

Project Acronym:DMwCI
 Version: Final Draft
 Contact: Tim Hitchcock
 Date: 14 Feb. 2010



Project Document Cover Sheet

Project Information			
Project Acronym	DMwCI		
Project Title	Using Zotero and TAPoR on the Old Bailey Proceedings: Data Mining With Criminal Intent		
Start Date	1 January 2010	End Date	31 March 2011
Lead Institution	University of Hertfordshire		
Project Director	Professor Tim Hitchcock		
Project Manager & contact details	Michael Pidd, University of Sheffield, Humanities Research Institute, 0114 222 6116, 0114 222 9894, m.pidd@sheffield.ac.uk		
Partner Institutions	University of Sheffield		
Project Web URL	http://withcriminalintent.blogspot.com/		
Programme Name (and number)	<i>Digging Into Data</i> CCICP001		
Programme Manager	Alastair Dunning		

Document Name			
Document Title	<i>Project Plan</i>		
Reporting Period	<i>for progress reports only</i>		
Author(s) & project role	Tim Hitchcock, Project Director		
Date	17 Feb. 2010	Filename	CriminalIntentProjectPlan
URL	<i>if document is posted on project web site</i>		
Access	<input checked="" type="checkbox"/> Project and JISC internal	<input type="checkbox"/> General dissemination	

Document History		
Version	Date	Comments
Final	17 Feb. 2010	Seen by all project participants

Project Acronym:DMwCI
 Version: Final Draft
 Contact: Michael Pidd
 Date: 14 Feb. 2010



JISC Website Template for Projects

To be completed by the projects	
Project Title	Using Zotero and TAPoR on the Old Bailey Proceedings: Data Mining With Criminal Intent
Project website address	http://www.criminalintent.org
Start date	January 2010
End date	31 March 2011
Overview	This project 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).
Aims and objectives	<p>Aims:</p> <ul style="list-style-type: none"> • To deliver a tested model for how large-scale tools can interoperate. • 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. <p>Objectives:</p> <ul style="list-style-type: none"> • To showcase the integration of online textual resources with bibliographical and analytical tools emerging from Digital Humanities. • And provide an exemplar of the use of datamining in historical research.

Project methodology	Tim Hitchcock and Robert Shoemaker will direct the British side of the project, in collaboration with Michael Pidd at the Humanities Research Institute at Sheffield, who will act as project manager. The HRI will implement the Old Bailey API. The CHNM will modify Zotero, and University of Alberta team will create the tools for quantitative analysis and visualization. Tim Hitchcock and William Turkel, with Robert Shoemaker will undertake the exemplar project in the history of crime.
Anticipated outputs and outcomes	A Web API (Application Programming Interface) which will permit third-party tools such as Zotero to data mine the Old Bailey proceedings dataset.
Technology / Standards used (if applicable)	C++, REST, SOAP, TEI P5 XML, RDF XML
Project Manager & Team	Michael Pidd, University of Sheffield, Humanities Research Institute, 0114 222 6116, 0114 222 9894, m.pidd@sheffield.ac.uk
Project Team	Jamie McLaughlin, University of Sheffield, Humanities Research Institute, 0114 222 9892, 0114 222 9894, j.mclaughlin@sheffield.ac.uk Kathy Rogers, University of Sheffield, Humanities Research Institute, 0114 222 6110, 0114 222 9894, k.m.rogers@sheffield.ac.uk
Lead Institution	University of Hertfordshire
Project partners	University of Sheffield - http://www.shef.ac.uk/hri
To be completed by Programme Managers	
JISC programme	
JISC theme(s)	
JISC Programme Manager	
JISC Programme Director	
Related projects	<i><list of related projects/links></i>

Project Acronym:DMwCI
Version: Final Draft
Contact: Tim Hitchcock
Date: 14 February 2010



JISC Project Plan:

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

Overview of Project

1. Background

1.1 Over the past few decades scholars have increasingly used court records to illuminate historical themes in novel ways. 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, and to the analysis of interpersonal relationships and human behaviour more generally. But, despite the fact that 120 million words of court transcripts published in the *Proceedings* are now available online in a structured and searchable form, historians and humanist scholars continue to use these legal records in an essentially iterative and traditional manner; and largely failing to take full advantage of the variety of forms of analysis the *Proceedings*'s online format allow. At the same time, in Zotero, a popular environment for managing online scholarship has been created that allows humanists to collect, index and manipulate large amounts of text; while in TAPoR Tools, a range of facilities for the quantitative analysis of text, has been piloted and tested. By bringing together in one seamless online environment, the text of the *Proceedings*, the functionality of Zotero and the tools created by TAPoR, this project will allow scholars to take new approaches to this old source.

1.2 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. 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 three things. First, a whole new way of charting changes in crime reporting and prosecution will be created; second, a new methodology for the consistent discovery of related descriptions will be benchmarked, and finally a working model of how large corpora can be handled online and in a distributed fashion, will be demonstrated. 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.

2. Aims and Objectives

2.1 This project aims to demonstrate that greater historical rigour can be achieved, and new insights gained through the application of data mining and statistical analysis to large bodies of primary sources such as the *The Proceedings of the Old Bailey*. 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. In the process, it will allow the end user, scholars and students, to experience the three separate components of this project (the *Proceedings*, Zotero and TAPoR tools) as a single seamless resource. To achieve this aim, we need to reach three specific objectives:

2.1.1. The creation of *Newgate Commons*: a new form of interface for the *Old Bailey Proceedings* that supplements the current 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 identify patterns in the text (and underlying account of prosecutions and punishments) of interest to the researcher.

2.1.2 The modification of *Zotero Virtual Collections*, 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.

2.1.3. *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 *Zotero* 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.

3. Overall Approach

3.1 This project will create an exemplar for modern text mining techniques, using the *Proceedings* as the basic text object. Each of three different facets will be refined or developed to allow the analysis of the *Proceedings* to work. This will be undertaken in parallel, with the common standards and levels of interoperability tested through joint development meetings. The three components are:

3.1.1 *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.

3.1.2. 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.

3.1.3. 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.

3.2 The main issue for this project is the establishment and implementation of common technical standards that allow each of the three components to communicate with each other, and to pass data between them. In order to achieve this, the *Newgate Commons* is being built from scratch, while Zotero and the Voyeur Toolset are being modified to take advantage of the *Newgate Commons*.

3.3 The project will deliver a tested model for how large-scale text mining can be achieved using current tools and facilities. It will make three model tools at different levels (corpus, collection management, and analytics) that can be used in new configurations by other projects. And a preliminary study in the history of criminality in Britain that exemplifies how the new research environment can be used will also be researched, written and published. The project does *not* seek to define a single set of standards for work in this area, or to create a framework for direct expansion to include other sources (although all its work will be recorded, archived and openly available).

3.4 The implementation of shared standards, and seamless passing of data from the *Newgate Commons* to Zotero, and from Zotero to TAPoR Tools, must work. Reliable, repeatable analysis that can be referenced with confidence are necessary to ensure the site is used extensively by the wider scholarly community, and seen to reflect a form of academic practise that can be adopted in other fields. As a result, the most significant success factor for this project lies in its basic functionality. In addition, the preliminary studies in the history of crime generated must also address issues that a wider historical and scholarly community consider important. This element of the project needs to demonstrate that datamining and visualisation methodologies add something important to current academic practise both in the history of crime and the humanities more generally.

4. Project Outputs

4.1 The specific deliverables associated with this project include:

- A tested model for how large- scale tools can interoperate.
- 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, published in the form of an academic article in a peer reviewed journal.

5. Project Outcomes

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. 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.

6. Stakeholder Analysis

Stakeholder	Interest / stake	Importance
Social historians and historians of crime.	The <i>Proceedings</i> represent perhaps the single most significant source in print or online, for the history of crime. By applying new strategies to the analysis of this source, historians of crime will be able to re-interpret, re-analyse and newly understand this source.	High
HE Humanities research community (academics, post graduates)	This facility will illustrate a different way of working with the sources of humanities scholarship, improving both the discovery of relevant primary source material, and allowing new types of structured searching, analysis and visualisation.	High
HE academic researchers in non-humanities disciplines.	The ability to apply data mining techniques and analytical tools to difficult textual objects is a problem shared by the humanities with other disciplines. This project will contribute to making these techniques more useful across all disciplines in which text forms either an object of study, or a primary medium of record.	Medium
The digital humanities community, in particular those creating digital resources.	'Data Mining with Criminal Intent' will demonstrate a new way to analyse large text corpora. It will also illustrate how the tools of datamining, visualisation and quantitative linguistics can be brought to bear on a wide variety of texts.	High

7. Risk Analysis

Risk	Probability (1-5)	Severity (1-5)	Score (P x S)	Action to Prevent/Manage Risk
Illness or unavailability of project manager	2	1	2	Should Pidd become unavailable, the HRI will appoint a new Technical director to undertake his roles.
Illness or unavailability of project directors	2	1	2	Should Hitchcock or Shoemaker become unavailable, the other could increase their level of participation.
Organisational, ie. Staffing - illness or unavailability	2	1	2	<p>The HRI employs three technical officers each of whom could undertake the implementation of this project.</p> <p>The Centre for History and the New Media, and the TAPoR project both employ large numbers of staff capable of undertaking their elements of the project should the initial named programmmers and project leaders become unavailable.</p>
Technical, ie. Failure to design a working API that can exchange data with Zotero.	1	5	5	<p>The basic methodology for the creation of the API is well understood and straightforward. The main risk is that poor communication between the various partners leads to separate developments that fail to communicate one to the other. Clear lines of communication, continuous testing and regular contact between all parties, will prevent this happening.</p>
Failure to create a publicly available API with full access to Old Bailey materials.	1	5	5	The process of designing and implementing an API is clearly understood, and subject to minimal risk.
Failure to work effectively with International partners.	1	4	4	Communication between the different partners at all levels, including both technical developers, and academic leads is

				central to the success of this project. A regular pattern of phone conferences, a series of workshops and collaborative meetings between developers, and a clearly articulated series of technical standards will alleviate this issue.
--	--	--	--	---

8. Standards

Name of standard or specification	Version	Notes
Java	6.0	Java will be used for the server-side processing of API requests
REST		REST will be used as the client-server communication protocol
MySQL	5.x	MySQL will receive the queries from the API and interrogate the dataset and indexes
XHTML	1.0	XHTML is one of the possible formats for returning the search results to the client.
TEI XML	P5	TEI is one of the possible formats for returning the search results to the client.
RDF		RDF is one of the possible formats for returning the search results to the client.

9. Technical Development

Overview of the API

The project will create a free standing API giving comprehensive access to the Old Bailey dataset for tools such as Zotero and TAPoR. The server-side functionality of the API (responding to requests) will be written in Java. The client-side functionality (making requests from Zotero and TAPoR) sent to the API using REST.

There will be two aspects to the API:

1. Search, enabling external web applications such as Zotero to interrogate the Old Bailey dataset using the Old Bailey's existing search criteria (an example of the existing search criteria is available here: <http://www.oldbaileyonline.org/forms/formMain.jsp>). It will also be possible for external web applications to interrogate the Old Bailey dataset using search criteria not currently available within the Old Bailey's own website using the API's accompanying SQL model.
2. Retrieval, enabling external web applications to retrieve the resulting Old Bailey data items in native TEI XML or optionally in RDF XML and XHTML, in addition to

retrieving statistical information such as the overall number of hits which were found and the Old Bailey website's existing statistical functions (existing statistical functions can be seen here: <http://www.oldbaileyonline.org/forms/formStats.jsp>).

The data process flow will be as follows:

1. The client-side tool (eg. Zotero or TAPor) will send a data mining request to the Old Bailey server using the REST communication protocol.
2. The Old Bailey server will receive the request and turn this into a SQL query statement.
3. The server will send the SQL query statement to the MySQL database which manages the Old Bailey dataset.
4. The MySQL database will perform the query and return the resulting records.
5. The server will transform the records into the client's requested format (the native format is TEI P5 but RDF and XHTML formats will be available).
6. The server will then forward the resulting records to the client-side tool using the REST communication protocol.

Development Methodology

The development methodology for the API will be the HRI's version of Agile Development within a SCRUM relationship:

1. The client (Hitchcock and Shoemaker, in consultation with Zotero and TAPoR) provides the HRI with requirements.
2. The HRI models the requirements as a technical specification.
3. A prototype is developed in accordance with the specification.
4. The prototype is tested internally, refined and re-tested (back to stage 3).
5. The client tests the prototype (using Zotero and TAPor tools in a live, online environment).
6. The HRI and the client discuss whether refinements are required of the HRI's specification or the client's specification (Zotero and TAPor).
7. The next cycle begins.

We envisage three cycles, although in reality we will conduct as many cycles as are necessary in order to deliver the solution.

10. Intellectual Property Rights

10.1 A collaboration agreement to cover shared IPR and the management of the project as a whole, again created in consultation with JISC Legal, will be completed by the end of month one of the project.

10.2 All code and technical developments generated by the project will be available as an Open Source, and governed by a Creative Commons Licence.

10.3 The *Newgate Commons* and associated web content (i.e. background pages and web design) will be jointly owned by the University's of Hertfordshire and Sheffield. Intellectual Property developed by the US and Canadian partners will remain with the HEIs that develop it.

10.4 As a matter of principal, the IP developed by the three national sub-projects collaborating in this project will retain ownership of the IP they develop, and that in the case of disputes regarding IP, it will be understood within a local national legal framework of the most fully effected sub-project.

10.5 A process of dispute arbitration is laid out in paragraph 9.3 of the UK Collaboration agreement.

10.6 The IP in the original Old Bailey transcripts are jointly owned by the University of Sheffield, the University of Hertfordshire, and the Open University. This material is free for non-commercial use, and is use in this project in accordance with this proviso.

Project Resources

11. Project Partners

University of Hertfordshire, lead institution.

Main contact: Professor T Hitchcock

Roles: UK Project Director; Project Management and historical analysis of data developed using the facilities created.

t.hitchcock@herts.ac.uk

University of Sheffield,

Main contact: Professor R Shoemaker

Roles: UK Co-Director; Project Management and historical analysis of data developed using the facilities created.

R.Shoemaker@sheffield.ac.uk

Humanities Research Institute, University of Sheffield

Main contact: Michael Pidd

Roles: Project Manager

(m.pidd@sheffield.ac.uk)

Center for History and New Media, George Mason University

Main contact: Daniel J. Cohen

Roles: US Project Director

(dan@dancohen.org)

University of Alberta

Main contact: Geoffrey Rockwell

Roles: Canadian Project Director

(geoffrey.rockwell@ualberta.ca)

University of Western Ontario

Main contact: William Turkel

Roles: Canadian Co-Director

(william.j.turkel@gmail.com)

The ‘Roles and Responsibilities’ agreed between the Universities of Sheffield and Hertfordshire are adumbrated in a separate schedule included with this Project Plan as Appendix C.

12. Project Management

12.1 For the purposes of supervising the carrying out of the Project including taking all major decisions on the direction of the Project, the Parties will establish a management committee (“Project Management Committee”).

12.2 The Project Management Committee will include 3 representatives (including the Project Director and one co-investigator) from each Sub-Project.

12.3 The Project Management Committee will have such composition, specific functions, duties and procedural rules (not in conflict with the Main Contracts) as the Parties may from time to time agree.

12.4 A Project Management Committee meeting will be chaired by the Project Director who has convened it.

12.5 Project Management Committee meetings may be held by teleconference, internet conferencing or other telecommunication means.

12.6 Each Project Director's functions include:

12.6.1 co-ordinating on a day-to-day basis the progress of work under the Sub-Project;

12.6.2 to act as a liaison between the Sponsor and Project Management Committee concerning the Project and Sub-Project;

12.6.3 following up decisions made by the Project Management Committee insofar as they affect the Sub-Project; and

12.6.4 convene and chair Project Management Committee meetings on the request of its Sub-Project collaborators or at reasonably convenient times and places as may be necessary or desirable.

The UK Project Management Team is:

- Professor Tim Hitchcock (P-I, Univ. of Hertfordshire, t.hitchcock@herts.ac.uk)
- Michael Pidd (project manager, m.pidd@sheffield.ac.uk).
- Jamie McLaughlin (HRI Technical Developer, j.mclaughlin@sheffield.ac.uk)

The Project Team

- Michael Pidd, UK Project Manager (5% of time spent on this project), HRI Digital Manager, Humanities Research Institute, University of Sheffield, (m.pidd@sheffield.ac.uk)
- Tim Hitchcock, UK Project Director, Professor of Eighteenth- Century History, University of Hertfordshire, (t.hitchcock@herts.ac.uk)
- Robert Shoemaker, UK Co-Director, Professor of Eighteenth- century History, University of Sheffield, (r.shoemaker@sheffield.ac.uk)
- Geoffrey Rockwell, Canadian Project Director, Professor of Philosophy and Humanities Computing, University of Alberta, (geoffrey.rockwell@ualberta.ca)
- Jörg Sander, Canadian Co-Investigator, Associate Professor, University of Alberta, (joerg@cs.ualberta.ca)
- Stéfan Sinclair, Canadian Co-Investigator, Associate Professor of Multimedia at McMaster University, (sgs@mcmaster.ca)
- Daniel J. Cohen, US Project Director, Director of the Center for History and New Media, George Mason University, (dan@dancohen.org)
- Sean Takats, US Co-Investigator, Acting Director of Research Projects, Center for History and New Media, George Mason University, (sean@takats.org)
- William J. Turkel, Canadian Co-Investigator, Associate Professor of History, The University of Western Ontario, (william.j.turkel@gmail.com)

Training Requirements

Training requirements to support technical staff at the HRI form a central part of its ongoing staff development programme.

13. Programme Support

13.1 Help in relation to legal services and advice; publicity and dissemination will all be sought from JISC.

14. Budget

See Appendix A. No changes have been made.

Detailed Project Planning

15. Workpackages

See Appendix B.

16. Evaluation Plan

Timing	Factor to Evaluate	Questions to Address	Method(s)	Measure of Success
Sept 2010	Functionality of Newgate Commons API	Does it allow the base text to be queried effectively?	The application of a series of distinct historical questions to the resource.	The ability to collect a body of research evidence which can be readily replicated..
Dec. 2010	Ability to develop robust research outcomes.	Does the application of datamining and visualisation tools give us new answers	Historical analysis, against current literature.	The ability demonstrate the existence of unforeseen patterns in OB text.
March 2011	Impact on historical community.	Will the wider historical community accept the methodologies developed as valid.	Submission to peer review.	Publication of article based on this resource in a peer reviewed journal.

17. Quality Plan

Output Timing	Newgate Commons API				
	Quality criteria	QA method(s)	Evidence of compliance	Quality responsibilities	Quality tools (if applicable)
March - September 2010	Functionality	Test against Old Bailey dataset on Server A using a mock web page on Server B.		HRI	
		Formal evaluation		All project	

				partners	
--	--	--	--	----------	--

Output					
Modified Zotero Virtual Collections					
Output Timing	Quality criteria	QA method(s)	Evidence of compliance	Quality responsibilities	Quality tools (if applicable)
March - September 2010	Functionality	Test against locally-held datasets in the first instance, then text against remotely-held Old Bailey datasets via the Newgate Commons API.		Centre for History and the New Media	
		Formal evaluation		All project partners	

Output					
Voyuer Analytics					
Output Timing	Quality criteria	QA method(s)	Evidence of compliance	Quality responsibilities	Quality tools (if applicable)
March - September 2010	Functionality	Test against locally-held datasets in the first instance, then text against remotely-held Old Bailey datasets via the Newgate Commons API.		The TAPoR Project	
		Formal evaluation		All project partners	

18. Dissemination Plan

18.1 Dissemination will need to target a number of different communities, in particular: 1) Historians of Crime in HE and more widely 2) the HE research community (academics, postgraduates, undergraduates), 3) the Digital Humanities community and resource creators.

18.2 To reach these groups, a multifaceted publicity programme will be implemented, relying in the first instance on the large user groups who already rely on the Old Bailey, Zotero and the TAPoR Portal. News about the project will be carried on the Old Bailey, Centre for History and the New Media, Connected Histories, and TAPoR Portal websites. Academic historians and scholars in related disciplines will be reached via professional newsletters and electronic notice boards such as H-Net; and discussion boards directed at developers.

18.3 At the conclusion of the project at least one article will be written for refereed history journal: by Hitchcock, Turkel and Shoemaker.

18.4 A wider audience will be reached via participation in JISC programme-level publicity; and equivalent programme activity in the US and Canada.

Timing	Dissemination Activity	Audience	Purpose	Key Message
Spring 2010 to Spring 2011	Posting news	Webusers and developers	To generate interest in the methodology and approach.	This is a new way of dealing with large bodies of text.
Spring 2011	Refereed Journal Article	Historians of Crime, wider HE humanities Community	Demonstrate the use of datamining to answer traditional humanist/history questions.	New tools can be effectively used within a humanist framework.

19. Exit and Sustainability Plans

Project Outputs (UK Only)	Action for Take-up & Embedding	Action for Exit
The <i>Newgate Commons</i>	This will be directly linked to <i>The Old Bailey Online</i> ; <i>London Lives</i> ; and <i>Connected Histories</i> websites - giving it purchase for a wider user community through association.	The API will form a facet of the the <i>Old Bailey</i> website. The protocol will be fully documented and made available for use by other web service developers.
Peer Reviewed Articles	These will be written for major peer reviewed journals in social and legal history, and designed to address historical issues directly - rather than issues of technical implementation.	Publishing in high level academic journals across the range of historical sub-disciplines.

Project Outputs (UK Only)	Why Sustainable	Scenarios for Taking Forward	Issues to Address
The <i>Newgate Commons</i> .	The <i>Commons</i> will form a second access point to the <i>Old Bailey Proceedings</i> , that have become an important part of the	If the example of the <i>Newgate Commons</i> demonstrates a need for API access to other large text corpora, the <i>Commons</i> could be expanded to	The precise relationship between the API and the current Old Bailey website. The extent to which the Newgate Commons is

	online historical environment, and which has a demonstrated and well tried strategy for sustainability into the foreseeable future.	include the <i>London Lives</i> material in the first instance, and other webresources created by the HRI, including the <i>Connected Histories</i> site in due course.	extensible, and how much background material for historical data needs to be developed for each new resource incorporated.

Appendixes

Appendix A. Project Budget

Appendix B. Workpackages

Appendix C. UK/JISC Roles and Responsibilities.

APPENDIX B

Project Acronym: DMwCI
Version: Final Draft
Contact: Tim Hitchcock
Date: 14 Feb. 2010



JISC WORK PACKAGE

WORKPACKAGES	Month	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1: Legal Framework		█	█													
2: Technical Standards			█	█												
3: Newgate Commons API				█	█	█	█	█	█	█						
4: Inter-operability						█	█	█	█	█						
5: Historical Research										█	█	█	█			
6: Dissemination													█	█	█	█
7: Sustainability and Consolidation										█	█	█	█	█	█	█

APPENDIX B

Project start date: 1 January 2010

Project completion date: 31 March 2011

Duration: Fifteen months

				Milestone	Responsibility
YEAR 1					
WORKPACKAGE 1: Legal Framework					
Objective: A comprehensive legal and management framework.					
1. Create a working legal framework for collaboration incorporating the Universities of Sheffield and Hertfordshire.	1 Jan. 2010	21 Feb. 2010	Collaboration Agreement between UK partners	Signed by participating HEIs	TH, RS
2. Create a clear management structure and series of defined milestones	1 Jan. 2010	21 Feb. 2010	Project Plan Submitted to JISC	Submitted to JISC	TH, MP
3. <i>Ensure that a working pattern of online project conferences has been established and is rolling forward</i>	1 Jan. 2010	21 Feb. 2010	Schedule of meetings, with collaboration methodology agreed.	Schedule of meetings is posted on the project blog.	TH, DC, GR, WT
WORKPACKAGE 2: Technical Standards					
Objective: Articulate shared technical standards					
4. Technical Standards agreed between UK, Canadian and US partners.	1 Feb. 2010	31 March 2010	Working sub-group of project directors and Programmers.	A clear list of technical specifications posted on the project blog.	TH, MP, DC, JM
WORKPACKAGE 3: Newgate Commons API					
			The purpose of the Newgate Commons API is to enable data mining and		

APPENDIX B

Objective: Development of the Newgate Commons API			visualisation tools such as Zotero and TAPor to query the Old bailey datasets.		
5. 1st, 2nd and 3rd Cycle - Specifications	1 March 2010	30 April 2010	The HRI will establish user requirements in consultation with the other project partners and model these as a formal technical specification. Requirements and specification will be modified at the beginning of each new cycle.	Written specification available	MP, JM, DC, GR, SS, ST, TH
6. 1st, 2nd and 3rd Cycle - Technical Development	1 March 2010	30 May 2010	Technical development will take place at the HRI. Development is iterative through items 13 and 14 (Internal Testing). Development iterations respond to internal testing (item 14) rather than changes in technical specification (item 12). As such, each cycle will involve more than one development iteration.	Prototype API finished	JM, MP
7. 1st, 2nd and 3rd Cycle - Internal Testing	20 May 2010	15 June 2010	The prototype is formally tested internally at the HRI. We will be testing each unit in the protocol's data flow sequence, the integration of each unit and finally the entire system. System testing will be conducted by building a simple web page on HRI Server A which sends search requests to the API on HRI Server B.	Prototype API successfully processing requests	JM, MP, TH, RS
8. 1st, 2nd and 3rd Cycle - Deployment for External Testing	1 June 2010	30 Sept 2010	Each cycle culminates with live testing in which the API is made available to other project partners. Project partners will be testing the API using Zotero and TAPoR tools in live environments. Each test phase will result in the need for technical modifications or adjustments in methodology. All partners will need to establish whether these modifications relate to the API or the Zotero and TAPoR Tools. Modifications to the API will trigger another cycle of specification	Prototype API successfully receiving and responding to requests from Zotero and TAPor	JM, MP, TH, DC, ST, WT, RS, SS

APPENDIX B

			adjustment and development. The final deliverable will be an API which satisfies the requirements of the partners.		
WORKPACKAGE 4: Interoperability					
Objective: Ensuring that Zotero and TAPoR Tools work with the Newgate Commons					
9. Technical development.	1 May 2010	30 June 2010	Technical Development will take place at George Mason University and the University of Alberta	Prototype Zotero Collections and Voyuer Toolset working.	DC, GR, WT, ST, SS
10. Testing the interoperability and functionality of the system and how its individual components work together.	20 May 2010	15 June 2010	The prototypes are formally tested internally	Prototypes successfully responding to specific research queries	DC, GR, WT, ST, SS
11. Deployment for Collaborative Testing.	1 June 2010	30 Sept. 2010	The three components of the system need to be tested as a working system, and modified as needed.	The receipt and passing of research queries normalised.	DC, GR, WT, ST, SS
WORKPACKAGE 5: Historical Research					
Objective: To develop a body of research data for historical analysis					
11. Exploring the OB and other large text sets using datamining and visualisation tools.	1 Sept. 2010	31 Dec. 2010	A series of basic models and queries, using each of the tools create, with re-iterated runs, against different samples of text.	Query results posted on the project blog.	TH, WT, RS, ST

APPENDIX B

YEAR 2					
WORKPACKAGE 6: Dissemination					
Objective: Promulgate the resource and approach.					
12. Writing an exploratory academic article on the history of crime based on datamining techniques.	1 January 2011	31 March 2011	Drafting, editing and revising a substantial academic article for submission to a peer-reviewed journal.	Article submitted	TH,WT,RS
13. Demonstrating the methodology to academic conferences.	1 January 2011	31 March 2011	Project demonstrations at conferences such as MLA, AHA, BSECS	conferences addressed	DC,GF,TH,RS, ST,SS, TW
WORKPACKAGE 7: Sustainability and Consolidation					
Objective: Consolidate the resource for sustainability and reuse					
30. Ensure all code is posted in an Open source format available for re-use	1 August 2010	31 March 2011	Record all technical decisions as they are made and confirmed.	Project webpage, and API site is populated with descriptions.	MJ, MP, TH

Members of Project Team: TH = Tim Hitchcock; RB = Robert Shoemaker; MP = Michael Pidd; JM = Jamie McLaughlin; DC = Dan Cohen; GR = Geoffrey Rockwell; WT= William Turkel; SS= Stefan Sinclair; ST = Sean Takats.

APPENDIX C

Data Digging into Data Initiative: Using Zotero and TAPoR in the Old Bailey Proceedings

Roles and Responsibilities of the Collaborators

1. Definitions

"Project" means the project entitled 'Digging into Data Initiative: Using Zotero and TAPoR in the Old Bailey Proceedings' as set out in the Project Plan.

"Lead Institution" means the University of Hertfordshire (UK); George Mason University (US) and University of Alberta (Canada)

"Project Director" means a representative of the Lead Institution with overall responsibility for a Sub-Project.

"Sub-Project" means the programme of work to be carried out by each Lead Institution and its collaborators.

"Main Contract" means the contract between each Lead Institution, its Sub-Project collaborator(s) and the relevant sponsor.

"Background Intellectual Property" means any intellectual property owned or controlled by a Party but not in connection with the Project.

"Foreground Intellectual Property" means individually and collectively all intellectual property which is developed by one or more Parties in the performance of the Project.

2. Collaboration on Project

2.1 Each Party agrees to collaborate with the others in the carrying out of the Project in accordance with the Project Plan and Main Contract governing its Sub-Project.

2.2 Each Lead Institution will be responsible for the completion of its own Sub-Project.

2.3 Each Party will immediately notify its Project Director in writing if it becomes aware of an unexpected or scientific problem which makes it impossible to achieve or is likely to cause a material delay to the achievement of any of the objectives of the Project. On receipt of such notification, each Project Director will immediately notify the other Project Directors.

2.4 Any changes to the Project Plan will be agreed between the Project Management Committee, subject to approval by the relevant sponsor.

3. Payments

Each Lead Institution will reimburse its Sub-Project collaborators for the actual expenditure they incur (up to any maximum budget) and in accordance with the terms as set out in its Main Contract. Payment arrangements will be agreed between members of each Sub-Project.

4. Project Management

4.1 For the purpose of supervising the carrying out of the Project including taking all major decisions on the direction of the Project, the Parties will establish a management committee ("Project Management Committee").

4.2 The Project Management Committee will include 3 representatives (including the Project Director and one co-investigator) from each Sub-Project.

APPENDIX C

4.3 The Project Management Committee will have such composition, specific functions, duties and procedural rules (not in conflict with the Main Contracts) as the Parties may from time to time agree.

4.4 A Project Management Committee meeting will be chaired by the Project Director who has convened it.

4.5 Project Management Committee meetings may be held by teleconference, internet conferencing or other telecommunication means.

4.6 Each Project Director's functions include:

4.6.1 co-ordinating on a day-to-day basis the progress of work under the Sub-Project;

4.6.2 to act as a liaison between the sponsor and Project Management Committee concerning the Project and Sub-Project;

4.6.3 following up decisions made by the Project Management Committee insofar as they affect the Sub-Project; and

4.6.4 convene and chair Project Management Committee meetings on the request of its Sub-Project collaborators at reasonably convenient times and places as may be necessary or desirable.

5. Confidentiality

5.1 In the event of any Party (the "Disclosing Party") making available to another Party (the "Receiving Party") confidential information relating to its business, scientific or other activities in the course of the Project, the Receiving Party shall maintain the confidentiality of such information, and shall not disclose it to third parties, members of its staff or students outside the research team working on the Project or include it in the results of the Project without the prior written permission of the Disclosing Party. If either Party uses the services of sub-contractors, consultants, agents or students to undertake part of the Project, advise on the Project or manage the Project, that Party shall promptly and diligently ensure that such sub-contractors, consultants, agents or students sign a written undertaking agreeing to abide by the same conditions of confidentiality as are set out herein.

5.2 The obligations in section 5.1 above will not apply to any such information which the Receiving Party is required to disclose by law or information which becomes available to the public generally, otherwise than through a breach of a duty of confidentiality.

6. Intellectual Property

6.1 For the avoidance of doubt all Background used in connection with the Project shall remain the property of the Party introducing the same. No licence to use any intellectual property is granted or implied, except the rights expressly granted in section 6.2.

APPENDIX C

6.2 Each Party shall grant to the other Parties an irrevocable royalty free licence to use such of its Background Intellectual Property, and any Foreground Intellectual Property belonging to it, as may be necessary for the performance of this Project.

6.3 All rights to Foreground Intellectual Property created solely by a Party in the performance of this Project shall belong to that Party.

6.4 All rights to Foreground Intellectual Property created jointly by the Parties in the performance of this Project shall be owned jointly by the Parties.

6.5 Each Party grants to the other Parties an irrevocable royalty free non-exclusive licence to use its Foreground Intellectual Property (including any that is jointly owned) for each Party's own academic, teaching and research purposes.

6.6 Should the grant of a licence of a Party's Background Intellectual Property be necessary in order for another Party to exploit any Foreground Intellectual Property belonging to it, then the owner of the Background Intellectual Property may on request consent to grant a non-exclusive licence, such consent not to be unreasonably withheld or delayed, and on agreement of a reasonable royalty rate.

6.7 Upon the commercial exploitation of any Foreground Intellectual Property, each Party agrees to pay the other relevant Party or Parties a fair and reasonable royalty rate on the value of any products or processes commercially exploited by it which incorporates any Foreground Intellectual Property owned partly or solely by another Party or Parties, taking into consideration the respective financial, intellectual and technical contributions of the relevant Party or Parties to the development of the Foreground Intellectual Property, the expenses incurred in securing intellectual property protection thereof and the costs of its commercial exploitation and the proportionate value of the Foreground Intellectual Property in any such product or process.

7. Publicity and Publication

7.1 No Party will use the name, logo, or trade mark of any of the other Parties, its employees or affiliates in any publicity, advertising or news release without the prior written approval of that Party.

7.2 It is the intention of the Parties that the results may be published in accordance with normal academic practice. Before such publication, each Party will be notified with a copy of any proposed publication. Should a Party believe that publication should be delayed in order to enable any intellectual property rights arising from the results to be registered then it shall notify the other Parties within thirty (30) days of the date of the notification and the publishing Party will refrain from publication in order to enable such rights to be registered. Such registration will be undertaken expeditiously and the registering Party will notify the other Parties when registration has been filed, and in any event the publishing Party may publish such publication after 60 days from the date of the notification.

7.3 No Party may include confidential Background Intellectual Property belonging to any of the Parties in any publication without prior written consent.

APPENDIX C

7.4 Nothing shall prevent a student from submitting for a degree a thesis based on the results obtained during the Project, the examination of such a thesis by examiners appointed by a Party, or the deposit of such a thesis in a library of a Party in accordance with the relevant procedures of the Party. A student's supervisor shall be notified in writing within twenty-one (21) days of any access restrictions or other special requirements required by a Party following disclosure of a manuscript for dissertation or thesis to that Party under section 7.2 above.

8. Termination

8.1 Any Party may withdraw from the Project by giving the other parties 3 months' written notice, subject to the consent of the relevant sponsor.

8.2 The Project Management Committee may decide to terminate a Party's involvement within the Project if the Party shall commit any material breach of any of its obligations under its Main Contract or this Project Plan and shall fail to remedy such breach (if capable of remedy) within 30 days after being given notice in writing by a member of the Project Management Committee to do so. This is subject to the consent of the relevant sponsor.

8.3 The remaining Project Management Committee will make all reasonable attempts to reallocate the obligations of the breaching or withdrawing Party amongst the remaining Parties or to a third party acceptable to the remaining Parties and the relevant sponsor(s).

9. Dispute Resolution

9.1 If any dispute arises between the Parties under or in connection with this Project either Party may serve notice upon the other setting out brief details of the dispute that has arisen ("Notice of Dispute") and the Parties shall use all reasonable endeavours to resolve the dispute by good faith negotiations.

9.2 If the dispute is not resolved within 2 months from the date of the Notice of Dispute, then the matter shall be referred to the Project Management Committee to assist in resolving the dispute.

9.3 Any dispute arising out of or in connection with this Project which cannot be resolved within three (3) months from the date of the Notice of Dispute, shall be finally resolved under the Rules of Arbitration of the International Chamber of Commerce.

10. Liability

10.1 In respect of any information or materials (including Foreground Intellectual Property, Background Intellectual Property and/or deliverables) supplied by a Party belonging to one Sub-Project to a Party belonging to another Sub-Project, no warranty or representation of any kind is made, given or implied as to the sufficiency or fitness for purpose nor as to the absence of any infringement of any proprietary rights of third parties. It is therefore agreed that any Party using such materials is fully responsible and liable for any loss, costs, claims or demands arising from that use.

APPENDIX C

10.2 Each Party is solely liable for any loss, damage or injury to third Parties (including its sponsor) and any third party claims resulting from the performance or non-performance of its obligations under this Project or the Main Contract.

10.3 A Party's aggregate liability towards the other Parties collectively in connection with any activities undertaken pursuant to or for any purpose related to this Project (whether arising in contract, tort, negligence, breach of statutory duty or otherwise) shall not exceed the Party's share of the total costs of the Project.

10.4 Section 10.3 shall not apply in the case of death or personal injury caused by negligence, fraudulent misrepresentation or in other circumstances where liability may not be so limited under any applicable law.

11. General

This collaboration is not intended to create a partnership or joint venture or legal relationship of any kind that would impose liability upon one Party for the act or failure to act of any other Party, or to authorise any Party to act as agent for any other.