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

The Proceedings of the OLD BAILEY A London's Central Criminal Court, 1674 to 1913

Home Search About The Proceedings Historical Background The Project Contact



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 <u>What's New</u>.

Ordinary's Accounts

The website now includes the texts of all <u>Ordinary of Newgate's Accounts</u> 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
Key	word(s)
Refe	erence No.
1	rch In Text> 🗸
	SEARCH
	More Search Options
CE	NTRAL CRIMINAL COURT

SESSIONS PAPER.

Search Home

Personal Details

Ordinary's Accounts, 1676-1772

Proceedings by date

Ordinary's Accounts by date

Statistics

Custom Search

Associated Records, 1674-1834

Place and Map Search, 1674-1834

Security Specialists

Get A home security system In Grimsby. www.abatis-fs.co.uk

Home & Family Protection



Ads by Google

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)		Ŷ
Surname		Ŷ
Given Name		Ŷ
Alias		{
Offence	<all offences=""></all>	- 🍫
Verdict	<all verdicts=""></all>	
Punishment	<all punishments=""> 🛛 🛛 🍫</all>	
Search In	<all text=""></all>	× 🍫
Time Period	From (month/year) To (month/year) Any Any Any Any Any Any Any	
Reference Number		?



www.londonlives.org

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

Home Search Browse Lives Historical Background Wiki The Project



Home

Search
Login / Register
Background
Lives
Project
Legal
Wiki
Contact

Advertisements

[Google code goes here]

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 <u>person search page</u> allows you to assemble a set of documents related to a single person, and the <u>project wiki?</u> 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 <u>Lives section</u>.

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

SEA	ARCH
Keyword	
Surname	Given Name
Reference ID	
From	То
1680 💌	1820 👱
SEA More Sear	RCH rch Options

Featured Life

Sonhie 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

- <u>St Botolph Aldgate</u>
- <u>St Clement Danes</u>
- <u>St Dionis Backchurch</u>

Criminal Records

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

Coroners' Records

- <u>City of London Coroners</u>
- Middlesex Coroners
- <u>City of Westminster Coroners</u>

Hospital and Guild Records

- <u>Carpenters' Company</u>
- <u>St Thomas's Hospital</u>

	<u>surname</u>	<u>given</u>	<u>role</u>	occupation	<u>place</u>	<u>age</u>	<u>div</u>	<u>date</u>	<u>certainty</u>	
568 🗖 3170	Walker	Charlotte				0	<u>OB PSpage</u>	Jan 1 1784	2	<u>search</u>
568 🗖 3171	Walker	Charlotte				0	<u>WJ PSpage</u>	Oct 1 1794	2	<u>search</u>
568 🗖 3172	Walker	Charlotte				0	<u>SL PSpage</u>	Dec 13 1779	2	<u>search</u>
568 🗖 3173	Walker	Charlotte				0	<u>SL PSpage</u>	Mar 13 1784	2	<u>search</u>
568 🗖 3174	Walker	Charlotte				0	<u>SM PSpage</u>	Feb 1 1778	2	<u>search</u>
568 🗖 3175	Walker	Charlotte				0	<u>OB PSpage</u>	Jan 1 1788	2	<u>search</u>
568 🗖 3176	Walker	Charlotte				0	<u>SM PSpage</u>	Jan 1 1781	2	<u>search</u>
568 🗖 3177	Walker	Charlotte				0	<u>HO CRpage</u>	Sep 1 1793	2	<u>search</u>
568 🗖 3178	Walker	Charlotte				0	<u>OB PSpage</u>	May 12 1785	2	<u>search</u>
568 🗖 3179	Walker	Charlotte				0	<u>SL PSpage</u>	Mar 13 1784	2	<u>search</u>
568 🗖 3180	Walker	Charlotte				0	<u>SM PSpage</u>	Apr 1 1789	2	<u>search</u>
568 🗖 3181	Walker	Charlotte				0	<u>SM PSpage</u>	Oct 1 1796	2	<u>search</u>
568 🗖 3182	Walker	Charlotte				0	<u>OB PSpage</u>	Aug 13 1787	2	<u>search</u>
568 🗖 3184	Walker	Charlotte				0	<u>SM_PSpage</u>	Jan 1 1795	2	<u>search</u>

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 <u>copyright</u> 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

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

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)

Other Sources

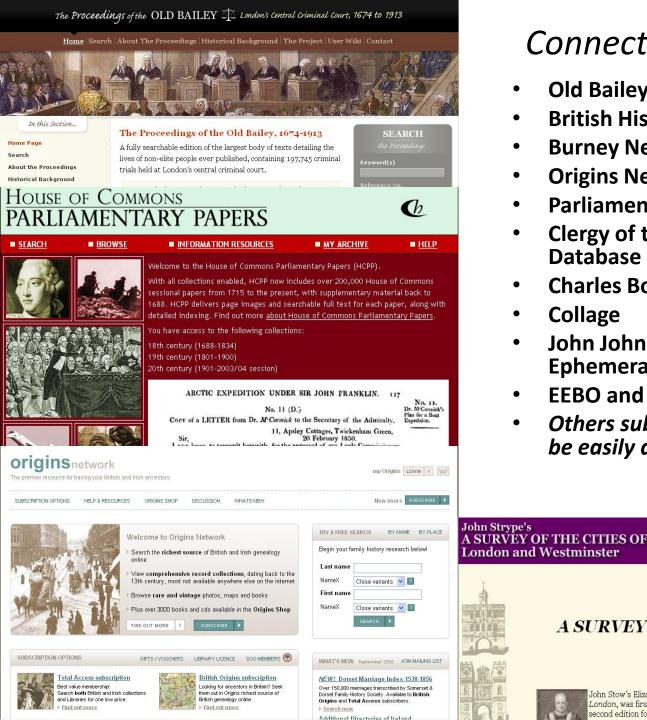
- Prerogative Court of Canterbury Wills (PCC)
- Boys Recruited into the Marine Society, 1770-1804 (MAR)
- Fire Insurance Policy Registers, 1777- 1786 (FIR)
- British East India Company: Salaries Paid to 'Clerks', 1760-1820 (EIC)
- The London and Westminster Directory, 1774 (DIR)
- Old Bailey Associated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



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

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)

A SURVEY OF THE CITIES OF London and Westminster

An electronic edition of John Strype's

home | transcriptions | parishes | illustrations | search | help | abo

Version 1.0 (ISBN: 0-9542608-9-9)



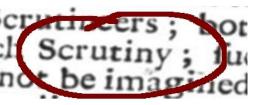
John Stow's Elizabethan classic, A Survey of London, was first published in 1598, with a See also: second edition following in 1603. Stow (c.

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

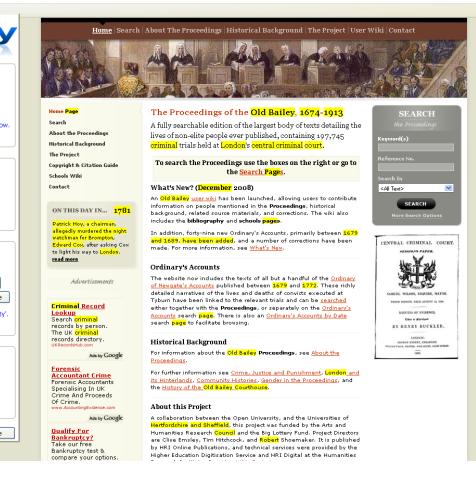
Connected Histories Outcomes

- Search engine on Connected Histories website (<u>www.connectedhistories.org</u>)
- 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

	Scruhiny	
	Scrutin	
Schliber		Ξ
	Step 1: Select dictionaries	
A REAL PROPERTY OF THE RE	Current custom dictionary. Click "Open" to choose a different o	ne.
	C:/Documents and Settings/humqth/Application Data, Open	
	• Current base dictionary. Click "Open" to choose a different one	
	C:/Documents and Settings/humqth/Application Data Open	
	Create a new dictionary. The path will appear on the text box	, bela
	Create	1
A Chillen Act III	• • • • • • • • • • • • • • • • • • • •	<u> </u>
	Step 2: Look for entities	
Children and Children and Children	Get Document Number of lines in the document:]
	Analyze from line: to: to	
	 The button will be blue while it is scanning the web page. 	
	Find Refine	
	Step 3: Add Entities/Types	
	Box To add, ignore or remove types from the dictionaries.	
	type1 Ignore	
Telleville	type2 📃 🗌 Add Typ	e
	testing	_
	Rem	.ove
a.05 8	Select the type, write the entity in the text box and click 'Add E	Intit
	Types V Add Entity	
no IFR	Step 4: Delete Entity	
urt hpt		
A CHL	Entities found in the web page.	
Der wit	Delete the entities to obtain a more refined search.	
SCHLENT	Delete	
Alle		
1901		
		_
		lide
	S Amazon.co.uk Down and = Y	
	Beggar on Delicious	



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

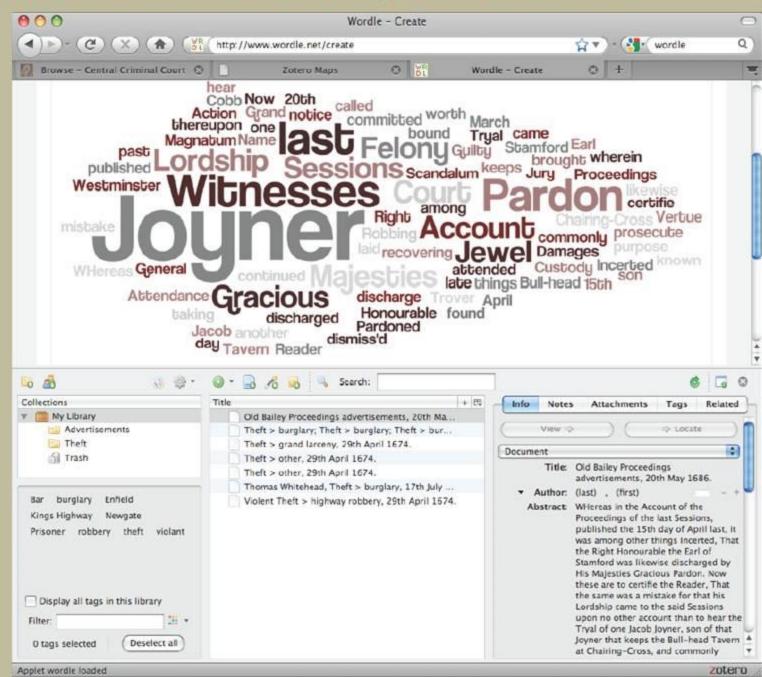
Data Mining with Criminal Intent

- 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

000	TAPoRware Text Analysis Tools		
	http://taporware.mcmaster.ca/~taporware/cgi-bin/prototype/tlistwordstem.i 😭 💎 🔹 🚷		۹)
TAPoRware Text Analysis Tools	+		=
those in the st than those in th	e are 390 unique words other than top list, there are 928 words other e stop list. There are 2037 words in ncluding the stop words.		C
Words	Distribution Counts		
persons	. hil hild 17		
cures	. <u>dhhu</u> 14		
england			
diseases	. 1.1		
teeth	<u> </u>		
water	8		
city	8		
mouth	8		×
Collections	Image: Organization of the second	6 🗔	0
My Library Advertisements Theft Trash Bar burglary Enfield Kings Highway Newgate Prisoner robbery theft violant Display all tags in this library Filter:	Old Bailey Proceedings advertisements, 20th Ma Theft > burglary; Theft > burglary; Theft > bur Theft > grand larceny, 29th April 1674. Theft > other, 29th April 1674. Theft > other, 29th April 1674. Thomas Whitehead, Theft > burglary, 17th July Violent Theft > highway robbery, 29th April 1674. No items selected		
0 tags selected Deselect all Done		zote	, סחו

	Words in the Entire	e Corpus	▲ @ ⊟ ?
	Word	Count 👻	Trend
	said	20,303	rowm 1
V	<mark>prisoner</mark>	17,443	mm
	mr	10,032	mm
	went	8,499	mont
	saw	7,961	mmm
	street	6,543	mount
	came	5,887	mound
	told	5,785	mmm
	know	5,769	mmm
	road	5,606	moun
	man	5,275	mm
	asked	5,195	mm
	money	4,766	mmm
	got	4,647	Mr.M.
	Page 1 of 710	Q Search -	× P >>

Hord Trends

Voyeur Tools: Reveal Your Texts

±

±

±

±

±

▲ @ ⊟ ?					• = ?		Words w	ithin each	Document				
	Summary	Corpus					Documer	Count -	Difference				
mm n						∃ Type: prisoner							
www		e 41 documen and 35,475 ur	ts in this corpus wi nique words.	th a total of 2,53	5,564		27) T191	712	-7.64				
mm			(by words ~~~	MMm):			3) T1912	642	-7.02				
mh U	T191010	11CLR_SUP_D	ONE (111,217), T cuments: T191207	19120227NW_SUP			4) T1911	614	-6.05				
mm			KP_SUP_DONE (30				40) T191	592	-6.80				
mm			ensity (~~~~				19) T191	564	-6.89				
mm			ONE (127.1), T1911 y: T19120227NW				9) T1912	554	-7.47				
m			ONE (63.0). All	_ 、			26) T191	547	-7.92				
mm			in the corpus: said		er 🛛		33) T191	545	-8.62				
han		mr (10,032), went (8,499), saw (7,961). More th notable peaks in frequency across the corpus: will (5) T1912	535	-6.46							
Num			(mm				21) T191	529	-8.69				
mm		(mhmm), november (handhh).		7) T1912	504	-7.46							
mm	More	in a second of the second of t	annear the the sec	t of the course)			32) T191	500	-7.53				
vmm 🛔			ompared to the res IP_DONE: muller (7			C	+++			i			
× P »			ltman (52), stever DONE: ullman (6		stevens	4	Page 1	of 2 🕨	1				
	· @ = ?												
Document		Keywo	rds in Context	Collocates									
				Left	H	Keywo	ord	Right					
		Documen	t: T19101011CLR	_SUP_DONE.xm	I								
		🗄 Guardian'	Newspaper Comp	pany.19101011000	06 <mark>P</mark>	Prison	ier	had been t	wice convicted. He	,			
		±	that system. The F	Recorder sentence	ed <mark>p</mark>	prison	er	to 20 mont	hs' hard labour, wa	Irr			
			Frederick	Cook and Co. Ead	ch <mark>r</mark>	prisoner confessed to a previous			to a previous conv	ic			

.d., with intent to defraud

,goods of N. Thierry, Limited

previous convictions were proved against

previous convictions were proved against

had obtained employment.Sentences: Each

Prisoner

Prisoner

prisoner

prisoner

prisoner

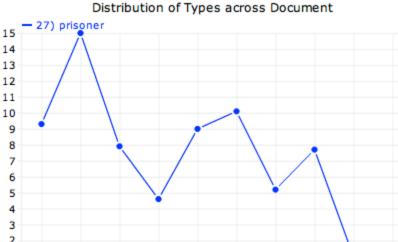
bore an excellent character, an

confessed to a previous convic

(not for coining offences) dating

, 15 months' hard labour; recon

(not for coining offences



				Voyeur Tools	s: Reveal Your Texts					¥ 8 ?
Words in the En	ire Corpus	▲ @ 🗄 ?					Words within each	Document		• • • • •
Word	Count 👻	Trend	Summary	🛄 Corpus			Documer Count -	Difference	Relative	Trend
election	18						Type: election			-
selections	2				atal of 2,535,564 words and 35,475 unique words . ///): T19101011CLR_SUP_DONE (111,217),	-	37) T191 7	-0.02	1.62	L
electioneering	2		T191202	27NW_SUP_DONE (103,304). Shortest	t documents: T19120722G5_SUP_DONE (22,965),		40) T191 5	-0.05	0.74	N
				235KP_SUP_DONE (30,150). All cocabulary density (>): T19120722G5_SUP_DONE (127.1),		32) T191 2	-0.08	0.30	Λ
			T191105	235KP_SUP_DONE (116.6). Lowest dei 35NW_SUP_DONE (63.0). All	nsity: T19120227NW_SUP_DONE (62.8),		35) T191 1	-0.09	0.15	Λ
					303), prisoner (17,443), mr (10,032), went (8,499), saw (7,961).		21) T191 1	-0.07	0.10	Λ
			More				8) T1911 1	-0.09	0.13	Λ
			Words w	th notable peaks in frequency acr	ross the corpus: <mark>will</mark> (~~~~~~, <mark>day</mark> (~~~~~, ^, ^, ^, ^, ^, ^, ^, ^, ^, ^, ^, ^, ^,		3) T1912 1	-0.08	0.12	1
			<u> </u>	ive words (compared to the rest of the res	··· ·		41) T191 O	-0.10	0.00	
			1. T191	107175KP_SUP_DONE: muller (74), br	ooks (66), mackenzie (54), <mark>coltman</mark> (52), stevens (53). More lafent (57), stevens (62), pearson's (52), hayes (53). More		39) T191 O	-0.11	0.00	_
			3. T191	20109NW_SUP_DONE: <mark>mirini</mark> (39), <mark>mar</mark>	rtin (57), rothwell (35), <mark>follett</mark> (30), <mark>clifford</mark> (31). More		Type: electioneering			
			5. T191	206115KP_SUP_DONE: woods (100), :	<mark>letcher</mark> (57), <mark>deceased</mark> (95), <mark>cabin</mark> (43), <mark>searle</mark> (37). More stone (62), hole (53), wallet (44), titheradge (33). More		40) T191 1	-0.07	0.15	٨
			Next	5 of 36 remaining			29) T191 1	-0.08	0.15	
							41) T191 0	-0.09	0.00	<u></u>
							39) T191 0	-0.10	0.00	-
							38) T191 0	-0.09	0.00	-
						<	00)1101 0		0.00	
4 Page 1 of 1 ▶	🔍 Search 🔻 elec	tion × P »				4	Page 1 of 3 🕨	، ا		
🖋 Word Trends		▲ @ 🗄 ?								• • = ?
Distribution of	Types across Do	cument 🛛 🕅 Key	words in Context	Collocates						
1 - 32) election		Collocat	e	_11 1	Raw Frequency	R	atio 🔻			
0.9		course			1 (8)	6	2			8
0.8		liberal			1 (4)	6	.2			
0.7		circums	tances		1 (2)	6	2			
0.6		despate	hed		1 (1)	6	2			
0.5		candida	te		1 (1)	6	.2			
0.4		barker			1 (1)	6	2			
0.3		mr			2 (313)		2.0			
0.2		george			2 (88)		2.4			
0.1		lloyd			2 (25)	1	2.5			
	0 4 4 6 6 F	4 4	Page 1 of 1	🕨 🕅 🛛 🧬 🛛 Context 🕶				Displa	aying 1 - 13 of 1	3 💛 🍕

Contact Details

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

Websites

www.oldbaileyonline.org

www.londonlives.org

www.connectedhistories.org

http://criminalintent.org/