Data Mining for Historians

Professor Bob Shoemaker
University of Sheffield
12 May 2010

"instead of looking for a needle in the haystack, an effective text mining tool will try and show you the shape of the haystack and tell you the words you might want to find"

—Geoffrey Rockwell

Home Page
Search
About the Proceedings
Historical Background
The Project
Copyright & Citation Guide
Contact

ON THIS DAY IN... 1736

William Rine allegedly pulled a man off of his horse, robbed him, stripped him naked, and tied him to a tree. **read more**

The Proceedings of the Old Bailey, 1674-1913

A fully searchable edition of the largest body of texts detailing the lives of non-elite people ever published, containing 197,745 criminal trials held at London's central criminal court.

To search the Proceedings use the boxes on the right or go to the Search Pages.

What's New?

Trials between November 1834 and April 1913 and the **Ordinary of Newgate's Accounts** between 1690 and 1772 have now been added. See What's New.

Ordinary's Accounts

The website now includes the texts of all Ordinary of Newgate's Accounts published between 1690 and 1772. These richly detailed narratives of the lives and deaths of convicts executed at Tyburn have been linked to the relevant

SEARCH
the Proceedings

Keyword(s)

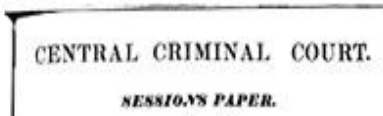
Reference No.

Search In

<All Text>

SEARCH

More Search Options


CENTRAL CRIMINAL COURT.
SESSIONS PAPER.

In this Section...

Search Home

The boxes below allow you to search the whole of the **Proceedings** and all published **Ordinary Accounts** (for the period 1679 to 1772). You may combine keyword searches with queries on ta information including **surname**, **crime**, and **punishment**. The default setting allows you to sea full text of all the documents available on this website. This page should be used for basic and searches. Please refer to the other pages listed to your left for more search options.

Keyword(s)

Surname

Given Name

Alias

Offence <All Offences>

Verdict <All Verdicts>

Punishment <All Punishments>

Search In <All Text>

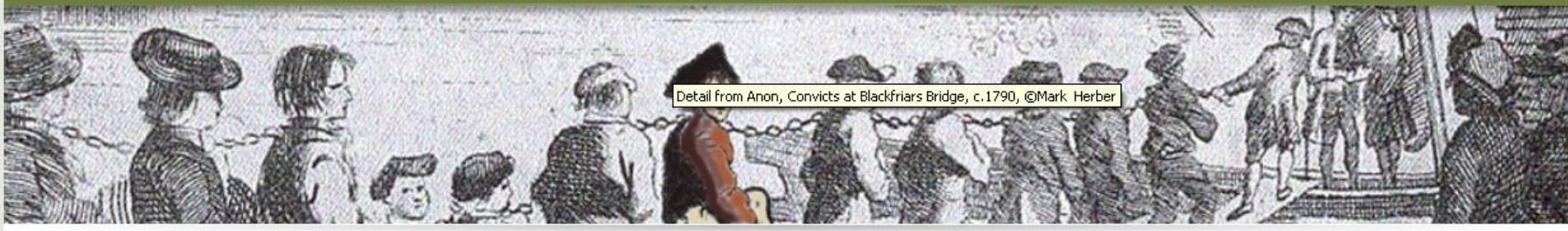
Time Period *From (month/year)* Any Any *To (month/year)* Any Any

Reference Number

SEARCH

www.londonlives.org

London Lives 1690 to 1800 ~ Crime, Poverty and Social Policy in the Metropolis



Home
Search
Login / Register
Background
Lives
Project
Legal
Wiki
Contact

Sources for London Lives

A fully searchable edition of 240,000 manuscripts from eight archives and twelve datasets, giving access to over 3.4 million names.

To search London Lives use the boxes on the right or go to the Search Pages.

Help us Construct London Lives

You can use this site to create biographies of eighteenth-century Londoners. The person search page allows you to assemble a set of documents related to a single person, and the project wiki? allows you to write and share their biography. You will need to register and login in order to make use of these features. See the Featured Life on this page, and the Lives section.

London Life in the Eighteenth Century

What was it like to live in the world's first million person city? Crime, poverty, and illness; apprenticeship, work, politics and money; how people voted, lived and died; all this and more can be found in these documents. For more

SEARCH

Keyword

Surname | Given Name

Reference ID

From 1680 | To 1820

SEARCH

More Search Options

Featured Life

Sophie Pringle, Executed 1787

Browse by Document Type

London Lives includes 39 individual **Document Types** from fourteen archives, and 15 **Additional Datasets**. By clicking on an individual **Archive**, you will be taken to a page which lists all the document types from that archive. By then clicking on a decade for which those documents are available, you will be taking to a listing of that document type year by year.

Parish Archives

- St Botolph Aldgate
- St Clement Danes
- St Dionis Backchurch

Criminal Records

- Bridewell Royal Hospital
- Home Office
- Old Bailey Sessions
- Old Bailey Proceedings
- Ordinary's Accounts
- City of London Sessions
- Middlesex Sessions
- Westminster Sessions

Coroners' Records

- City of London Coroners
- Middlesex Coroners
- City of Westminster Coroners

Hospital and Guild Records

- Carpenters' Company
- St Thomas's Hospital

	surname	given	role	occupation	place	age	div	date	certainty	
568 3170	Walker	Charlotte				0	OB_PSpage	Jan 1 1784	2	search
568 3171	Walker	Charlotte				0	WJ_PSpage	Oct 1 1794	2	search
568 3172	Walker	Charlotte				0	SL_PSpage	Dec 13 1779	2	search
568 3173	Walker	Charlotte				0	SL_PSpage	Mar 13 1784	2	search
568 3174	Walker	Charlotte				0	SM_PSpage	Feb 1 1778	2	search
568 3175	Walker	Charlotte				0	OB_PSpage	Jan 1 1788	2	search
568 3176	Walker	Charlotte				0	SM_PSpage	Jan 1 1781	2	search
568 3177	Walker	Charlotte				0	HO_CRpage	Sep 1 1793	2	search
568 3178	Walker	Charlotte				0	OB_PSpage	May 12 1785	2	search
568 3179	Walker	Charlotte				0	SL_PSpage	Mar 13 1784	2	search
568 3180	Walker	Charlotte				0	SM_PSpage	Apr 1 1789	2	search
568 3181	Walker	Charlotte				0	SM_PSpage	Oct 1 1796	2	search
568 3182	Walker	Charlotte				0	OB_PSpage	Aug 13 1787	2	search
568 3184	Walker	Charlotte				0	SM_PSpage	Jan 1 1795	2	search

Additional Datasets

In addition to the sources fully digitised for the **London Lives** project, this website also includes a range of databases created for other projects. Each of these **Additional Datasets**, created over a 30 year period, was designed with a specific project in mind. As explained in our webpages, these databases are in different formats, and use different conventions and forms of coding. The copyright in this material remains with the original projects.

Many of the databases consist of **parish** and **taxation** records, but there are also databases of wills, apprenticed boys, fire insurance policies, clerk's salaries, an urban directory, and the Old Bailey associated records.

Parish Records

- St Botolph Aldgate Parish Registers, 1681-1709 (PAR)
- St Martin's in the Fields Settlement Exams, 1725-1793 (SET)
- St Martin's in the Fields Workhouse Register, 1725-1819 (SWR)
- St Luke's Chelsea Workhouse Registers, 1743-69 and 1782-99 (WHR)

Records from Taxation

- Poll Tax Assessments, St Botolph Aldgate, 1690-98 (POL)
- Four Shillings in the Pound Aid, 1693/94 (FSP)
- Marriage Duty Assessments for St Botolph Aldgate, 1695 (MDA)
- Westminster Historical Database, 1749-1820 (WHD) (including voting records)
- Income Tax Payments, 1799-1802 (ITP)

Edward Collier, trompe l'oeil, c.1699. Wikimedia Commons. The original of this artwork is held by the Victoria and Albert Museum.

Other Sources

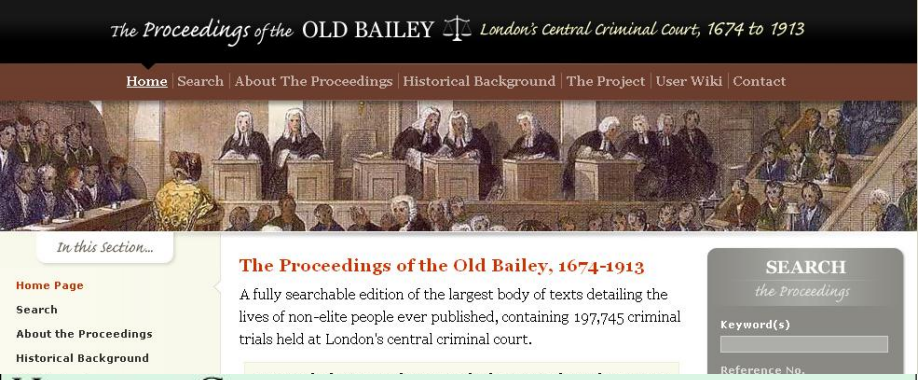
- Prerogative Court of Canterbury Wills (PCC)
- Boys Recruited into the Marine Society, 1770-1804 (MAR)
- Fire Insurance Policy Registers, 1777- 1786 (FIR)
- British East India Company: Salaries Paid to 'Clerks', 1760-1820 (EIC)
- The London and Westminster Directory, 1774 (DIR)
- Old Bailey Associated Records (ASS)

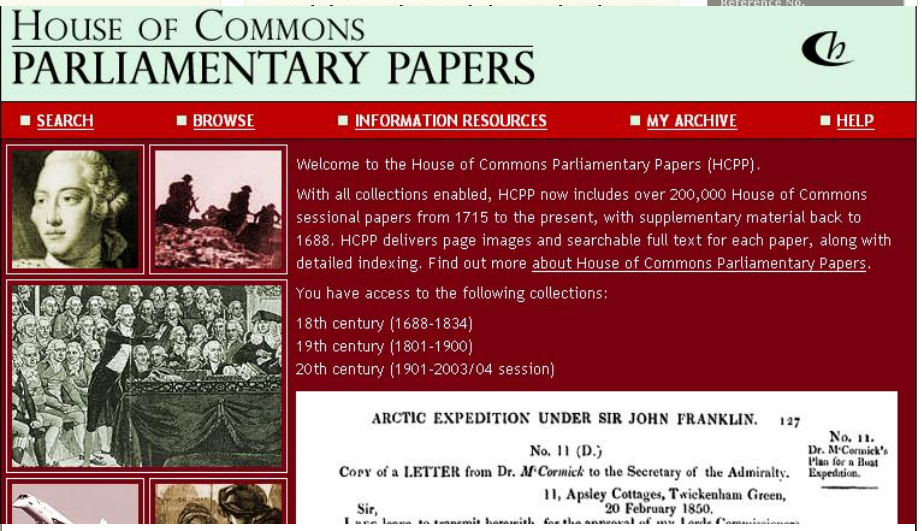
Connected Histories: Sources for Building British History, 1500-1900

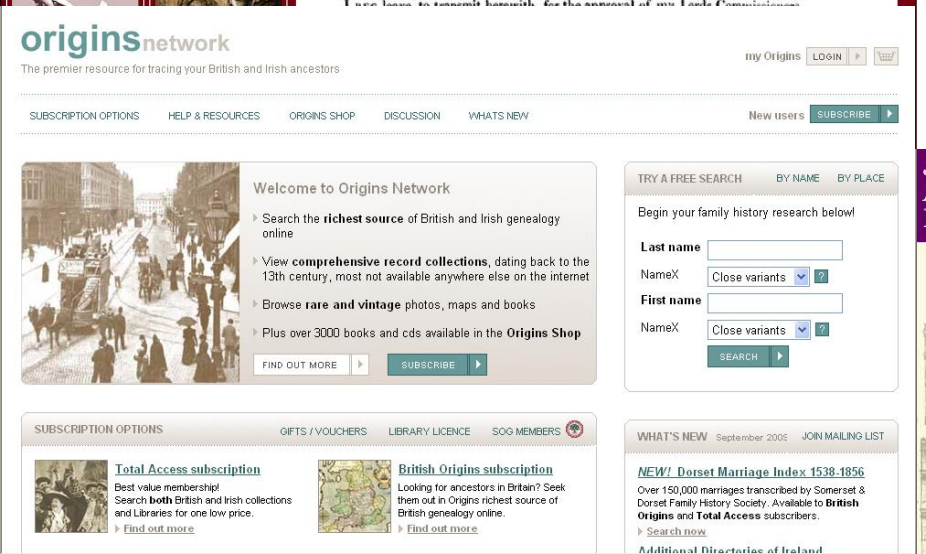
- Funded by JISC (UK) e-Content Capital Programme
- Partnership between:
  - University of Sheffield (Humanities Research Institute)
  - University of Hertfordshire
  - Institute of Historical Research, London

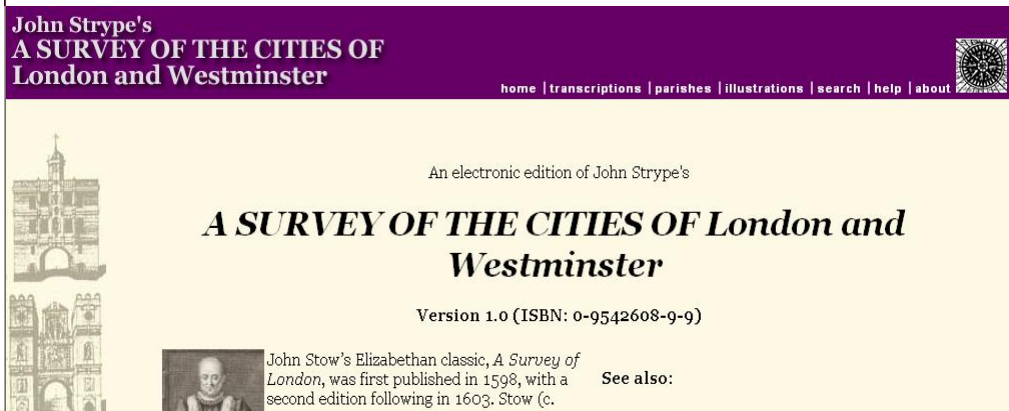
Connected Histories Sources

- **Old Bailey Online and London Lives**
- **British History Online**
- **Burney Newspaper Collection**
- **Origins Network (genealogical database)**
- **Parliamentary Papers**
- **Clergy of the Church of England Database**
- **Charles Booth Online Archive**
- **Collage**
- **John Johnson Collection of Printed Ephemera (provisional)**
- **EEBO and ECCO (provisional)**
- ***Others subject to negotiation (more can be easily added)***







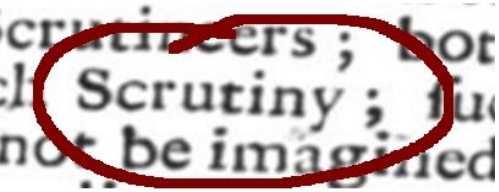


Connected Histories Methods

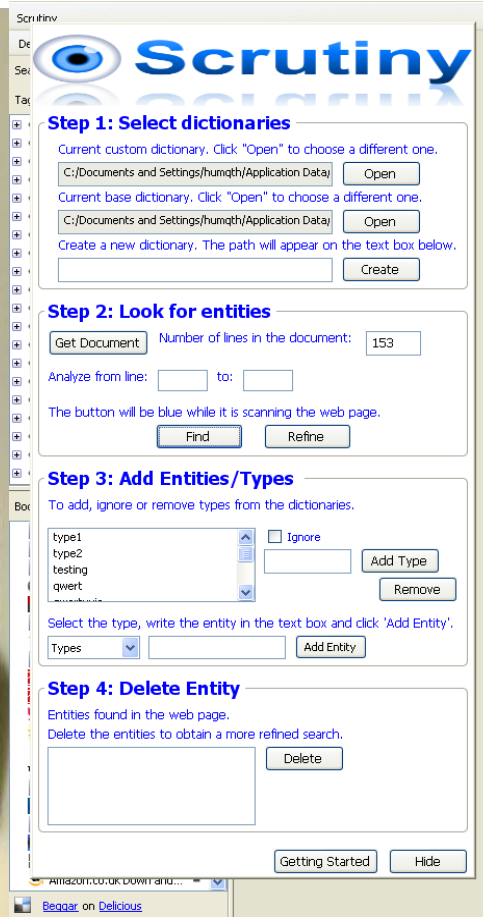
- **A federated search facility, sitting as umbrella over the distributed sites**
- **Search by keyword, name, place, date**
- **Use existing metadata and tagged information where available**
- **Where information not tagged, data will be obtained by web crawling and natural language processing will be used to identify key entities**
- **Data saved as RDF files held centrally**

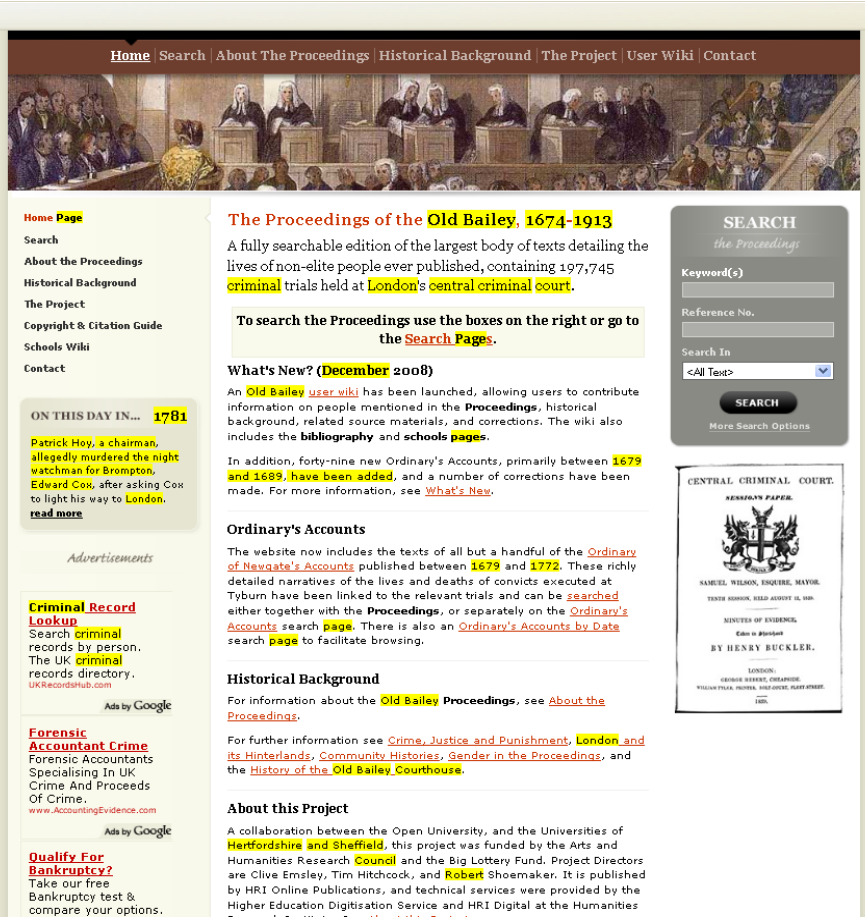
Connected Histories Outcomes

- **Search engine on Connected Histories website (www.connectedhistories.org)**
- **Search results point users to original sites (with login for commercial sites)**
- **Background information**
- **User engagement**
- **Website launch early 2011**



Scrutiny: A Firefox Extension for Entity Recognition within Research Data





Available from: https://addons.mozilla.org/en-US/firefox/addon/75085

Data Mining with Criminal Intent

- Project Partners
  - Tim Hitchcock (University of Hertfordshire)
  - Robert Shoemaker (University of Sheffield)
  - Michael Pidd (Humanities Research Institute, University of Sheffield)
  - Daniel Cohen (Center for History and New Media, George Mason University)
  - Geoffrey Rockwood (University of Alberta)
  - William Turkel (University of Western Ontario)

Data Mining with Criminal Intent

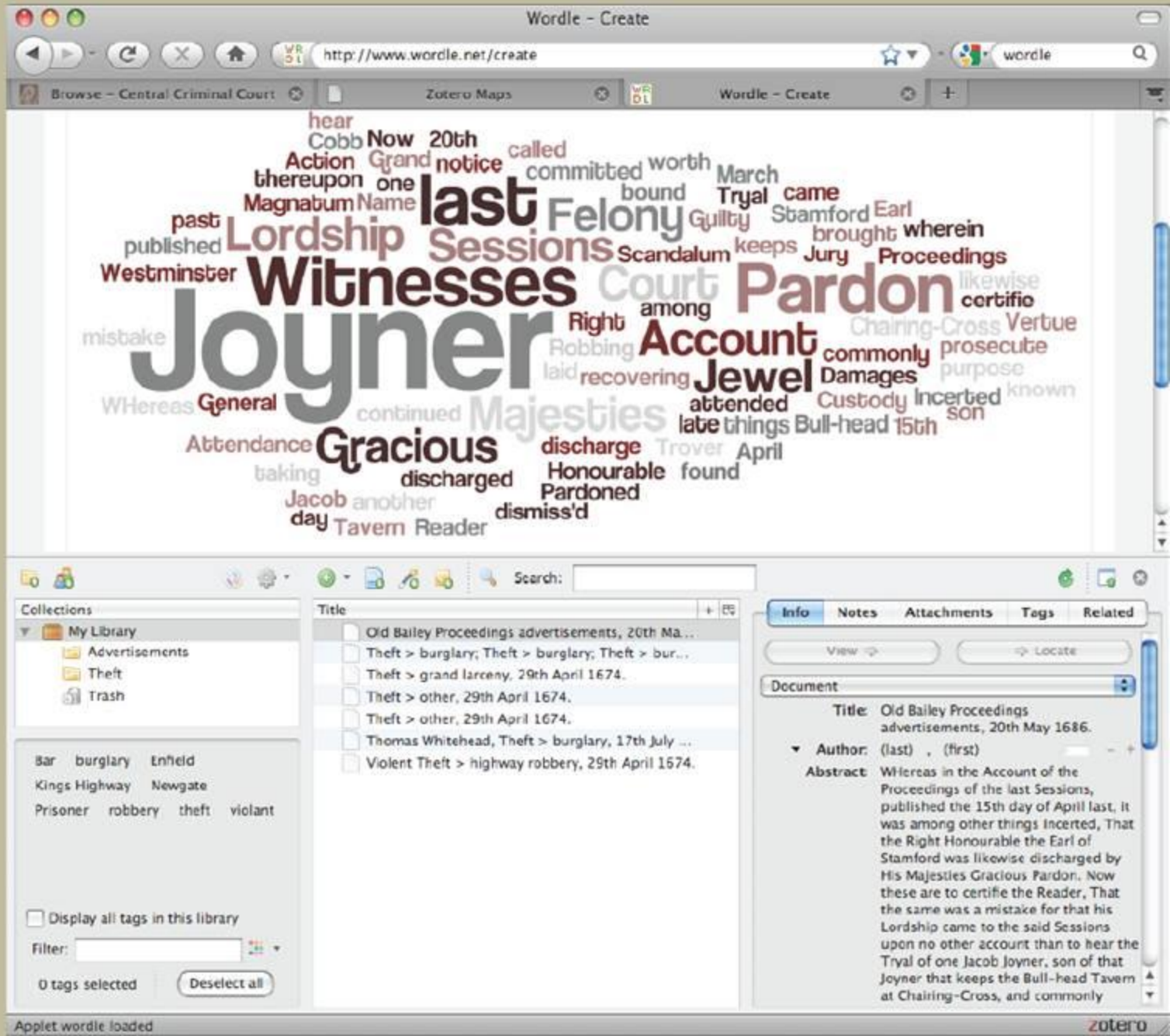
- Tools:

  --Zotero (Firefox Plug-in)

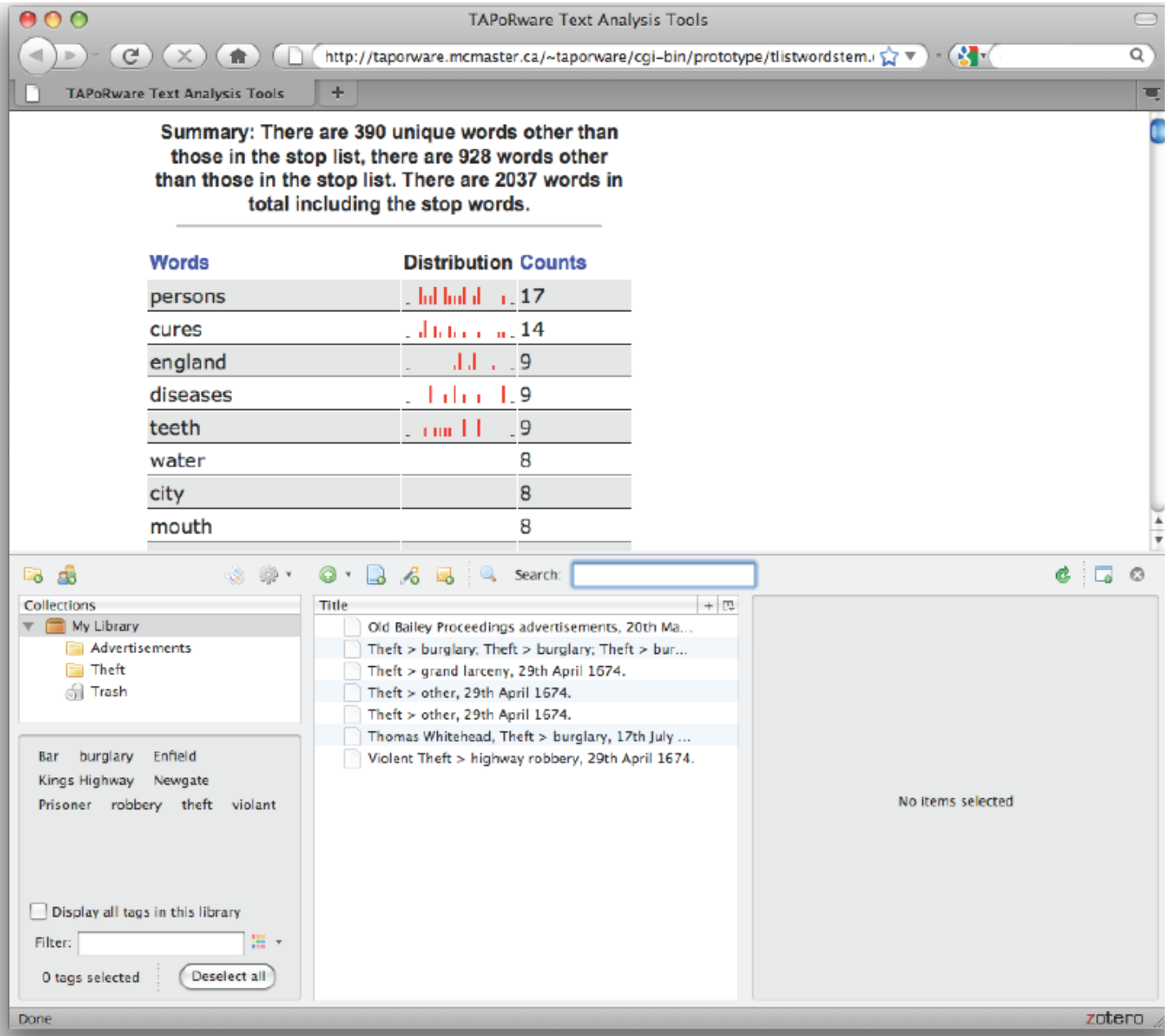
  -- TAPoR toolkit (including Voyeur)

  -- Compression Analysis

Word Cloud Visualization of Old Bailey text via Zotero



Invoking the TAPoR Frequency Tool with the Old Bailey via Zotero



Voyeur Tools: Reveal Your Texts

Words in the Entire Corpus

Word	Count	Trend
said	20,303	
prisoner	17,443	
mr	10,032	
went	8,499	
saw	7,961	
street	6,543	
came	5,887	
told	5,785	
know	5,769	
road	5,606	
man	5,275	
asked	5,195	
money	4,766	
got	4,647	

Page 1 of 710 | Search

Summary | Corpus

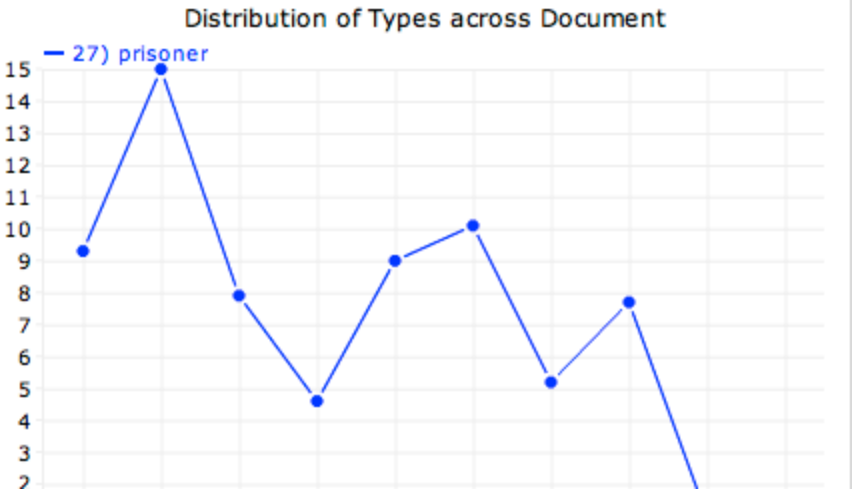
- There are 41 documents in this corpus with a total of **2,535,564 words** and **35,475 unique words**.
- **Longest documents** (by words): T19101011CLR_SUP_DONE (111,217), T19120227NW_SUP_DONE (103,304). Shortest documents: T19120722GS_SUP_DONE (22,965), T19110523SKP_SUP_DONE (30,150). All...
- Highest **vocabulary density** ( ): T19120722GS_SUP_DONE (127.1), T19110523SKP_SUP_DONE (116.6). Lowest density: T19120227NW_SUP_DONE (62.8), T19111205NW_SUP_DONE (63.0). All...
- Most **frequent words** in the corpus: said (20,303), prisoner (17,443), mr (10,032), went (8,499), saw (7,961). More...
- Words with **notable peaks in frequency** across the corpus: will ( ), day ( ), they ( ), wanted ( ), november ( ). More...
- **Distinctive words** (compared to the rest of the corpus)
  1. T19110717SKP_SUP_DONE: muller (74), brooks (66), mackenzie (54), coltman (52), stevens (53). More...
  2. T19120319eg_SUP_DONE: ullman (61), gallafent (57), stevens

Words within each Document

Documer	Count	Difference
Type: prisoner		
27) T191	712	-7.64
3) T1912	642	-7.02
4) T1911	614	-6.05
40) T191	592	-6.80
19) T191	564	-6.89
9) T1912	554	-7.47
26) T191	547	-7.92
33) T191	545	-8.62
5) T1912	535	-6.46
21) T191	529	-8.69
7) T1912	504	-7.46
32) T191	500	-7.53

Page 1 of 2

Word Trends



Keywords in Context | Collocates

Left	Keyword	Right
Document: T19101011CLR_SUP_DONE.xml		
Guardian" Newspaper Company.191010110006	Prisoner	had been twice convicted. He
that system.The Recorder sentenced	prisoner	to 20 months' hard labour, warr
Frederick Cook and Co. Each	prisoner	confessed to a previous convic
.d., with intent to defraud	Prisoner	bore an excellent character, an
,goods of N. Thierry, Limited	Prisoner	confessed to a previous convic
previous convictions were proved against	prisoner	(not for coining offences) dating
previous convictions were proved against	prisoner	(not for coining offences
had obtained employment.Sentences: Each	prisoner	, 15 months' hard labour; recon

Voyeur Tools: Reveal Your Texts

Words in the Entire Corpus

Word	Count	Trend
election	18	
selections	2	
electioneering	2	

Page 1 of 1 | Search election

Summary

Corpus

- There are 41 documents in this corpus with a total of **2,535,564 words** and **35,475 unique words**.
- **Longest documents** (by words): T19101011CLR_SUP_DONE (111,217), T19120227NW_SUP_DONE (103,304). Shortest documents: T19120722GS_SUP_DONE (22,965), T19110523SKP_SUP_DONE (30,150). All...
- Highest **vocabulary density** ( ): T19120722GS_SUP_DONE (127.1), T19110523SKP_SUP_DONE (116.6). Lowest density: T19120227NW_SUP_DONE (62.8), T19111205NW_SUP_DONE (63.0). All...
- Most **frequent words** in the corpus: said (20,303), prisoner (17,443), mr (10,032), went (8,499), saw (7,961). More...
- Words with **notable peaks in frequency** across the corpus: will ( ), day ( ), they ( ), wanted ( ), november ( ). More...
- **Distinctive words** (compared to the rest of the corpus)
  1. T19110717SKP_SUP_DONE: muller (74), brooks (66), mackenzie (54), coltman (52), stevens (53). More...
  2. T19120319eg_SUP_DONE: ullman (61), gallafent (57), stevens (62), pearson's (52), hayes (53). More...
  3. T19120109NW_SUP_DONE: mirini (39), martin (57), rothwell (35), follett (30), clifford (31). More...
  4. T19110905GS_SUP_DONE: august (156), fletcher (57), deceased (95), cabin (43), searle (37). More...
  5. T19120611SKP_SUP_DONE: woods (100), stone (62), hole (53), wallet (44), titheradge (33). More...

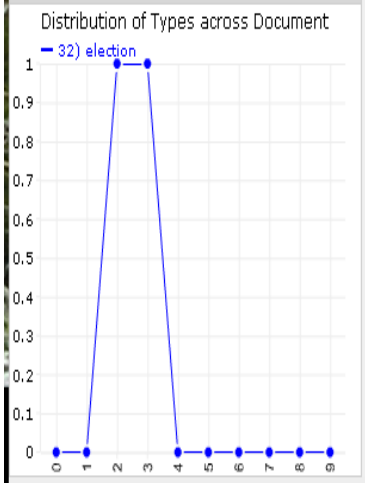
  Next 5 of 36 remaining

Words within each Document

Documer	Count	Difference	Relative	Trend
Type: election				
37) T191	7	-0.02	1.62	
40) T191	5	-0.05	0.74	
32) T191	2	-0.08	0.30	
35) T191	1	-0.09	0.15	
21) T191	1	-0.07	0.10	
8) T1911	1	-0.09	0.13	
3) T1912	1	-0.08	0.12	
41) T191	0	-0.10	0.00	
39) T191	0	-0.11	0.00	
Type: electioneering				
40) T191	1	-0.07	0.15	
29) T191	1	-0.08	0.16	
41) T191	0	-0.09	0.00	
39) T191	0	-0.10	0.00	
38) T191	0	-0.09	0.00	

Page 1 of 3

Word Trends



Keywords in Context | Collocates

Collocate	Raw Frequency	Ratio
course	1 (8)	6.2
liberal	1 (4)	6.2
circumstances	1 (2)	6.2
despatched	1 (1)	6.2
candidate	1 (1)	6.2
barker	1 (1)	6.2
mr	2 (313)	12.0
george	2 (88)	12.4
lloyd	2 (25)	12.5

Page 1 of 1 | Context | Displaying 1 - 13 of 13

Contact Details

- Bob Shoemaker (r.shoemaker@shef.ac.uk)
- Tim Hitchcock (t.hitchcock@herts.ac.uk)
- Michael Pidd (HRI) (m.pidd@shef.ac.uk)

Websites

www.oldbaileyonline.org

www.londonlives.org

www.connectedhistories.org

http://criminalintent.org/