Data Mining with Criminal Intent

Final White Paper

August 31, 2011

Dan Cohen (US Lead), George Mason University
Frederick Gibbs, George Mason University
Tim Hitchcock (UK Lead), University of Hertfordshire
Geoffrey Rockwell (Canadian Lead), University of Alberta
Jörg Sander, University of Alberta
Robert Shoemaker, University of Sheffield
Stéfan Sinclair, McGill University
Sean Takats, George Mason University
William J. Turkel, University of Western Ontario
Cyril Briquet, McMaster University
Jamie McLaughlin, University of Sheffield
Milena Radzikowska, Mount Royal University
John Simpson, University of Alberta
Kirsten C. Uszkalo, Independent Scholar

Executive Summary

The *Data Mining with Criminal Intent* (DMCI) project brought together teams from the United States, the United Kingdom, and Canada to create a seamlessly linked digital research environment for working with the *Proceedings of the Old Bailey*. This environment allows users to

1. Query the 127 million words of trial records available through the **Old Bailey** Online.
2. Save and manage the result sets in a **Zotero** account, and then
3. Send the selected texts and results for analysis and visualization to tools like **Voyeur**

More generally, the project explored the tools and infrastructure that will make it possible for the 'ordinary working historian,' not currently using these techniques, to integrate text mining and visualization into his or her day-to-day work.

Table of Contents

0. Introduction and Aims of the Project

The *Data Mining with Criminal Intent* (DMCI) project (http://criminalintent.org) brought together three teams from three different countries. Each team provided a complete and complementary digital research resource – the online trial records of the Old Bailey (UK), the bibliographical management software (USA), and the text analysis portal (Canada) – with the shared goal of creating a linked research environment where it would be possible to:

1. Query the **Old Bailey** site
2. Save different result sets directly to a **Zotero** library where they could be managed
3. Send the results from Zotero to text analysis and visualization tools like **Voyeur**

In each case, existing tools were augmented and enhanced to allow for seamless integration. More generally, DMCI aimed to discover what tools and infrastructure would enable the 'ordinary working historian' (OWH) to integrate text mining into her or his day-to-day work.

The Old Bailey contains about 127 million words of text related to past crimes. Knowing that a myriad of unusual and compelling stories can be found in this resource, especially using tools that go beyond reading individual trials, we aimed to construct a way to zoom out to look at statistically significant patterns in the data and to zoom in to consider cases.

This white paper begins with a discussion of our technical developments, moves to textual discoveries, considers the challenges and opportunities this model of research presents to historians, and reveals some of the larger historical trends these tools have illuminated in the data.

1. The Old Bailey API

The Old Bailey website (http://www.oldbaileyonline.org) was conceived in 1999, the architecture was originally designed in 2001, and the full body of trial records covering the period 1674-1913 was launched in stages in 2003, 2005, and 2008. It makes available 127 million words of accurately transcribed historical text, recording the details of 197,000 trials held at the Old Bailey or Central Criminal Court in London, between 1674 and 1913 (Hitchcock & Shoemaker 2006). Having been double-entry rekeyed (transcribed) to an accuracy of about one character error in 3000, it is one of the largest bodies of accurately transcribed historical text currently available online. It is also the most comprehensive thin slice view of eighteenth and nineteenth-century London available online. All manner of people, across all ages, genders, and classes appeared in London's court. Their words, collected in the evidentiary statements (comprising 85% of the *Proceedings*), detail crimes, their ancillary activities, their perpetrators, and their victims. All of human life is here (Huber 2008).

The architecture for the 2001 site design constrained most users to a single (if complex) series of search pathways and pre-determined forms of analysis. It used keyword searches and tag queries that called up features such as 'offence type,' 'verdict,' or 'punishment.' Although the format restricted searches to the text and tag information architects and designers deemed most useful, scholars actively engaged with the *Proceedings*. With an invested user base in place, we sought to leverage alternate and emergent modes of exploration.

DMCI was conceived of as one way of escaping restricted research pathways and ensuring that the Old Bailey material could be explored with the full range of analytical and visualization tools evolving on the web. For this to happen, we needed an application programming interface (API) that would allow large numbers of trials and texts to be either queried at a distance, or freely downloaded for remote use and analysis.

The Old Bailey API (OBAPI) was designed and implemented in the summer of 2010 by Jamie McLaughlin. The demonstrator interface, publicly available on the main web site, provides an alternative search form for users to create subsets of texts and files, and allows keyword and structured searches (see http://www.oldbaileyonline.org/obapi/).

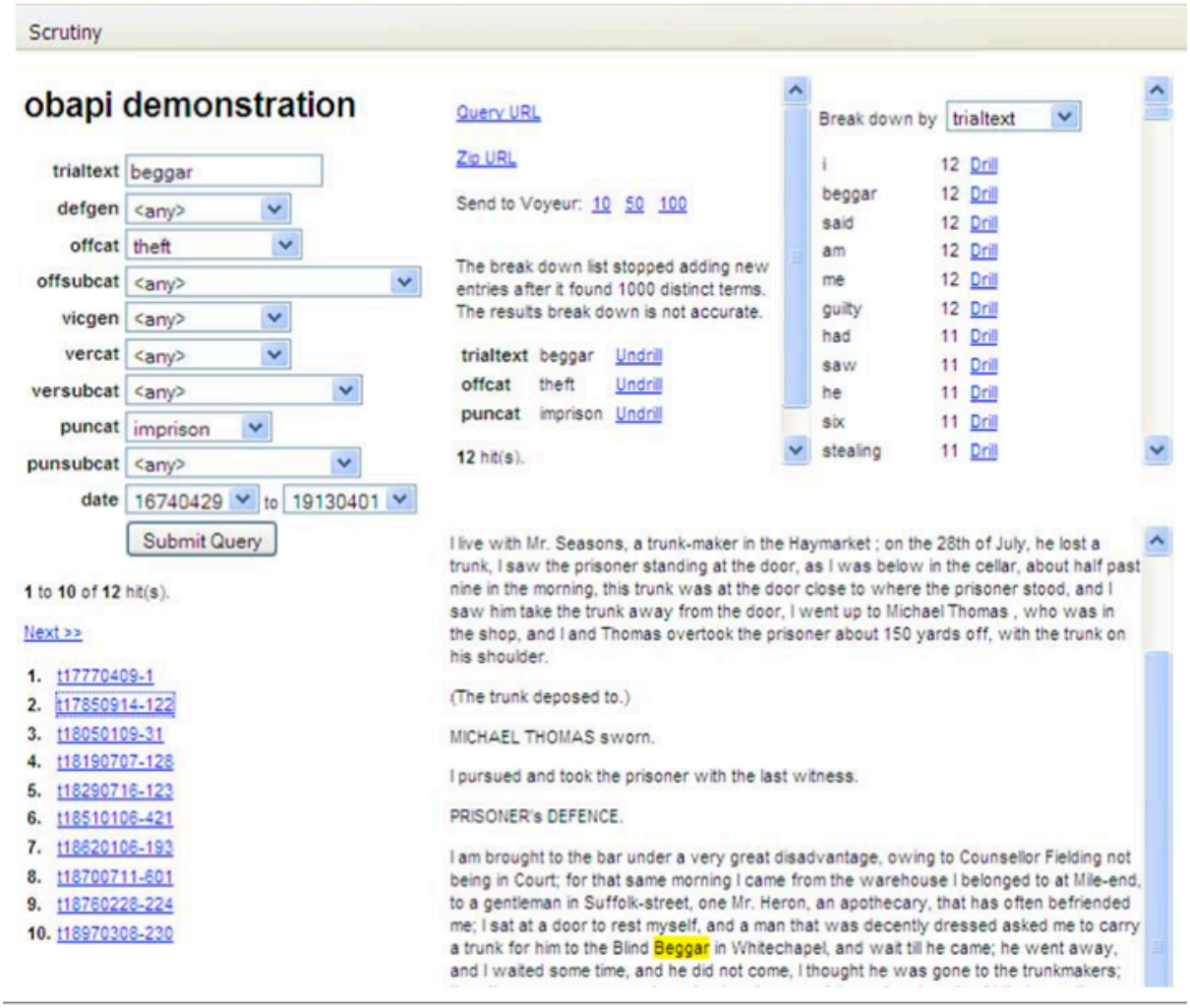

Figure 1: The Old Bailey API (OBAPI) Demonstrator

There are a number of direct and immediate advantages to this API. On the server side, it is faster than the current Apache Lucene-based search facility. Keyword and structured searching is better integrated into the site, making it easier for the user to perform searching and statistical analysis. A frequency table that illustrates how many of its hits contain certain 'terms' (i.e. categories of controlled vocabulary contained in the XML tags or words in the unstructured trial text) can accompany search result sets. The associated frequency table also contains links beside each term which allow the search to be further refined.

The process of developing an API has changed and extended how the *Proceedings* can be used. Rather than searching by 'defendant,' 'offence,' or 'victim,' users now search by 'trial.' Understanding the impact this shift has on process and result has been one of the challenges of this project. A researcher working with Old Bailey material now choses to investigate the *Proceedings* as a vast collection of 197,000 generically similar trial reports (as a collection of

"texts") or look at 'offences' and 'defendants' as units of interest.. The OBAPI makes it easy to maintain this balance.

Moreover, the OBAPI also makes it easier to understand the *Proceedings* as constituting a 'massive text object' that can be beneficially passed onto tools optimized for managing texts, tools like Zotero and Voyeur.

2. Zotero As Intermediary

Zotero (http://www.zotero.org), a personal research environment, was brought into the project to act as an intermediary layer between the digital collection of the Old Bailey and the text mining and visualization capacities of Voyeur. The goal was to allow the OWH to take finer-grained control of the sources she or he wished to analyze. The Zotero intermediary linking facilitated this nuanced analysis, analysis which lead to some promising new research methods and scholarly discoveries about crime and British history. To do this, however, we first needed to create the necessary infrastructure and connectors.

Fred Gibbs of the Roy Rosenzweig Center for History and New Media at George Mason University (http://chnm.gmu.edu/) wrote, maintained, and updated a Zotero 'translator' for the Old Bailey site. Translators are the technology that allows Zotero to recognize semantic objects on the web such as books, articles, manuscripts, and letters, and allows the user to save those objects to their personal library by clicking on an icon in the browser address bar. Translators allow Zotero users to hand-select scholarly items from the web and combine them in *ad hoc* collections for further research. The Zotero translator, used on an experimental basis during the prototype phase of this project and shipped to all Zotero users with the release of Zotero 2.1 in early 2011, enables Zotero users to import individual cases into their personal research library, and to search, sort, and combine them into sub-collections.

Since articles and books are the most frequently used types of scholarly evidence, the team exploited an apt and underused Zotero item type, the court case, to create a collection of data derived from the Old Bailey. Following the deployment of the OBAPI users can also use the Zotero translator to save pointers to 'slices,' or large numbers of legal cases that match defined criteria. Rather than saving individual cases, which might number in the thousands, Zotero saves a URL that points to the search results, results which can be utilized by Voyeur.

Gibbs wrote a plugin for Zotero that packages any text from an *ad hoc* collection of Zotero items and beams it to Voyeur. This plugin takes the full texts and metadata of cases,

concatenates them, and sends them via a HyperText Transfer Protocol (HTTP). As one of the deliverables for this Digging into Data project, the plugin is extensible and can also send the same text package to other analytical services. For more information about the Zotero plugin, see Appendix 1.

With the infrastructure in place, we explored a few research streams. As shown in the following figures, Gibbs used Zotero on the Old Bailey site, via the regular Old Bailey search pages and via the OBAPI, to extract records that contain the word 'poison.'

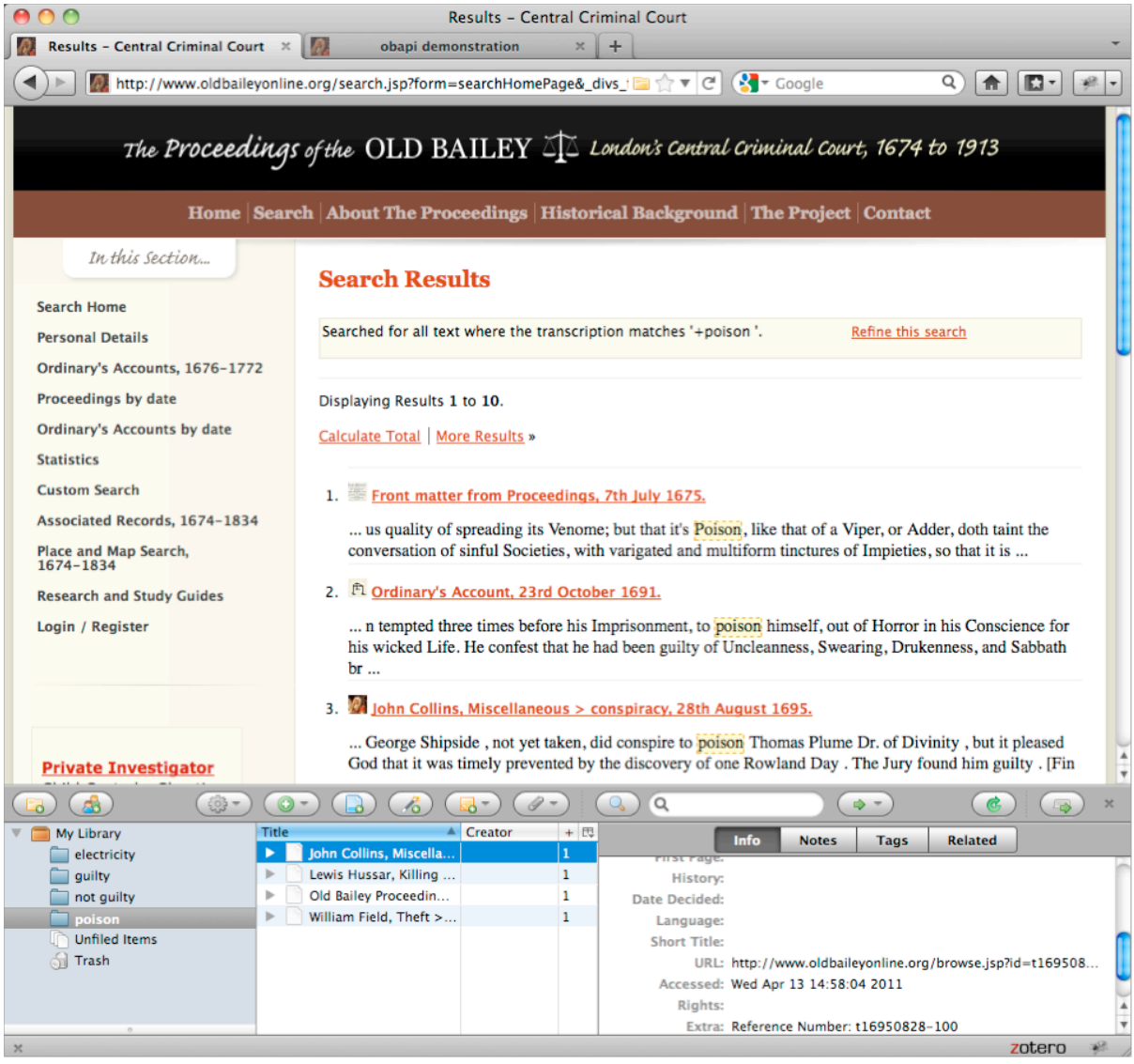

Figure 2: Searching the *Proceedings* for 'Poison'

Within his Zotero collection, Gibbs used the new plug-in to send selected cases to Voyeur, in order to visualize the particularities of specific legal records. Filtering out the stop words and legalese (e.g. 'prisoner'), he noted the common appearance of the verb 'drank.' Zooming in,

he observed the routine appearance of the nearby noun ‘coffee.’ Was this strong, bitter drink a known medium for administering poison? Or did cheap coffee taste so foul that someone, having fallen ill after drinking coffee, could feel they had coffee-poisoning? This preliminary research suggests that coffee, like other beverages, may have been a vehicle for London’s poisoners; the term ‘poison’ rarely occurs near the verbs ‘ate’ or ‘eaten.’

Zotero also empowers users in selecting relevant documents for textual analysis, freeing them from the constraints of predetermined categories provided by website search forms or archive metadata.

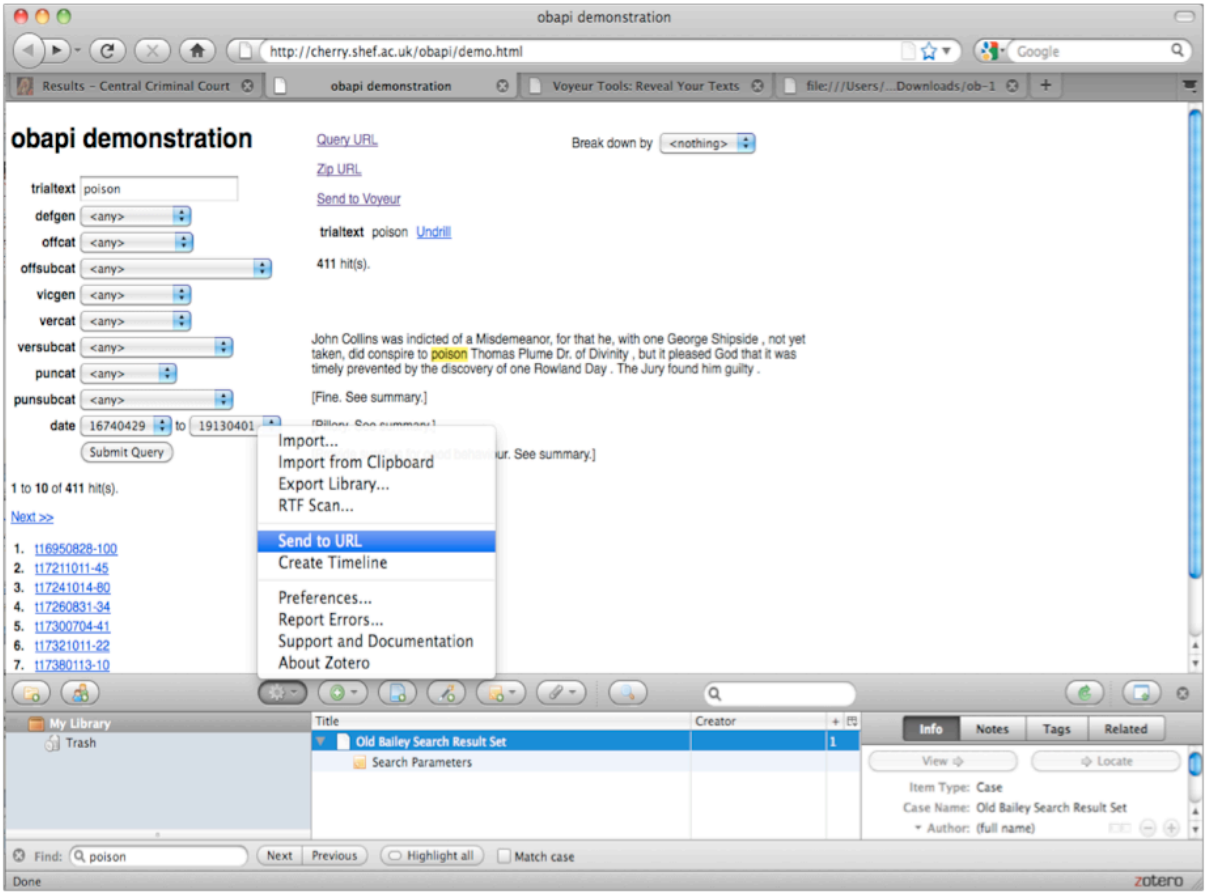

Figure 3: Exporting ‘Poison’ Search Results from the Old Bailey API (OBAPI) to Zotero

For example, one might be interested in charting references to ‘electricity’ in forgery and coining offences, but not cases of larceny or murder. Researchers can quickly scan through search results for ‘electricity,’ add the most relevant cases to a Zotero collection, and forward them to Voyeur for further analysis.

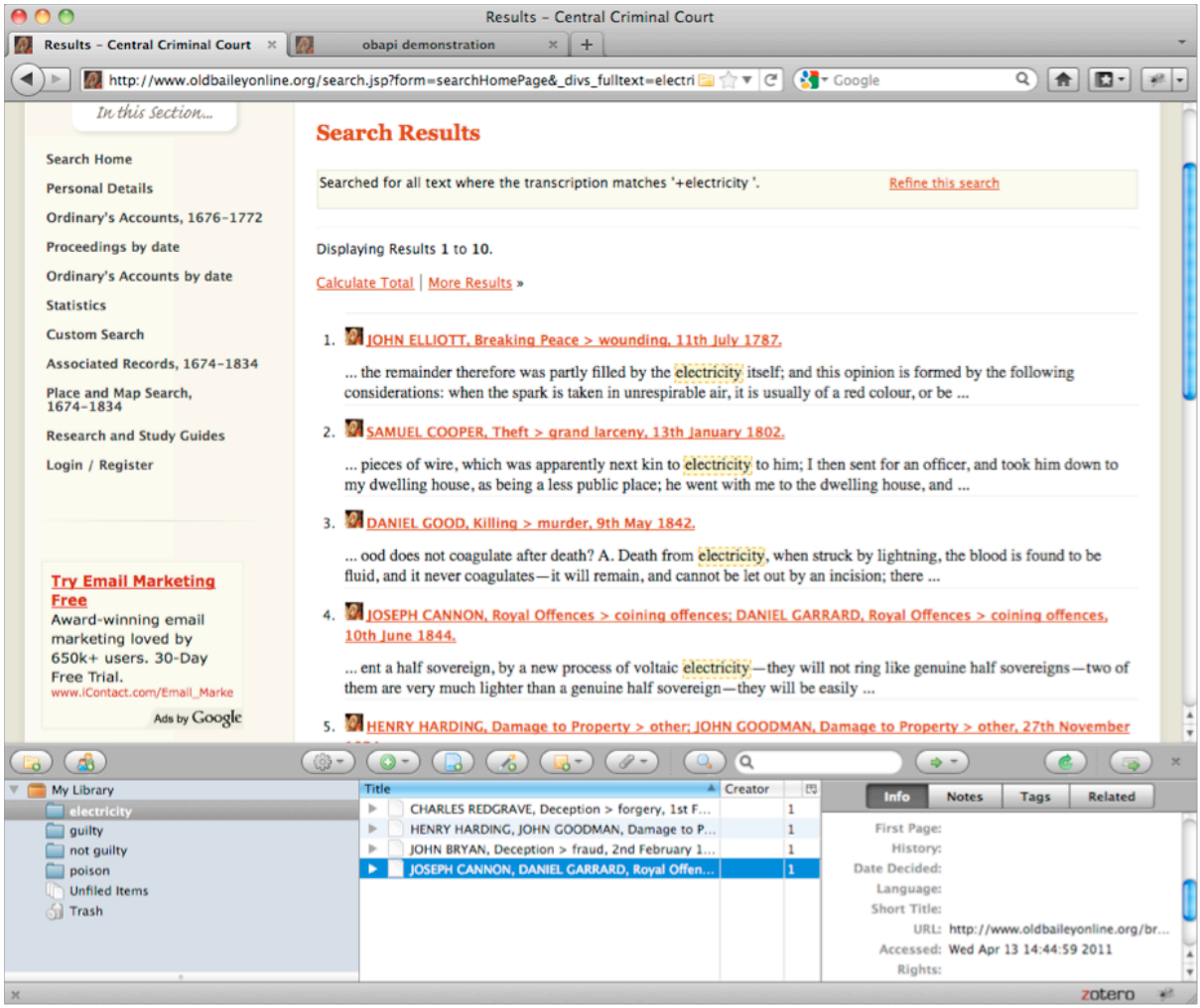

Figure 4: Using the Old Bailey to Search Contemporary Use of the Word 'Electricity' Across Different Crimes

3. Voyeur Tools

Voyeur Tools (http://voyeurtools.org) is a web-based environment for doing text analytics and visualization on a user-provided corpus of texts. Some of its notable characteristics include:

- the ability to work on documents in various formats such XML, PDF, MS Word, and HTML
- the ability to work on larger-scale document collections; the underlying architecture can support hundreds of megabytes of text in a single corpus (latency across some slower networks may create delays in uploading data into Voyeur).
- a wide range of tools to perform linguistic analysis and to visualize data
- a modular design of tools that can be combined to interact in pre-defined and user-defined skins

- functionality to export data from Voyeur Tools in various formats (text, comma-separated values, XML, images, etc.)
- persistent corpora that can be shared and recalled with URLs
- an export feature that allows Voyeur Tools (a specific tool or a skin) to be embedded in remote sites, much like a YouTube clip.

Several refinements and additions were made to Voyeur Tools to enhance interoperability with the Old Bailey and Zotero.

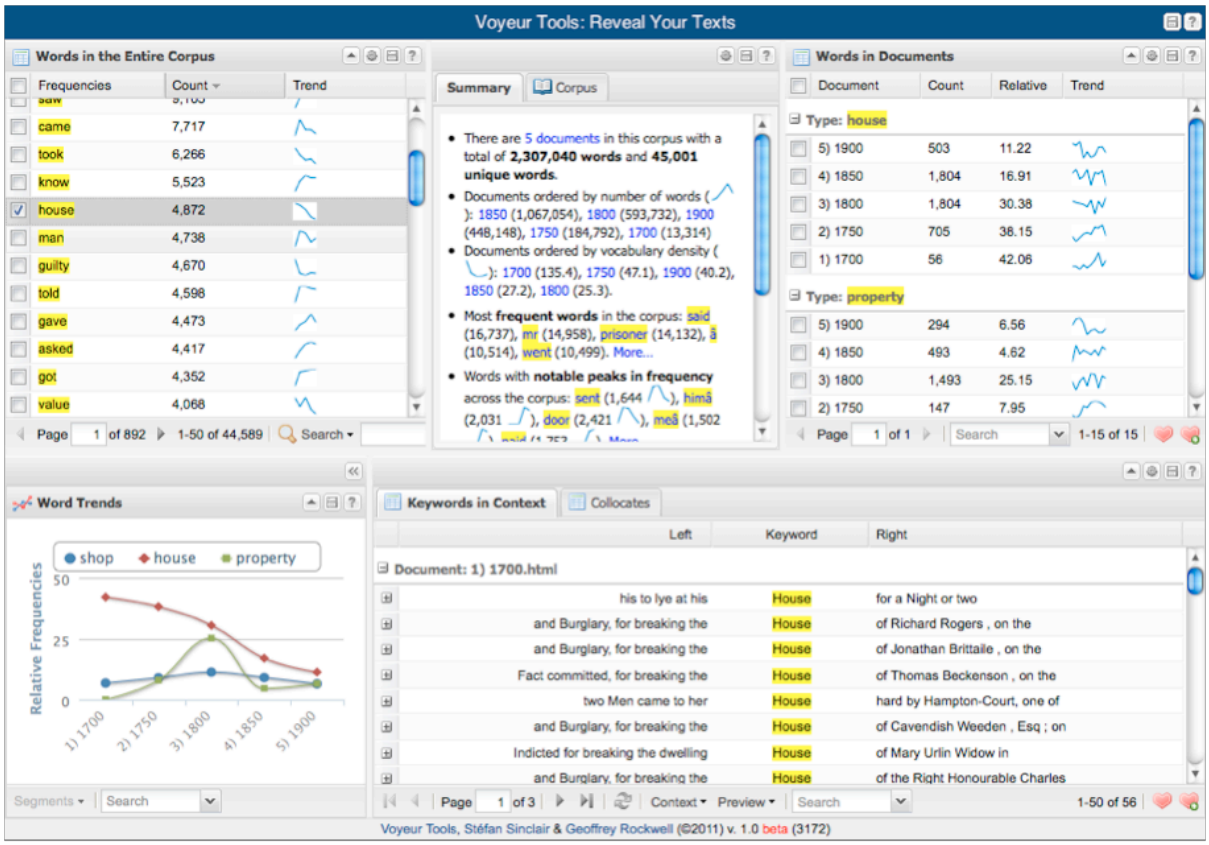

Figure 5: Voyeur as Skinned B**efore DMCI**

Stéfan Sinclair designed a new, simplified skin (a combination of tools) to optimize Voyeur's visual ease-of-use. Whereas after preliminary interaction, the previous default skin featured six panels of tools (most of which provided tabular data), the new default skin has been stream lined include two visualizations and a text reader in four tool panels. This redesign and redevelopment was heavily influenced by a group of twelve OWHs who experimented with the functionality that connected the Old Bailey API to Voyeur Tools. Their influence pointed to the need for prominent placement of clearly articulated search results: a plan to include a Wordle-like visualization of the top frequency terms, was shifted after several users

insisted on the greater need for a tabular view of all terms in the corpus.

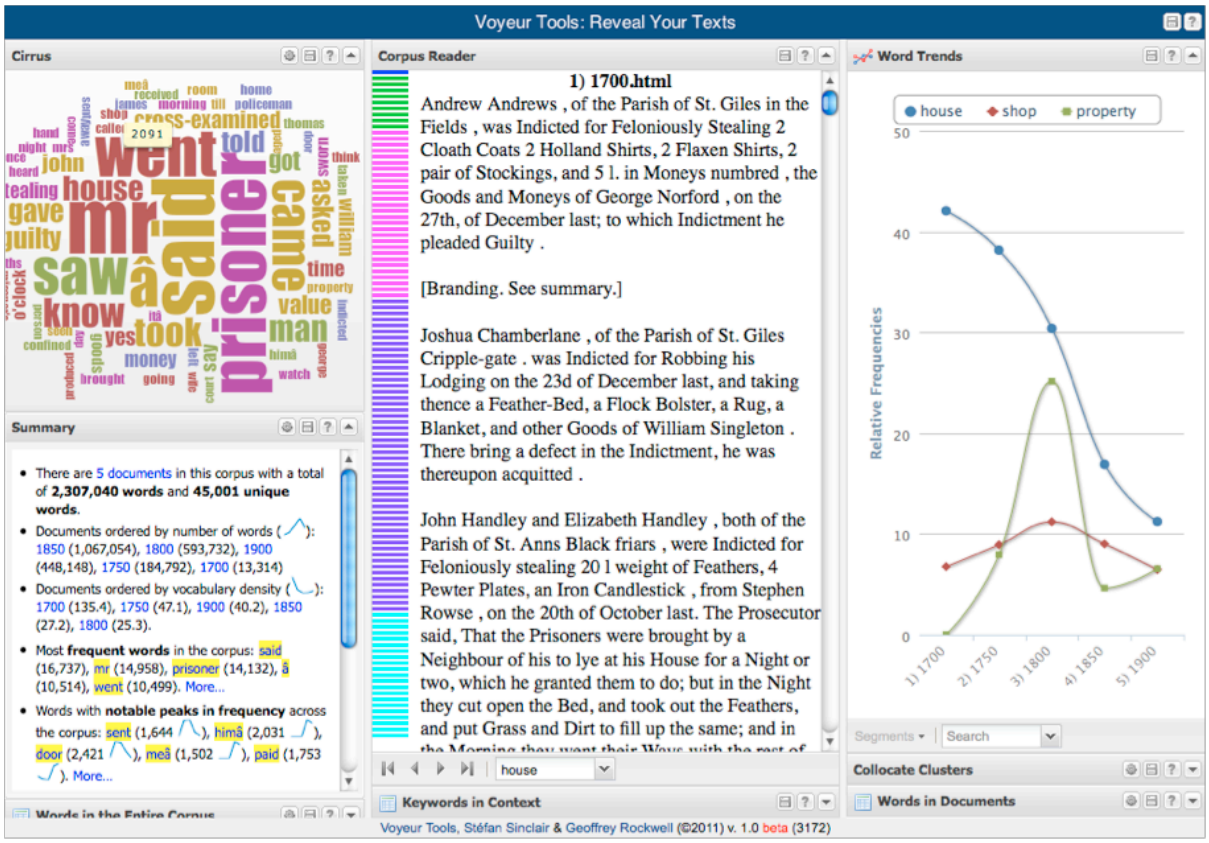

Figure 6: Voyeur as Skinned after Modifications were made for DMCI

In conversation with other development streams across DMCI, Sinclair then refined a correspondence analysis tool that features a scatter plot (a visualization that illustrates the top frequency terms as clustered around individual documents). These graphs can be enormously helpful in seeing trends between documents selected by keyword or constructed according to chronological segments (by year or decade, for instance).

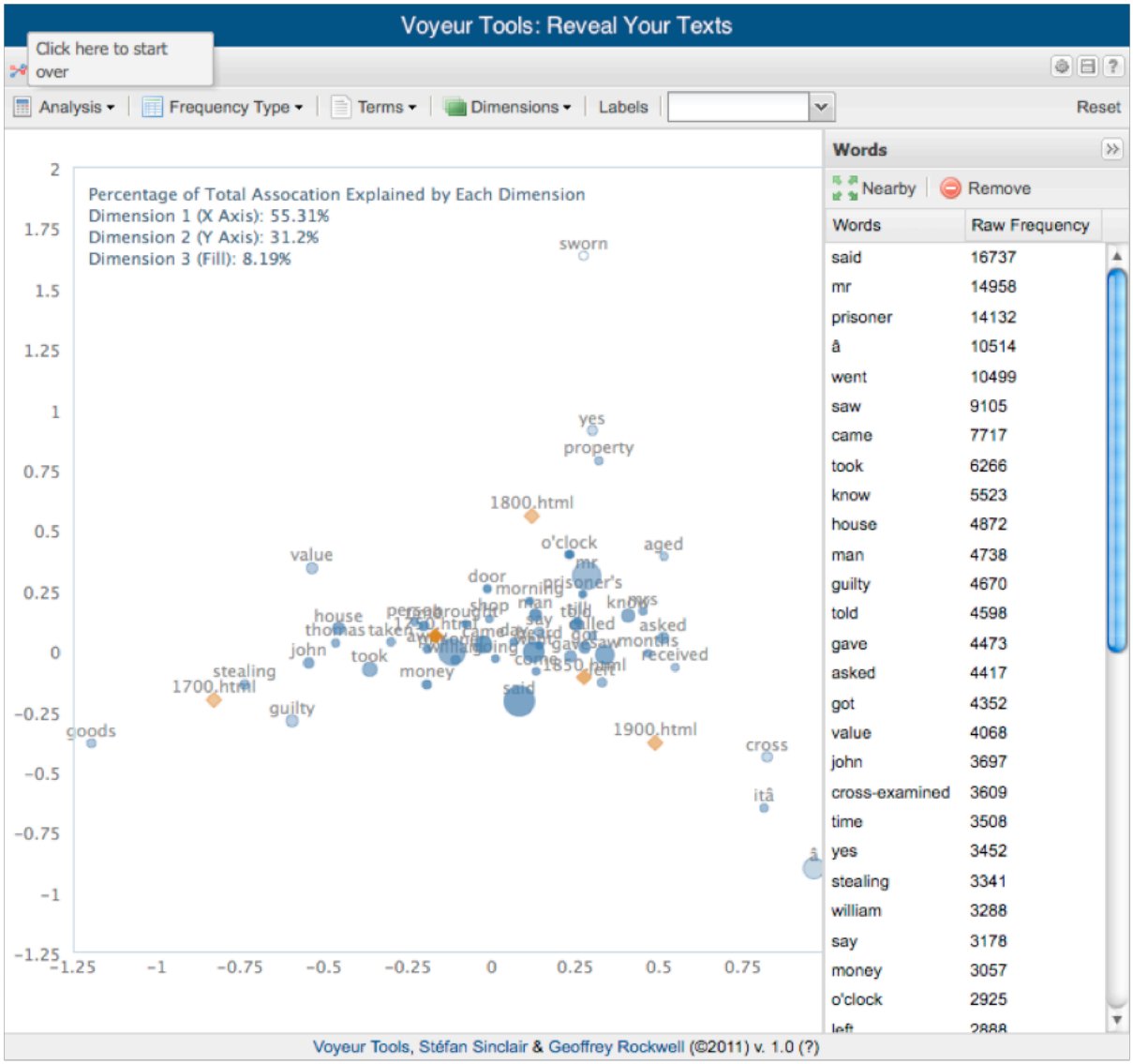

Figure 6: The Correspondance Analysis Tool, Highlighting the Scatter plot

As a whole, the Voyeur team created a local store of the *Proceedings* to facilitate and speed the creation of arbitrary subsets of records. The OBAPI can be used to fetch individual and compressed archives of documents, however, this procedure adds time to processing and processing time should be minimized for real-time web applications. Whereas additional processing time may not be perceptible when retrieving a small set of approximately one dozen documents, as the amount of data exceeds this range, processing time slows: the more documents, the more untenable the wait-time becomes. To address this issue, we now have a full mirror of the raw XML trial accounts on the Voyeur server enabling individual documents to be read as local files. A series of simple tests were done to test this optimization strategy: load times were cut by factor of four, facilitating an approximate speed increase of 75%. More importantly, this strategy allows larger real-time corpora to be constructed,

corpora that would otherwise cause a network timeout.

***

Having explained the basic architecture that joins the Old Bailey, Zotero and Voyeur, we will now discuss our experiments in text mining and data warehousing.

4. 'Show Me More Like This'

During the design of the OBAPI, we also wanted to embed new forms of search that moved beyond representing keywords and tagged data, to allow a variety of forms of machine learning. In particular we wanted to create tools that allowed users to navigate through all 197,000 trials rapidly and intuitively. Analyzing this kind of data by decade, or trial type, or defendant gender etc., can re-enforce the categories, the assumptions, and the prejudices the user brings to each search and those applied by the team that provided the XML markup when the digital archive was first created.

In the spirit of exploration rather than interrogation, we wanted to provide users numerous measures of similarity, models for refining the data, and modes of selecting the textual. objects. In essence, we sought to use the OBAPI demonstrator to allow the user to 'search by example,' or to use a 'more like this' functionality. The 'more like this' tools should help researchers define what they are looking for and possibly confound their expectations about what results might be returned. We believed this approach would expand the range of potentially relevant evidence to the questions historians have already identified.

Bill Turkel suggested a number of ways of finding 'more like this,' options that were explored at the Mind the Gap workshop held in Edmonton, at the University of Alberta, Canada, on May 10-14, 2010 (http://ra.tapor.ualberta.ca/mindthegap/Home.html). Computing the Normalized Compression Distance (NCD) for every pair of trials in the *Proceedings* was the most ambitious of these options.

The NCD is a domain-neutral similarity measure based on Kolmogorov complexity (see Cilibrasi and Vitanyi 2005). Computable because it makes use of standard compression algorithms like the data compressor Bzip, we hoped the NCD would provide a measure of the 'distance' (i.e. similarity or dissimilarity) between transcriptions of any two trials. The advantage of doing this could be two-fold. It would expose groups of trials that were textually similar even though they did not share legal characteristics such as 'offence,' 'verdict,' or 'punishment.' In doing so it would reveal and map behaviors found in witness statements

outside of the taxonomy of those embedded within the XML tagged legal descriptions. For example, this would identify all texts in which an assault is described, regardless of the eventual charge against, reprieve, or release of the suspect. Secondly, that the NCD tables would create a straightforward look-up facility that could move from any one trial, or group of trials, to similar trials.

Preliminary implementation of the NCD measure was stymied by the size of the relevant table; our set of 20 billion measurements (half a matrix of 197000^2 elements) is too large to be fully analyzed with our existing tools. We now anticipate that calculating the full set of distance measures on a typical desktop machine would take several weeks of continuous computing.

Using the Shared Hierarchical Academic Research Computing Network's (SHARCNET) High Performance Computing architecture to run the process in parallel with Apache Hadoop enabled us to make some progress. This platform shift necessitated a considerable redesign of Voyeur Tools to allow the underlying processes to support very large datasets (e.g. the *Proceedings*), as well as performing efficient NCD calculations. When complete, this work will allow us to generate similarity measures for any corpus in Voyeur Tools, regardless of size. The NCD is only one of the machine learning options we are currently exploring.

Based on the success of earlier research, we are also implementing a Naïve Bayesian machine learner. In 2008, Turkel implemented a Naïve Bayesian machine learner, trained it to recognize the offence categories of the 1830s, and tested its performance on the trials of the 1830s and the 1840s (Turkel 2008). He then analyzed the false positive errors and showed that the system tended to generalize beyond category boundaries in an interesting way. The offence categories for 'trials' that the system falsely identified as assault are shown in the pie chart below. Most researchers would agree that the trials Turkel has identified using this methodology, would be recognized as forms of assault, even though they were not coded as such.

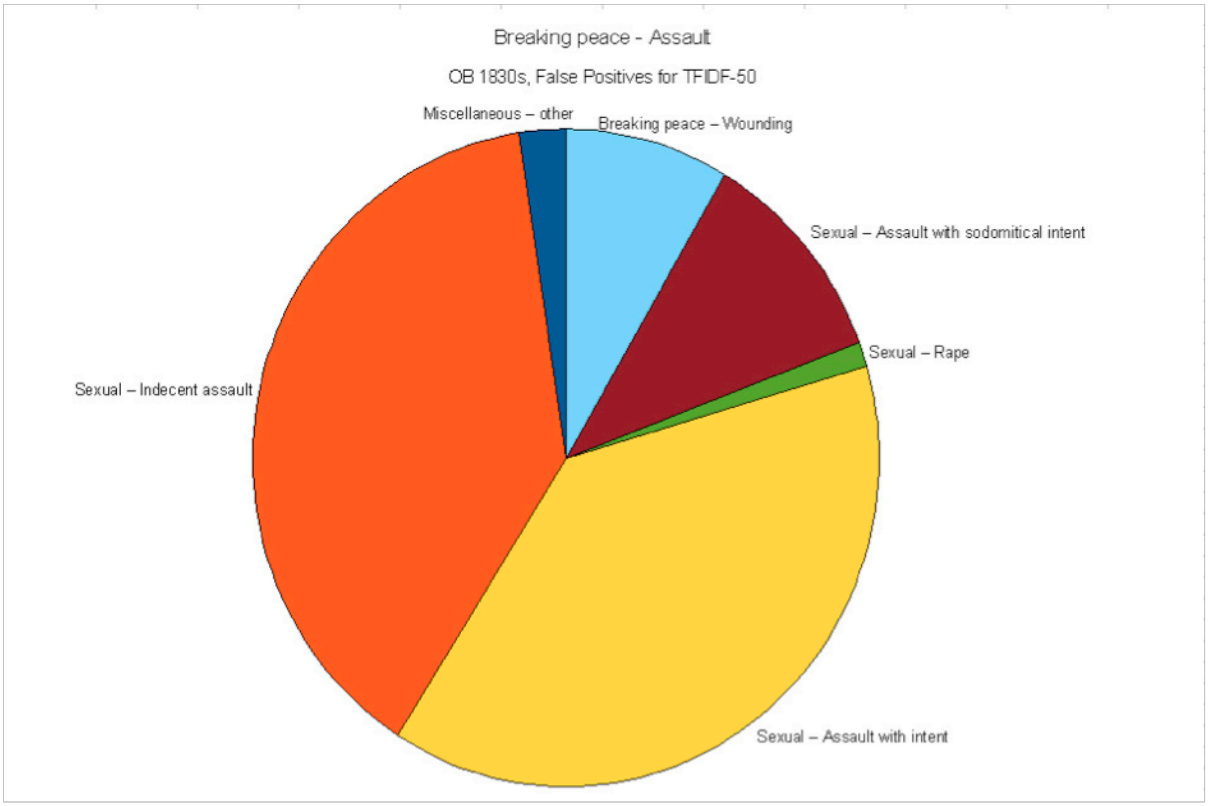

Figure 7: Accounts of Criminal Acts of Assault Identified with Naïve Bayesian Machine Learner

Turkel is currently re-implementing the learner in *Mathematica* (see below) for a new round of experiments. Naïve Bayesian machine learners can be operated in 'online mode' – iteratively trained, tested – and easily be introduced into tools that 'watch' the researcher as he or she decides which sources are relevant to a particular task; this is the basic method deployed by an email spam filter (watching to see which messages are mail and which are not). The machine learner can then look through records that the historian has not seen yet, and suggest there are 'more like this.'

In Turkel's experiments, the machine learners use the standard term frequency / inverse document frequency (TF/IDF) to determine what a document is about (Manning et al 2008). In related work, Stéfan Sinclair used the very fast TF/IDF-based similarity functionality of Lucene to allow a user to experiment with measures of textual difference between trials directly at the search interface. The Apache Solr implementation can be found at:

http://tapor-dev.mcmaster.ca/~sgs/ob/similar/compare/t18721028-764/

The alternative Lucene based TF/IDF 'more like this' functionality has been implemented by Jamie McLaughlin and is available directly in the Old Bailey API demonstrator

(www.oldbaileyonline.org/obapi). This facility uses an index of every word in the *Proceedings* to determine where words appear and how common they are. Starting with an individual trial, this version of the 'more like this' function counts all the words in a particular trial, and ranks them from the most frequent to the least frequent (excluding 'stop' and two character words). The number of appearances for each word is then multiplied by a measure of how rare each word is in the *Proceedings* as a whole: or its Inverse Document Frequency (IDF). This allows a 'score' to be calculated for each word in the trial. The twenty-five highest scoring words in the resulting list then generates a query that locates similar trials.

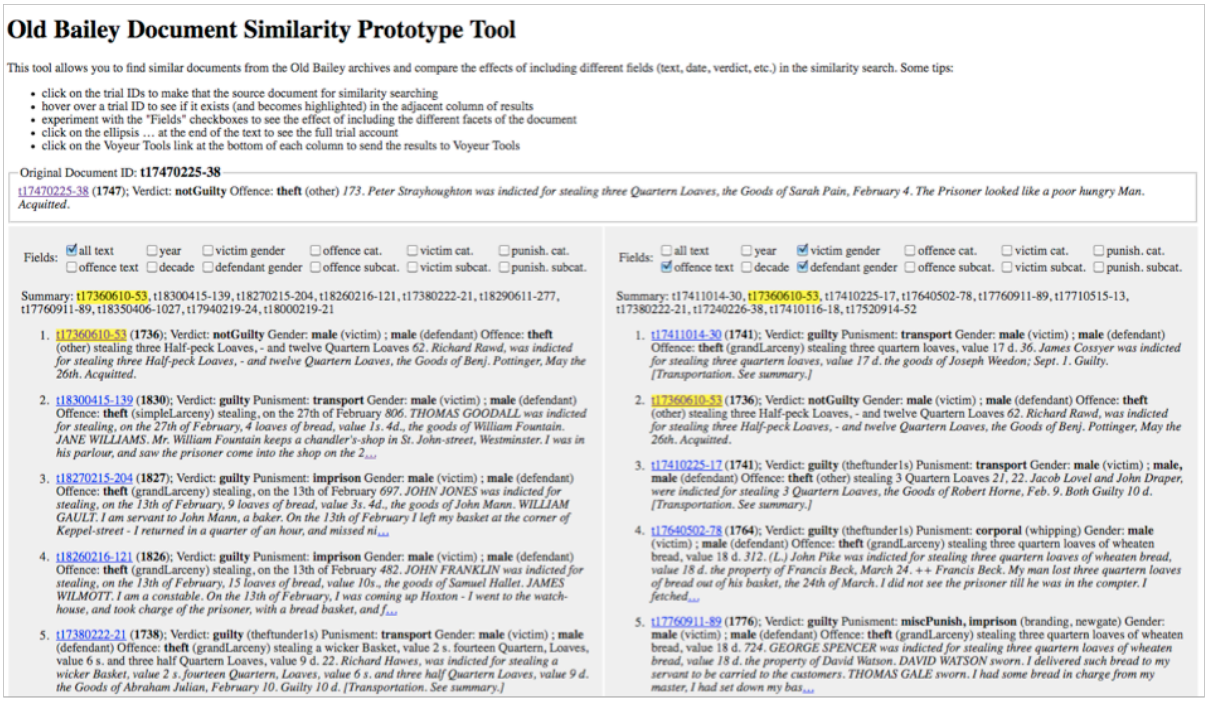

Figure 8: Similarities Located by Calculating the IDF in Trials where the Accused has Allegedly Stolen Bread

These two iterations of a 'show me more like this' tool, in addition to the machine learning and NCD methodologies, are being refined and implemented on the Old Bailey site to enhance its functionality. They are also being applied in a range of ongoing projects to, for instance, to identify trials in which in which legal counsel may have been present but not explicitly mentioned, or in which particular linguistic constructions can be associated with ethnic or regional groups. DMCI looks to find new ways of delivering results. We continue exploring possibilities.

5. Data Warehousing

Geoffrey Rockwell and Jörg Sander also considered a number of data and text mining models

that could be beneficially applied to the *Proceedings* during the May 2010 Mind the Gap workshop. The data warehousing model proved to be the most promising.

Basing their design after industry standard look and feel, John Simpson, under the supervision of Rockwell and Sander, developed a data warehouse with a user-friendly interface (e.g., Chaudhuri & Dayal 1997). Data warehouse models facilitate multi-dimensional analysis of data; their tools group, summarize, and display time-varying data at different aggregation levels (when dimensions are hierarchically organized).

Data derived from the *Proceedings* can be meaningfully modeled using a data warehouse. In the data warehouse user-specified keywords and XML tag values such as the 'offence,' 'verdict,' and types of 'punishment,' as well as the number of defendants, defendants' gender, and trial date become 'dimensions' in a multidimensional data cube. Each combination of tag values in the given dimensions and specified keywords uniquely determines a cell in the multidimensional cube. Different 'measures' can be defined for a cell, such as the number of trials that fall into it. Some of the dimensions are also hierarchically organized, allowing the user to view cells at different levels of granularity. Offences, for instance, have subtypes. In a data warehouse, they can be viewed as a two-level hierarchy, the time dimension can be viewed as a day-month-year-decade-century hierarchy, and so on.

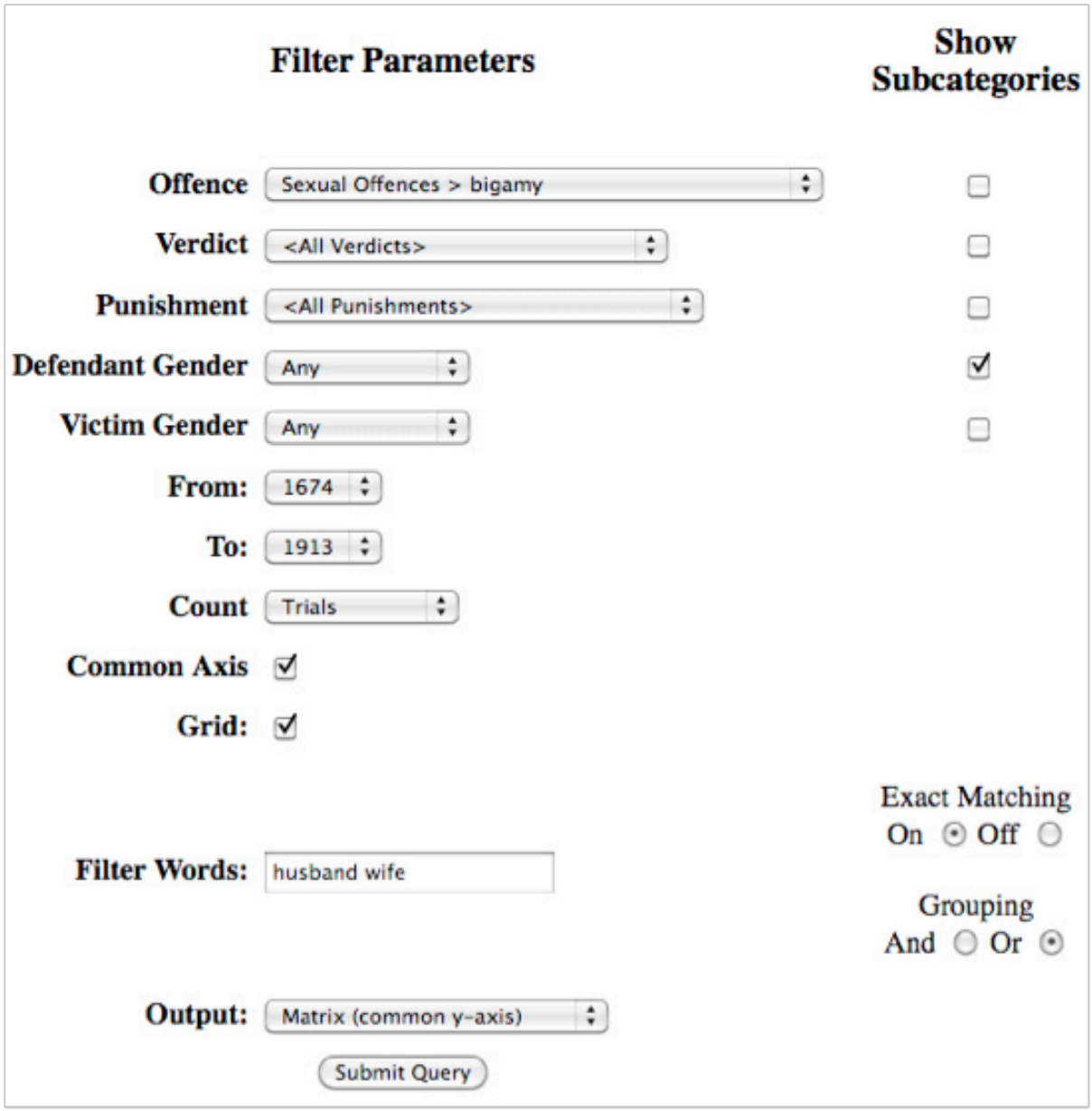

Figure 9: Preliminary Design of the Data Warehouse

To support this model we built a 'fact table' that stores the 'facts' of a trial, collected from the XML tag values and the word frequencies in the textual part. The fact table represents the multidimensional cube where the dimensions/axes are the tag categories (e.g. 'offence' subtypes, 'verdict' type), at the lowest level of granularity. The measures associated with the cells are different kinds of counts based on the trials that fall into a given cell, including the count of trials, defendants, offences, verdicts, and punishments. We are also currently implementing word frequencies in the data warehouse.

The user interface designed for the Old Bailey data warehouse allows the specification of:

- different dimensions that define possible groups of data

- filters that only include a certain subset of the data in the analysis
- the measure to be aggregated and displayed for each of the groups in the output of the corresponding query (e.g., count trials or count defendants).

Query results are displayed as a plot or set of plots, which can be organized and visualized in a few different ways.

This model allows researchers to look at the trials through many different lenses. It generates a myriad of graphs that follows different variables enabling OWHs to compare, for example, male bigamy against female bigamy, as shown above. By allowing multiple queries to be graphed against one another, this facility goes substantially beyond the current statistics function available on the Old Bailey site. Generating a matrix of graphs encourages visual comparison across variables.

We have also developed a version of the data warehouse that superimposes the same information onto a single chart know as a digraph. This tool allows for data streams to be turned on, turned off, and zoomed. It provides interactive annotations, and displays exact values.

Most recently, we have implemented a keyword filter that allows users to display only those trials that contain the specified words (or specific variants) in the full text of the trials. The matrix and digraph samples shown here included exact matches for the word 'husband' and for the word 'wife' in the filter criteria. The next iteration of the tool will include an option to see display word frequencies from the filtered trials.

The data warehouse tool was well received by the historians on the project. This has encouraged us to integrate it back into the Old Bailey website. Our designer, Milena Radzikowska, is working to improve the interface, the presentation of results, and to allow the data warehouse tools to be visually integrated into to the Old Bailey.

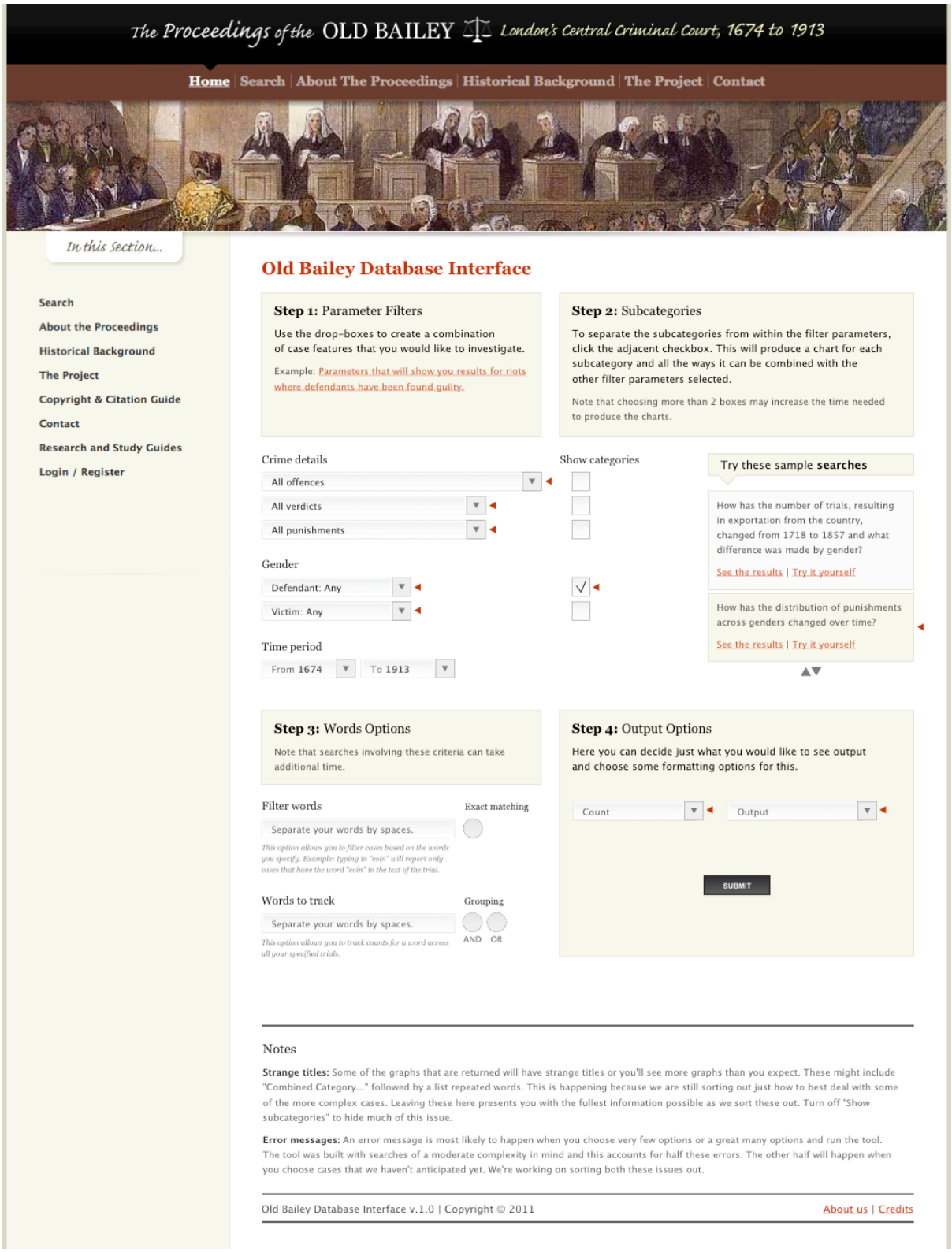

Figure 10: First Draft of Proposed Datawarehouse Redesign

Users on the Old Bailey site in the UK will be able to call up the University of Alberta hosted data warehouse tool seamlessly. The most current prototype is online at:

http://ra.tapor.ualberta.ca/~digging2data/cgiTestForm6.html.

***

Having explained our experiments in text mining and data warehousing, we will now move on to considering the opportunities and challenges we have encountered over the course of this project. We will then address the benefit of using agile programming as an approach to rapid-development.

6. Challenges, Academic Reaction, and Usability

This project began with the understanding that one of its major challenges would involve encouraging 'ordinary working historians' (OWH) to adopt its research methods. As with many of the resources available in the Digital Humanities, DMCI makes fantastic tools and striking visualizations available to scholars. However, the impetus for uptake needs to come from the OWHs (and by extension literary critics, philosophers, art historians, and classicists). When these scholars regularly engage with digital tools as part of their normal methodology, projects like Data Mining with Criminal Intent will become transformative elements in the public cultures of the humanities in all the ways they can and should.

In pursuit of this end, we built up a body of a dozen professional historians interested in the city of London, criminal issues, and of course, specifically interested in using *The Proceedings of the Old Bailey.* We asked them to sort through the contents of the archive to discover texts they found exciting or provocative, to create libraries of these texts using Zotero, and to apply Voyeur's corpus linguistics tools to their textual analysis. Informally and iteratively, at each stage our development, we asked them to provide feedback.

Collecting trials in Zotero came naturally; it was clearly within the comfort zone of our OWHs. Zotero is the citation system of choice among more methodologically-aware historians, in our experience. We concluded that Zotero's ability to allow end-users to build and share libraries and to organize and annotate collections, needs to be built into distributed web resources.

Cconcurrently and predictably, our OWHs were more challenged by the task of using corpus linguistics to perform analysis. While Zotero allows historians to do familiar things better, Voyeur is about doing very unfamiliar things. Voyeur was designed with textual scholars in mind, so it is perhaps no surprise that we found that most historians, when asked to choose

frequency tables, context lists, and trends, tended to rely on relatively traditional forms of enquiry – looking for collections of words. By looking for collections of related words, they turned Voyeur into a search engine for slightly more complex keyword queries, rather than leveraging its capacity as an environment in which a new measure of relatedness might be analyzed.

This observation led us in two directions. First, we realized that we need tutorials and walkthroughs that demonstrate the process of going from the Old Bailey API to Zotero to Voyeur. These pedagogical materials need to include serious and usable research results. For all their enthusiasm, most historians need very clear and easy to follow instructions when first beginning to preform these tasks. To this end we are creating documentation for historians on the Criminal Intent site (http://criminalintent.org) that are being integrated by Kirsten C. Uszkalo into a tutorial that crosses the different tools. And second, we concluded that we need to publish serious research that uses these tools to answer historical questions. We need to mash up traditional forms of article writing that engage with historical debates, with evidence drawn from what would have been overwhelming sums of text objects. We need to lead by example, or no one will follow.

***

In the next section we present a few examples of the data mining discoveries being built into our own more traditional forms of historical arguments. These examples will demonstrate how textual analysis of the infinite archive can be used in conjunction with established forms of scholarly practice.

7. Prototyping and Literate Programming

Attempting to anticipate the needs of historians and other humanists who do not currently do this kind of work has been one of the exciting challenges of this project. We have explored a range of interdisciplinary use cases. We imagined the OWH interacting with the Old Bailey website through his or her web browser, digitally aware researchers familiar with the kinds of tools that a program like Voyeur provides, and those able to program and to interact directly with APIs. We were resolute in our efforts to use our own software to accomplish our own research goals.

To determine the measures, manipulations, and visualizations that could be usefully built into the Old Bailey website or into tools like Voyeur, we have used a strategy based on the

methodology of agile/extreme programming. Working in pairs, we turn around revisions on a very short timescale, throwing out what does not work and refactoring what does, a method successfully by Sinclair and Rockwell, who used Voyeur to do a form of "extreme text analysis" (Rockwell 2008). Turkel and Tim Hitchcock then extended the practice into the domain of historical research programming using *Mathematica*.

Although it is probably not recommended for most OWHs, *Mathematica* is perfect for Turkel and Hitchcock for a number of reasons:

- **Literate programming.** One of the key advantages of working in *Mathematica* is that coding can be done in the form of notebooks: enriched documents that include prose, data, simulations, running code, visualizations, and arbitrarily complex interface elements like buttons, sliders, and live graphs. These can be shared with people who do not have a license for the software via the use of free reader software (as PDFs can be shared via an Acrobat Reader). They can also be shared online at the Wolfram Demonstrations Project. To illustrate this we have uploaded one of our examples of computing TF-IDF using one of the Old Bailey trials (Turkel 2010).
- **High level Succinctness**. Programs written in *Mathematica* can draw on language constructs that include elaborate pattern matching and functional programming. This makes it possible to write very powerful programs in a few lines of code. For example, if `textstr,` is a text string, one can generate all of the five-grams with the command `Partition[StringSplit[textstr], 5, 1]`. Importing a webpage and converting it to plain text is as easy as:
  `Import["http://criminalintent.org", "Plaintext"]`
- **End-to-end.** In our work, we have gathered online material via the Old Bailey API and by spidering and scraping, have done extensive text and string processing, written probabilistic machine learners, done some computational linguistics, exchanged information with SQL databases, created and measured some relatively extensive networks, and made hundreds of visualizations. *Mathematica* has extensive built-in support for practically every branch of pure and applied mathematics, so none of these operations required any additional libraries.

There is no existing literature on the characteristics of the *Proceedings* as a whole, although researchers like Magnus Huber (2008) have analyzed a substantial subset of the trials. Turkel and Hitchcock have been using *Mathematica* and the methodologies described above, to explore various kinds of trial measures across the entire span of records. Preliminary results are very encouraging; they promise to revise many of the claims that past researchers have implicitly or explicitly endorsed about the history of the criminal trial, largely on the evidence of the *Proceedings* (see, for example, Langbein 2003). In many cases, data in the Old Bailey are distributed according to apparent power laws. These distributions still need to be characterized carefully, but the upshot is that measures that may be familiar to historians (such as the range, mean or standard deviation of a set of data) are simply misleading. For example, there is no such thing as a typical trial length.

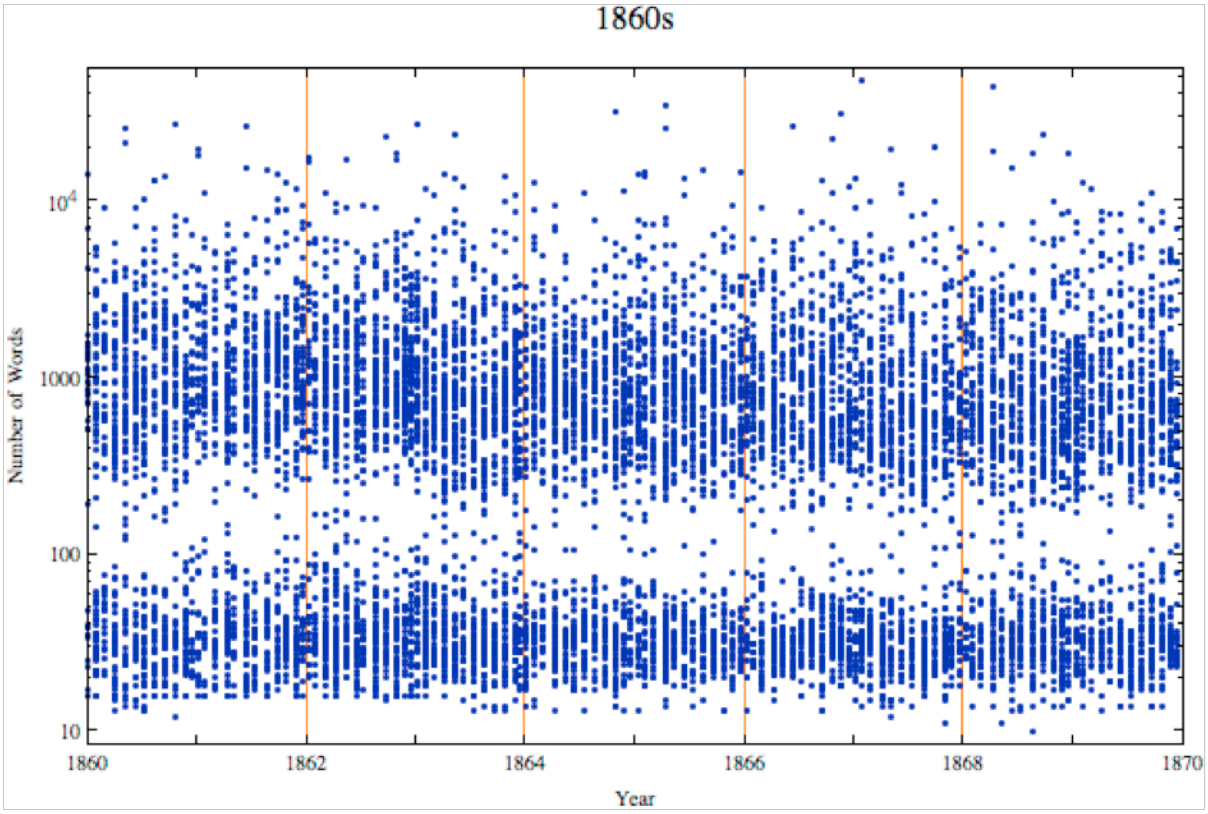

Figure 11: A Scatterplot illustrating the Lengths of Trials in the 1860s as Measured in Words

As far as we know, no one has ever observed that the printed trials in this decade and a number of others were either shorter than about one hundred words, or considerably longer, but almost never around 100 words long. We are currently investigating the reason(s) why this may have been the case.

Using tools like these to re-examine the *Proceedings* as a whole, we have realized that the historiography of the modern Anglo-American trial needs to be revised. Most historians have associated the major developments in the history of the modern 'adversarial trial' with eighteenth century. We believe that the trial evolved much more dramatically during the course of the nineteenth century. Nineteenth-century developments transformed a system of community-based judicial theatre in which the outcome was substantially uncertain, and which was gradually taking on an adversarial character, into a new form dominated by plea bargaining, increasing bureaucratic control by professional lawyers, and an ever-rising conviction rate (Hitchcock & Turkel, ms).

8. Pulling It All Together

Using these integrated tools make broader historical phenomena likewise visible. Dan Cohen, for instance, looked at female agency in Britain in the nineteenth-century, extracting cases

that involved women and analyzing them in Voyeur. He leveraged the scale of the *Proceedings* and the organization and visualization capabilities of Zotero, Voyeur, and the separate data warehouse to trace the rise of a greater latitude in female behavior in the late Victorian period. By taking the slice of the Old Bailey involving bigamy and using the data warehousing charting tool, for instance, he could see a significant rise in women taking on other spouses when their husbands had left them (often for their own escapades abroad or in a distant town). This is shown in the figure below.

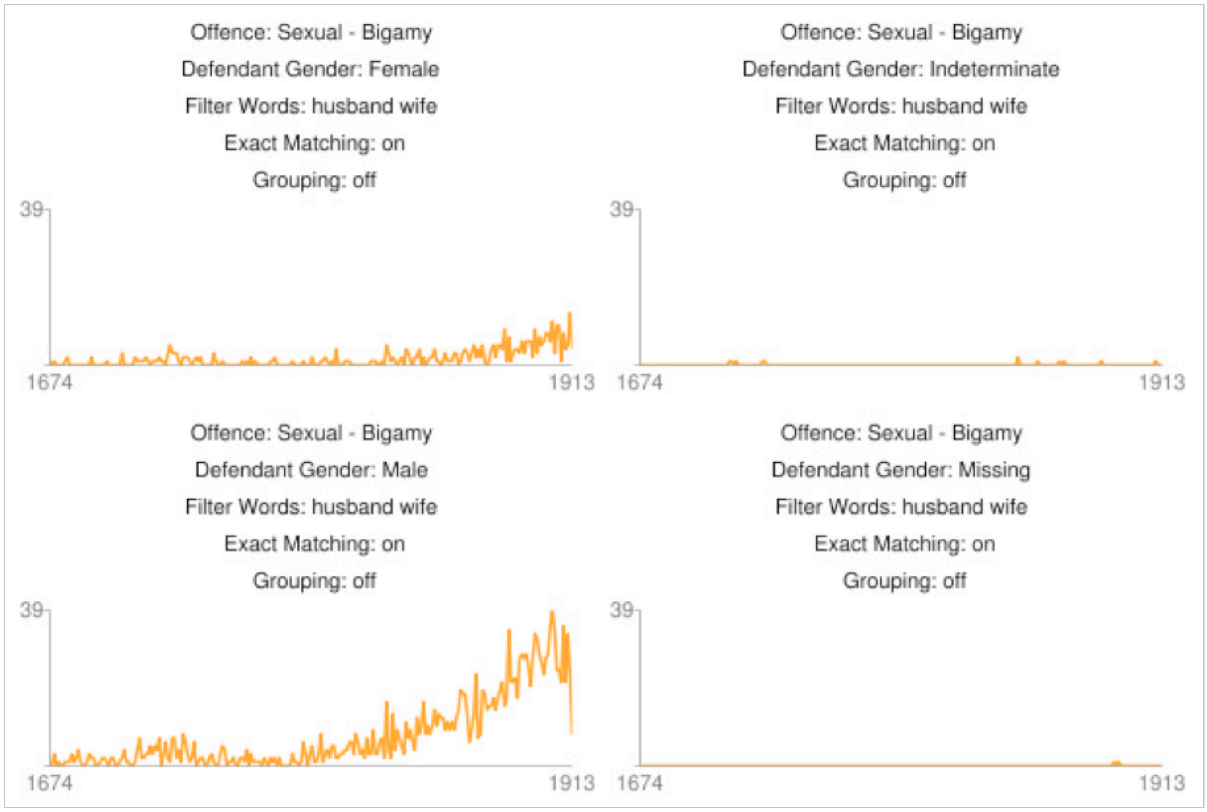

Figure 12: Graphing the Presence of Female Bigamists

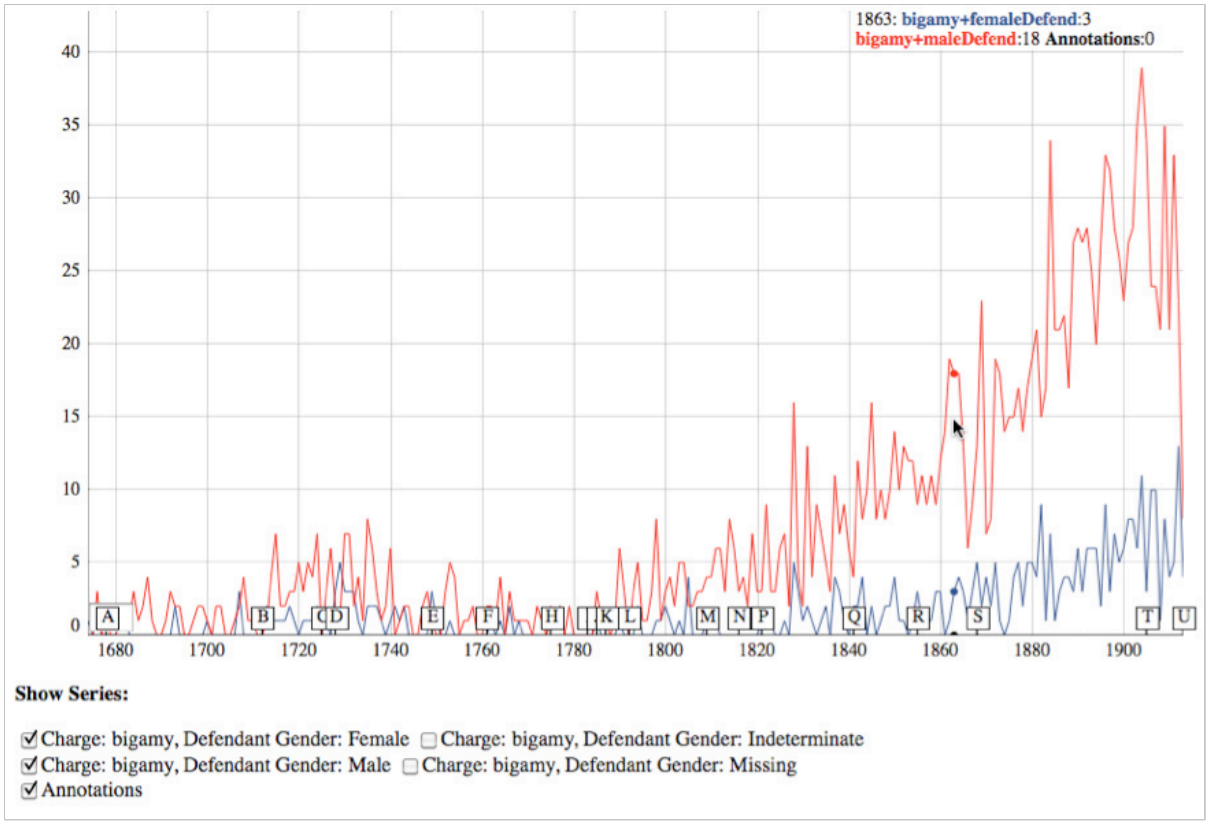

Figure 13: Displaying the Presence of Female Bigamists Using a Dygraph

By looking at specific trials in Zotero, it also became very clear that these late-Victorian women were not receiving significant punishment for this action, as they did earlier in the century. This successfully matched the results produced in Gibbs and Cohen's NEH-funded project on text mining in history, which has shown a sharp rise in mentions of 'loveless' marriages after around 1870.

***

Developing research in the digital humanities can require divided commitment. On the one hand you are committed to the subject and the particularity of the data – in this case the rich historical data we have about criminality. On the other hand you are committed to developing increasingly sophisticated tools that can help you reinterpret the data. One of the benefits of a project like this is that we can iterate between different ideas about new tools/interfaces and hypotheses about what we can learn from the data given new tools. We developed methods of agile analysis and programming, both in smaller groups and for the team as a whole. New questions led to new tools and interfaces. New tools and interfaces provoked new questions. The adaption of tools to texts and back to tools, whether categorized as literate programming

or agile analytics, is new to historical studies. Methodology is part the message. These approaches encourage us to ask different questions, not only about our own tools, but about how to approach historical research with a wide view lens and a fine grained brush. They demonstrate how great things happen when the techies and the OWHs work closely together. Moreover, they demonstrates the real benefits already enjoyed by the ordinary working historians who have chosen to datamine with criminal intent.

As a way of concluding, we reference Stephen Ramsay's response to the first draft of our White Paper,

> If the creators of this project are unabashed in their use of scientific tools and methods, they are likewise unapologetic in their description of why they are doing so. The Old Bailey, like the Naked City, has eight million stories. Accessing those stories involves understanding trial length, numbers of instances of poisoning, and rates of bigamy. But being stories, they find their more salient expression in the weightier motifs of the human condition: justice, revenge, dishonor, loss, trial. This is what the humanities are about. This is the only reason for an historian to fire up *Mathematica* or for a student trained in French literature to get into Java. (Ramsay, 2011)

9. References

Agile / Extreme Programming *<http://agilemanifesto.org/>*

Apache Hadoop <*http://hadoop.apache.org*>

Apache Lucene <*http://lucene.apache.org*>

Apache Solr <*http://lucene.apache.org/solr*>

Chaudhuri, Surajit & Umeshwar Dayal, "An overview of data warehousing and OLAP technology," *ACM SIGMOD Record*, Volume 26, Issue 1 (March 1997).

Cilibrasi, Rudi & Paul Vitányi, "Clustering by Compression," *IEEE Transactions on Information Theory*, Volume 51, no. 4 (2005): 1523-45.

Dygraphs <*http://dygraphs.com*>

Extreme Text Analysis *<http://hermeneutica.stefansinclair.name/>*

Hitchcock, Tim & Robert Shoemaker, "Digitising History from Below: The Old Bailey Proceedings, 1674-1834," *History Compass,* Volume 4, Number 2 (2006), 193-202.

Hitchcock, Tim & William J. Turkel, "The Old Bailey Proceedings, 1674-1913: Text Mining for Evidence of Court Behaviour," manuscript (2011).

Huber, Magnus, "The Old Bailey Proceedings, 1674-1834. Evaluating and annotating a corpus of 18th- and 19th-century spoken English," *Annotating Variation and Change (Studies in Variation, Contacts and Change in English 1)*, vol.10 (2008). <*http://www.helsinki.fi/varieng/journal/volumes/01/huber*>

Langbein, John, *The Origins of the Criminal Trial*. Oxford: Oxford University Press, 2003.

Manning, Christopher, Prabhakar Raghavan & Hinrich Schütze, *Introduction to Information Retrieval*. Cambridge: Cambridge University Press, 2008.

Mathematica <http://reference.wolfram.com/mathematica/guide/Mathematica.html>

Normalized Compression Distance *<http://www.complearn.org/ncd.html>*

Old Bailey API *<http://www.oldbaileyonline.org/obapi>*

Ramsay, Stephen, "Prison Art." Paper presented at the Digging Into Data Conference in June, 2011. <*http://lenz.unl.edu/papers/2011/06/10/prison-art.html*>

Rockwell, Geoffrey, "What is Extreme Text Analysis?" *Text Analysis Developers Alliance* (May 2008). <*http://tada.mcmaster.ca/Main/WhatIsExtremeTextAnalysis*>

TAPoR: Text Analysis Portal for Research *<http://portal.tapor.ca>*

Turkel, William J. "A Naïve Bayesian in the Old Bailey," *Digital History Hacks (2005-08)* [weblog]. (May-July 2008). *<http://digitalhistoryhacks.blogspot.com/search?q=%22naive+bayesian%22>*

Turkel, William J. "Term Weighting with TF-IDF," Wolfram Demonstrations Project (2010). <*http://demonstrations.wolfram.com/TermWeightingWithTFIDF*>

Voyeur Tools: Reveal Your Texts <*http://voyeurtools.org*>

Zotero <*http://www.zotero.org*>

Appendix 1: Connecting Zotero to Voyeur

As part of the *Criminal Intent* project, we have created a Zotero plug-in that allows users to create *ad hoc* collections of text, attachments, or URLs from within their Zotero libraries and send them to analytical tools like Voyeur. The plug-in can be downloaded and installed from the Zotero plug-in page, <*http://www.zotero.org/support/plugins*> or from the Criminal Intent website <*http://criminalintent.org*>

To use the plug-in, select items from your Zotero library and click the gear icon. Select: "Send to URL"

You will then see a dialog box that prompts you to define what you want to do with the items in your Zotero library. It is easy to select one of the predefined services, like Voyeur (the default selection). You will also need to select which Zotero item fields you would like to use At this time, users familiar with creating custom URLs can make one to send to other web services.

The "Send To URL" plug-in and Old Bailey translator are designed to work seamlessly together. The Zotero translator saves search results from the OBAPI page as a reference URL stored in the "Extra" field of the Zotero item. This reference URL can be used by other services, like Voyeur, to locate and use those results. When sending a search result set (or several of them) obtained from the OB API page to Voyeur, only include the "Extra" field.

You can also use the Zotero translator to save the text of individual cases from the *Proceedings* and send them to Voyeur. When using the translator to save the trial text as a "Note" (a kind of attachment to Zotero items), only check the "Notes" box on the plug-in dialog box.

After clicking OK, your data is beamed to the selected service or the specified URL via HTTP. Results will appear in a new browser tab.