# Gresham College Main logo

12 November 2014

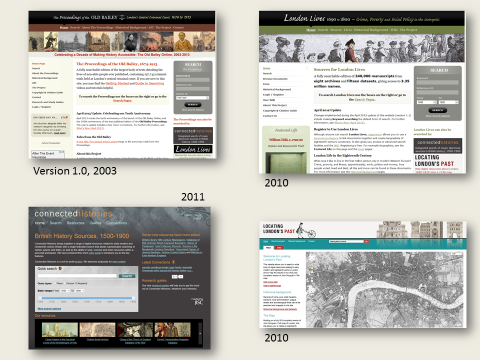
**Making History Online**

Professor Tim Hitchcock  
Professor Robert Shoemaker



*Tim Hitchcock*

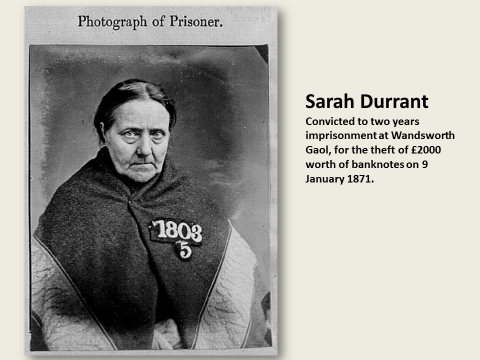
It is sometimes hard to remember just how much has changed in the last twenty years - the extent to which how we do historical research has been transformed by digitisation and the internet. Between Jstor, which has provided online access to a vast archive periodical articles since 1995; and the online British census since 2001; and Google Books since 2004; that traditional journey in to the library and the archive and back out again, has been reshaped. I think of this as the creation of the Western print archive – second edition.



And Bob Shoemaker and I have been hugely privileged to be allowed to contribute to that larger project; through the Old Bailey Online, launched in 2003, London Lives and Locating London’s Past in 2010, and Connected Histories in 2011.

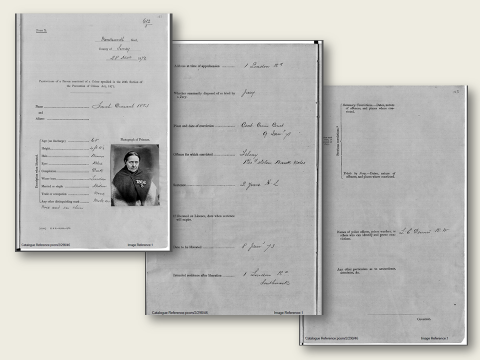
Through these projects, and innumerable others, emerging from the private sector, from museums and archives, as well as from the academy and higher education, British history in particular has been made newly available at the click of a mouse, to anyone with an internet connection. Of course, there are limitations - issues of coverage and what is left un-digitised, of access and paywalls, and of OCR quality and copyright - but in less than a generation, the British past, particularly the British past prior to the twentieth century when copyright kicks in, has become the most digitised where and when in the world.

And it has let historians - of all stamps - do remarkable things. And just by way of a simple example, I want to introduce you to a single person.

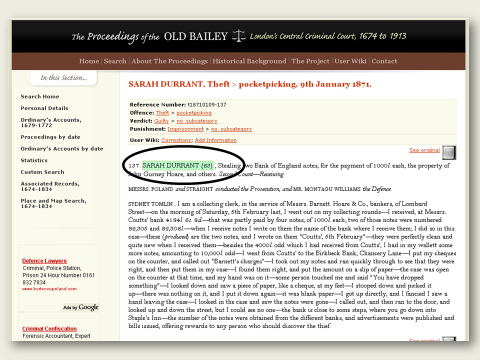


Her name is Sarah Durrant, and we have chosen her almost at random. She is not important, her individual story does not change anything, but quite suddenly that experience is available to us in a new way.

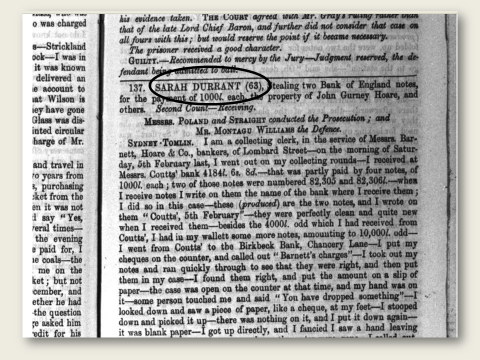
Sarah claimed to have found two bank notes on the floor of the coffee house she ran in the London Road, in 1871; at which point she pocketed them. In fact they had been lifted from the briefcase of Sydney Tomlin, at the Birkbeck Bank, Chancery Lane a few days earlier.



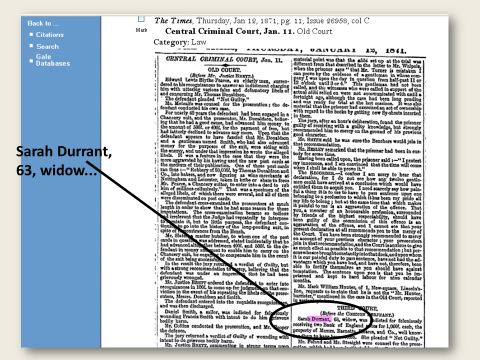
We know what Sarah looked like. We have her image, her details, her widowed status, the existence of two moles on her face - one on her nose and the other on her chin. We have her scared and resentful eyes staring at us from a mug shot.



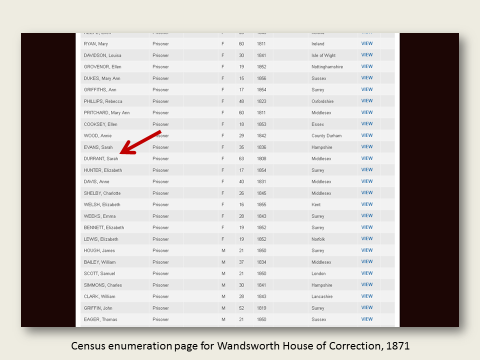
But, we also have the words recorded at her trial. From which we know that Sarah, moles and all, was convicted of receiving stolen goods; and that she had been turned in by a Mrs Seyfert - a drunk, to whom Durrant had refused a hand-out.



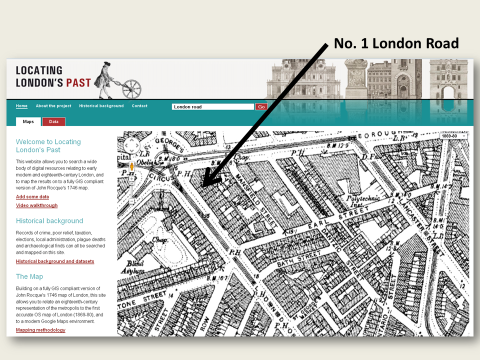
And, of course, we have an image of the original page on which that report was published.



And just in case, we can also read the newspaper report of the same trial.



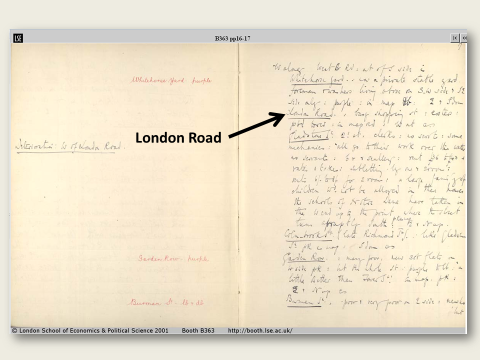
So far, so much text, with a couple of layers of coding, and the odd image. But we also know who was in Wandsworth Gaol with her on the census day in 1871.



And we know where Durrant was living when the crime took place – in Southwark, at No 1 London Road. We know that she was a little uncertain about her age, and we know who lived up one flight of stairs, and down another.



From which it is a small step to The Booth Archive site posted by the London School of Economics in 2001, which in turn lets us know a bit more about the street and its residents.

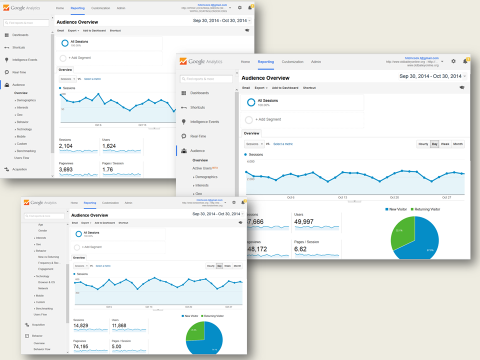


According to this policeman, London Road was ‘a busy shopping street', with the social class of the residents declining sharply to the West – and coded red for lower middle class.

In half an hour’s search we can put together a life; an experience - an emotional and empathetic contact with the dead.

And in the process of making this kind of research possible, digitisation and the internet have helped spirit into being new audiences for history, new practitioners of history writing, and new forms of history. When in 2001 the first British census was made available online – and immediately crashed through overwhelming demand; and when, in launch after launch similar web resources, attracted the lustful attentions of an eager, non-academic audience - it became increasingly clear that something new was happening.

The simple act of putting a surname into Google, will generate a slew of ads for paid genealogical services, leading back to the massive, private as well as public sector initiatives, giving unprecedented access to the details of the lives of the long dead. Between Ancestry, Find-My-Past, Adam Matthew, and Gale Cengage, – not to mention Google Books - the non-academic historian has at their fingertips more real data than can be found in any single archive or hard-copy library.



In an average week, the Old Bailey Online attracts around 15,000 visits from dozens of countries around the world, and of these the vast majority are from private individuals accessing the site from outside of higher education. The first academic URl to appear in our lists, normally comes in around thirty places from the top.



And it is not just usage of the internet that has changed. The four million plus people who watched episodes of the television series *Garrow’s Law* over three seasons, the six million who tune in to each episode of *Who Do You Think You Are?*; and the three million who sought out the first screenings of *Secrets from the Workhouse* – evidence the vastly expanded audience for history – both online and on TV. And many of these ‘viewers’ are not just passively consuming history. They are doing their own research and writing.



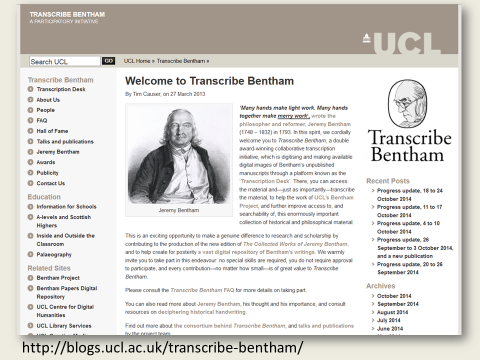
My favourite example of the involvement of a wider public in research and, via crowdsourcding, in the creation of new historical materials, is *Trove*. I suspect *Trove* is unfamiliar to many of you. But, it is the single most successful historical crowdsourcing project in the world, and gives direct public access, and input, to around 400 million items held in the libraries of Australia – most notably newspapers published up until the 1950s. On an average day, users of Trove make around 100,000 corrections to the newspaper archive – combining good citizenship with research of their own. In the process *Trove* has built a community of historically interested users that dramatically helps shape Australian public culture.

Less dramatically, the same could be said of the Old Bailey and Connected Histories, and more importantly, the online resources created by the National Archives and British Library. A new audience of consumers and producers of history has evolved – many of them co-creating the sources of historical research, in the process of undertaking their own research.

This is a fantastic thing, but it is not without its problems. The digital threatens to deracinate the leavings of the past, and allow them to be used with little sense of context or meaning – all flashy quotes, located using key-word searches. But in our estimation, there has not been such a vibrant audience for history writing since the heydays of Macaulay and Gibbon.

*Robert Shoemaker*

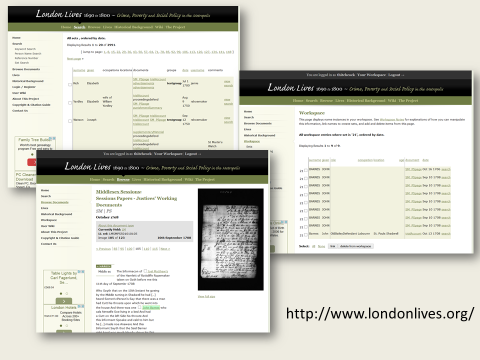
Academic historians are obvious beneficiaries of this sea of change, but, with some exceptions, they have been rather reluctant to embrace this new public of practicing historians. One way of doing this, as with Trove, is through crowdsourcing, which has been used by a few academic projects. But in contrast to Trove -- a project run by the National Library of Australia, an institution with public engagement embedded in its mission -- academic projects have had some difficulty recruiting the help of a wider audience.



One of the most successful examples is *Transcribe Bentham*, which has invited public volunteers to transcribe the voluminous and often difficult to read papers of the utilitarian philosopher Jeremy Bentham, with the specific aim of breaking down traditional barriers between the public and academic research.

In four and a half years, some 11,000 manuscripts have been transcribed or partially transcribed by volunteers, about a quarter of the manuscripts which remained un-transcribed when the project started. This is a tremendous achievement, but as I am sure the project directors would admit, it has been a long and difficult journey. At this pace the project will need another twelve years or so to complete the task. And although many people have participated, most have only transcribed a few pages, while the vast majority of the pages have been transcribed by a small number of volunteers. The ‘crowd’ turns out to be rather small.

Our own experience with crowd sourcing was even less successful--when we invited people to contribute content to our *Old Bailey* and *London Lives* sites through specially created wiki pages, the response was underwhelming, and eventually the wiki pages were discontinued.



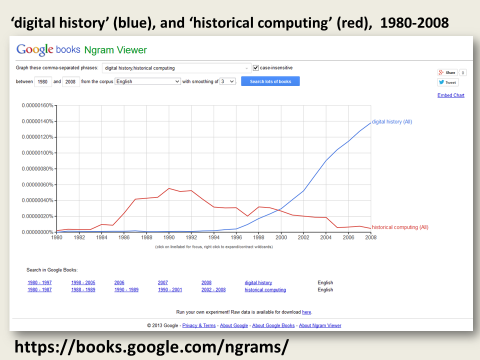
We have had more success with a very specific task, which takes advantage of the way users are already interacting with the London Lives website. The site provides access to over 240,000 manuscripts about poverty and crime in eighteenth-century London, and includes over three million separate name instances. Registered users can link records together which they think concern the same individual, and so far we have some 3000 of these sets. Admittedly many of these sets were created by the two of us, but the vast majority were created by the site’s 4,000 registered users.

The point is that crowd sourcing and public engagement are difficult, and require considerable time and effort. We cannot just assume that the ‘public’ will do what we ask them to do. In part people are making their own histories, and do not necessarily want to be led by the academy. As both *London Lives* and the *Transcribe Bentham* projects have discovered, there needs to be a substantive dialogue between project staff and the volunteers--a community of people working on the project needs to be developed--so that both feel that they are getting something out of it, and the resulting resource is CO-created. Unfortunately, current structures for funding projects, and allocating academic workloads, do not normally make the level of human resource necessary for this work available.

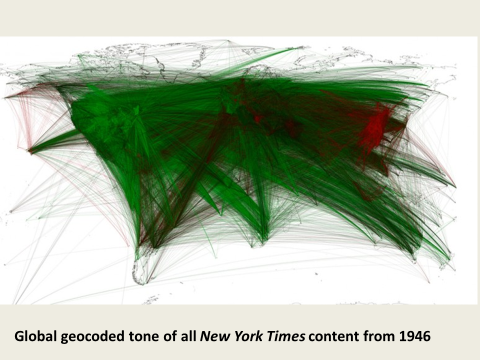
*Tim Hitchcock*

It is not just that academics have had only limited success in creating a dialogue with family, local and public historians; there are also moves from within academic digital history which have had the effect of driving a wedge between a wider readership, and historians working in higher education. One of the virtues of digital history is that it has attracted scholars from different disciplines, with different skills, to look afresh at the evidence now that it has been transformed from mere words and objects - into data. But the results of these interdisciplinary collaborations are often problematic.

Perhaps the best known example is the Culturomics movement, emerging from the Cultural Observatory at Harvard, using Google’s Ngram viewer to analyze word frequencies in Google books.



If anyone is unfamiliar with the Ngram viewer it is wild and wonderful. It essentially allows you to chart the relative importance of words and phrases, year by year, in the full body of Google Books. At its best this forms a powerful way of exploring the content of what is now some 14 million volumes from Google Books. But while the Ngram viewer is available to all, and easy to use, some of the academic history that is being written on the basis of this tool can seem increasingly arcane.



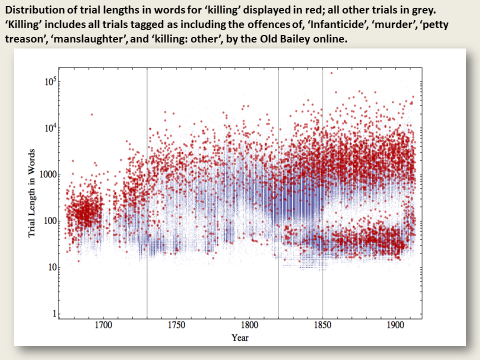
The Ngram viewer is used by its creators, Jean Baptiste Michel and Erez Lieverman Aiden, to generate what they consider to be a newly ‘scientific’ reading of the past, that privileges varieties of history that are highly technocratic. Their most powerful claim to date is that the Ngram viewer demonstrates that irregular verbs in English have declined steadily over the last four hundred years. This is a potentially significant (if contentious) finding that implies language change is not subject to human agency. But whatever else it is, it is not the stuff of popular history, and tends to generate varieties of ‘shock and awe’ graphics, that are designed more to bludgeon readers into submission, rather than to convince through detailed argument.



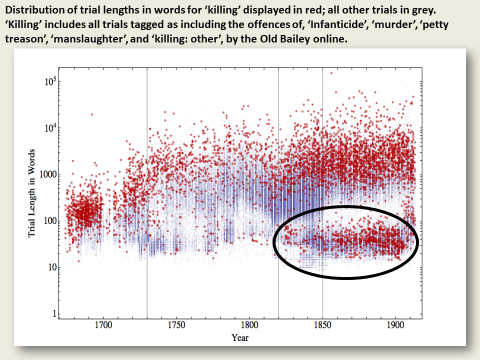
At its best, this sort of work is wonderful - I would point to something like Ben Schmidt’s ‘prochonism’ projects, which take the individual words in modern cinema and television scripts that purport to represent past events and compares them to every word published in the year they are meant to represent. In the process, he illustrates all the anachronism in *Downton Abbey*, and more impressively, the subtle changes in the presentation of masculinity, decade by decade, in the evolving world of *Mad Men*.

But this type of history moves the focus resolutely away from people like Sarah Durrant, and towards a variety of cliometrics - a ‘scientific’ approach to history, that has little relevance for the new audience for history evidenced in popular culture.

We are guilty of this too. Some of our own projects, such as *Data Mining with Criminal Intent*, have taken us in precisely this direction. In collaboration with Bill Turkell, at the University of Western Ontario, we started to treat the Old Bailey text - not as a collection of individual dramas - but as a ‘massive text object’.

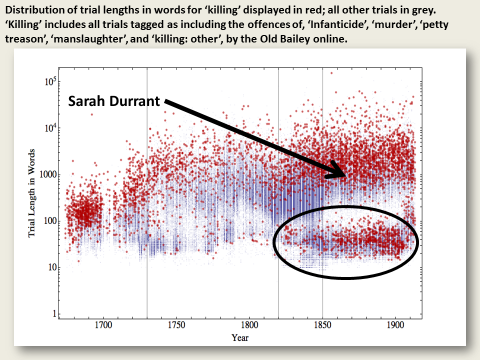


This graph, for instance, represents all 200,000 trials in the Old Bailey online, divided between those for forms of ‘killing’ and all other offences, and distributed according to how many words each trial contained - from the shortest, at 8 words, to the longest at 157,000 words.



By doing this we discovered that the evolution of the nineteenth-century trial, and criminal justice system was marked by large numbers of very short trials, resulting from the rise of plea bargaining – even for those accused of serious crimes like killing. For many ‘justice’ had moved from the courtroom to the police cell.

Figuring out what this graph meant was an illuminating journey, but again, it did not result in popular history. And the same could be said of many of the important projects that are beginning to use sophisticated techniques such as ‘formal network analysis’; ‘Topic modelling’; ‘Text Frequency/Inverse Document Frequency’ measures; and most influentially, approaches derived from Bayesian Probability--all generally thought of as forms of ‘big data’ analysis.



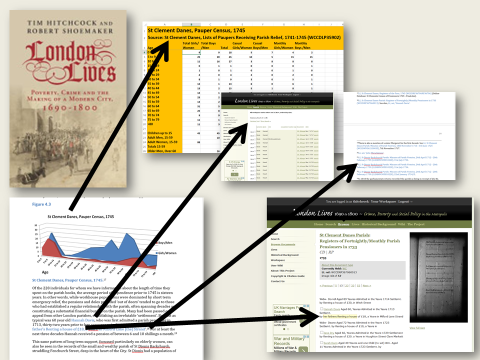
The challenge is to attempt to link the individual to the complex patterns of data we can generate—to put Sarah Durrant back into the picture.

*Robert Shoemaker*

While academic *research* is in danger of turning its back on the rich opportunities digital history presents for bridging academic and popular history, academic history *writing* has been even slower to change. We continue to prioritise our traditional publication genres--the monograph, the journal article, the chapter in an edited collection. These are what we submit to the Research Excellence Framework (REF), and what we list on our CVs, and these are what we focus on when making decisions about promotion.

The ‘e-book’ is a good example of how slowly the academic history world, and its partners in the publishing sector, is changing. While academic monographs are now frequently published in an e-book edition, alongside the traditional hardback and sometimes paperback, the e-book as currently conceived is little more than an online pdf file--taking advantage of only the most basic of the numerous opportunities the internet offers for disseminating knowledge and fostering dialogue. This is particularly frustrating in the case of published editions of primary sources, such tax and legal records and diaries. Despite the fact these editions often make little sense when simply read in book form and they require very detailed indexes to be useful, they are still rarely delivered in an electronic format which would enable the keyword and structured searching necessary to maximise their usefulness.

In a book due out in the new year, we have tried to push the boundaries of the e-book, by designing it so that it is best read online, with thousands of hyperlinks which will take the user from relevant places in the text directly to the free online editions of the primary sources we have cited or quoted from, or the complete databases which underlie our tables and graphs, or catalogue entries of the printed primary source texts we cite, or e-book editions in Google books of the secondary sources.



This is intended to make real the traditional purpose of a footnote - to allow the relevance of a specific piece of evidence to be confirmed by the reader, and the research journey of the authors to be made explicit. No doubt this is all very foolish, tempting readers to check our footnotes and catch us out on the inevitable errors, but it is designed to transform the way books are read. This is not a book that needs to be read sequentially from page 1 to 350 (how many of us do that anyway?), but rather readers are encouraged not only to dip in and out of the book as they wish, but also to follow research threads back into the sources, where they can conduct their own research and consider other interpretations and lines of argument, that we have not even thought of.

Bringing our publisher along on this journey has proved difficult. While they seemed keen on the approach when we initially talked to them, the production processes publishers use to generate books are not readily adaptable to new formats. They are still framed in terms of the production of a printed book, with the pdfs generated as an afterthought. As we approach the proof stage, we still have not had any copy-editing or checking of the book’s electronic features, and the advance flyers which announce the publication of the book, and the advance sales offered through the publishers’ website and Amazon, fail to mention even the existence of the e-book. Perhaps everything will come right in the end, but our experience suggests that the structures of academic publishing are changing very slowly.

So, while the online and the digital have fundamentally reshaped the landscape of historical research - bringing in to being a new audience, and a new class of practitioners, the practices of academic research and writing threaten to ignore these new possibilities, heading in a different direction entirely. Developments in the analysis of big data (including text and data mining, and Bayesian probability) have tended to make digital history seem even more inaccessible, technocratic, and in some ways irrelevant for a wider audience, while academic publishing is changing very slowly.

So, there are very real challenges to overcome if we are to achieve the potential for the internet to bridge the divide between academic and public history - if we are both to let history serve its primary function as a form of social memory; while also taking advantage of new methodologies created by ‘big data’.

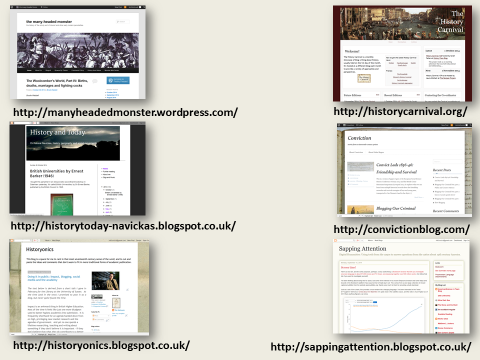
And in our estimation, the answer to this conundrum is provided by the internet itself.

*Tim Hitchcock*

First and most obviously, ease of publication means that historians can put the results of our research into the public domain almost instantly, and in innovative formats which are more readable by a wider audience.

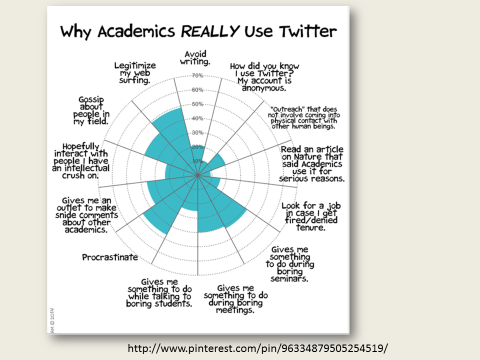


The academic blog is a good example of this. While some blogs explicitly address a public rather than academic audience (the ‘History Matters’ blog at the University of Sheffield is a good example of this), other historians use blogs to disseminate findings, or try out emerging arguments, addressing an audience of both academics and the public. Many, particularly younger scholars, are beginning the use Twitter and blogs as part of the process of developing ideas, doing research, and perhaps most importantly, ensuring that once complete, the history they have written, actually has an audience of eager readers who have followed the research process from day one.



Perhaps the best example of this is Ben Schmidt, and his hugely influential blog: *Sapping Attention*. His blog posts analysing nineteenth-century word frequency, and authorship were re-purposed as part of his doctorate, and will form part of his first book. Or Helen Rogers, who maintains two blogs: *Conviction: Stories from a Nineteenth-Century Prison* - on her own research; and also a collaborative blog, *Writing Lives*, created as an outlet for the work of her undergraduates. These blogs bring together research and teaching, and in the process are building a substantial community of interest.

The list could go on. The *Many Headed Monster*, the collective blog authored by Brodie Waddell, Mark Hailwood, Laura Sangha and Jonathan Willis is rapidly emerging as one of the sites where seventeenth-century British history is being re-written. While Sharon Howard has been overseeing the *History Carnival* for over a decade.

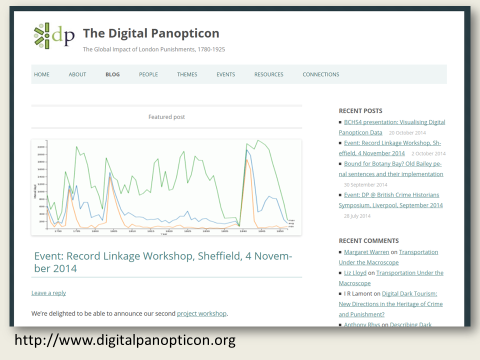


An analysis of the types of social media academics use recently revealed that Twitter was actually where a lot of the ongoing discussion, in at least some corners of history, is taking place. This is Jorge Chad, of PhD Comic’s take on the data. And following @Twitterstorians for a few days, will reveal a wild world of debate and engagement.

The relationship between these ‘publications’ and more formal academic outputs and reward procedures is unclear. At the moment, blogs listed on an academic CV are unlikely to be taken seriously as research publications, though they may well be seen as evidence of public engagement, which IS starting to be recognised within promotions procedures.

The biologists are taking this further than most historians would be comfortable with, developing new metrics such as the ‘Kardashian Index’, - or K-Index - which was initially proposed as a joke, but is rapidly finding new advocates. This compares the number of citations achieved by a person’s academic outputs, against the number of followers they have on Twitter -- how you evaluate the resulting statistic would be an interesting topic for debate.

*Robert Shoemaker*



One way of viewing the research blog is, as it were, as the first draft of history; one can use one’s blogs, and any comments received, as a starting point for writing more formal publications, testing ideas, polishing text, and building an audience along the way. This is the approach we are adopting on our latest project, the Digital Panopticon.

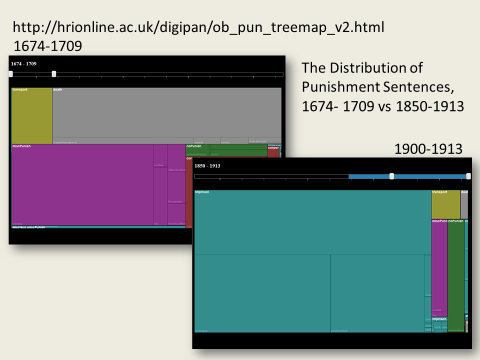
This leads us to the dreaded topic of Open Access, the emerging requirements to make research publications freely available through the internet. You will be relieved to hear that we do not propose to rehearse the various problems that Open Access raises, except to observe that it is hard to resist the conclusion that academics have lost the plot on this issue. By focusing on economic costs and business models, and insisting on ring-fencing forms of dissemination suitable for traditional methods of scholarship, we are in danger of undervaluing the opportunities Open Access provides to widen the audiences for academic writing, rethink the content of that writing, and to develop new styles and genres--not just blogs, but more interactive or iterative forms of writing, where there is a dialogue between the writer and a broad audience.



Many experiments in this area have failed to gain real traction, - but journals such as *Digital Humanities Now*; and the Open Access online journal - *Law, Crime and History* from Southampton, are pioneering new forms of co-operative Peer Review, pre-review publication, and Open Access that point the way towards a more useful, re-usable, transparent and Open form of scholarly communication.

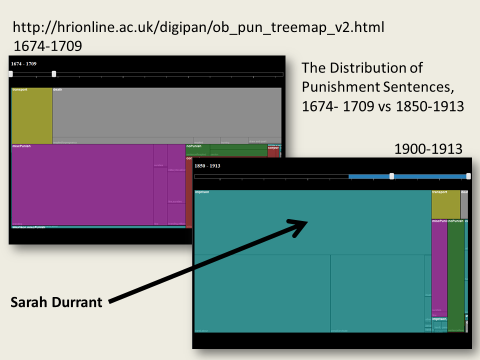
The internet also opens up the possibility of addressing the problem of the growing sophistication and lack of transparency of computer-based forms of analysis, in other words, the increasing divide between statistically and computer literate academics and everyone else interested in history. There are new presentational methods that can be used to summarise large bodies of data and the complex relationships identified through data analysis, and present them in an accessible form. A very straightforward example is any graph produced by the Ngram viewer of the sort we have already seen. Even more user-friendly possibilities can be found via the in-vogue term ‘visualisations’-- pick up any newspaper today and you will find numerous examples, presented with varying degrees of effectiveness. Used properly, visualisations, and more precisely, infographics, provide an accessible form of ‘distant reading’, allowing all of us to quickly see broad patterns, identify trends, and note anomalous cases--all of which can then be investigated in greater detail using a variety of research methods, including more traditional forms of ‘close reading’.

For example, in our latest project, the Digital Panopticon, we are using visualisations to document patterns of judicial and penal experiences in the lives of felons convicted at the Old Bailey--to identify commonalities in the 1000s of life stories we are tracing of those convicted from 1780 to 1865.

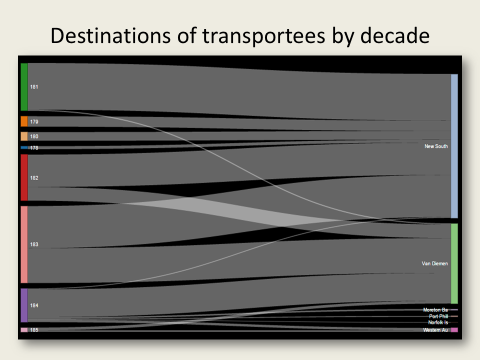


<http://hrionline.ac.uk/digipan/ob_pun_treemap_v2.html> [digipan; password PAVii712]

The project is still in its early phases, but here is an example, in the form of a ‘tree map’, of an interactive visualisation of punishment sentences at the Old Bailey over the entire period of the Proceedings, demonstrating more effectively than a table or graph ever could, the dramatically changing composition of Old Bailey punishment sentences.



And we can rapidly move from this ‘distant reading’ to individual stories – eventually, simply by clicking through from the relevant box to individual trials (though this feature is not yet enabled).

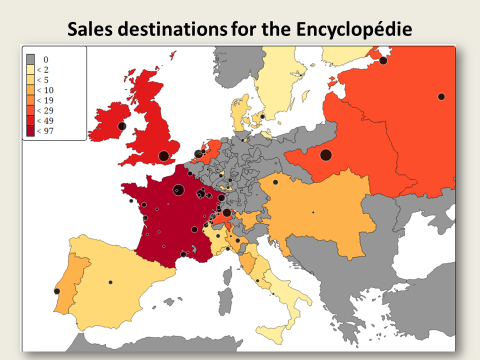


And here’s an example of our findings concerning a key stage in the convict story--from sentence at the Old Bailey to transportation to Australia, showing how the date of conviction determined, but not entirely, the place in Australia to which convicts were sent.

Ultimately, we will be charting the entire lives of these convicts (or as much as we can document), and our visualisations will chart various life courses from birth, through pre-Old Bailey convictions, the Old Bailey trial and sentence, to the actual punishment experienced, and subsequent life events (reoffending, marriage, death). When you are tracing the lives of some 90,000 convicts, the only way of summarising and analysing this evidence will be through visualisations like this; and they should also be useful for demonstrating patterns to a wider audience.



Another form of visualisation which is increasingly popular is mapping, using Geographical Information Systems (GIS). This kind of display has contemporary resonance through the widespread use of google maps in day to day life, but it is also a widely accessible research tool. There are many possible examples, but one some people may be familiar with is the French Book Trade in Enlightenment Europe project, which has created an interactive database of the detailed records of a Genevan publisher and bookseller, the Société Typographique de Neuchâtel, from1769-94.



The Société sold books to purchasers all over Europe, and their mapping function allows one to map the distribution of particular titles and authors, —analysing a total of 70,000 sales transactions. Here we have a map of the sales of the most famous Enlightenment publication, the Encyclopédie, which shows at a glance some surprising facts about at least this bookseller’s customers for the Encyclopédie --outside France and Switzerland, the biggest sales (in descending order of importance) went to the UK, Poland, Russia, and the United Provinces, while sales in what is now Italy and the Iberian peninsula, and particularly what is now Germany were surprisingly low.

Users can also interact with the site, creating their own maps according to book titles, authors, subject matter (theology, history), types of publication, language, and client’s profession, and limit by time period, thus allowing them to define their own research questions.

*Tim Hitchcock*

By way of a conclusion, there are a couple of terms that have recently been used in relation to the internet - ‘affordances’ and ‘disruption’. The first comes from the world of design, and refers to the uses that any bit of technology - any object - can be put. In writing history online, we are confronted with a remarkable set of ‘affordances’, in terms of genres of publication and communication, and new forms of analysis - from big data and distant reading, to the simple but magical power of keyword searching on the infinite archive. But as an ever growing list of industries has discovered, every new affordance brings in its train ‘disruption’. New career paths, open access, MOOCS, and the simple and wonderful empowerment of people outside the academy to produce their own histories, cut to their own cloth, challenges the historical profession to re-think how we do history online. It is not a challenge we can afford to ignore.

Our belief is that we are in a fantastic age of new and popular historical engagement, and while it is not being led by academic historians (nor should it be); we need to be actively involved - and make sure that we add our tuppence to the pot. Academics should do our bit to ensure that academic history is remade more open, more democratically accessible, and ever more able to do the business of allowing society to question itself, to question its values in light of its past, its politics and its inherited principles. Despite the ‘disruptions’ as long as we keep in mind these underlying purposes of history writing we can’t go far wrong.

