources of newspapers and periodicals can be mined to reveal things about print culture more broadly. The neglect of the tools and techniques of computational linguistics by those in literary studies has often been noted, but Franco Moretti's call for 'distant reading' has usefully turned the attention of the discipline to the

---

Blackwell, 2004), pp. 3-19. Online. Available HTTP: <http://www.digitalhumanities.org/companion/view?docId=blackwell/9781405103213/9781405103213.xml&chunk.id=ss1-2-1> [accessed 20 August 2011].

[10] See for instance Simon Eliot, 'Some Trends in British Book Publication, 1800-1919', in *Literature and the Marketplace*, edited by John O. Jordan and Robert L. Patten, Cambridge: Cambridge University Press, 1995, pp. 19-43; William St Clair, *The Reading Nation in the Romantic Period*, Cambridge: Cambridge University Press, 2004.

bulk of material that necessarily remains unread.[11] If literary texts offer a rich ground for statistical analysis, then journalistic texts, published systematically in serial parts, provide an even better data set.[12] As many scholars have noted, newspapers and periodicals negotiate between novelty and familiarity in order to satisfy the demands of their readers. Each issue offers something new, but this novelty was tempered by familiarity: what was provided must be more of the same. As Margaret Beetham has put it, each 'number is different, but it is the same periodical'.[13] The repetition of various features, article to article and issue to issue, created a formal identity for a publication that enabled it to transcend its particular instantation, reassuring readers that they were still reading the same title despite changing content. Recurring features such as mastheads, layout, typeface created a visual consistency that linked articles and issues and the reappearance of certain types of articles in regular positions, whether on the page or in the issue, further entrenched the identity of the publication. Although these repetitions were intended to mark the identity of an individual title, they were themselves part of larger patterns, identifying the style of contributing authors, situating articles within textual genres, and publications within the wider market. Statistical analysis allows us, for the first time, to map the interplay of print and textual genres that enabled newspapers and periodicals to function as commodities in the competitive market for nineteenth-century print.

---

[11] Franco Moretti, 'Conjectures on World Literature', *New Left Review*, 1, 2000: 54-68. Online. Available HTTP: <http://www.newleftreview.org/A2094> (accessed 10 August 2011).
[12] John Burrows, 'Never Say Never Again: Reflections on the Numbers Game', in *Text and Genre in Reconstruction: Effects of Digitalization on Ideas, Behaviours, Products and Institutions*, edited by Willard McCarty, Cambridge: Open Book Publishers, 2010, pp. 13-36.
[13] Margaret Beetham, 'Towards a Theory of the Periodical as Publishing Genre', in *Investigating Victorian Journalism*, edited by Laurel Brake, Aled Jones and Lionel Madden, Basingstoke: Macmillan, 1990, 28.

The digitization of the archive has produced a corpus of data large enough to reveal this systemacity while compensating for errors contained in the transcripts. At the moment, scholars can interrogate the transcripts and readily get a sense of how textual repetition operates within the archive, but only one search at a time and with no easy way to visualize the results. These resources are designed to delimit the archive, providing the 'right' article for the individual user. However, the archive itself constitutes a large data set, and would be profitably approached as such. Whereas individual scholars can search for phrases that reveal how content was reproduced across the press, or how articles or sections recur issue after issue, such relationships could be easily mapped and then visualized over time and space. There are a number ways this data could be usefully explored. The metadata, predominantly used to structure search queries and assist users browsing lists of results, might be interrogated for what it can reveal, for instance, about the amount of articles published per issue (or issues per volume, or volumes per year); periodicities; characteristic titles for articles, sections, publications; structure; relative page spans; etc. This data could be mapped in order to provide a better understanding of different genres in the marketplace, as well how new publications were situated with regards to their competitors. Or the textual information within individual articles might be compared to examine the operation of genre, exposing how different types of article were deployed in particular publications and how these varied according to print genre. Given that issues of periodicals and newspapers are marked with a date, it would be straightforward to explore how publications reacted to changes in the marketplace, perhaps due to the introduction of new technology, significant public events, or shifts in the market. Equally, as nearly all publications are also marked with a place of

publication, it would be easy to examine the print trade in different places, as well as the way these different markets depended on one another for content. These sorts of analyses, crucial for an understanding of the interconnected nature of print culture and the operation of the market, can only be accomplished by manipulating processable data.

At present, scholars are restricted to mining the data, query after query, in order to read articles, one by one. There has, however, been some work on processing the textual transcripts in order to refine the data that they contain. The *Nineteenth-Century Serials Edition* (*ncse*), of which I was one of the editors, used text mining techniques to identify the names of persons, places and institutions within the textual transcripts and apply rudimentary semantic tags to articles.[14] This resource published six nineteenth-century newspapers and periodicals – *Monthly Repository* (1806-1838), *Northern Star* (1837-1852), *Leader* (1850-1860), *English Woman's Journal* (1858-1864), *Tomahawk* (1867-1871), *Publisher's Circular* (1880-1890) – constituting a corpus of just over one hundred thousand pages and organized into around five hundred thousand individual textual components. Recognizing that this represented too much to read, the process of marking up the content of the articles (and other textual components) had to be passed to the machine. In an example of the sorts of collaborations required to do this work, the project was a partnership between experts in print culture at Birkbeck College and King's College London; the British Library, who provided the bulk of the material; a private software company, Olive Software, who delivered a web application and server-side architecture; and the Centre for Computing in the Humanities (CCH, now Department of Digital Humanities) at King's College London, who oversaw the implementation of the

[14] *Nineteenth-Century Serials Edition* (*ncse*), 2008. Online. Available HTTP: <http://www.ncse.ac.uk>

Olive product while carrying out experimental work in data mining. As the project unfolded, it became apparent that there was too much material to process by hand and so computational methods were adopted instead. Scholars at CCH used GATE (General Architecture for Text Engineering) to identify and extract named entities from the transcripts and then a set of gazetteers and further post-processing to refine the proper nouns and resolve possible contradictions. This produced lists of persons, places and institutions from the transcripts that could be appended as processable metadata, providing a searchable field that complemented free-text queries with a more reliable and easily refined dataset while also permitting navigation through cross-reference.[15] The application of semantic tags was more experimental. In collaboration with the University Centre for Computer Corpus Research on Language (UCREL) at Lancaster University, semantic tags were applied to articles according to the frequency and deployment of words and multi-word expressions.[16] This produced a list of tags, derived from a hierarchy of 232 category labels organized under 21 major discourse fields, for each article.[17] As the tags were derived from this ontology, it was possible to present them in a faceted browse interface, allowing users to navigate the materials by adding and removing terms from the hierarchy. The results were mixed: for instance, delimiting the articles by 'the body and the individual', 'health and disease', and 'disease' provides 23452

[15] 'Technical Introduction', *Nineteenth-Century Serials Edition* (*ncse*), 2008, unpaginated. Online. Available HTTP: <http://www.ncse.ac.uk/about/technical.html> (accessed 20 August 2011). For details of the research see Manolis Christodoulakis and Gerhard Brey, 'Edit Distance with Combinations and Splits and its Applications in OCR Name Matching', International Journal of Foundations of Computer Science (IJFCS), 20, 2009: 1047–1068; Manolis Christodoulakis, GerhardBrey, Rizwan Ahmed Uppal, 'Evaluation Of Approximate Pattern Matching Algorithms For OCR Texts', *Proceedings of the 4th Advances in Computing and Technology Conference (AC&T)*, edited by Roy Perryman et al., London: ISGES, 2009, pp. 35–42; Manolis Christodoulakis and Gerhard Brey, 'Edit Distance with Single-Symbol Combinations and Splits', *Proceedings of the Prague Stringology Conference*, edited by Jan Holub and Jan Zdárek, Prague: Czech Technical University, 2008, pp. 208–217.

[16] 'Technical Introduction', unpaginated.

[17] 'UCREL Semantic Analysis System (USAS)', *UCREL Home Page*, University of Lancaster. Online. Available HTTP: <http://ucrel.lancs.ac.uk/usas/> (accessed 20 August 2011).

hits, the first ten of which are for various proprietary medicines advertised in the press. If the role of the interface is to provide articles about disease, then users need to exploit the granularity of the data and drill down to something relevant. However, as a way of discovering something about print culture, the interface is very useful, in this case exposing the reliance of the press on advertisements for drugs and treatments of various kinds.

A more recent example is *Connected Histories*, a JISC-funded resource produced in collaboration between the Universities of Hertfordshire, London and Sheffield, launched in 2011.[18] *Connected Histories* is a portal and research space that allows cross-searching of eleven different resources (at the time of writing) dedicated to British history, 1500-1900. By treating the contents of these different resources as processable (textual) data, the project was able to generate its own indices and so link them together. Like *ncse*, *Connected Histories* used techniques from computational linguistics to identify named entities and then sort them into lists.[19] The interface allows users to construct searches across four separate indices: one consisting of all the words from the constituent resources; and three processed lists of names of people; places; and dates. *Connected Histories* overcomes the divisions between the different digital resources, allowing users to navigate their diverse historical content, but it does so by interrogating something they

[18] *Connected Histories*, 2011. Online. Available HTTP: <http://www.connectedhistories.org/> (accessed 20 August 2011).
[19] The project used ANNIE <http://www.aktors.org/technologies/annie/>, an information extraction system and GATE plugin, to identify named entities and then various gazeteers to refine and sort the indices. See 'About this project', *Connected Histories*, 2011. Online. Available HTTP: <http://www.connectedhistories.org/about.aspx> (accessed 20 August 2011). Tim Hitchcock, 'Towards a New History Lab for the Digital Past', Institute of Historical Research, 2011. Online. Available HTTP: <http://sas-space.sas.ac.uk/2854/1/Hitchcock_-_Towards_a_new_History_Lab.pdf> (accessed 20 August 2011).

all have in common, a set of processable data that, with differing degrees of accuracy, represents their verbal text.

These projects exploit the properties of data so that users can access and manipulate historical information in ways that otherwise would be too laborious or outright impossible. They join the many other digital resources more explicitly oriented around historical data.[20] For instance, The French Book Trade in Enlightenment Europe is in the final stages of creating a database derived from the business records of the Société Typographique de Neuchâtel, a Swiss publishing house that operated from 1769 to 1794.[21] Although these records have been used for studies in the past, as processable encoded data they can be queried in a variety of ways, scaled up or down, and presented as visualizations or lists of figures. Further examples can be found at Stanford's Spatial History Project, where historical tabulated data is combined with geographical information in order to model cultural space.[22] Like all historical knowledge, that produced from encounters with these resources depends upon the manipulation of evidence; however, the power of these resources comes from the imposition of a layer of processable data that allows evidence to be repurposed, often in radical ways. Not only does this allow the historian to work with datasets that might be otherwise too large, complex or distinct, but it is also generative, producing new bodies of evidence as they

---

[20] A good overview can be found in Douglas Seefeldt and William G. Thomas III, 'What Is Digital History? A Look at Some Exemplar Projects', *Perspectives on History*, 2009: 1-7. Online. Available HTTP: <http://digitalcommons.unl.edu/historyfacpub/98/> (accessed 20 August 2011).

[21] 'About the Project', *The French Book Trade in Enlightenment Europe, 1769-1794: Mapping the Trade of the Société Typographique de Neuchâtel*, 2011. Online. Available HTTP: <http://chop.leeds.ac.uk/stn/about.html> (accessed 20 August 2011).

[22] *Stanford Spatial History Project*, 2007-. Online. Available HTTP <http://www.stanford.edu/group/spatialhistory> (accessed 20 August 2011).

are transformed for analysis, and iterative, as datasets are adapted, supplemented, and transformed anew.[23]

It is here that interdisciplinary collaboration becomes vital. Historians must account for the transformation of the evidence base in their analysis, and this necessitates understanding the methodologies and technologies responsible for these transformations. In nineteenth-century media history, such collaboration is undermined by a publishing model that positions the researcher in the role of passive client. Rather than collaborate and innovate, experimenting with data and working it against its grain (what Stephen Ramsay calls the 'hermeneutics of screwing around'[24]), the researcher must be content with programmatic, goal-oriented resources that guard their data and mask their methodologies. Scholarship in media history has tended to generalize about the press from the close reading of select publications. The success of these generalizations has rested on a growing scholarly consensus, recorded in books and journals and reaffirmed, implicitly, by delegates at numerous conferences. Approaching digital resources as data-processing devices (of considerable power) rather than delivery mechanisms for facsimile reproductions will allow us to model and explore this consensus, to probe our assumptions about the field and, certainly, prompt questions that, at the moment, cannot be thought. However, at present media historians remain content to be seduced by the

[23] See, for instance, Peter White, 'What Is Spatial History?', *Spatial History Lab: Working Paper*, 2010, 36. Online. Available HTTP: <http://www.stanford.edu/group/spatialhistory/cgi-bin/site/pub.php?id=29> (accessed 20 August 2011).
[24] Stephen Ramsay, 'The Hermeneutics of Screwing Around; or What You Do with a Million Books', unpublished book chapter, 2010. Online. Available HTTP: <http://www.playingwithhistory.com/> (accessed 20 August 2011).

simulations onscreen, bewildered by the riches of a new archive that continues, resolutely, to serve old methodologies.

Aren't these just tools?

What is at stake in the shift towards history 2.0 is the status of the archive. No serious historian would deny that history is a process and its findings contingent, but often the admission of history's dynamism depends on the tacit assumption that the archive remains static, a fixed point of reference through which history corrects itself. Yet this interpretation depends upon locating the primary materials of history solely within the historical objects that survive. This material is obviously central to historical study, but its significance does not lie locked within the objects and documents on the shelves of libraries and archives. Rather, what appears to be latent significance is the product of critical engagement, where scholars return to objects with revised analytical frameworks, whether derived from new theoretical positions, different sets of data, or simply from a different historical moment. This is not to deny that some frameworks are more durable than others, but that historical significance is a product of discourse rather than intrinsic to anything we inherit from the past.

It is overconfidence in the integrity of historical artefacts that results in the common accusation that digital resources are simply tools. Such comments have dogged digital scholarship for years, judging its outputs by a set of utilitarian criteria that are seldom applied to more traditional scholarly publications. The UCLA manifesto (2.0) has a

section addressed 'to the great **diminishers**', those who disparage digital scholarship (practice or product) as '*just* a tool […] *just* a repository […] *just* pedagogy', but doesn't offer a critique of this criticism in return. From the entrenched perspective of disciplines whose products are narrative accounts, published in stable (because familiar) print publications, it can be easy to overdetermine the division between primary and secondary material. If primary material is imagined as stable, curated in libraries and archives, and impervious to the changing interpretations of scholars, then digital resources created from this material will always be secondary, useful for access or analysis but unworthy of study in their own right. It can be tempting to rely on this distinction to enforce disciplinary boundaries, with history concerned with the objects and documents mediated by digital resources and the digital humanities in the digital aspects of the resources themselves. Yet if digital resources (and what researchers do with them) are understood as constitutive parts of the framework through which historical objects become primary sources, then digital technologies and methods become part of historical studies more broadly. There is still disciplinary space for the digital humanities, but given the widespread digitization of our cultural heritage, none of the established disciplines of the humanities can afford to ignore the digital – whether in terms of resources, technology, methodology or pedagogy – or designate it the sole intellectual terrain of this emerging discipline.

The study of digital data does not take history away from primary sources but rather provides a new context in which these sources might be encountered. This idea is a common one in textual scholarship, a discipline concerned with the transmission of texts

and, because of its concerns with scholarly editing, closely connected with the digital humanities. Textual scholarship might be committed to transmission, passing text from one generation to another, but is nevertheless always interpretive and generative, revealing new things about the text even as it remains putatively the same.[25] To produce the iterative text – a text that declares its prior existence in older print and manuscript forms – it must be carefully produced, its previous documentary witnesses sifted, and its final (but contingent) presentation carefully controlled. Print provides an often unremarked field of continuity for textual transmission, helping to support textual features through the recurrence of certain formal and material conventions. In *Radiant Textuality*, Jerome McGann argues that all editions make embodied arguments about their contents but digital editions, because of the different way in which they model text, can lead editors to imagine what they did not know.[26] For McGann digital publication can expose hitherto unthinkable aspects of textuality as modelled by the printed codex because digital editions 'can be designed for complex interactive transformations.'[27] However, this works both ways: digital publication might liberate editors from the demands of the codex, but it imposes its own material conditions upon textuality that, while opening up / closing down possibilities of representation, reveal hitherto unsuspected aspects of both types of media, print and digital, and their respective relations to textuality. This knowledge might be dialectical, generated through difference but, as McGann notes, it can only be realized in practice.

---

[25] See Peter Shillingsburg, *From Gutenberg to Google: Electronic Representations of Literary Texts*, Cambridge: Cambridge University Press, 2006, pp. 12-24.
[26] Jerome J. McGann, *Radiant Textuality: Literature After the World Wide Web*, Basingstoke: Palgrave, 2001, p. 81-2.
[27] McGann, *Radiant Textuality*, p. 81.

McGann suggests that textual editing illustrates the 'pragmatics of theory', arguing that editions constitute a form of 'poesis' rather than a more speculative, conceptual 'gnosis'.[28] Edited works thus embody a form of applied theory, like works of art or engineering projects. At the same time as providing access to content, these editions reflexively interrogate the problems of mediation while nevertheless recognizing its necessity. For McGann, one of the main contributions digital textualities offer humanities disciplines is this 'poesis-as-theory': the recognition of the intellectual (and creative) work of modelling, mapping, reconstructing and editing.[29] All of these processes are in some way transformative, situating whatever is being represented – whether a document, object, set of historical data, or event – in a digital environment in order to learn something about it. Our cultural heritage, as it survives, is always already abstract, separated from the historical culture within which it was produced and had significance. What digital technologies allow scholars to do is provide new contexts within which this material can function. As programmable, dynamic and responsive environments, they permit scholars to study the emergence of different, unsuspected properties as they emerge in response to changing conditions, or the relationships between different entities as they unfold in time. These digital environments might be considered abstract or artificial, but only if we respect the surviving condition of historical objects as somehow natural. It is the role of history to make absent contexts tangible, to make the imagined virtual, in order to reconstruct the significance of material from the past. Digital technologies provide powerful instruments that do just this, transforming material so that it can function in new

[28] McGann, *Radiant Textuality*, p. 83
[29] McGann, *Radiant Textuality*, p. 83

environments, exposing both unrealized aspects of this material and the unthought assumptions that have hitherto structured our engagement with it.

Digital tools and techniques make apparent the changing condition of historical evidence. Even though scholars are prepared to acknowledge the constitutive role of cultural relations, there is a tendency to consider the archive as a hermetically-sealed space in which historical material can be preserved untouched. The aura of authenticity is cherished as it promises an illusory historicity: by respecting the integrity of historical objects they appear to offer direct access to the past; yet this can only ever be achieved indirectly, by an engagement with the object in the present that, necessarily, changes what it means. This paradox is enacted in the architecture of the museum, where one department imposes stasis while another reinterprets content for changing social conditions. Tim Hitchcock, one of the creators of *Connected Histories*, engaged with precisely this division in his lecture at the launch of the resource in 2011. *Connected Histories* is a direct response to the problem of the 'silo': digital resources that republish historical content in such a way that it is not interoperable with other resources, restricting access to their own respective interfaces.[30] For Hitchcock, though, the most insidious silo is the one that 'suggests that information itself is something to be consulted and collected; that it is an unchanging object of study, rather than a pool of constantly changing stuff that can be interrogated from any angle, and pursued along any trajectory.'[31] *Connected Histories*, he argues, addresses the division between 'traditional

[30] See Jerome McGann, 'Culture and Technology: The Way We Live Now, What Is to Be Done', *New Literary History*, 36, 2005: 71–82; Presner, 'Digital Humanities 2.0: A Report on Knowledge', unpaginated.
[31] Hitchcock, 'Towards a New History Lab for the Digital Past', unpaginated.

forms of criticism and scholarship that assume we can contain data in an internally structured and divided, "library"; and the emerging world of text and data mining, that sees data as a process – something to be played with and analysed on a massive scale, across boundaries of genre and type.'[32] As described above, the resource fully achieves this aim, using the techniques of computational linguistics to provide an added layer of functionality to a set of resources that could otherwise only be consulted individually. It is perfectly possible for users to treat its content as surrogates for the historical material it republishes, carrying out fairly traditional research as if the resource was not there (but hopefully citing it nevertheless). However, the project's plans to publish an application programming interface (API), a piece of software that will make its contents machine-readable, demonstrates its commitment to the idea of history as practice and evidence as dynamic. The API means that others will be able to interrogate the *Connected Histories* indices, reconceiving the data in ways unimagined by the creators of *Connected Histories* and its contributing resources. Not only does this recognize that the objects in the constitutive resources can mean different things in juxtaposition, with *Connected Histories* offering itself as a 'work site' through which these objects can establish themselves and their relation to one another, but it also acknowledges that the presentation of these objects in *Connected Histories* is not the final or definitive representation of this content.[33] By opening up the data within the resource to other uses, the creators of *Connected Histories* imply that this material is not finished, its potential for meaning not restricted to this particular configuration of resources in this particular digital environment.

---

[32] Hitchcock, 'Towards a New History Lab for the Digital Past', unpaginated.
[33] See Paul Eggert, 'The Book, the E-text and the "Work-Site"', in *Text Editing, Print and the Digital World*, edited by Marilyn Deegan and Kathryn Sutherland, Farnham: Ashgate, 2009, pp. 63-82.

Connected Histories provides a good example of a central trend of digital scholarship that has been adopted from the culture of the web more broadly: publish openly and rapidly and then iterate to perfection.[34] Digital resources might provide a rich environment within which to manipulate data, but it is only one environment and will have been designed to model data in particular ways. By publishing the data, especially in machine-readable formats, it can be taken up and reused by other resources, placing it in new contexts that can reveal unexpected properties and relationships. These transformative uses will inevitably provide new perspectives on the data, perspectives currently unimaginable because the environments within which data becomes meaningful do not yet exist. The challenge for the digital historian is to understand these uses and reuses and account for them. The digital historian 1.0, using a digital resource to access representations of historical objects or documents, must be able to understand why data performs as it does, why certain material is returned and what might be done with it. This is a process of reconstruction, of compensating for the way the digital resource misrepresents the authentic original. The digital historian 2.0 requires a more advanced understanding of the affordances of the digital in order to perform more advanced research. In manipulating data from multiple resources, modelling their relationships and so exposing facets hitherto unrealized, the historian moves from simulation to simulacra, to validating representations against reified originals to producing analyses of phenomena, objects and relationships that belong to the past. History concerns the

[34] Dan Cohen, 'The Ivory Tower and the Open Web: Introduction: Burritos, Browsers, and Books [Draft]', *Dan Cohen's Digital Humanities Blog* (26 July 2011). Online. Available HTTP: <http://www.dancohen.org/2011/07/26/the-ivory-tower-and-the-open-web-introduction-burritos-browsers-and-books-draft/> (accessed 01 December 2011).

evaluation of evidence, using objects to posit their relationships in a past that is unaccessible to us. The historian's traditional skills are still necessary, but the focus on practice – on doing things with data – extends their application, forcing a recognition of the constructed nature of evidence and its relation to the absent past. Necessarily speculative, the historian must bring his or her expertise to bear on these digital environments and evaluate the plausibility of what they both embody and imply.

Conclusion: Documenting the data

The first draft of the UCLA manifesto claimed that the first wave of the digital revolution 'replicated a world where print was primary and visuality was secondary, while vastly accelerating search and retrieval.'[35] The identification of print with verbal text betrays a bias towards the verbal in scholarly accounts of print culture even while claiming to move towards a more sensitive treatment of media fostered by the second digital revolution. Print has always been a visual medium and layout and typography, not to mention the printed image, were (and continue to be) central to print culture. In the second iteration of the manifesto, quoted as the epigraph at the head of this chapter, the sentence has been changed to 'replicated the world of scholarly communications that print gradually codified over the course of five centuries: a world where textuality was primary and visuality and sound were secondary (and subordinated to text), even as it vastly accelerated the search and retrieval of documents, enhanced access, and altered

[35] Todd Presner and Jeffrey Schnapp, 'A Digital Humanities Manifesto' (2008). Online. Available HTTP: <http://manifesto.humanities.ucla.edu/2008/12/15/digital-humanities-manifesto/> (accessed 20 August 2011).

mental habits.'[36] The substitution is a telling one, distinguishing between print and text (while acknowledging the former privileges the latter) but inserting 'scholarly publication' as the paradigmatic print genre. Given that the manifesto addresses the digital humanities as a (revolutionary) academic discipline this insertion is sensible; yet it also makes a more subtle change, moving the discussion from the republication of primary materials to the scholarly publication of secondary materials. This shift is telling as it recognizes the interpretive work of the edition.

Scholarship has always been uneasy with contingency, preferring the myths of the definitive edition or monograph, the finished output over the work in progress, to acknowledging the integral role played by provisionality in advancing debate. The entire apparatus of the academy – from the way work is reviewed and published to how it is archived and referenced – is oriented towards finished works, even if these are to be superseded by the other finished works they prompt.[37] What is never finished is our understanding of the past. As scholarly debate moves on, output by output, our sense of the past changes as we revisit the evidence anew. The status of this evidence – belonging to the past, and so finished and appropriately archived – is not fixed, but changes as we approach it in new ways. Digital publications make this mutability explicit by encoding it in the performance of resources. Manipulating the properties of data, these resources make it easy for historical objects to function in new contexts, demonstrating unexpected behaviour and allowing us to test suspected relationships. This practice – experimental,

[36] Presner and Schnapp, 'A Digital Humanities Manifesto 2.0', unpaginated.
[37] See Susan Brown et al., 'Published Yet Never Done: The Tension Between Projection and Completion in Digital Humanities Research', *Digital Humanities Quarterly*, 3, 2009. Online. Available HTTP: <http://www.digitalhumanities.org/dhq/vol/3/2/000040.html> (accessed 29 August 2011).

speculative, concerned with data, but nevertheless historical – must be written up and disseminated. Traditional scholarly outputs such as monographs and journal articles will continue to serve a purpose, providing an institutionally-validated and accessible way for this research to reach a wider (and hopefully interested) audience. Where such work will really become important is in digital-first scholarly publications that can handle the visualizations necessary to narrate data. These publications, usually open access, are poised to respond to the dynamic world of digital research, often providing useful data of their own for reuse elsewhere. Finally, of course, the resources themselves must be curated. These are both archives of primary material, sites of scholarly practice, and arguments in their own right. They demand curation not just to preserve their content but to enable continued exploration, reuse and reconfiguration. Libraries and archives must also enable practice, not just memorialize product.[38]

Archivists and librarians are used to thinking about data and have considerable expertise in responding to the requirements of diverse sets of objects. Nevertheless, there are challenges to digital history 2.0. An important barrier to this type of scholarship is in the way resources are constructed. The Linked Data movement makes it easy for the creators of resources to share their content, encoding it in such a way that is machine-readable and redistributable.[39] Yet there are those who are resistant to the idea of reuse: the emphasis on output in the humanities has encouraged scholars to be secretive, hoarding evidence until they are prepared to publish; commercial vendors also have an interest in

[38] See Bethany Nowviskie, 'a skunk in the library', *Bethany Nowviskie*, 28 June 2011. Online. Available HTTP: <http://nowviskie.org/2011/a-skunk-in-the-library/> (accessed 9 September 2011).
[39] See, for instance, *Linked Data: Connect Distributed Data Across the Web*. Online. Available HTTP: <http://www.linkeddata.org> (accessed 9 September 2011).

intellectual property, and will not publish anything that might jeopardize their place in the market. Yet what the digitization of our cultural heritage has made clear is that the past is processable and, with the tools and technologies developed by the digital humanities, often in collaboration with scholars from across the academy, we can model these processes, building them into the sites where we carry out historical practice. The 'Digital Humanities Manifesto' is iterative: so too is history.