

Digging Into Data Application Cover Sheet

Project Information

Title: Using Zotero and TAPoR on the Old Bailey Proceedings: Data Mining With Criminal Intent

Project Description: The With Criminal Intent project will create an intellectual exemplar for the role of data mining in an important historical discipline—the history of crime—and illustrate how the tools of digital humanities can be used to wrest new knowledge from one of the largest humanities data sets currently available: the Old Bailey Online. It will create a seamlessly connected environment, the Newgate Commons, in which scholars can use data mining techniques to select themed texts from the 120 million words of trial records contained in the Old Bailey, and employ these texts as the basis of a study collection in Zotero where they will in turn be available for analysis using TAPoR tools (including quantitative text analysis and visualization). In the process, this project will showcase the integration of online textual resources with bibliographical and analytical tools emerging from Digital Humanities.

NEH Application

PROJECT DIRECTOR

Daniel Cohen
dcohen@gmu.edu
703.993.4524
Robinson B 359
4400 University Drive, MSN 3G1
Fairfax, VA 22030
USA

INSTITUTION

George Mason University
Office of Sponsored Programs
4400 University Drive, MS 4C6
Fairfax, VA 22030
USA
Duns Number: 07-781-7450

Funding Request: \$100,000

GRANT ADMINISTRATOR

Ryan Pryke
rpryke@gmu.edu
703.993.4106
Office of Sponsored Programs
4400 University Drive, MS 4C6
Fairfax, VA 22030
USA

JISC Application

PROJECT DIRECTOR

Tim Hitchcock
t.hitchcock@herts.ac.uk
+44 01707 284 5650
De Havilland Campus
University of Hertfordshire
Hatfield, AL10 9AB
UK

INSTITUTION

University of Hertfordshire
The Research Office
University of Hertfordshire Hatfield,
AL10 9AB
UK

Fund Req (GBP): 60,195

GRANT ADMINISTRATOR

Liz Nolan
e.a.nolan@herts.ac.uk
+44 01707 284 3280
The Research Office
University of Hertfordshire
Hatfield, AL10 9AB
UK

SSHRC Application

PROJECT DIRECTOR

Geoffrey Rockwell
geoffrey.rockwell@ualberta.ca
780.248.1209
3-67 Assiniboia Hall
University of Alberta
Edmonton, AB T6G 2E7
Canada

INSTITUTION

University of Alberta
222 Campus Towers
8625 - 112 Street
Edmonton, AB T6G 2E1
Canada

Fund Req (CAN): 99,280

GRANT ADMINISTRATOR

Kerri Calvert
kerri.calvert@ualberta.ca
780.492.8059
222 Campus Towers
8625 - 112 Street
Edmonton, AB T6G 2E1
Canada

Statement of Significance

Over the past few decades scholars have increasingly used court records to illuminate historical themes in novel ways. Topics as varied as the changing status of women in the seventeenth century, the organization of species in the nineteenth century, and the rise of oppositional teen culture in post-War America have all found new interpretations through the detailed analysis of legal documentation. The published [Proceedings of the Old Bailey](#) have been a fertile source for scholars working in these varied traditions, allowing them to use both qualitative and quantitative approaches to the evolution of the criminal justice system, of interpersonal relationships and human behaviour more generally (cf. Beattie 2001; Gatrell 1994; Linebaugh 1991; Shoemaker 2004; Trumbach 1998; Hitchcock 2004; Wise 2004).

Even though the 120 million words of court transcripts published in the *Proceedings* are now available online in a structured and searchable form, the recent use of these legal records remains essentially iterative and traditional. It is now possible, by bringing together in one place, the “Newgate Commons”, the text of the *Proceedings*, the functionality of [Zotero](#) and the tools created by the [TAPoR](#) project, to take a new approach to this old source. Some of the questions that can be newly addressed have already been asked by historians, but never fully answered because of the limitations of traditional research methods. Other interpretations of the vast legal corpus await a new toolkit that allows for intelligent prospecting at scale, the discovery of hidden patterns, and detailed computer-assisted analysis of selected portions of the record. What if scholars, including those with few technical skills, had at their disposal a way to mine the *Proceedings of the Old Bailey* and similar digitized archives?

This project will create an intellectual exemplar for the role of data mining in an important historical discipline—the history of crime—and illustrate how the fundamental conundrums of historical research on large bodies of text that have dogged humanist research over the last forty years might be addressed. Historical trial reports are both a constantly changing generic form (evolving from a few lines of text intended for a popular readership in the 1670s to tens of pages of formal legal reporting by the mid-nineteenth century), and at the same time represent the end point of a continuously changing and complex legal process of arrest, accusation and conviction. As a result the final trial text incorporates a series of patterns that respond to both the workings of the criminal justice system, and the evolution of a genre. Because of their sheer volume, these records also contain incidental descriptions and references to a wide range of non-criminal behaviour. Historians of crime are aware of these issues and opportunities, but have no way of mapping the patterns embedded within the text, or of discovering 'non-criminal' descriptive elements in a consistent way.

By allowing the analysis and statistical representation of the types of language used in court and how it changed over time, and by comparing these ‘data mined’ patterns to those found in tagged data *With Criminal Intent* will achieve two things. First, a whole new way of charting changes in crime reporting and prosecution will be created; and second a new methodology for the consistent discovery of related descriptions will be benchmarked. The significance of this project therefore runs beyond the discipline of the history of crime, and addresses historical scholarship more broadly, and scholarly engagement with large corpora. It will allow scholars to move beyond the Enlightenment methodology of textual collection and comparison to a structured analysis of layers of patterned meaning as found in specific and distinctive forms of historical representation.

Project Teams

Proceedings of the Old Bailey (UK): Tim Hitchcock (Project Director) and Robert Shoemaker

Zotero (USA): Daniel Cohen (Project Director) and Sean Takats

TAPoR (Canada): Geoffrey Rockwell (Project Director), Stéfan Sinclair, William J. Turkel, and Jörg Sander

Table of Contents

Statement of Significance	1
Table of Contents	2
List of Participants	3
Narrative	4
Partnership	5
History and Future of the Project	6
Environmental Scan	7
Wielding the Tools	8
The Old Bailey Dataset and Standards	9
Technologies and Techniques.....	10
Conclusions	11
Budgets	12
Resumes	29
Letters	43
Appendix A: References.....	60
Appendix B: Mocked-up Screenshots of Old Bailey, Zotero, and TAPoR Integration	61

List of Participants

Alphabetical List of Principal and Co-Investigators, Stakeholders and Collaborators

*Project Directors

†Co-Investigators

Carter, Paul, UK Stakeholder, Special Projects Manager, the National Archives

*Cohen, Daniel J., US Project Director, Director of the Center for History and New Media, George Mason University

Couchman, Hugh, Canadian Stakeholder, Scientific Director, SHARCNET, McMaster University

Emsley, Clive, UK Stakeholder, Professor of History, The Open University

*Hitchcock, Tim, UK Project Director, Professor of Eighteenth-Century History, University of Hertfordshire

Hunt, Margaret, US Stakeholder, Professor of Women's & Gender Studies, Amherst College

Pidd, Michael, UK Collaborator, Research and Development Officer, Humanities Research Institute, University of Sheffield

*Rockwell, Geoffrey, Canadian Project Director, Professor of Philosophy and Humanities Computing, University of Alberta

†Sander, Jörg, Canadian Collaborator, Associate Professor, University of Alberta

†Shoemaker, Robert, UK CI, Professor of Eighteenth-century History, University of Sheffield

Sienna, Kevin, Canadian Stakeholder, Associate Professor of History, Trent University

†Sinclair, Stéfan, Canadian CI, Associate Professor of Multimedia at McMaster University

†Takats, Sean, US CI, Acting Director of Research Projects, Center for History and New Media, George Mason University

†Turkel, William J., Canadian CI, Associate Professor of History, The University of Western Ontario

Narrative: Large Scale Historical Scholarship

The [Old Bailey Proceedings](#) comprise 120 million words of structured text, reflecting evidence given in 197,000 individual trials held over 240 years. As such, it represents the largest coherent body of printed descriptions of behaviour ever published, either in print or online. Naturally scholars have gravitated toward this corpus as a rare and expansive storehouse of human activity, made more important by its unique record of the activities of non-elite actors who are otherwise absent from the official record.

Given its emphasis on crime the *Old Bailey Proceedings* are inevitably filled with unusual and compelling stories. Historians have picked through these stories to illustrate social, cultural, and intellectual history; half a dozen book-length academic case studies have been based on events most fully documented in single trials recorded in the *Proceedings* (cf. Wise 2004; Bondeson 2000; Brewer, 2004; Andrew 2001; Moore, 1994). And yet, despite this wealth of scholarship, the use of these legal records by historians remains little different to how a solicitor or attorney might have approached them a century ago. Scholars still sift through these documents one at a time, hoping to discover an unusual or indicative case, or rely on a rough count of keywords and trial outcomes to discern patterns of criminal behavior that might support a thesis.

This project will show that greater historical rigour can be achieved, and new insights gained by moving from the single trial or narrow run of relevant examples to an analysis of statistically significant textual patterns found in this source as a single, massive whole. Given the availability and power of modern text mining techniques and the fact that the *Proceedings* have already been optimized for use with these techniques, we believe that by building on the success of previous work, this project will change the research paradigm by tying together three separate facilities into a seamless whole:

1. **Newgate Commons:** the project will create a new form of interface for the *Old Bailey Proceedings* that supplements traditional search interfaces. The Newgate Commons will allow scholars to use mining and clustering techniques to identify, collect and work with sets of relevant trials and related texts, and to extract them for further study with other tools. The interface will also make it easy for users to train machine learning ‘agents’ to help discriminate patterns of interest to the researcher.
2. **Zotero Virtual Collections:** the project will extend the [Zotero](#) bibliographic reference management tool so it can be used to manage the collections of documents created within the Newgate Commons and call upon full texts only when needed.
3. **Voyeur Analytics:** the project will connect Zotero to analytical tools designed by the [TAPoR](#) project to work on large collections, including the [Voyeur](#) toolset for analysis and visualization.

The emphasis throughout will be on extending existing tools as needed to allow researchers to navigate between them seamlessly and to use them as a hub from which to manage large study collections. In the process we will create the potential to analyze and visualize change over time in a way that goes beyond current historical methodologies, illuminating the relationship between text and event in new ways. For the first time scholars wanting to examine the criminal record over several centuries will be able to do so, not just quickly, but with a powerful facility to identify patterns that go deeper than a simple count of tagged information. There is also an enormous potential for visualization tools to help in identifying and studying aspects of the corpus that would otherwise resist analysis.

In the short term, the project will deliver:

- **A tested model for how large-scale tools can interoperate.**

Using Zotero and TAPoR on the Old Bailey Proceedings: **Data Mining With Criminal Intent**

- **Three model tools at different levels (corpus, collection management, and analytics) that can be used in new configurations by other projects.**
- **A preliminary study in the history of criminality in Britain that exemplifies how the new research environment can be used.**

Partnership

Criminal Intent brings together three projects with different strengths, along with researchers with the background needed to develop and profit from this integrated set of resources and tools.

Old Bailey Online (Hitchcock and Shoemaker): The Old Bailey Online team and the Humanities Research Institute (HRI) have worked together for ten years, creating and publishing large data sets, including the Old Bailey Online, and developing Natural Language Processing techniques to automate tagging and refine search methodologies. The directors of the Old Bailey (Hitchcock and Shoemaker) are both leading historians of eighteenth- and nineteenth-century Britain, who have consistently pursued a historical approach informed by the possibilities created by the use of new technologies.

Zotero (Cohen and Takats): The Center for History and the New Media ([CHNM](#)), where Cohen is Director and Takats is Director of Research, has fifteen years of experience building educational and scholarly resources, collections, and tools online. In particular for this project it has shown great strength in producing and maintaining end-user software such as Zotero, which has been downloaded over 2 million times. CHNM has a staff of fifty programmers, designers, historians, and researchers that it can leverage for this project.

TAPoR (Rockwell and Sinclair): TAPoR is a consortium of 6 Canadian universities that has provided infrastructure for some 55 projects involving over 150 researchers. One of the most significant outcomes of TAPoR has been the development of a unique Portal for scholars working with digital texts, as well as a diverse array of web-based text analysis tools that can be invoked on the texts ([TAPoRware](#), [HyperPo](#), and [Voyeur](#)). Development of the Portal and tools has been led by Rockwell and Sinclair, who have gained valuable expertise in the interoperability of web-service tools and responding to the needs of humanities scholars. More recent efforts have focused on enabling large-scale text analysis (from a humanities perspective), leveraging the potential of high performance computing. Joerg Sander joins this team to help develop text mining modules for Voyeur.

The partners have worked together before. Zotero co-leads Dan Cohen and Sean Takats have been communicating and collaborating for years with Bill Turkel, Geoffrey Rockwell, and Stéfan Sinclair. As part of an existing major text mining grant from the National Endowment for the Humanities, the Zotero team has already begun integration between the Zotero client and the TAPoR text analysis tools. Cohen and Rockwell serve on the Steering Committee for [centerNet](#), the international network of digital humanities centers. Bill Turkel has worked with Tim Hitchcock on a prototype that uses a Naive Bayesian machine learner to classify criminal trials from the Old Bailey; this work in turn has been developed by the HRI at Sheffield as the basis for a JISC-funded Rapid Innovations project to create a Firefox extension for entity recognition for use with humanities data. For over ten years Tim Hitchcock and Robert Shoemaker have collaborated in partnership with the [HRI](#) at Sheffield and the Higher Education Digitisation Service at Hertfordshire to create the Old Bailey Online and [Plebeian Lives](#) projects. Members of the Canadian team have also worked together at a 2009 workshop that Turkel organized on "[Hacking as a Way of Knowing](#)."

Using Zotero and TAPoR on the Old Bailey Proceedings: **Data Mining With Criminal Intent**

Each individual member of this team has a strong track record of delivering innovative technical projects on time and within budget, many which directly relate to the proposed project. In addition to the NEH text mining grant, Takats and Cohen have marshaled nearly US\$2 million in support from the Institute of Museum and Library Services, and the Andrew W. Mellon Foundation for the Zotero Project. The Old Bailey Online was originally conceived by Hitchcock and Shoemaker in 1999 and following three separately funded development programmes of nearly £1 million in total was completed in March of 2008. In addition the two have worked on the successful 'Plebeian Lives and the Making of Modern London' project. The TAPoR Project, started in 2002, had a total budget of 6.8 million Canadian dollars, a large portion of which was provided by the Canadian Foundation for Innovation.

History and Future of the Project

Just as this project has grown from series of innovative initiatives in related areas, the partners in this project continue to work on the cutting edge of data mining and digital humanities. Turkel received SSHRC ITST funding to host an October 2009 workshop on the subject of application programming interfaces (APIs) for the digital humanities. The workshop will gather researchers and programmers from some of the key online projects in Canada, the US and the UK, to develop a strategy for providing APIs that mesh seamlessly with one another, expose data that has remained inaccessible until now, and provide a platform for a new generation of online research. If *Criminal Intent* is funded the workshop will be extended to focus on the APIs needed for Old Bailey, Zotero and TAPoR to interoperate. As part of the NEH text mining grant, Cohen and Takats have already completed a comprehensive survey of researchers to discover what they would like out of text mining tools applied to collections like the Old Bailey. In 2009 Hitchcock and Shoemaker received a grant from JISC for '[Scrutiny](#): A Firefox Extension for Entity Recognition within Research Data' that utilizes both NLP and Naive Bayesian logic to assist in locating discipline-specific information in large bodies of text.

The team also brings sophistication in high-performance computing. In April 2008 Geoffrey Rockwell and Hugh Couchman, the scientific director of [SHARCNET](#), an HPC consortium in Southern Ontario, organized a symposium to help bridge the gap between digital humanities and High Performance Computing research. Several concrete initiatives emerged from the workshop including the development of a high-resolution visualization interface for text analysis, called the Big See. Finally, Joerg Sander, an Associate Professor in the Department of Computing Science at the University of Alberta, has research experience in data mining and similarity search in databases.

This project is a natural extension of this work and experience. The following provides a brief timeline for the project:

1. **[Pre-Grant] October 2009 - Development of APIs.** Building on the previously SSHRC-funded API workshop, the team will specify ways that study collections can be gathered, called and passed among tools like Zotero and TAPoR. This will include specifications for passing virtual collections of metadata and pointers to full text items held elsewhere.
2. **Beginning January 2010 – Development of Tools.** Building on the defined APIs the three teams will develop the three major components.
 - a. The Newgate Commons will be developed by the UK team with support from Turkel and Sander in relation to mining techniques for document clustering.
 - b. The USA team will extend Zotero so it can query the Newgate Commons directly and import study collections made up of metadata and pointers.

- c. The Canadian team will adapt the Voyeur environment so that it can take a study collection (or subset) as described within Zotero, call the Newgate Commons for the full text, index it, and provide the user with a rich analytical environment. They will also build new tools that cluster associated collocating words for visualization in this and other large study collections.
3. **Beginning September 2010 – Research Testing.** Historians in all three countries will use the project tools to study criminality in 18th and 19th century Britain. These research projects will not only aim to answer pressing historical questions, but also be used to test and refine the system and to develop documentation so that other projects can work with any of the three components and build on them.

Environmental Scan

Scattered attempts to allow scholars to perform computational research on digital corpora have shown promise. The National Center for Supercomputing Applications ([NCSA](#)) at the University of Illinois at Urbana-Champaign (UIUC) has produced Text to Knowledge ([T2K](#)), a complex Java package that provides a number of interchangeable text analysis modules (T2K has recently become part of the Software Environment for the Advancement of Scholarly Research ([SEASR](#)) project, funded by the Mellon Foundation). The [MONK](#) project (formerly NORA), of which Sinclair is one of the principal investigators, has used T2K as well as several other text-mining tools to perform tests such as identifying words of affection in Emily Dickinson's poems. Open-access visualization tools like IBM's [Many Eyes](#) provide exciting new opportunities to create graphs and charts of the words in a text or texts.

Although pioneering, these tools remain unlikely to become part of most scholars' tool kits. First, they generally require a significant amount of technical knowledge to install and run. For instance, [Philomine](#), the University of Chicago's text-mining extension to their full-text search tool [Philologic](#), requires knowledge of the programming language Perl, command-line installation and configuration skills. Indeed, a majority of text-mining tools require knowledge of a specific programming language and an understanding of how to connect the program to existing collections of text. Second, even technically sophisticated users will find that each program has a different interface and preferred input type that often requires preprocessing texts (i.e., putting texts into a particular format before analysis can begin). Third, with rare exceptions these tools are generally the products of software developers working towards specific research aims and have not been field tested. Fourth, existing text mining tools, like Monk, are limited for end-users to static, pre-indexed corpora, but as Zotero has shown, scholars want to build and manage their own collections of texts. In short, while these tools make new forms of analysis possible, they do not make it easy.

One of the goals of this project is to marry techniques developed in computer science to research tools and interfaces that humanists are comfortable with. In computer science, text mining based on both unsupervised and supervised methodologies have formed the basis for a great deal of work. But published work usually describes research prototypes that aren't easily used by people other than the originators. For this project we will be using promising methods such as Association Rule Mining to enable the visual exploration of a document collection based in the co-occurrence of sets of terms with one or more user-specified reference terms. Lopes et al 2007 suggest that association rules can represent "themes" drawn from selected sets of documents to create a visual representation of that collection. *Criminal Intent* will adapt this and other techniques to research contexts already familiar to humanist scholars, such as that provided by Old Bailey Online (a research corpora where users can

search and read), Zotero (bibliographic resource management), and TAPoR (concordance tools that always allow users to drill down to the full text.)

Wielding the Tools

Who will use the system? Social and cultural historians of eighteenth and nineteenth-century Britain and scholars working on the histories of crime and the law will be most immediately affected. But the resource that we create will be applicable to practically every other English-language historical dataset. The creation of the Newgate Commons will allow historians to use the *Proceedings* in a more subtle and powerful way; but the simple existence of 120 million words of text in a form that can be used to train a machine learning 'agent' also provides a starting point in the creation of tools applicable across the broader digital landscape. In the first instance, the histories of technology and personal appearance, of migration and social interaction, of gender, sexuality and identity, of literature and art are all available through the *Proceedings*, and the Newgate Commons will empower researchers in all these fields in a new way. And as the technology is transferred to other datasets through documented APIs, we anticipate an ever-growing series of users drawn from an ever growing number of disciplines, applying this methodology to a growing body of corpora.

All historians seek to re-use and re-interpret our inherited archive in ways that were never intended by its creators. This project allows the process of re-interpretation to be pushed one step further. Take the example of a historian working on a topic such as domestic violence in eighteenth- and nineteenth-century Britain, whose main interest is to gain a greater understanding of the changing nature of relations between men and women through an exploration of those moments when such relations fail violently. The existence of sources such as the Old Bailey at first sight do not help much. While inter-gender violence was and is commonplace, in the past it was not illegal, except in the rare cases in which death or brutal injury made it intolerable. As a result, through structured searching of the Old Bailey a historian can identify only a relatively small number of cases in which inter-gender violence resulted in a named charge - most frequently murder. Although over 197,000 trials are recorded in the *Proceedings*, less than a thousand involve inter-gender murder (743 men murdering women, and 245 women murdering men), and in only 264 cases was inter-gender assault the focus of a legal case. As a result a gender historian working with the Old Bailey would normally be both directed to an unrepresentative sample of extreme violence (predominantly murder), and forced to work with a relatively small sample (only 38 assaults of men by women, for instance, in almost 240 years).

By using the Newgate Commons a historian of domestic violence could both access non-criminal forms of inter-gender violence, and locate such violence in trials where the named crime was not the violence itself. Here are the steps our gender historian would take:

1. Establish an account to use the Newgate Commons environment, giving them access to data mining tools and the ability to create study collections within the Old Bailey.
2. Start their exploration by creating a machine learning agent that 'watches' as they look through the *Proceedings* for trials of interest. When the historian identifies the 264 inter-gender assault cases as relevant for the study at hand, the agent uses the XML tagging to find the types of language used to describe violence between men and women. As these pre-selected cases are explored, the machine learner builds a cluster of concepts.
3. The researcher can then filter and correct the mined results, providing both positive and negative feedback for the learner, which, in turn, can suggest further cases from the wider archive of Old Bailey trials. In this way, the agent is trained as the historian builds a Study Collection.

4. Our historian can then do a number of things with this Study Collection. First, as it has been imported into Zotero, it can be managed as a 'library' or Virtual Collection. Zotero can distil the metadata attached to each trial, allowing new ways of sorting and searching the collection. The researcher can also use Zotero to access the full records on the Old Bailey site, ensuring that the subset of trials in the Study Collection can be fully and continually contextualized.
5. The researcher could simply go through and read the collection one trial at a time, but as the body of material is likely to have grown beyond an easily digested literary meal, they will want to look for patterns in the descriptions of violence brought into Zotero. To do so, TAPoR tools such as Voyeur can now be used to analyze the Study Collection. These tools could be accessed via the Newgate Commons, but more commonly will be launched from Zotero on a subset of the study collection. This process will be supported with online help, descriptions and research recipes.
6. Behind the scenes Zotero will allow the Study Collection to be passed to Voyeur, which in turn, communicates directly with the Newgate Commons site, allowing it to access the full text of the *Proceedings*. Once it has indexed the collection, Voyeur can then present our historian with both a trend summary for their own Study Collection, a contextual summary of the *Proceedings* as a whole, and visualizations of clusters of co-occurring words of interest using Association Rule Mining. Our historian can then use the interface provided by Zotero to open panels allowing navigation both around the texts they have selected, and away from these to texts on either side.
7. The agent created in the process of sifting for a particular kind of violence from the text of the *Proceedings*, can also be applied to other libraries held in Zotero; so the descriptive markers for inter-gender violence found in court records might be applied as a preliminary filter to libraries of eighteenth and nineteenth century literature or newspaper reports, or to a broader corpus of digitized historical material. Since the agent will represent what it has learned in a human-readable form, its internal representation can be inspected at any time and used as a source of new hypotheses. It can also be shared with other researchers who have similar interests.

The Old Bailey Dataset and Standards

The [Old Bailey Proceedings Online](#) is a fully searchable, digitized collection of the complete text of all surviving editions of the *Old Bailey Proceedings* from 1674 to 1913, and of the *Ordinary of Newgate's Accounts*, 1679 to 1772. These form a comprehensive record of all trials held at Britain's foremost criminal court. The site allows access to over 197,000 trials and through the *Ordinary's Account*, biographical details of approximately 2,500 men and women executed at Tyburn. It also provides digital images of all 190,000 original pages of the *Proceedings*, 4,000 pages of *Ordinary's Accounts*, advice on methods of searching this resource, information on the historical and legal background to the Old Bailey court and its *Proceedings*, and descriptions of published and manuscript materials relating to the trials covered. Contemporary maps and images have also been provided. In total it allows direct access to 120,000,000 words of text.

The text included in the site has been 'double entry rekeyed', resulting in an accuracy rate of well over 99%. An image of each page, in either a GIF or JPEG format is also associated with the relevant text. The digitized text has been comprehensively tagged in XML. All tagging is fully TEI compliant, and the website as a whole complies with W3C HTML/HTTP standards, and is W3C Web Accessibility Initiative (WAI) compliant. Trials tend to have a regular structure (though with considerable minor variations) and certain aspects of the text were tagged to reflect the meaning of particular words or phrases, for example names and crimes. In order to create meaningful and consistent statistics, certain subcategories of information were also identified, such as types of verdict. The site also allows keyword and phrase searches to be combined with statistical queries based on tagged information. The following categories of information have been marked up: Crime (divided into 9 general categories and 56 specific

Using Zotero and TAPoR on the Old Bailey Proceedings: **Data Mining With Criminal Intent**

types), date, and location; Defendant name, status or occupational label, gender, age and location; Victim name, status or occupational label, and gender; Judges' names, Jury names, Other person names, Verdicts (divided into 4 general categories and 23 specific types), Punishments (divided into 6 general categories and 26 specific types), and Advertisements. Where appropriate tagged fields can be tabulated statistically.

Technologies and Techniques

This project will build on the Old Bailey corpus by integrating and extending two tools widely used by digital humanities scholars: Zotero, and TAPoR tools.

[Zotero](#) is a powerful research tool that helps scholars gather, organize, and analyze sources and lets them share the results of their research in a variety of ways. Zotero already has preliminary support for the Old Bailey site, but we plan to significantly enhance the support for Old Bailey metadata and tagging. We will also continue our prototype work on providing analytic functionality from within Zotero – users will be able to select documents from their library to be packaged by Zotero for analysis within the TAPoR Portal.

The [TAPoR](#) project provides access to text analysis tools for both micro-analysis of small corpora ([TAPoRware](#) and [HyperPo](#)) and macro-analysis of larger corpora ([Voyeur](#)). We have already experimented with connecting TAPoR tools to Zotero. In this project we will focus on allowing Voyeur to call study collections from the Newgate Commons, integrating more text mining capabilities into Voyeur and adding visualization of text mining results.

The project will also create a freestanding set of APIs giving comprehensive access to the Old Bailey dataset, Zotero and Voyeur. There will be four aspects to the APIs:

1. Enabling Zotero and other web applications to interrogate the Old Bailey dataset, using both existing search criteria (e.g., www.oldbaileyonline.org/Main) and the API's accompanying SQL model.
2. Enabling Voyeur and other web applications to retrieve both data items in native TEI XML and statistical information such as the overall number of hits which were found (existing statistical functions can be seen here: www.oldbaileyonline.org/Stats) Enabling external web applications to retrieve the Old Bailey data items in a variety of display formats, such as RSS, RDF and XHTML.
3. Providing a space to allow tools created by independent developers using Old Bailey material to be posted, re-used and distributed.

In the first instance, standard machine learning and text/data mining algorithms for clustering, entity extraction, and filtering will be made available. These will include, but not be limited to, statistical NLP, Bayesian and k-nearest neighbor approaches, non-negative matrix factorization and clustering based on compression (as a way of estimating Kolmogorov complexity).

One of the most promising areas for innovation in working with the Old Bailey corpus will be in the integration and development of visualization interfaces. Several types of visualizations will be available immediately through TAPoRware and Voyeur, including the following:

- **Word clouds:** a common representation of relative frequency of words or themes
- **Visual collocations:** a dynamic tool for exploring words or themes that occur in close proximity

- **Weighted centroids:** a technique for visualizing word distribution information in a circular area
- **Mandala Browser:** a rich prospect interface for studying associations between multiple search items
- **LAVA:** a multi-layer browsing interface for exploring very large datasets

New visualizations available through Voyeur will include Document Clustering: which creates a visualization of the inherent groups of documents in the collection with respect to a similarity measure that can be user-specified based on the raw text term frequencies and/or the metadata. To create the visualization, the data will first be clustered using a hierarchical density-based clustering algorithm. The cluster hierarchy can be directly represented as a tree, a so-called "dendrogram", or it can be mapped in two dimensions so that the similarities between documents (represented then as 2d points) are preserved using some form of dimensionality reduction such as principal component analysis. When the mapping works well, this second visualization represents the density contours of the existing clusters and their containment relationships in a more direct way that is easier to perceive than in a cluster tree.

We will create a Voyeur panel that will allow visual exploration of a document collection with respect to sets of terms that frequently co-occur with one or more user-specified reference terms. For this panel, the user will be able to specify one or more terms/words as reference terms of interest. The system will then find sets of other terms that co-occur together with the reference terms with high frequency within a specified neighborhood in the given document collection. The concept of word co-occurrence patterns as conceived here goes beyond the currently-available feature of listing frequent word collocations. The panel would provide the following functionality: (1) display a list of co-occurrence patterns that can be ranked by different criteria such as the frequency or the size of the patterns; (2) select a subset of patterns for further analysis: (2.1) allow interactive inspection of the "support set" of each selected pattern, i.e., the set of documents that contain the combination of terms that make up the pattern; (2.2.) find frequent co-occurrence patterns in each of the support sets in order to characterize and explore these subsets of documents in more detail; (2.3) allow the selection of one of the patterns, as the basis for new reference terms and compute and display the frequent co-occurrence patterns with respect to this new set. The user will be able to iterate these steps.

Conclusions

The significance of this project lies in three distinct characteristics:

1. It will create a generic template for the integration of online textual resources with the disparate tools for analysis rapidly emerging from Digital Humanities – creating a space, to be called the 'Newgate Commons', where the object of study (tagged historical full-text records) can be flexibly interrogated and represented in an abstract way that goes beyond the practices of post-Enlightenment scholarship.
2. It will also provide a benchmark for the opportunities and pitfalls of a range of data-mining tools. Because the Old Bailey Online was developed using double-entry rekeying rather than an OCR methodology (resulting in a remarkably accurate transcription), because it is a highly structured source that lends itself to statistical representation, and because it has been heavily tagged for information such as crime, gender, verdict and punishment, the application of data mining techniques to this corpus can be used to benchmark these techniques against results based on tagged information, allowing us to measure the extent to which the patterns being discovered are in the text, rather than in the OCR.
3. In the process it will provide the type of measurable assurance of the strengths of these methodologies of the sort required to assure their adoption by scholars in the humanities.

“Using Zotero and TAPoR on the Old Bailey Proceedings: Data Mining with Criminal Intent”

Letters of support

Universities and Institutes:

Arts Resource Centre (University of Alberta)
George Mason University
Humanities Research Institute (Sheffield University)
McMaster University
Sharcnet (University of Western Ontario)
University of Hertfordshire
University of Sheffield
University of Western Ontario
WestGrid (University of Alberta)

Stakeholders:

Dr Paul Carter, the National Archives UK
Clive Emsley (Professor of the History of Crime and Policing, Open University)
Margaret Hunt (Chair of Women's & Gender Studies, Amherst College)
Kevin Siena (Professor of Early Modern European History, Trent University)

Appendix A: References

- Andrew, Donna T., *The Perreaus and Mrs Rudd: forgery and betrayal in 18th century London*. Berkeley (CA): University of California Press, 2001.
- Beattie, John M., *Policing and punishment in London 1660-1750: urban crime and the limits of terror*. Oxford: Oxford University Press, 2001.
- Bishop, Christopher M., *Pattern recognition and machine learning*. New York (NY): Springer-Verlag, 2007.
- Bondeson, Jan, *The London Monster: a sanguinary tale*. Philadelphia (PA): University of Pennsylvania Press, 2000.
- Brewer, John, *A sentimental murder: love and madness in the eighteenth century*. London: Harper Collins, 2004.
- Cilibrasi, Rudi and Vitanyi, Paul, "Clustering by Compression," *IEEE Transactions on Information Theory*, 51, no. 4 (2005): 1523-45.
- Gatrell, V. A. C., *The hanging tree: execution and the English people, 1770-1868*. Oxford: Oxford University Press, 1994.
- Hitchcock, Tim, *Down and out in eighteenth-century London*. London: Hambledon & London, 2004.
- Huber, Magnus, "[The Old Bailey Proceedings, 1674-1834, evaluating and annotating a corpus of 18th- and 19th-century spoken English](#)" in Meurman-Solin, Anneli and Nurmi, Arja (eds.), *Annotating variation and change. (Studies in Variation, Contacts and Change in English 1, 2007)*.
- Lee, Daniel D. and Seun, H. Sebastian Seung, "Algorithms for non-negative matrix factorization," <<http://hebb.mit.edu/people/seung/papers/nmfconverge.pdf>>
- Linebaugh, Peter, *The London hanged: crime and civil society in the eighteenth century*. London: Allen Lane, 1991.
- Lopes, A. A., Pinho, R., Paulovich, F. V., Minghim, R., "Visual text mining using association rules," *Computers & Graphics* 31 (2007): 316-326.
- Manning, Chris and Hinrich Schütze, Hinrich, *Foundations of statistical natural language Processing*. Cambridge (MA): MIT Press, 1999.
- Moore, Judith, *The appearance of truth: the story of Elizabeth Canning and 18th century narrative*. Newark (DE): University of Delaware Press, (1994).
- Segaran, Toby, *Programming Collective Intelligence*, Sebastopol (CA): O'Reilly Media, Inc., 2007.
- Shoemaker, Robert Brink, *The London mob: violence and disorder in eighteenth-century London*. London: Hambledon & London, 2004.
- Trumbach, Randolph, *Sex and the gender revolution, vol. 1: heterosexuality and the third gender in enlightenment London*. Chicago (IL) and London: Chicago University Press, 1998.
- Weiss, Sholom M., Nitin Indurkha, Nitin, Tony Zhang, Tony and Fred J. Damerau, Fred J., *Text Mining: Predictive Methods for Analyzing Unstructured Information*. New York (NY): Springer-Verlag, 2005.
- Wise, Sarah, *The Italian Boy: murder and grave-robbery in 1830s London*. London: Jonathan Cape, 2004.

Appendix B: Mocked-up Screenshots of Old Bailey, Zotero, and TAPoR Integration

Figure 1:
Enhanced
Integration of
Old Bailey and
Zotero

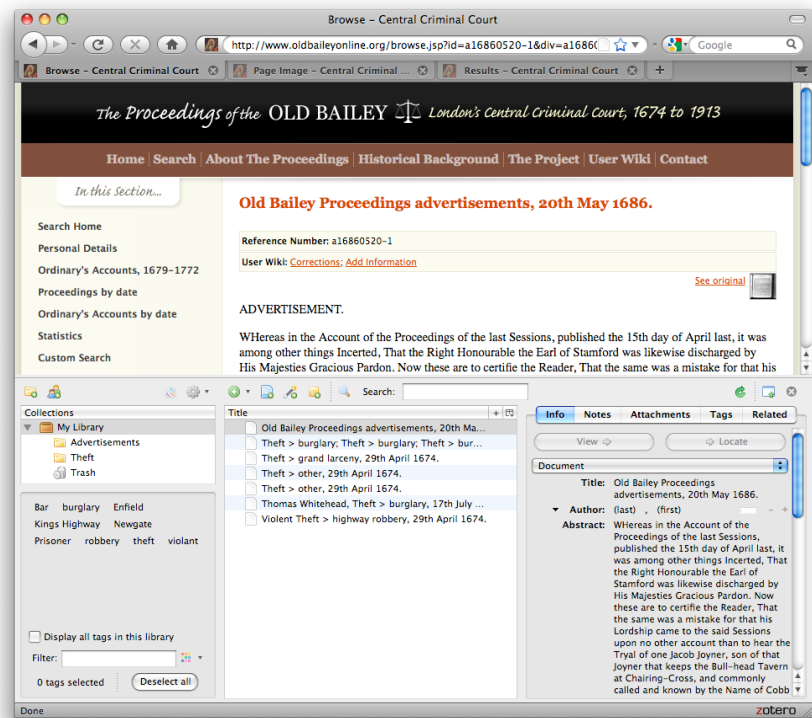
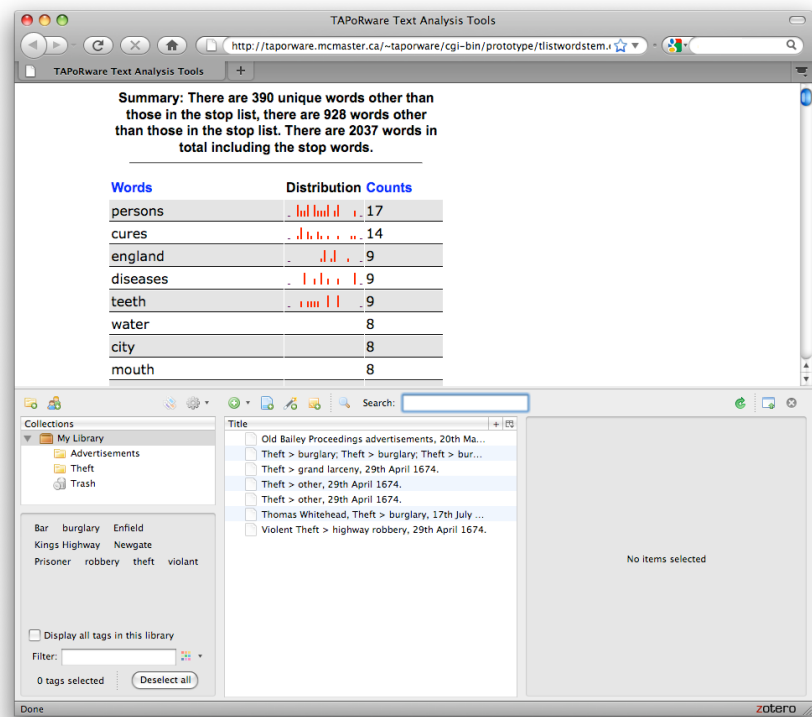


Figure 2:
Invoking TAPoR
frequency tool of
Old Bailey data
via Zotero.



Using Zotero and TAPoR on the Old Bailey Proceedings: **Data Mining With Criminal Intent**

Figure 3:
Word Cloud
visualization of
Old Bailey data
via Zotero.

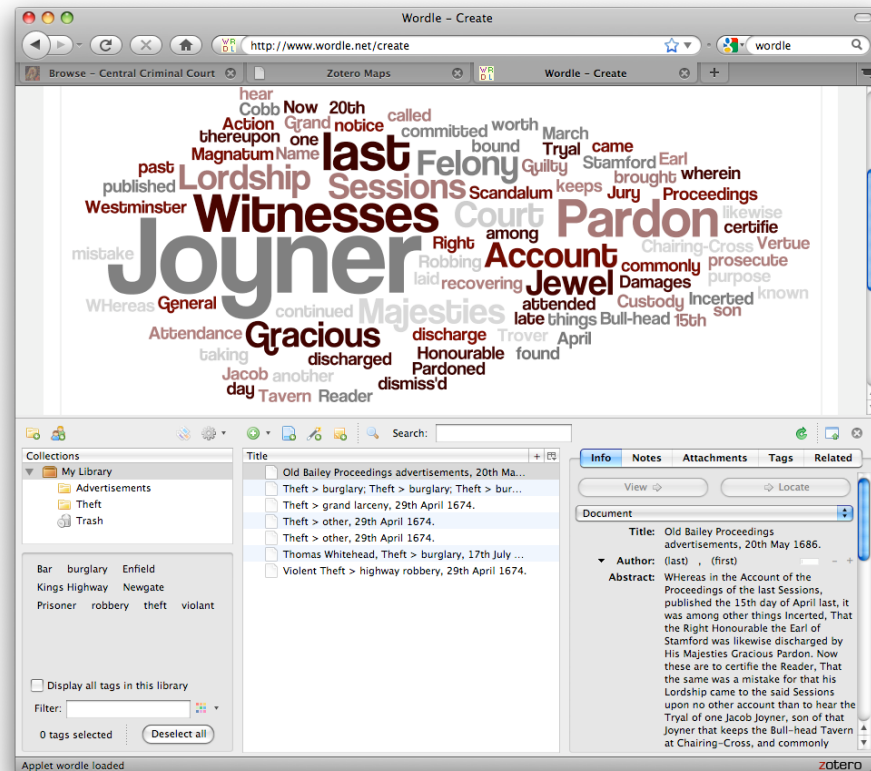


Figure 4:
Map mash-up of
Old Bailey data
via Zotero.

