Cultural and Social History
The Journal of the Social History Society

ISSN: 1478-0038 (Print) 1478-0046 (Online) Journal homepage: http://www.tandfonline.com/loi/rfcs20

Confronting the Digital

Tim Hitchcock

To cite this article: Tim Hitchcock (2013) Confronting the Digital, Cultural and Social History, 10:1, 9-23

To link to this article: http://dx.doi.org/10.2752/147800413X13515292098070

DEBATE FORUM

CONFRONTING THE DIGITAL

OR HOW ACADEMIC HISTORY WRITING LOST THE PLOT

Tim Hitchcock
University of Hertfordshire

ABSTRACT This discussion piece argues that the design and structure of online historical resources and the process of search and discover embodied within them create a series of substantial problems for historians. Algorithm-driven discovery and misleading forms of search, poor OCR, and all the selection biases of a new edition of the Western print archive have changed how we research the past, and the underlying character of the object of study (inherited text). This piece argues that academic historians have largely failed to respond effectively to these challenges and suggests that while they have preserved the form of scholarly good practice, they have ignored important underlying principles.

Keywords: digital humanities, digital history, standards, scholarship, referencing, OCR, search

We are halfway through what has frequently been described as a revolution. In the last fifteen years trillions of words of printed text have been digitized and delivered to an eager audience in a keyword searchable form. Google books alone has digitized some seven million of what it estimates are 1.3 billion volumes,[1] and in the next ten years the pre-1900 archive of printed materials in most major languages will be available for keyword searching.[2] We are witnessing the creation of the Western print archive, second edition. Even now it is possible to research and write credible, evidence-based history on many topics using exclusively online sources. This is not to imply that the process is complete, or even nearly so. We have only just begun to digitize manuscript materials, ephemera, images and objects. There also remain serious issues about what should be digitized next, and what impact that selection has on the direction of scholarship.[3] But what has been achieved is nevertheless remarkable. Britain has been at the forefront of the international campaign to make this happen, and its academic community – and the wider community of scholars working on British subjects – has been its greatest beneficiary. Historians working on early modern and nineteenth-century British history in particular have been gifted the most thoroughly digitized period and place in the world. Between Google Books, Early English Books Online (EEBO) and Eighteenth Century Collections Online (ECCO), *The Times* Digital

Address for correspondence: Professor Tim Hitchcock, University of Hertfordshire, Hatfield, Hertfordshire, AL10 9AB, UK. E-mail: T.Hitchcock@herts.ac.uk

Cultural and Social History, Volume 10, Issue 1, pp. 9–23 © The Social History Society 2013
DOI 10.2752/147800413X13515292098070

Archive, the British Library's Newspaper Collections, Project Gutenberg and the Million Book Project, Parliamentary Papers, the Nineteenth Century Censuses, the Clergy of the Church of England Database and the Old Bailey Online, among a host of other similar resources, British and more particularly English history has benefited more comprehensively than any other humanist subject from the advent of the 'infinite archive'.[4]

Of course, academic historians did not ask for these resources, and nor for the most part have they been directly responsible for their creation.[5] In terms of the overall quantity of digitized historical material available, only a tiny fraction has been produced in a university environment under academic leadership. Instead, we are possessed of these electronic riches because Google saw an opportunity to capture users and content, and because Cengage/Gale, ProQuest, BrightSolid and Ancestry.co.uk were quick to recognize unmet demand and to forge new business models in a rapidly changing publishing sector. And we are possessed of these resources because the institutions charged to preserve our archives and memories and make them available to a wider public have been willing to collaborate with the private sector to ensure their timely creation. These online collections are available largely because of the work of librarians, archivists and entrepreneurs, rather than academics.[6] Certainly, large amounts of public money, much of it distributed to academics following a rigorous process of peer review, have been dedicated to digital scholarship, but to relatively little effect. David Thomas estimates that between the JISC, the AHRC, the Wellcome Trust and the New Opportunities Fund, a minimum of £137.5 million has been spent in the archive, public and university sectors directly on the creation of digital resources (excluding the large amounts spent on digital creation in the context of essentially non-digital projects).[7] But even the small number of 'academic' websites this funding helped to create have enjoyed very mixed fortunes. Of the 155 projects bankrolled with 50 million pounds by the New Opportunities Fund between 1999 and 2004, only 30 show evidence of development following their formal completion, while 83 have languished unchanged. A further 25 have disappeared entirely.[8] The driving force that led to the creation of the major resources historians are now reliant upon in order to undertake their day-to-day research and teaching came from beyond the academy, and this has resulted in web resources that have been designed and implemented for other purposes and other audiences.

In parallel with these developments, there has been a second digital revolution that impinges just as directly on how we practise academic history writing. Just as the primary sources we use have become digital, so the history we create has itself been turned into a new digital form. From our grant applications, to our notes and bibliographies, to our prose, to our submitted drafts, peer review reports, proofs and off-prints, what was once a mechanical system marching to the rhythm of a printing press has become a semi-magical process of silent production and reproduction. To the irritation of librarians we still produce doorstop monographs and hard-copy versions of academic journals, but these are the empty skins sloughed off by a long-departed animal.

Just as with our newly electronic primary sources, this digital process of publication has been created elsewhere, and for other purposes. In a British context, the primary

mechanism for applying for funding to a research council (the Je-S system) was created at the behest of the STEM subjects and their funding organizations. For its first years of operation it was exclusively concerned with the work of physicists, engineers and biological scientists. The relatively junior role of humanist scholars in the wider world of research funding in the UK and their late inclusion in the Je-S system, long after its basic format had been set in stone, have ensured that this and similar systems around the world have been designed around the needs of STEM rather than the humanities.[9] The categories of assessment, labels and processes that underpin Je-S have necessarily either been designed to satisfy the most active users of the system (STEM academics) or else ground down to the lowest and most banal common denominator – encompassing the least bad form that could be agreed upon by all the research councils. And just as our primary sources have formed the basis for an innovative new business model, publishers have been quick to see the opportunities for absorbing and 'monetizing' our newly digital production process. Our scholarly journals, in particular, have been gradually wrested from academic hands and embedded within an international and near monopolistic system of commercial publishing. The science journals have been more fully concentrated in the hands of a small number of large commercial presses.[10] But in the field of history, Sage, Maney, Elsevier, OUP and CUP have aggressively courted and consumed a large number of journals, taking over the production of many publications that were traditionally wholly managed by small groups of academics on a co-operative basis. There remains something of a mixed economy of production, and the growing significance of Open Access publishing (once more led by STEM) helps to mitigate the direct impact of these developments, but large commercial publishers currently operate a near monopoly of sales and distribution. By packaging small humanities journals (the sub-prime mortgages of this particular market) into large flabby collections with one or two 'must have' publications at their heart, the publishers have effectively found a new way of selling our own scholarship back to us at a profit. Only the large publishers have the facilities in place to create the online editions which most of us now read and which form the basis for citation metrics, and all the hierarchies and marks of prestige used to sift the few from the many. As authors and peer reviewers, trustees and editors, we have been fully implicated in this process but cannot claim to have had much influence over it.

*

Perhaps surprisingly these developments seem to have had little real impact on the kinds of history we write, and have gone largely unmarked by the profession. Each new web resource has been greeted with appreciative sighs and the happy realization that we no longer need to leave our warm desks on a cold winter's day, while the wonders of 'track changes' have made the querulous tasks of proof-reading and copy-editing almost enjoyable. Many historians will ask why we should worry if we are the inadvertent beneficiaries of a wonderful new resource created for some other purpose. Why care if international commercial publishers use our journals as loss leaders and come-ons if we are still able to research and write history that can be read by other historians? A quick

scan of the publishers' catalogues and recent journal articles might reinforce this conclusion. In the short term, these developments seem to have had little negative effect on the academy or the practice of academic history writing. We continue to produce books and articles that look very much like those written thirty years ago, making almost no concession to the dramatically changing way in which we now work. We continue to publish single-authored hard-copy monographs that silently seek to demonstrate that we have each been on the arduous physical journey into the archive and back again and, even more confusingly, multi-volume collections of 'primary sources', many of which are also available online. And we then use these artefacts of scholarship, designed to meet the technological needs of movable type (including that beautiful but essentially antediluvian finding aid, the index), as part of an extended system of scholarly judgement and debate. It will be hard-copy monographs that are consistently double weighted in the Kafkaesque process of peer review that is the Research Excellence Framework (REF) in the UK, and which form the meat and blood of the massacre of young minds that is tenure in North America. Although most academics use the internet as a finding tool, their footnotes reference a direct and physical consultation of a printed edition. The vast majority of both journal articles and early modern and nineteenth-century printed sources are now accessed online and cherry-picked for relevant content via keyword searching. Yet references to these materials are still made to a hard copy on a library shelf, implying a process of immersive reading. By persevering with a series of outdated formats, and resolutely ignoring the proximate nature of the electronic representations we actually consult, the impact of new technology has been subtly downplayed. History as a discipline, largely uninvolved in the production of digital resources and apparently uninterested in changing how it illustrates its scholarship to accommodate the digital, has put its head in the sand and tried to ignore the whole issue.

This is not a posture we can afford to maintain. History as an academic discipline in its professional and post-Rankean formulation is built on a series of practices that are intended to ensure the critical use of evidence and the clear and citable development of argument. What sets academic history writing apart from populist accounts that seek only to entertain is that it provides a critical reader with the tools to trace evidence back to its origin and to unpack the blocks and shards of detail that make up an argument. Ironically, by persevering with traditional forms of publication we are appearing to claim 'authority' and traceability but are failing to live up to these claims. If the discipline of academic writing is designed to allow evidence to be verified and the process of research to be re-created, then we have abandoned the standards we inherited while maintaining their empty form.

To take a single example of this disconnect between research process and representation, many of us use and cite eighteenth and nineteenth-century newspapers as simple hard-copy references without mention of how we navigated to the specific article, page and issue. In doing so, we actively misrepresent the limitations within which we are working. Because ECCO, Google Books and the other commercial providers of historical print are anxious to deliver large amounts of electronic text quickly, they rely on a methodology for capturing text known as Optical Character

Recognition, or OCR. For the simpler forms of most post-1840s print (particularly of the more expensive varieties) this methodology works reasonably well. But it does not produce reliable results when applied to early modern publications, the cheap and quotidian print of the nineteenth century, or any form of complex formatting including tables, lists and advertisements. OCR has then been combined with a system of graphical mapping that associates strings of OCRd text with particular locations on an image of the original page, allowing search results to appear to be highlighted.[11] It is this combination of graphical mapping and OCR that is used by the Burney Collection of Eighteenth Century Newspapers, Google Books and *The Times* Digital Archive, along with the vast majority of similar sites. But the Burney Collection's OCR has a character accuracy rate of 75.6 per cent – almost one in four characters is wrong, giving an overall word accuracy rate of 65 per cent, which drops to a rate of 48.4 per cent when looking at 'significant words' of the sort that historians use for keyword searching.[12] The search engines associated with websites such as the Burney Collection have adopted a strategy of 'fuzzy' searching that helps by guessing what is a name and what is not, by recognizing common OCR issues, and by effectively expanding the scope of a search to include possible variants. But there comes a point after which 'fuzzy' searching simply produces increasingly random results as OCR errors are compounded by ever widening search criteria. And because the user is only ever presented with an image of the original, there is no way of judging the quality of the results. Ironically, the flexibility and sophistication of our human ability to interpret poor quality text have been used to hide the limitations of the digital and the computer assisted. An entirely random selection of raw OCRd text from the Burney Collection reflects the quality of what most of us are searching, most of the time:

> It is frippofed that we have actually 40 Frigates at Sea for the ProteEaion of our Trade. r The St. Domingo Fleet is fafely arrived at La Rochelle vith a rich t Cargo. ltsArrival is the more agreeable to the Merchants, as there was not a Ship of tde whole Fleet infui ed." Extrolg of a Lette?-from Mr, Coxvwood, Afate of tThe Wfter, ViWlsaller, from Cork fir Ne-w-rork, dated P3qlon, March So. "On the z6th of January we fell in with an American Privateer, comnmanded by CQrptain Bailey, called the Flower of the Sea, of 26 Guns and sSo Men, to which l'Velil we were obliged to itrike after difcharging our Guns, fix tour Pounders. We 'were towed into this Place by the Privateer the 23d infiant, after a rely da1n- gerous Paffage, having carried away all our Malts.[13]

This means, first, you have no way of checking what you are searching, as there is no easy way to discover what the raw text looks like. By downloading a PDF of the original page image, and running a local OCR system over it, you can approximate the underlying text that was used for your search, but the actual text on the original site is universally hidden on all the major commercial resources that include historical texts. Second, there is an automatic bias in relation to what you can find, leading you always and ineluctably to the same subset of text over and over again, with no way of knowing what that subset actually comprises. Essentially, 52 per cent of the Burney Collection and a similar proportion of other resources are entirely unfindable, and as

importantly it will always be the same 52 per cent, determined by typeface, layout, bleed through and a host of other factors no one has thoroughly investigated. If you want to use these materials to trace tabular data, or advertisements that include graphical elements, or any text normally represented in italics, you are largely out of luck. At a more fundamental level, the problem is that while we think we are searching newspapers, we are actually searching markedly inaccurate representations of text, hidden behind a poor quality image. And even more damning, by citing a hard copy of the original we are then refusing to document our research path, making it difficult for others to critique the process. Many historians would argue that this simply reproduces the happenstance of traditional hard-copy research. The vagaries of survival, the biases of the catalogue, and the intellectual taxonomies within which we work, form integral components of our research methodology and have always shaped our conclusions. But whereas the technologies of the hard-copy library and archive are intelligible to most humanist scholars and form the primary object of study of most post-modernist scholarship, we are not similarly forearmed for working with digital sources.

Our uncritical approach to OCR is just one instance of a series of wider issues that are raised by our growing reliance on keyword search itself. We all do it. Whether it is a quick search on a name or a place or a descriptive word – our first point of call is no longer the library and certainly not a hard-copy encyclopaedia or dictionary, but Google or Wikipedia. If you want to know the area of Hertfordshire in statute acres, you rely on Wikipedia.[14] If you want to look up the legal context for the prosecution of vagrancy in the 1760s, you probably go to ECCO and read one of the several editions of Burn's justicing manual published in that decade and available online, navigating to the relevant text using a keyword search (discovering along the way several discussions of vagrancy not in the index or table of contents).

The advent of keyword searching has been fantastically liberating. But it has also resulted in the substantial deracination of knowledge, the uprooting, or 'Googleization', of the components of what was once a coherent collection of beliefs and systems for discovering and performing taxonomies on information. Embedded within the Dewey Decimal and Library of Congress systems of classification (and in all their less successful imitators) are clear disciplinary boundaries which constrain how a reader imagines their topic and the intellectual landscape through which they navigate. Academic history was built on this lattice-work of understanding and has traditionally been constrained within it – if we are experts in anything, it is in how to use a library and archive. The advent of keyword searching lets us escape this post-Enlightenment knowledge system, but it also removes the framework of source criticism and classification that we have come to rely upon, and which we silently assume is shared with our readers, when we reference a specific edition or archive. As a result, the advent of keyword searching challenges us to be even more honest than is required by the form of a traditional footnote, about how we are searching evidence, what it is we are searching, and how those searches sift the relevant from the unfindable. A traditional footnote has always been a shorthand that can be understood only by applying a shared knowledge of the technology of the archive and the library. Keyword searching

threatens to make our footnotes essentially unintelligible. And it is difficult to avoid the impression that most academic historians have chosen not to engage with this process or question their own scholarly practice.

This is perhaps understandable since the search engines actively militate against good academic standards. In an ideal world, we would cite the searches we have used to build our body of evidence (perhaps in the form of secure and repeatable links). But because the search engines we rely upon were not built with an academic audience primarily in mind, they do not make this possible and instead effectively misrepresent and obscure the character of each search. There is a large literature on the workings of the Google algorithm and how to make your own website rise up the list of results to grab the attention of an eager audience.[15] But this literature does not address what is being excluded from results, or how the results are represented on screen.

There are two things in particular that raise serious problems from the perspective of academic history writing. First, since 2009, Google has made its 'personalized search facility' standard for all users.[16] This builds on either your own search history, or that of the computer you are currently using, to prioritize results for you – bringing to the fore the sites you visited in the past and privileging results for resources geographically near you (privileging British results if you are searching from an IP address in Britain). With some effort you can turn this facility off, but very few people do so as it works as an extreme filter in a context where most people feel overwhelmed by the volume of online data. In other words, Google's innovations mean both that researchers are increasingly being directed to bodies of online material they have used before at the expense of a wider knowledge landscape and, more significantly, that two researchers are unlikely to find the same results returned from the same query. We could specify the search criteria that have led us to a particular body of evidence, but this would still not allow us to perform that critical, scholarly function of allowing another researcher to follow in our footsteps. Or if you wanted to footnote the search process that led you to a particular site or piece of information, it would require you to include both the precise date and the full search history of the computer being used.

The other issue is that Google does not actually count results, although it appears to do so. Instead, it estimates the numbers of hits on the basis of the speed with which it locates the first few instances. In other words, the search engine really just counts how quickly it can locate the first ten or a hundred 'hits', and then extrapolates a number from these two measures.[17] If one was to search for 'eighteenth-century crime' as a single phrase (and depending on the search history of the computer you are using), you could be returned what claims to be 'About 66,700 results' in '(0.30 seconds)'. Ten minutes later, the same search could easily come up with 79,000 hits in 0.23 seconds.

But that is only a fragment of the problem. If you start clicking through the line of results at the bottom of the page, moving to the final result, it turns out that this initial search on 'eighteenth-century crime' only actually found 199 relevant sites. The more general the search term the greater the variance is likely to be between the number of 'hits' claimed and the actual number of links returned. A keyword search on a phrase such as 'cleaning products' can easily claim a return of 22 million results but only produce 953 links.[18]

Arguably, this does not matter as long as you can find what you are looking for – as long as the results are 'good enough'. But we need to remember that these claimed results effectively shape and distort our intellectual worldview, that the shape of the archive and the library we are working with is being misrepresented. We are led to assume there is more material on eighteenth-century crime, or indeed cleaning products, than actually exists, and are blinded with an image of unlimited data on a scale that we could never digest. Intellectually, 66,700 websites is a very different proposition than 199: the first is beyond human ken, available only to a high priesthood of programmers and analysts, while the second is a week's hard work for a good scholar.

Of course, the simple answer is to ignore the number at the top of the screen and do what we all do: check out the sites listed on the first or second page, relying on the Google algorithm to prioritize them and deliver the most relevant first. And if we do not find what we are looking for, we can reformulate our search and try again. As long as we positively avoid any attempt to build the search process into how we represent our scholarship, this is fine. By citing whatever physical book or journal underpins the electronic image we actually read, and by ignoring the search process completely, we can avoid the issue. But this strategy is predicated on the belief that the algorithm Google uses to hierarchize its results will actually work both to deliver what you want and to represent what is actually out there, and that what is 'out there' is actually a known quantity, rather than the unknown object of the research itself. Many of us who were trained up in traditional library and research skills and who work on materials produced before the advent of the online probably feel reasonably comfortable with our knowledge of what it is we are searching. Most scholars over a certain age spent years familiarizing themselves with the physical form of the textual leavings of previous generations, and our intellectual worldview was irreformably shaped by the systems of knowledge embedded in libraries and archives. But younger scholars have been left to swim (or drown) in a sea of deracinated text located through keyword searching, while even older scholars are using these methods of discovery in an uncritical way. One of the great advantages of the old-fashioned library catalogue of the sort that took pride of place in the British Museum's Round Reading Room was that it allowed you to assess the sum of the library's holdings at a single glance. Its simple physicality allowed the knowledge it imparted about the extent of the library's holdings to be conveyed silently and accepted uncritically. That your doctoral thesis or latest book involved just a few pages of entries out of 400 hundred volumes provided a secure measure of its importance (or irrelevance). We have *chosen* not to translate that unspoken knowledge into our online search facilities.

Historians working on early modern British materials have some advantages. The online newspaper collections reflect reasonably accurately the underlying microfilm projects from which they have been created, and ECCO and EEBO, while not nearly comprehensive, have the great advantage of having been catalogued in line with (and against the backdrop of) the English Short Title Catalogue (ESTC). Despite the crippling issue of OCR quality, you can find most items in ECCO by reference to the cataloguing details generated by the ESTC (or else be reasonably certain that a

particular item is actually not there, rather than simply mis-catalogued). But once you move beyond the parochial world of early modern Britain, the situation becomes less satisfactory. The largest collection of digitized material on the web is Google Books, which promises to represent our inherited body of literature in a formal and scholarly manner – seven million volumes so far, focused primarily on works in English from the nineteenth century. Google Books essentially purports to catalogue its collection in a way that reflects traditional forms of intellectual taxonomy but, as with OCR issues and search, this misrepresents the object of study.

The first problem is that Google Books suffers from an issue of quality control. There are unreadable pages, pages scanned in the wrong order and pages with bits left off that short of a programme of retrospective corrections by Google (which will not happen) will never be found again. These materials are now essentially lost to scholarship. This is not a criticism of Google, but rather part of the normal error rate suffered every time a new 'edition' of text is created (which is why historians worry about 'editions').

But a much bigger issue is poor 'metadata'.[19] Metadata is information about a text or work, or a webpage. The index card for a book from a physical card catalogue is metadata: the title, the author, the publisher, the date of publication, the number of pages and so on. Online metadata can include file types, measures of textual complexity and language. Conceivably, it could also include other information, such as how many people have accessed it, and what else they have downloaded. The difficulty is that if you get this wrong in an online environment (just as much as in a physical library), you can no longer find things.

On the face of it, through Google's Advanced Book Search page, you can use the metadata to restrict your results to a remarkable degree to precisely target a subject and theme. You can use all the power of a Boolean search (AND OR NOT); you can do a phrase search on the full text; you can select the language and the subject. For most of us it moves well beyond the point where technology becomes magic.

But the moment you start looking hard at results, problems begin to emerge. Google Books has been catalogued using a system called BISAC (Book Industry Standards and Communication) instead of the Library of Congress or Dewey Decimal system. BISAC was designed by the Book Industry Study Group (BISG) on behalf of publishers to help book shops manage and display their stock, and its use has led to profound systemic errors which make many volumes in Google Books extremely difficult to locate. In the words of the BISG, it 'categorize[s] books based on topical content' and maps them on to 'merchandising themes'.[20] If you use an Advanced Google Book Search for the subject 'history', or even the suggested exemplar subject label 'medieval history', not a single volume that is recognizably either 'history' or 'medieval history' is returned on the first page (medieval history only returns two results out of seven million books). In itself, this is not a serious problem as few people approach their searches using these broad subject classifications, but it reflects a wider issue with the quality and character of the metadata used in creating Google Books, which effectively misrepresents what is there. More generally, an unappetizing smorgasbord of metadata standards have been applied to historical materials online

that makes their discovery in any context an information turkey shoot of dubious value.

The taxonomies of knowledge we have inherited continue to have real value – they form the most sophisticated engagement with inherited culture we currently possess, and it is these systems of knowledge that underpin the work of all historians. In other words, we should preserve and continue to apply older systems of understanding the past. But to do that we need to engage much more directly than we have so far with the technologies we are rapidly coming to rely upon. At the moment we are using them to make our lives easier, while pretending that they do not exist. We are lost in a world of reference and evidence which most of us do not understand and which we do not currently have the critical tools to represent. There have always been issues with the quality of the technologies we rely on – whether that technology was an encyclopaedia or an online database. Indeed, the process of digitization has substantially highlighted the inconsistencies and poor quality of many classic works of reference.[21] But because most of us were at least introduced to library science as a part of our training, we were partially forearmed. The hidden character of the digital equivalents of the card catalogue, created using a language and tools few of us understand, exacerbates the problem. I normally refer to this as the IKEA effect, because we are being led through a system to a particular conclusion, without signposts or a map. In a phrase coined by Bill Turkel, it is the 'Las Vegas syndrome' – all invitations to play, with no exit signs and no indication of the odds – an intellectual crap shoot in which hope overrides experience and quiet judgement. We use the Burney Collection and Google Books regardless, playing a form of research roulette dressed up as traditional scholarship.

*

Thus far this article has concentrated on what is wrong with the resources we use and has implied that we were sold a faulty product by librarians, multinationals, computer programmers and mathematicians. But that would be to blame the mechanic and the road builders for the actions of the drunk at the wheel. Our own lack of engagement is just as significant. We have not established the necessary new systems of reference and validation that would make our use of these resources transparent and repeatable.[22] I have yet to see a piece of academic history that is explicit about its reliance on keyword search and electronic sources. As editors and authors, we accept and write footnotes that misrepresent the research process. As teachers, we largely fail to instil a critical engagement with the real electronic text we ask our students to read, and instead encourage them to pretend they are sitting with a dusty book in their hands. And as intellectuals paid from the public purse, we ignore our obligation to question the quality of the data we encounter in the wider world.

In part, the answer is simple. What is required is a much clearer understanding on the part of us all of what precisely we think we are doing with the sources, and how they map onto what we are producing as scholars. We need to step back and explicitly interrogate how older taxonomies of knowledge relate to newer iterations of representation, and to be clear in our own minds what forms of scholarly representation

and citation are most useful, and most fully satisfy the requirements of reproducibility and traceability. We need to interrogate the sifting and selection that is embedded within OCRd text; we need to record our searches as a part of our research methodology (and train our students to do so); and we need to cite the actual websites and online resources we use, rather than pretending they do not exist. In the mature scholarly environment we inherited, many of these kinds of tasks were outsourced to archivists and librarians, leaving us to build on their unquestioned foundations. But as those foundations have shifted, we need to re-survey the whole structure.

This at least would tie us to a higher standard of scholarship, but it would not address the opportunity that the online and the digital also present. Since we are all using keyword searches to locate texts in a sea of deracinated knowledge, it is time to interrogate more seriously what this new knowledge environment makes possible, at the same time as being more honest about what it makes difficult.

In 1962, Louis Mumford observed that 'minds unduly fascinated by computers carefully confine themselves to asking only the kind of question that computers can answer'.[23] In the last couple of decades historians who continue to represent their research as reading 'books and articles' have restricted themselves to asking only the kind of questions physical books and articles can answer. Fifty years is a long time in computer science (and historical scholarship), and it is time we found out if a critical and self-consciously scholarly engagement with computers and the infinite archive might not now allow us to do something more. It is a wonderful thing when a keyword search returns a killer quote and powerful example. The observation from Louis Mumford above was located in just this way. But it is not research.

As well as challenging our systems for representing and conveying scholarship, the digitization of the infinite archive provides a way to do more than has ever been possible before. We can now engage with that new resource in new ways, but to do so we need scholars who can work with old sources in their new digital guise, and who recognize that the digital makes them different. This is not only about being explicit about our use of keyword searching – it is about moving beyond a traditional form of scholarship to data modelling and to what Franco Moretti calls 'distant reading'.[24] We need to recognize that we are no longer reading books and articles, but are working with massive bodies of digital text – the difference between one volume and another little more than a line or two of metadata.

Despite the caveats over OCR and metadata retailed above, the existence of a digital edition of the Western print archive allows us to start modelling its content in new ways. With a minimum of effort we can simply count words, and how they are distributed between years, sites of publication, genre and format, and compare the results to the version of this same information embedded within post-Enlightenment scholarship (every cataloguing system is an attempt to map precisely this). Did novels get longer as well as more numerous? How were words distributed across genre? What changes when you measure text instead of 'books'? Or more ambitiously, we could use the tools of data mining and corpus linguistics to categorize inherited material according to its textual and semantic content, tracing collections of emotions, ideas or descriptive forms across works that have long been separated by older forms of

classification. Mapping the constantly changing language of affect, for instance, across genre, measured against textual density, would give us new ways of characterizing and connecting disparate collections of text – crime reporting and novels; poetry and legal precept. We have the opportunity both to test and to revise the taxonomies of knowledge that we inherited and to extend them from crude piles of books to subtle maps of meaning.

Perhaps counter-intuitively, this would not undermine traditional forms of scholarship but would strengthen them and give them greater intellectual purchase. Creating this new infinite edition of the Western print archive as an object of study in its own right and mapping the distribution of language across its surface would raise the critical standard for how we read all kinds of evidence. It would provide a new and subtle context for each word and phrase. In the process it would make more compelling both the kind of 'thick description' that has proved so productive of new insights over the last three decades and the use of quotes and examples in synthetic description. If we could just represent the research process that led us there in an explicit and scholarly manner, and if we could commit to training our students in the critical skills they need to navigate this sea of data, we could have a subject that looks more like a 'discipline' than it does at the moment. This is about not simply living up to the standards of scholarship we inherited, but going beyond them.

It has been a dozen years since the first major historical web resources were launched on the World Wide Web, and a decade since the posting of the first census demonstrated beyond reasonable doubt the existence of profound popular demand for history online, and the distance academic history had inadvertently travelled from that wider audience. But academic historians have yet to effectively address the implications of the online and the digital for their scholarship, or to rise to the challenge that these resources present. We need to know about OCR and metadata, and we certainly need to learn how to use the tools of data mining, GIS and corpus linguistics; and we need to be able to wield the tools of large-scale visualization, as spearheaded by the hard sciences, network theory and 'big data' analysis of the sort implemented in the Google NGram viewer.[25] And we need to do so at the same time as we preserve the values and practices that underpin traditional academic history, while going beyond the standards of scholarship we inherited. We need to insist that the resources our universities pay for, and the journals we all review for and edit, rise to the challenge of the digital, while at the same time ensuring that our students are forearmed with the skills and experience they need to navigate this new sea of data.

NOTES

1. In a corporate blog post, on 5 August 2010 (8:26 a.m.), Leonid Taycher, a software engineer employed by Google, claimed there were 129,864,880 books in the world. See Leonid Taycher, 'Books of the World, Stand up and Be Counted! All 129,864,880 of You' (http://booksearch.blogspot.co.uk/2010/08/books-of-world-stand-up-and-be-counted.html); Google Books Search (http://booksearch.blogspot.co.uk/).
2. The legal mire that has overtaken the Google Books Settlement in the United States is likely to have a significant impact on the development of this new digital environment, but moves

in both Europe and the US to create 'national digital libraries' and continuing commercial and state interest in the wider project suggest that there will be no substantial let up in the pace of digitization in the near future. For the issues associated with the Google Books Settlement, see James Grimmelmann, 'The Elephantine Google Books Settlement', 24 August 2011 (http://works.bepress.com/cgi/viewcontent.cgi?article=1031&context=james_grimmelmann).

3. The relatively paucity of non-Western materials available in a digital format is particularly noteworthy, suggesting that the very process of digitization is effectively reproducing a kind of Western cultural hegemony that would not be acceptable if it was a product of self-conscious policy.
4. See Google Books (http://books.google.co.uk/bkshp?hl=en&tab=wp), Early English Books Online (EEBO) (http://eebo.chadwyck.com/home) and Eighteenth Century Collections Online (ECCO) (http://gale.cengage.co.uk/product-highlights/history/eighteenth-century-collections-online.aspx), *The Times* Digital Archive (http://gale.cengage.co.uk/times.aspx/), the British Library's Newspaper Collections (http://www.bl.uk/welcome/newspapers.html), Project Gutenberg (http://www.gutenberg.org/) and the Million Book Project (http://archive.org/details/millionbooks), Parliamentary Papers (http://parlipapers.chadwyck.co.uk/marketing/index.jsp), Nineteenth Century Censuses (http://www.nationalarchives.gov.uk/records/census-records.htm), the Clergy of the Church of England Database (http://www.theclergydatabase.org.uk/index.html) and the Old Bailey Online (http://www.oldbaileyonline.org/). Even the small subset of primary sources for British history from 1500 to 1900 available through the Connected Histories site (http://www.connectedhistories.org/Default.aspx), which currently (6 December 2011) excludes all books and includes only larger collections that meet a minimum standard of academic presentation, runs to over ten billion words of text and 140,000 images. It should be noted that the author has spent the last twelve years helping to create four substantial web resources: the Old Bailey Online; London Lives, 1690–1800, (http://www.londonlives.org/); Connected Histories, and Locating London's Past (http://www.locatinglondon.org/index.html) and is fully implicated in the process and project.
5. This is not to denigrate the many, primarily small-scale, digitization projects led by academics, and generally directed towards addressing specific research questions. Transcribe Bentham (http://www.ucl.ac.uk/transcribe-bentham/), The Newton Project (http://www.newtonproject.sussex.ac.uk/prism.php?id=1) and The Electronic Enlightenment (http://www.e-enlightenment.com/), to take just three examples among many others, represent a high academic standard, created in pursuit of answers to serious academic questions. But, by their nature, and in comparison to the large commercial digitization projects, they represent a small part of a wider online landscape of historical materials.
6. Many of the librarians and archivists involved in the wider process of digitization are employed by universities or major research libraries and would consider themselves 'academics', but the widespread adoption of the term 'alt-ac' (alternative-academic) to describe the careers of the people involved reflects the distance between the professional academic discipline of history and the people most fully engaged in the wider project. See Bethany Nowviskie, '#alt-ac: Alternate Academic Careers for Humanities Scholars', 3 January 2010 (http://nowviskie.org/2010/alt-ac/, accessed 4 May 2012).
7. Valerie Johnson and David Thomas, 'New Universes or Black Holes? Does Digital Change Anything?', in Toni Weller (ed.), *History in the Digital Age* (Oxon. and London, forthcoming 2012). This excludes figures for the British Library (which were not available), and money generated for organizations such as the National Archives by participating in commercial digital projects. It should be noted that even this ten-year total is dwarfed by

the amount spent each year supporting more traditional forms of humanist scholarship.

8. See Nancy L. Maron and Matthew Loy, 'Funding for Sustainability: How Funders' Practices Influence the Future of Digital Resources', June 2011, p. 9, note 3 (http://www.jisc.ac.uk/media/documents/programmes/contentalliance/sandrfundingforsustainability.pdf). Of the NOF projects, approximately 15 per cent were located in universities and a further 20 per cent in recognized academic institutions such as the National Archives and British Library, but projects across the programme suffered a similar rate of failure.
9. The Je-S system went live in May 2003, and the first use of the system by the AHRB/C was recorded in June 2005. See Je-S Help (https://je-s.rcuk.ac.uk/Handbook/index.htm#pages/JeSHelpdesk.htm).
10. It is difficult to access local and subject-specific information in this area, but for a general overview of the patterns of concentration of journal publishing across the academy see Corinne M. Flick (in cooperation with the Oxford Internet Institute), *Geographies of the World's Knowledge* (Oxford, 2011), http://www.oii.ox.ac.uk/publications/convoco_geographies_en.pdf, pp.14–19. Internationally the largest players are Elsevier, Wiley, Springer and Taylor and Francis.
11. The most common system used for this 'graphical mapping' is called Olive. See http://www.olivesoftware.com/.
12. Simon Tanner, Trevor Muñoz and Pich Hemy Ros, 'Measuring Mass Text Digitization Quality and Usefulness', *D-Lib Magazine*, 15(7/8) (2009), http://www.dlib.org/dlib/july09/munoz/07munoz.html. The Burney Collection is being highlighted here only because of the existence of a detailed assessment. It is largely typical of other sites providing OCRd text, and better than many.
13. Seventeenth and Eighteenth Century Burney Collection of newspapers (http://www.bbk.ac.uk/lib/elib/databases/newspapers/burney): *Daily Advertiser* (London, England), Saturday 16 May 1778; Issue 14784, Gale Document Number: Z2000162519. This selection of raw OCR was constructed by stringing together snippets of text exposed as part of the search results in Connected Histories. It is impossible to link directly to this text as the system of licensing used by Gale Cengage embeds local access data within the URL. An accurate transcription of this text would read: '"It is supposed that we have actually 40 Frigates at Sea for the Protection of our Trade. The St. Domingo Fleet is safely arrived at La Rochelle with a rich Cargo. Its Arrival is the more agreeable to the Merchants, as there was not a Ship of the whole Fleet insured." *Extract of a Letter from Mr. Coxwood, mate of the Ulster, Victualler from Cork for New-York, dated Boston, March 30.* "On the 26th of January we fell in with an American Privateer, commanded by Captain Bailey, called the Flower of the Sea, of 26 Guns and 180 Men, to which Vessel we were obliged to strike after discharging our Guns, six four Pounders. We were towed into this Place by the Privateer the 23d instant, after a very dangerous Passage, having carried away all our Masts."'
14. According to Wikipedia, the statute acreage of the administrative county of Hertfordshire in 1891 was 406,932. The relevant article footnotes the census report for this date. See 'Administrative counties of England' in Wikipedia (http://en.wikipedia.org/wiki/Administrative_counties_of_England#Area_and_population, accessed 6 December 2011). Interestingly, more sophisticated users who knew the digital landscape and original sources more fully might have relied on the highly academic and comprehensive edition of this material available through Histpop: The Online Historical Population Reports Project (http://www.histpop.org/ohpr/servlet/Show?page=Home), a site that gives full access to the original census reports. But if they had done this, they would have been confronted with three different areas for Hertfordshire in 1891 – for the ancient county, the administrative

county (used in Wikipedia) and the registration county.

15. See, for example, Arel Dornfest and Tara Calishain, *Google Hacks: Tips and Tools for Smarter Searching*, 3rd edn (Sebastopol, CA, 2006) and David Viney, *Get to the Top on Google: Tips and Techniques to Get Your Site to the Top of Google and Stay There* (London, 2008).
16. For an accessible account of these developments, and of the evolution of the Google algorithm more generally, see Steven Levy, 'Exclusive: How Google's Algorithm Rules the Web', *Wired*, 22 February 2010 (http://www.wired.com/magazine/2010/02/ff_google_algorithm/).
17. Google's description of this process makes a positive virtue of its ever changing results: 'When you perform a search, the results are often displayed with the information: About XXXX results (X seconds). Google's calculation of the total number of search results is an estimate. We understand that a ballpark figure is valuable, and by providing an estimate rather than an exact account, we can return quality search results faster. In addition, when you click on the next page of search results, the total number of search results can change. Google's search index is constantly changing, and your second search results page may come from a slightly different version of the index than the first page.' (http://support.google.com/webmasters/bin/answer.py?hl=en&answer=70920).
18. These searches were undertaken from my home computer located in Muswell Hill, UK, on 8 December 2011 and, as explained in the text, cannot be repeated or effectively cited.
19. See Geoffrey Nunberg, 'Google Books: A Metadata Train Wreck', *Language Log*, 29 August 2009 (http://languagelog.ldc.upenn.edu/nll/?p=1701).
20. See the BISG webpage advertising the BISAC subject headings, 2011 edition (http://www.bisg.org/publications/product.php?p=149, accessed December 2011).
21. It is perhaps invidious to point out any single example, but it was only when the Clergy of the Church of England Database project attempted to use the detailed information contained in the Royal Historical Society's *Guide to the Local Administrative Units of England* (Frederic A. Young, vol. 1, 1979, vol. 2, 1991) that the inconsistent character of many entries became apparent.
22. One looks to the Royal Historical Society, the Historical Association or the British Academy, or for a clear steer from the major academic presses and journals.
23. Lewis Mumford, 'The Sky Line "Mother Jacob's Home Remedies"', *New Yorker*, 1 December 1962, p. 148 (http://www.newyorker.com/archive/1962/12/01/1962_12_01_148_TNY_CARDS_000269697). This observation was made as an aside directed at urban planners in a review of a book by Jane Jacobs. It was located following a half-remembered reference in an article in a popular newspaper I have since thrown away, and a Google search on the phrase 'questions that computers can answer'.
24. Franco Moretti, *Graphs, Maps, Trees: Abstract Models for a Literary History* (London, 2005).
25. Jean-Baptiste Michel*, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, William Brockman, The Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak and Erez Lieberman Aiden*, 'Quantitative Analysis of Culture Using Millions of Digitized Books', *Science* (published online ahead of print:, 16 December 2010, http://www.sciencemag.org/content/331/6014/176.full).