2015

# Bloody Code: Reflecting on a Decade of the Old Bailey Online and the Digital Futures of Our Criminal Past

Howard, Sharon

# BLOODY CODE: REFLECTING ON A DECADE OF THE OLD BAILEY ONLINE AND THE DIGITAL FUTURES OF OUR CRIMINAL PAST

***Sharon Howard***[1]

## Abstract

The Old Bailey Online was conceived at the turn of the millennium and has been online since 2003. In this article, I reflect on its evolution and its impact on crime history and digital history, and I explore some key themes and challenges for the next decade: improving digital and online access to archival crime records; funding and sustaining digital resources; and building skills to make the best use of these resources. I emphasise the importance of sharing and re-using digital content, and of building partnerships within and beyond the academy, with public and commercial institutions, and with the huge non-academic audience which has been a key factor in the success of the Old Bailey Online.

**Keywords:** Old Bailey Online, digitisation, crime and criminal justice archives and sources

## Introduction

My academic apprenticeship, at the University of Aberystwyth, was spent engrossed in two things: firstly, the criminal court archives of early modern Wales and northern England, and secondly, the potential of the Internet for research and teaching, and simply opening up early modern history to as many people as possible. The latter was not an entirely respectable academic interest in 1999, and even now it still feels remarkable that I have been able to spend several years shamelessly indulging in that obsession as part of a team creating a series of large digital history projects mainly focused on the history of crime, beginning with *The Old Bailey Online* (OBO), which have more than fulfilled my early hopes for history online.[2] OBO in particular has already made its mark on academic research: I know of more than 300 scholarly publications citing it as a source, and it is widely and often creatively used in university teaching.[3] Beyond academia, it has directly contributed to countless family histories, as well as published novels, radio series and TV dramas and documentaries.[4]

---

[1] Sharon Howard is currently project manager for the AHRC-funded project, The Digital Panopticon (2013-17), Humanities Research Institute, University of Sheffield, Sharon.Howard@sheffield.ac.uk

[2] http://www.oldbaileyonline.org.

[3] See 'Publications that Cite the Old Bailey Proceedings Online', http://www.oldbaileyonline.org/static/Publications.jsp; Drew Gray, 'Putting Undergraduates on Trial: Using The Old Bailey Online as a Teaching and Assessment Tool', *Law, Crime and History* 4(1) (2014), http://www.pbs.plymouth.ac.uk/solon/hjournal2014Vo4.html.

[4] Notable broadcasting examples include *Garrow's Law,* BBC TV 2009-11 and *Voices from the Old Bailey,* BBC Radio 4, 2010-14. OBO was also the focus of a recent creative writing project at Sheffield, *Beyond the Bailey*: http://beyondthebailey.wordpress.com/

But what of the archives, and the peripheries, the first of my obsessions? In May 2013, the *Financial Times* told us that more tower cranes have been erected in London in the past three years than the rest of the UK put together: when I review digital crime history I fear that I unwittingly contributed to a similar metropolitan distortion.[5] There have often been concerns that OBO might skew researchers' attention disproportionately towards London and the higher courts.[6] More than ten years since OBO's launch, vast swathes of British crime and punishment archives - especially the local and provincial - remain entirely untouched by digitisation. It is not simply a lack of searchable websites; there are surprisingly few historical crime and justice datasets deposited in the UK Data Archive.[7] OBO, for all its impact in other ways, has not led the way in open access digitisation of primary sources for British histories of crime and punishment. While there *has* been large-scale digitisation of some sources that crime historians use, the bulk of it is not freely accessible, and even less has been done *by* or *for* us.

There are several reasons for this. Not least, funding for such projects has become harder to obtain than it was in the early 2000s. There is more competition, and funders are more demanding. OBO was in the right place at the right time, a beneficiary of the New Labour government's enthusiasm for the digitisation of Britain's cultural heritage, and its willingness to provide large amounts of money with relatively few strings attached.[8] Moreover, early modern court archives are not at all like OBO's consistently packaged, readable trial reports. They are unwieldy, often dirty, fragmentary, and intimidating in overall scale. Documents can vary hugely in size, structure, handwriting, materials and condition, defying any 'one size fits all' approach to digitising. They are frequently written in heavily abbreviated Latin, or legalised English, or an unholy mix of both. They are rarely indexed in any detail. Moving into the nineteenth century, they become perhaps more standardised - some are even printed, to the early modernist's amazement - but even more voluminous.

How much easier is it to turn to Old Bailey Online for immediate digital rewards than to start new digitisation projects with such awkward and intractable material? And equally, from the

[5] 'Cranes lift London to towering height over rest of UK', *Financial Times*, 5 May 2013.
[6] Drew D. Gray, 'Review of The Old Bailey Proceedings Online', *Reviews in History*, no.897 (May 2010), http://www.history.ac.uk/reviews/review/897; Steve Poole, 'Tales from The Old Bailey: Writing A New History from below', *History Workshop Journal* 59, no.1 (2005) 282–284, doi:10.1093/hwj/dbi027 (these concerns, though only touched on in Poole's review, were a recurring theme in discussions at the conference).
[7] A search for 'crime OR punishment OR justice' covering dates to 1900 at http://discover.ukdataservice.ac.uk/ returns just 41 results, and many of these are not British records or specifically focused on crime and punishment. (Search at 30 August 2014).
[8] Andrew White, 'Digital Britain: New Labour's digitisation of the UK's cultural heritage', *Cultural Trends* 20, 3-4 (2011) 317–325, doi:10.1080/09548963.2011.589712.

point of view of the researcher, why struggle with that kind of source material when they can turn to something like Old Bailey Online instead? I am always delighted to learn of young researchers who chose history of crime because of Old Bailey Online. But as the archives grubber who used to complain about the over-representation of a few small counties in the history of early modern British crime, I am now rather more ambivalent.

I was asked to introduce themes and challenges that I think are important for the future of digital crime history. So here is the first challenge: improving digital access to the hundreds of thousands of crime records in our local as well as national archives. A second challenge: as always, how to pay for it and sustain it in the long term. A third is improving digital skills: in particular, better understanding of data management and of the online tools we use as a matter of course, and how to work with people who have more advanced and specialised technical skills. And then there are two themes I want to emphasise that might help us to face these challenges: the need to re-use, recycle and share digital content; and the importance of collaboration and partnerships in order to make that happen.

## 2 Re-use and Collaboration: Old Bailey Online as a Case Study

Part of the challenge of creating Old Bailey Online, apart from its sheer volume, was the need to capture two different kinds of information. For much of its existence the Proceedings had been a quasi-official record in a consistent format of (nearly) all the criminal trials held at the court; thus, it was ideal for quantitative analysis, which required structured data. But at the same time the trial reports possessed many rich witness narratives that could only be truly represented by full text transcription. This dual identity was resolved by transcribing the texts using a double-rekeying process that is less accurate than traditional standards for scholarly editions, but more accurate than automated Optical Character Recognition (OCR) techniques, and then tagging the transcriptions with eXtensible Markup Language (XML) for structured elements that can be extracted into a database.[9]

There were downsides to this: it was expensive and time-consuming to create, and unwieldy in practice to work with. But the benefits have been huge and far-reaching: accuracy, completeness and *versatility*. The resource was readily appreciated by a wide range of different website users: family historians, teachers and students, crime and legal historians, historians of material culture, Londoners who simply found reading the stories of their city's past addictive, and many more. But there is more to it than that. What had been created was not simply a digital reproduction of a primary source on a website, with all the constraints of

[9] For more detail on the process, see the section on 'Technical Methods' at http://www.oldbaileyonline.org/static/Project.jsp

browser-based online search. It is also, quite independently, a dataset that can be processed by machines in many ways: converted into other formats; indexed in different ways for different kinds of search; transformed with new XML markup for different purposes. There have been uses of OBO data that no one predicted in 2003, far beyond the creators' research agendas and ambitious visions for opening up access to 'history from below'.[10] Moreover, many of those re-uses have involved active collaboration with both creators and users of other digital resources; they have taught us how to work with people and organisations that do not approach problems in the same way as we do, and have different priorities and goals.

We ourselves have re-used the OBO data in a growing number of projects and resources, each providing new contexts and connections, the possibility of seeing trials and the people involved in them in different ways. The first of these was *London Lives 1690-1800: Crime, Poverty and Social Policy in the Metropolis*, launched in 2010. London Lives is still concerned with crime, but its research agenda was to investigate interactions between poverty, crime and formation of the modern state. As such, it places the original Proceedings data alongside a range of eighteenth century manuscript sources - parish records, court records, hospital records - and a selection of available datasets created by other projects.[11]

Our second re-use came in *Connected Histories* (2009-11), which was funded as part of a JISC programme explicitly focused on 'clustering and enhancing' digital content and breaking down online silos.[12] Connected Histories is a 'federated search' platform, which enables users to search the content of a variety of separate online resources in one place and make connections between them.[13] It required the formation of a number of important new partnerships: our project collaborators, the Institute for Historical Research in London; the other 20 owners of digital resources that can currently be searched in Connected Histories; Kings College London, who carried out invaluable quality evaluation work; the web design company Mickey and Mallory, experts in eliciting user feedback in the design

[10] Tim Hitchcock and Robert Shoemaker, 'Digitising History From Below: The Old Bailey Proceedings Online, 1674–1834', *History Compass* 4, no.2 (2006) 193–202, doi:10.1111/j.1478-0542.2006.00309.x.
[11] http://www.londonlives.org.
[12] See *Content Clustering and Sustaining Digital Resources* (JISC, 2011), http://www.jisc.ac.uk/publications/generalpublications/2011/08/ContentClusteringAndSustainingDigitalResources.aspx.
[13] http://www.connectedhistories.org.

process; the participants in focus groups, our academic advisors, and the programme managers from JISC.[14]

We began another project, *Datamining with Criminal Intent*, at about the same time as Connected Histories. This international collaboration under the Digging into Data programme, rather than creating another big resource in itself, aimed to facilitate 'big data' textmining approaches to our data by integrating it with two other platforms, Zotero (information management) and Voyant Tools (textual analysis).[15] (Our first tentative efforts in this area began considerably earlier, with a collaboration between the OBO team and the University of Sheffield Computer Science department on a prototype textmining/semantic web tool called *Armadillo*.[16])

We joined forces again with the IHR and Centre for Metropolitan History (CMH), and with the Museum of London Archaeology team (MOLA), for the next re-use and re-combination of resources, in a mapping/GIS project: *Locating London's Past* (2011-12).[17] This needed all of MOLA's expertise to tackle the daunting task of creating 'a version of John Rocque's 1746 map of London that could be used to accurately illustrate the distribution of data drawn from other sources'; and all our own to begin to bring order to the wealth of place names marked up in our XML in Old Bailey Online and London Lives. In contrast to MOLA's archaeological datasets, which are already fully georeferenced, or even CMH's structured hearth taxes data, in our XML the same place name may be written down in an array of different ways, and converting this profoundly heterogeneous information into usable data was a major headache. It is still something of a work in progress.[18]

And finally, our most recent project is *The Digital Panopticon: The Global Impact of London's Punishments 1780-1925*, a collaboration with universities in the UK and Australia, in which we are attempting to trace the lives and experiences of thousands of defendants in the Old Bailey Online during the period of transportation to Australia. This involves automated record linkage on a much larger scale than we have done previously, as well as the formation of

[14] Kings College, London, Department of Digital Humanities: http://www.kcl.ac.uk/artshums/depts/ddh/index.aspx; Mickey and Mallory: http://www.mickeyandmallory.com/.
[15] http://criminalintent.org/; http://zotero.org; http://voyant-tools.org/.
[16] http://www.hrionline.ac.uk/armadillo/index.html
[17] http://www.locatinglondon.org/.
[18] Peter Rauxloh, 'Mapping Methodology', *Locating London's Past*, 2011, http://locatinglondon.org/static/MappingMethodology.html; Jamie McLaughlin, 'The Geocoder: A Tool for Geo-Referencing Place Names', *Locating London's Past*, 2011, http://locatinglondon.org/static/Geocoder.html.

new partnerships with commercial and non-commercial data creators, and learning new big data and visualisation techniques.[19]

Beyond our own work, OBO data is to be found in a growing number of important resources and projects. It has been included in two large related resources for literary texts: *18thConnect* for the eighteenth century and *NINES* for the nineteenth century. These two belong to an expanding digital aggregation initiative, a collaboration among several US universities, linked by a shared interface and software tools. They are very similar to Connected Histories' federated search in conception, but they also have considerably more ambitious collaborative platforms, encompassing crowdsourced correction and peer review. It will be interesting to see the results of placing a source we normally think of in social history terms into its literary-historical context.[20]

The second field that I want to highlight is the use of OBO data by historical linguists. This dates back to 2004, when Magnus Huber, a linguist based at the University of Giessen, looking online for sources, stumbled on OBO and immediately recognised its potential. The process of transforming the XML dataset into a linguistic corpus involved identifying and tagging direct speech in the XML files, and has culminated in *The Old Bailey Corpus Online*, which includes '407 Proceedings, ca. 318,000 speech events, ca. 14 million spoken words, ca. 750,000 spoken words/decade'. [21] The OBO data has transformed eighteenth to nineteenth century English from a relatively neglected period in historical linguistic study to a flourishing field of enquiry, something that was certainly not anticipated at the outset.

Finally, there is the role that OBO has played in educating and training students, since its first launch. The availability of high-quality online sources has dramatically expanded the opportunities for school and university students to practise primary source analysis, and OBO is a key resource for teachers of the history of crime as well as other social and cultural history topics, and, for example, using its statistical functions to introduce students to quantitative methods. In addition to large-scale projects mentioned, we have made the data available to PhD students (and other researchers) for whom the website resources were insufficient. And, beyond crime and social history, OBO has also become a key resource for teaching advanced digital humanities skills, including wrangling data, building databases, and writing code for textual analysis.[22]

[19] http://www.digitalpanopticon.org.
[20] http://www.18thconnect.org/; http://www.nines.org/
[21] http://www.uni-giessen.de/oldbaileycorpus/index.html
[22] OBO makes a number of appearances in the digital history tutorials at *The Programming Historian:* http://programminghistorian.org/

I argue that there are two lessons to be drawn from OBO for everyone, whatever kind of project or source they have in mind. The first, most important, lesson is that you should aim to digitise in a way that most effectively captures the information in a particular source. This emphatically does not mean that you must always provide full text transcriptions and images, just as OBO does. From a practical point of view, for many archival records indexing key information may be a more viable form of digitisation than full-text with images, and almost as useful. Many important crime history sources do not contain the rich stories of the Proceedings. Verbatim transcription, especially if it fails to pay careful attention to document structure, is not always the best approach: for example, many English crime history sources before 1733 are in Latin and heavily formulaic, while nineteenth century sources are often registers in quite complex tabular formats.

I think that a rather good example of how *not* to digitise a crime history source for crime history is the ten years or so of criminal registers that we digitised as part of London Lives.[23] In our defence, that was a different kind of project: crime was only part of the focus, and we had to digitise a very diverse assortment of manuscript source materials - from tiny scraps of paper to massive bound volumes, in varying formats, from different kinds of institution. Of these, the criminal registers - and, other documents in tabular format - were quite a small component. Following the same digitisation method as Old Bailey Online worked well overall. It was ideal for the richer narrative sources: the depositions in sessions papers, the coroners' inquests, the settlement examinations. But it was far less well suited to tabular registers. When we decided to re-use the registers in the Digital Panopticon project for more systematic record linkage, we were severely hampered by our cavalier approach to their original structure.

All this brings me to my second lesson, which is that you should do your best to facilitate future re-use and collaboration. Of course, as we have seen, you often cannot know exactly what future re-uses may look like. But the basic principles to aim for are those of good data management practice more generally: the creation of data that is as accurate and *consistent* as possible and pays attention to the structure and meaning of the original source; and good documentation so that others can understand its meaning, how and from what sources it was collected, and what it may be missing as well as what it includes. This approach may well cost more at the beginning, and it may require more of an investment in technical skills and management, but it will make your efforts worth more in the long run.

[23] These are The National Archives HO26, Home Office Criminal Registers, Middlesex; see http://www.londonlives.org/static/CR.jsp.

## 2 Collaborations: Institutions, Publishers and Researchers

What kind of collaborations and partnerships do we need, and why are they so important? In theory, historians of crime are in the fortunate position of using sources that are of considerable interest to the booming online business of family history, marked by the current programme to digitise a huge swathe of important nineteenth century records of crime and punishment held at The National Archives.[24] In practice, as Andrew Prescott has recently commented, digitisation by commercial genealogical publishers has tended 'to fragment the availability and use of major categories of historical records', and crime records are no exception.[25]

These commercial online collections are problematic not simply because they are behind a paywall. In an ideal world, all these resources would be freely accessible to all. But digitisation is expensive; someone has to pay for it. The grim reality is that archives are under intense financial pressure and genealogical sources represent much-needed income. And yet this does not entirely justify the apparent zeal with which some institutions have rushed to exploit popular records, including the use of exclusive commercial agreements that lock digital collections behind paywalls for many years. Family history publishers create high quality, affordable resources for family historians, but terribly limited ones for academic crime historians. They are primarily intended for searching for names and places, and only a fraction of the information contained in many of the original records is indexed. The needs and priorities of family historians and academics overlap, but are not close enough that creating resources that can serve both groups well simply *happens*.

On the other hand, there is the business model aimed explicitly at universities and related institutions (rather than individual subscribers) as customers, that has created source collections like *Eighteenth Century Collections Online* or *British Newspapers 1600-1900*, both of which contain much valuable source material for historians of crime.[26] These collections, largely drawn from print sources rather than archives have been designed with academic audiences in mind, but they are virtually inaccessible to researchers outside academic institutions, as well as to academics whose institutions cannot afford the price tag. Even these 'academic' resources are still limited and often more problematic for academic

[24] The first tranche of these records has been published by Findmypast.co.uk, with more on the way: http://search.findmypast.co.uk/search-world-records/crime-prisons-and-punishment. A smaller group of similar records was earlier digitised by Ancestry.com.
[25] Andrew Prescott, 'I'd Rather be a Librarian: A Response to Tim Hitchcock, "Confronting the Digital"', *Cultural and Social History*, 17(3) (September 2014) 339, doi:10.2752/147800414X13983595303192.
[26] http://gdc.gale.com/products/eighteenth-century-collections-online/; http://gdc.gale.com/products/19th-century-british-library-newspapers-part-i-and-part-ii/.

research than many of their users realise. They have keyword search – but that is just about all they *do* have. This usually includes quite sophisticated fuzzy search options - but it needs to, since the quality of the underlying text can be so poor that not much is likely to be found without it.

One thing that they do share with the resources for family historians, however, is that they are black boxes that make it virtually impossible for a researcher to evaluate the quality of the data as a whole, and hinder any kind of use other than those the platform was specifically built for. A number of historians have expressed concern about the naive use of this kind of black box resource, and about the uncritical reliance on keyword search as a research methodology - whether in Google, ECCO, Old Bailey Online or anywhere else on the web that one can simply start typing terms into a box and receive an alluring list of search results, with no clear awareness of what may be *missing*.[27]

For example, a particularly acute issue for historians attempting systematic analyses using newspaper databases, created using OCR and presented as page images with no access to the underlying text, is not simply that there are many inaccuracies in the texts, but that the errors are far from evenly or predictably distributed throughout the dataset as a whole. All newspapers are not equal in the eyes of the OCR engine; entire newspaper titles can be more hidden from search than others and,

> much of that variability is associated with the historical and contemporary resources of publishing houses, meaning that major metropolitan papers typically sit at one end of the legibility spectrum and smaller, regional papers sit at the other.[28]

Beyond the limitations and problems of specific types of online resource, there are the divides between commercial rivals, and even more so between the two commercial digitisation models, that present further barriers to both re-usability and to collaborations among digitisers and users. The two groups of resource users, academics and family historians, tend to have rather different priorities; commercial publishers unsurprisingly focus

[27] Ted Underwood, 'Theorizing Research Practices We Forgot to Theorize Twenty Years Ago' (2014), https://www.ideals.illinois.edu/handle/2142/48906; Tim Hitchcock, 'Confronting the Digital: Or How Academic History Writing Lost the Plot', *Cultural and Social History* 10(1) (2013): 9–23, doi:10.2752/147800413X13515292098070.

[28] Carolyn Strange et al., 'Mining for the Meanings of a Murder: The Impact of OCR Quality on the Use of Digitized Historical Newspapers' 8(1) (2014), http://www.digitalhumanities.org/dhq/vol/8/1/000168/000168.html (para. 16). Hitchcock, 'Confronting the Digital', 12-14; Simon Tanner, Trevor Muñoz, 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, no.7/8 (July 2009), http://www.dlib.org/dlib/july09/munoz/07munoz.html; Kenning Arlitsch and John Herbert, 'Microfilm, paper, and OCR: Issues in newspaper digitization. The Utah digital newspapers program', *Microform & Imaging Review* 33, no. 2 (2004): 59–67, http://www.degruyter.com/view/j/mfir.2004.33.issue-2/mfir.2004.59/mfir.2004.59.xml.

on creating data that serves those priorities, and is less adaptable to other uses. That would not be such a problem if the data were not locked away and jealously guarded in its black boxes, where it can never be corrected or improved or enhanced – even though technologies and methodologies to enable improvement of existing data is continually developing. So publishers may lose out too, in the end.

But are there any alternatives to the black box? One useful model to consider is the federated search platforms already mentioned - the likes of Connected Histories, NINES and 18thConnect. In addition to bringing together separate resources for searching, these can enable existing data to be enhanced in various ways: for Connected Histories, we used automated techniques to extract names, places and dates from unstructured full text datasets such as newspapers for more structured searching options. They can facilitate at least limited access to subscription resources to aid evaluation, and their user collaboration facilities aid connection-making and scholarship.

A second model is the Text Creation Partnership (TCP), which is creating full, highly accurate transcriptions of content from major commercial page-image digital collections. The resulting texts are restricted to partnership members and resource subscribers for up to five years and are then released into the public domain (the images continue to be paywalled).[29] The TCP's flagship project is the transcription of *Early English Books Online* (EEBO), aiming to transcribe every unique text in the collection. In January 2015, a first phase of 25,000 EEBO-TCP texts will be released into the public domain, with another 45,000 to follow in the second phase in a few years' time. Proquest, the publisher of EEBO, benefits too, as it can also use the transcriptions to improve search facilities and user experience at the EEBO website.[30]

It ought not to be beyond our wit to translate that kind of public-private collaborative model for literary sources to historical crime records and other archival records with overlapping academic/family history user groups. Indeed, it should be noted that the commercial digitisation of The National Archives' records already takes place alongside extensive volunteer transcription and indexing projects, including TNA's own volunteering programmes and Ancestry.com's World Archives Project.[31] These projects receive rather less fanfare than the launch of subscription resources, and are often not as well known to researchers. In order to extend the reach of such initiatives, historians, archivists and publishers will need to

[29] http://www.textcreationpartnership.org
[30] http://www.textcreationpartnership.org/tcp-eebo/; http://eebo.chadwyck.com/home
[31] http://www.nationalarchives.gov.uk/get-involved/volunteering-projects.htm; http://community.ancestry.co.uk/wap/download.aspx.

build closer partnerships; academics need to be willing to compromise with commercial publishers on access; archives need to be willing to give up some short-term commercial exploitation opportunities for longer term benefits. Our experiences, from Connected Histories to the Digital Panopticon, have demonstrated that publishers are far from unwilling to work with academics to expand access where they can see potential benefits from giving away some of their data.

However, as the mention of volunteer projects suggests, if we are to make progress it is not enough simply to think about institutional collaboration or partnerships between academic projects and publishers. There will never be enough funding for entirely professional digitisation: can 'the crowd' make the difference? In recent years, there has been considerable research into collaborative user participation in digitisation – transcription, indexing, correction, tagging, annotation, linking, and so on, as well as practical projects and tools to facilitate such endeavours. Academics often express anxiety about the quality of work done by 'amateurs', and yet this is belied by the experiences of collaborative and crowdsourcing projects, from relatively simple correction of Australian newspapers to more difficult transcription of philosophical papers. [32] Importantly, what emerges from crowdsourcing experience is that typically the majority of contributions are not done by casual passers-by, but by a relatively small number of committed enthusiasts who bring significant expertise to their task and are amateurs only in the sense of being unpaid.[33] More than this, it is argued, engaging 'the crowd' in volunteer projects has a vital role to play in broader and deeper public engagement in cultural heritage and history: ultimately, the *process* is more important than the *product*.[34]

It should not be imagined that 'the crowd' is an easy option to get work done on the cheap. The OBO team have been trying to work this out for some years now with only limited success. Much of the problem is that we have tended to try to add in the crowd near the end

[32] *Trove* digitised newspapers, https://trove.nla.gov.au/newspaper; *Transcribe Bentham*, http://www.transcribe-bentham.da.ulcc.ac.uk/td/Transcribe_Bentham. Other transcription projects include The National Archives of Australia's *arcHIVE*: http://transcribe.naa.gov.au/ and *Marine Lives*, which is creating a digital edition of seventeenth century High Court of Admiralty records: http://www.marinelives.org/.

[33] Tim Causer and Valerie Wallace, 'Building A Volunteer Community: Results and Findings from Transcribe Bentham', 6, 2 (2012), http://www.digitalhumanities.org/dhq/vol/6/2/000125/000125.html; Rose Holley, 'Crowdsourcing: How and Why Should Libraries Do It?', *D-Lib Magazine* 16, 3/4 (March 2010), http://www.dlib.org/dlib/march10/holley/03holley.html; Nancy Proctor, 'Crowdsourcing—an Introduction: From Public Goods to Public Good', *Curator: The Museum Journal* 56, 1 (January 1, 2013): 105–106, doi:10.1111/cura.12010.

[34] Trevor Owens, 'Digital Cultural Heritage and the Crowd', *Curator: The Museum Journal*, 56, 1 (January 1, 2013) 121–130, doi:10.1111/cura.12012; Mia Ridge, 'From Tagging to Theorizing: Deepening Engagement with Cultural Heritage through Crowdsourcing', *Curator: The Museum Journal*, 56, 4 (October 1, 2013) 435–450, doi:10.1111/cura.12046.

of a project, and then hope something will magically happen while we go off to the next project; rather, it is important to start building an engaged community around your project from as early as possible. We have also struggled with user interfaces. We took too long to learn that we have to make participation simple and, again, build it into the design of the whole resource rather than attempt to tack it on at the end.

We may also have failed to understand the limits of what most users are willing to do: small-scale structured contributions, such as tagging or linking records, are much easier to ask for than larger investments of time and creative energy. So, for example, users of London Lives have been far more active using workspaces on the main website and doing nominal record linkage, than writing biographical articles on the project wiki. In 2011, when we added to OBO a number of new features including user accounts and workspaces with bookmarking facilities, we included a new, much simpler way for registered and logged-in users to report errors and issues (such as missing images) through a single click on the problem page itself. We have not particularly publicised it and submitted corrections are not displayed publicly. Without prompting, people have used the feature not simply to report errors in transcriptions but to add information from their research. The desire on the part of our site users to contribute what they know clearly exists: it is for us to find better ways to work with them to share that knowledge.

And finally, there is the potential role to be played by historians working in archives and libraries: we are all digitisers now, and have been for a long time. Unfortunately, most of us tend to do it rather haphazardly. (My own computer has folders of old databases, transcriptions, notes and so on, kept from public view because it is messy, and incomplete, and it would be thoroughly embarrassing to let people see all my past mistakes.) A decade ago, it was easy to get away with this but there will increasingly be mandatory requirements from funders to share research data. Ideally, this is something that historians should be looking to do not simply because a funder demands it but because of the potential benefits of sharing and re-using data among researchers: that is true of personal research datasets as well as large ones like the Old Bailey Proceedings.[35] However, for such activities to happen consistently, it is vitally important to provide better training in digital skills for students - not confined to students of 'digital history' - so they know how to create good, shareable data, how to look after it, *and* how best to re-use data shared by others.

[35] Seth Long, '"Re-purposing" data in the Digital Humanities: Data beg to be taken from one context and transferred to another', *Impact of Social Sciences*, http://blogs.lse.ac.uk/impactofsocialsciences/2014/04/02/re-purposing-data-in-the-digital-humanities/.

Digitisation and data creation do not have to be and *should* not be all about large funded projects. Indeed, that may not even be the best way to think about how to go about improving digital access to social history sources, including crime records. The overwhelming emphasis of academic digitisation has not been on this kind of source material, but editions of the works and papers of historical elites.[36] Historians of crime waiting for large freely accessible projects that will provide alternatives to the London-centric perspective of OBO can probably expect to be waiting for a very long time.

Academic historians need to think much more seriously about what they can do for themselves, how to work with family historians, and ways in which to make their work and the ad hoc digitisation historians have already been doing for years in the course of research more available and re-usable. At the very least, historians should be depositing those hard-won databases, spreadsheets, transcripts, in online repositories, as a matter of course.[37] Perhaps we can go further than that and listen to recent calls for 'participatory archives' initiatives that would bring together existing official sources with the research data of academic and family historians into a 'Digital History Commons'.[38] This would not *simply* create new digital collections that could be re-used, build on and added to over time; it could also function as a kind of digital history lab where historians would be able to work together and develop new skills, and also as a springboard for funding more conventional large scale projects. If we learn to digitise for re-use, and re-use to digitise, we can share and collaborate, and build partnerships that make some of the challenges of digitisation less intimidating.

[36] Tim Hitchcock, 'Textmining British Studies: an Overview of Recent Developments' (presented at the North American Conference on British Studies, Montreal, 2012), http://www.historyworkingpapers.org/?page_id=266.

[37] Options for depositing data are expanding; in addition to institutional repositories, there are initiatives such as the *Harvard Dataverse*: https://thedata.harvard.edu/dvn/ and *Figshare*: http://figshare.com/

[38] Mia Ridge, 'Creating a Digital History Commons through crowdsourcing and participant digitisation' (presented at the Herrenhausen Conference: (Digital) Humanities Revisited, Hanover, 2013), http://www.miaridge.com/herrenhausen/; Isto Huvila, 'Participatory archive: towards decentralised curation, radical user orientation, and broader contextualisation of records management', *Archival Science* 8 (2008): 15-36, doi:10.1007/s10502-008-9071-0.