Digital affordances for criminal justice history


This version is available from Sussex Research Online: http://sro.sussex.ac.uk/id/eprint/77732/

This document is made available in accordance with publisher policies and may differ from the published version or from the version of record. If you wish to cite this item you are advised to consult the publisher's version. Please see the URL above for details on accessing the published version.

Copyright and reuse:
Sussex Research Online is a digital repository of the research output of the University.

Copyright and all moral rights to the version of the paper presented here belong to the individual author(s) and/or other copyright owners. To the extent reasonable and practicable, the material made available in SRO has been checked for eligibility before being made available.

Copies of full text items generally can be reproduced, displayed or performed and given to third parties in any format or medium for personal research or study, educational, or not-for-profit purposes without prior permission or charge, provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way.
Digital Affordances for Criminal Justice History.

More than any other field of human endeavour, the criminal justice system has spawned a textual machine for knowing. One needs to look no further than the traditional law office or library, with its endless series of published law reports and series of statutes – both a visual claim to authority, and a form of working reference system - to understand that the criminal justice system is built on a foundation formed of its own archive. Its authority lies in recorded precedent and recorded statute; and on the secure record of arrest, trial and imprisonment. This authority in turn demands a uniquely sophisticated system of preservation and discovery. In recent decades the archive of these systems, and ever expanding body of reference associated with them, has been thoroughly reconstructed as a complex digital system – largely in service of the American and European legal and criminal justice professions; and secondarily, at the behest of family historians. In relation to the judicial side of this coin, for almost half a century, the entrepreneurs of knowledge exploitation have been creating ever more sophisticated digital representations of the law and legal process. Westlaw, Lexisnexus, HeinOnline, and so on, have created an ecology for legal research that has made the practise of the legal professions in the West synonymous with the use of digital search tools. In North America and Europe, to find a relevant statute or precedent any practising lawyer will turn to the internet. No argument is made to a higher court without recourse to online resources; and no legal document is signed or sealed, without first being checked electronically for errors. And for older records of arrest, trial, and punishment, Ancestry and FindMyPast, and a host of smaller companies, have diligently sought out and digitised lists of names to satisfy a largely Western audience seeking their criminal forebears. In future these digital resources will include the electronic records generated from the moment an individual comes in contact with criminal justice onwards. A crude measure of the significance of new technology to the legal professions is the 10,000+ attendees attracted to the annual Legaltech trade show held in January in New York and again in July in San Francisco.¹ And while historians of the crime and the law have

made relatively few direct contributions to the creation of these new digital resources, they have been more than happy to go along for the ride.

When combined with digitised newspapers, and parliamentary records; with Google Books, ECCO and EEBO; this digital archive provides historians of criminal justice in the West with a uniquely comprehensive body of sources. For eighteenth and nineteenth century US and UK, Anglo-phone print material, digital coverage of the judicial side of criminal justice is all-but complete. From case law, to public opinion via newspapers, to parliamentary debates, and drafting revisions, to implementation and reaction and repeal; the archive of the legal process is available. And the records of policing and punishment are rapidly catching up. These new resources seem to promise and demand new methodologies, new approaches, and new ambitions for the field. And certainly, criminal justice historians are producing ever more scholarship; and there are minor ‘turns’ away from the history of the law, towards histories of crime and punishment. But, perhaps surprisingly recent criminal justice history has been remarkably conservative – perhaps better researched, and certainly more rapidly produced, than the scholarship of previous generations, but nevertheless largely lacking in technical or intellectual innovation.2

In part, this intellectual conservativism reflects the nature of the criminal justice system, and the digital resources created to serve it. For all of its speed and apparent reliability, the digital representations of the archive of the law used by many historians runs to the rhythms of nineteenth-century librarianship, and are built on the labours of nineteenth and twentieth century librarians. And although many of the records of policing and punishment are now available via commercial sites, these are generally framed with the needs of genealogists in mind, making them difficult to exploit for more general analysis. The great series of British and American statutes and legal reports published in the nineteenth century, were among the first to be digitised. The materials chosen for republication as microfiche and microfilm in the mid-twentieth century – again filmed by commercial organisations in service of the legal profession – came next; followed eventually by long lists

---

of names scraped from the archives of policing and punishment. All of these materials have reflected the selection biases and categories of early modern, nineteenth and twentieth century thought, and commercial demand. They have been created to serve the needs of a rich Western legal system; and latterly, a rich community of genealogists. Both groups are perhaps understandably conservative, even reactionary. The authority of the law depends on continuity, and its promise of fairness is necessarily acted out across time; and the Western bias of Western data companies is perhaps to be expected. And for all of their fervour, family historians tend towards the train spotting end of the spectrum of historical analysis. But, as a result, these resources brings with them a number of substantial problems that have not been fully acknowledged by the historians who use them.

First and most obvious is the problem of selection. To find a nineteenth century legal precedent for a case in Illinois is a matter of seconds; but doing so for Ottoman case law is a very different matter – even though the records exist. Vast swathes of the legal records of Empire have been preserved in the United Kingdom, but they remain almost entirely unavailable in digital form; and will remain so, since they are of limited interest to the companies behind resources such as LexiNexis. And all those lists of names created to serve the needs of family historians, are equally biased towards Western sources. The records of the wider world are simply not relevant to the wealthy Western legal practitioners and family historians whose needs these publishers serve.

This is a problem that goes well beyond the history of the criminal justice system. Laura Puttnam has recently described the unintentional bias the digital revolution forces on historical scholarship more broadly.3 But, as criminal justice history is served so comprehensively by commercial information providers, this bias is a particular problem in this sub-field; and as importantly, has been largely ignored by the field’s scholars. It is also a problem made more acute by the ‘national’ character of legal systems, and the extent to which inherited reference materials incorporate a national perspective. We are encouraged to look first for a national precedent, at the expense of influences from beyond a perhaps porous border. Finding related materials produced just hundreds of yards apart, on either

side of the Canadian/US border, is difficult. It is near impossible to find cognate materials from either side of the US/Mexican border. Of course, this is not a new problem – it is a problem gifted to us by the nineteenth and twentieth century creators of the reference materials that now form the grist and gristle of the past online. ‘National’ histories, language barriers, bigotry and racism, effect all history writing. But, the ease of use associated with digital resources, creates an intellectual pool of gravity that subtly attracts us in the direction of conservative prejudice over clear-eyed innovation, and can only be escaped by self-conscious effort.

This new world of digital resources has a second, subtler, bias. Optical Character Recognition (OCR) is the technology of choice for most of the companies providing access to historical legal materials. It is also the technology of choice for commercial providers including Google, Cengage Gale and Adam Matthew. For poorly produced materials, such as eighteenth-century newspapers, OCR can produce results delivering an accuracy below 50% for semantically significant words.\footnote{Simon Tanner, Trevor Muñoz, and Pich Hemy Ros, ‘Measuring Mass Text Digitization Quality and Usefulness’, \textit{D-Lib Magazine} 15.7/8 (2009).} The results are much better for high quality print such as statutes. This largely inaccurate OCRd text is then used to navigate an image of the original page, giving the impression that the user is searching an accurate transcription.

For almost all purposes this is entirely reasonable; and most scholars apparently believe the results of an OCR search are ‘good enough’. In a recent article Carolyn Strange, \textit{et al}, have undertaken an analysis of the effect of poor OCR on cross comparisons of text in a legal corpus, and demonstrated that an error rate of up to 20% or even higher, does not substantially hinder linguistic analysis. But, what has not been investigated is the extent to which OCR in combination with keyword searching is fundamentally biasing our access to the underlying source material itself.\footnote{Carolyn Strange, Daniel McNamara, Josh Wodak, and Ian Wood, ‘Mining for the Meanings of a Murder: The Impact of OCR Quality on the Use of Digitized Historical Newspapers’, \textit{Digital Humanities Quarterly}, 8, no. 1 (2014). http://www.digitalhumanities.org/dhq/vol/8/1/000168/000168.html.} OCR errors (and the search errors they imply) make ephemeral materials more difficult to use; and make particular types of text and image difficult to locate. Finding tabular information in nineteenth-century newspapers or text in


italics, for example, is almost impossible. And yet, because we are presented with a digital image of an original page at the end of a keyword search, we are lulled into a false sense of accuracy, and a false sense of the comprehensive character of the results. It is remarkably telling that the vast majority of historical materials online, do not allow the user to interrogate the underlying OCR text. This just re-enforces the selection bias generated by the existence of commercial datasets designed for legal practitioners instead of historians, and the biases embedded as a result of their construction from nineteenth and twentieth century reference materials. Endless layers of prejudice and selection are laid one on top of another.

The design of these digital materials have one further, and arguably more damning, result. Legal materials are organised via metadata schema that seek to represent at face value the nature of the underlying text. Statutes and trial reports are divided one from the other according to criteria that have existed for half a millennium. In the first instance, this encourages historians to read these documents as generic formulations, demanding a specific, conservative reading. We read a trial account as representing a single drama, from the perspective of a juror, or a judge – always seeking to discern the logic of a conviction. We read a statute in light of its drafters’ intentions, and as the outcome of a particular sort of Parliamentary process (perhaps like a trial, as a drama). And we read an arrest, an arraignment, or prison transfer as part of a legal bureaucracy. In all these instances, the limits of our reading is shaped by the bureaucratic division that separates one ‘trial’ from the next; one statute, from its parliamentary bedfellow; one step in a legal dance of punishment, from the next. This imposed metadata serves to limit how we engage with what in reality is simply a massive body of inherited text.

Collectively, the new digital form of the sources we use to generate historical insight have had the effect of ensuring that the history of criminal justice remains Western-centric. It encourages us to reading the text ‘with the grain’ rather than against it. And it takes as it primary object of study the evolution of the legal process – from the rise of the adversarial trial to the education of lawyers; from the rise of the ‘police’ to the nature of the trial evidence the police present. If you were an old lawyer or policeman seeking a genealogy for
your profession this might seem enough. But, as historians interpreting the past to the present, and facilitating an ongoing conversation with the dead; it is simply not sufficient.

The criminal justice system – in policing, in trial and in statute – is where the state negotiates its relationship with the public; and where standards of public behaviour and private moralities are most fully exposed. It is where power is exercised – either with terror or rhetoric; and it is where power of the state and of community is most fully experienced. The archive of the law is where the full measure of human experience is most fully recorded. And yet, the focus of criminal justice history remains largely on the practise of law. This is not sufficient, nor is it necessary.

One remarkable aspect of the new digital guise in which we now encounter the remains of the dead, is that although materials are structured to reproduce long-standing systems and divisions, these systems are no longer necessary pre-requisites for their use. The metadata that describes one body of text as a ‘statute’ and another as a ‘trial’ may add knowledge about the text that follows, but it does not need to determine how we read it. When all text is data, we are no longer forced to read our inheritance in the ways its creators initially demanded. And this is not limited to ‘text’. When place and space, image and sound are all equally coded forms of data, the divisions between them become irrelevant. We can locate every word recorded in its precise place on earth. We can divide materials afresh – between types of language of our own definition. We can make images speak, and books dissolve into their constituent words. And we can interrogate more fully how all of these remains of the dead, relate to both a knowable past and a constantly evolving present.

One of the favourite observations of the punditocracy is that the digital revolution creates a series of ‘affordances’ leading to endless ‘disruption’. The ‘affordance’ created by computerisation to stream sound and video from your bedroom has ‘disrupted’ the music industry; and the smart phoned have allowed Uber to ‘disrupt’ the system of licensed cabbies. What follows are three suggestions for how the ‘affordances’ created by the digital archive might ‘disrupt’ the criminal justice history.
The first of these affordances would be created by temporarily ignoring all the metadata associated with inherited text. This would allow us to eliminate the distinctions between Parliamentary debates, statutes, newspapers articles, trial accounts and fiction; and to recognise that what is readily available to study is an almost infinite archive of ‘text’. Whether we then apply a Text Frequency/Inverse Document Frequency (TF/IDF) methodology, or some variant of ‘Topic Modelling’, we could then identify patterns between text that are not predicated on the assumptions of genre.6 This would allow us to see the law as part of a wider social system; and to read the law ‘against the grain’. Where does literature and statute use similar language? When do defendants ape the language of the newspapers, or the language of fiction? Stepping back from inherited categories of text would allow us to effectively critique the systems we have uncritically inherited from the past.

This exercise would also encourage us to think differently about how language ‘encodes’ the past. History, from its Rankean origins onwards, has placed textual comparison at the heart of its methodology; and yet we seldom ask explicitly whether and how text relates to a knowable past. Taking as our object of study an undivided pool of text in its fullest compass challenges us to revisit that relationship and interrogate how texts actually encode the things we read into them. And just by way of a starting point, we might consider dividing text differently. We could, for example, divide inherited text between that which purports to represent speech uttered across real space (in courts, churches and legislatures), and text that is written from mind to hand, to an imagined audience (forms filled and essays written). I have no idea what the outcome of such an exercise would be, but am certain that recorded speech has a different relationship to the past, than does ‘mind→text’. The former encodes a physical action in a material space, interpreted by a person other than its maker; while the latter can exist independently of its reception or intended meaning. In broader terms, interrogating this division would also force us to think more clearly about the role of the individual – it would effectively foreground human agency, and background institutional structure.

A second affordance rests on the equally simple observation that once place is encoded in the way that text has been, the boundaries between these forms of data evaporate. Every point on the globe is available via Google Earth and GIS. And words, particularly speech, happen in particular places – many of which either survive, or are possible to reconstruct. Combining words and place creates a new way of understanding the past. In many respects, place is already tacitly present in much of our inherited text. A trial report assumes a particular courtroom; a Parliamentary debate assumes a chamber, and a sermon, a specific church. But adding place to text more explicitly allows us to complicate the nature of the record and, as with speech, to complicate the divisions we apply in our analysis of text. Does text change with geography? Can we glean differences in the character of a text, neighbourhood by neighbourhood, from region to region? Or even if we start from inherited metadata; can regional and national boundaries help explain observed changes in text?

By combining apparently disparate types of data, we allow ourselves the opportunity to ask different questions about the past. In part this would again, effectively foreground the ‘experience’ of the law, but in the case of encoded locations it also foregrounds the distribution of historical knowledge across the globe – forcing us to at least acknowledge the global context of our Western and national narratives. And even if we choose to eschew these intellectual opportunities, adding new varieties of metadata, gives us a different test and measure of the nature of institutional change – over time, and through space.

A final affordance emerges from the digitisation of records by companies such as Ancestry and FindMyPast. Adding this growing prosopographical database directly to the records of the management of the criminal justice system, again changes how that archive of social ordering can be read and analysed. By the middle of the nineteenth century most lives were recorded somewhere; and criminal lives were being recorded almost everywhere. For many of the dead we have descriptions and photographs and trails of reference spanning multiple records series. We have lives in their fullest compass, available in unbelievable numbers. And the family historians, for their own purposes, are gradually making those

---

7 See [http://www.ancestry.co.uk/](http://www.ancestry.co.uk/) and [http://www.findmypast.co.uk/](http://www.findmypast.co.uk/).
lives coherent and discoverable. Gaining access to the underlying records and life archives, in a form that can be fully exploited is difficult, but recognising their value is easy. If we can interrogate them at scale; and map them against the bureaucracies of the criminal justice system, it becomes possible to add yet another mode of analysis to the wider project of understanding the role of the criminal justice system in a wider social history.

The bureaucrats of the criminal justice systems – its police, its lawyers, judges and clerks – knew perfectly well what they thought they were doing – and we normally accept their account. But, testing this against the collective experience of a wider population would reveal the unstated characteristics of the criminal justice system about which our predecessors were silent. Accent, gender, height, dress, all determined how you were (and are) treated by the criminal justice system. Recovering this element of the process at a statistically significant scale, changes what we can ask.

As with the recovery of speech and the addition of space; incorporating lives into the archives of the criminal justice system, in part, re-orientates our analysis towards the ‘experience’ of the criminal justice system; but it also allows us to interrogate more fully the narrative of institutional change that forms the basis for so much current work. Knowing that the evolution the trial process or the growing role of lawyers are reflected in the collected lives of individual defendants, allows us to understand the criminal justice system as a working component of a broader social system in a new and more sophisticated way.

Each of these individual ‘affordances’ is being exploited by one or more projects across the digital humanities, and digital criminal justice history. The Digital Harlem project exploits place to add nuance to our understanding of text. The Prosecution Project directed by Mark Finnane is stepping back from the structures of the inherited archive, to both interrogate the nature of legal text, and legal process. And Helen Rogers’ Writing Lives project is making serious use of individual records, to reconstruct individual, criminal lives. And finally, the Digital Panopticon project, led by Barry Godfrey and Robert Shoemaker, is tying together dozens of record series, to reconfigure our understanding of criminal

8 http://heuristicscholar.org/digital_harlem/.
9 https://prosecutionproject.griffith.edu.au/.
10 http://www.writinglives.org/.
transportation and imprisonment through a process of collective biography. The list could go on, but there is much more to be done.

These projects challenge us to reconsider the ‘object of study’ at the core of our practise. Is criminal justice history about the evolution of the law and its bureaucracies? Or is it about the role of the law in the evolution of society? Is it about social change as reflected in criminal justice? Or are we merely creating a genealogy of present day systems?

The digital reconstruction of the archives of criminal justice allows us to fundamentally change the questions we ask, and the methods we use to answer them. To date, this opportunity has been largely neglected by criminal justice historians, but it nevertheless challenges us to both more clearly define what it is we think we are doing; and further challenges us to have a bit of imagination about how we do it.

---

11 https://www.digitalpanopticon.org/.