Corpus Protocols:

digital transformations of commercial newspaper collections for text and data mining to support academic research

IFLA 2014 Pre-Conference: Digital Transformation and the Changing Role of News Media in the 21st Century

August 13-14, 2014

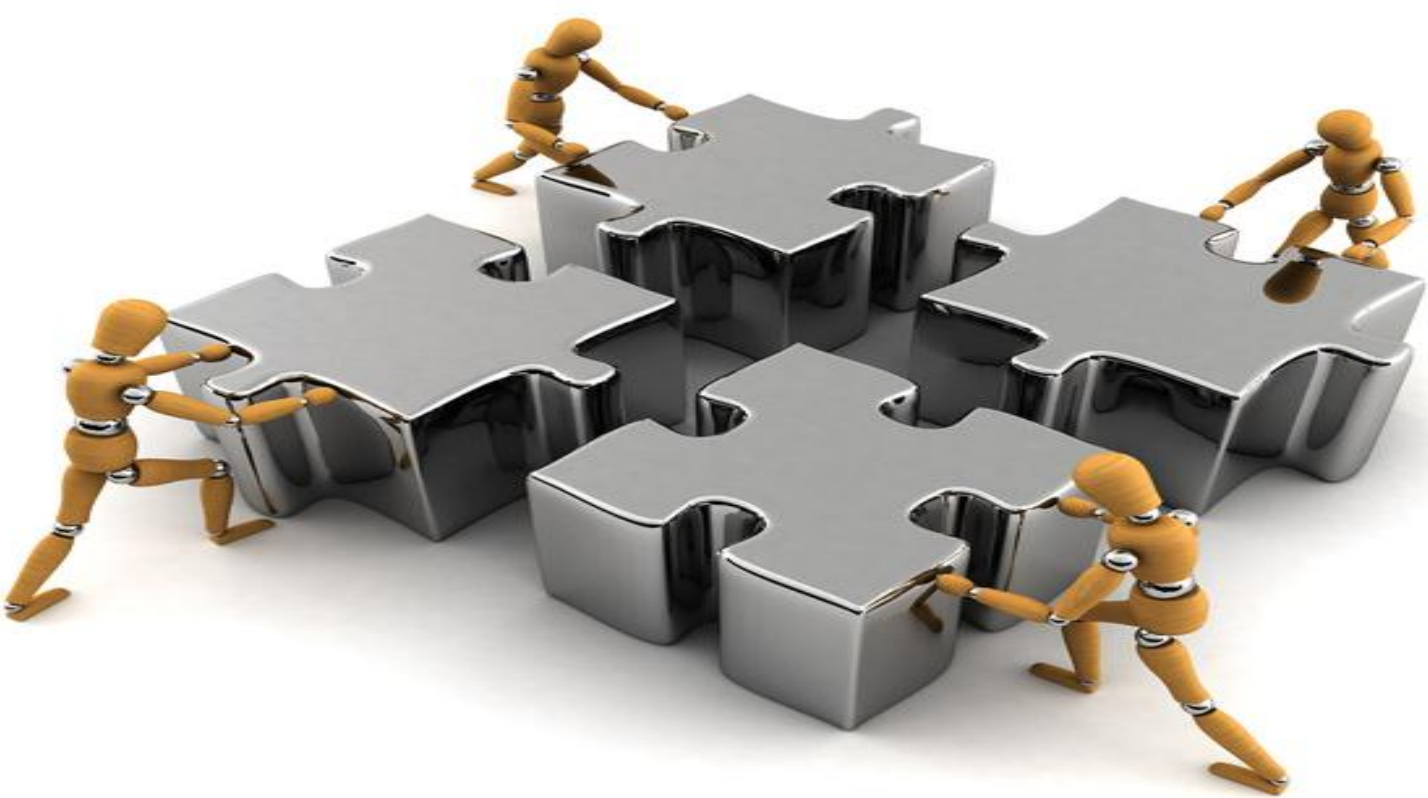
International Telecommunication Union, Geneva, Switzerland

Arts & Humanities Research Council

horizon
DIGITAL ECONOMY RESEARCH

The University of Nottingham
UNITED KINGDOM · CHINA · MALAYSIA

Working together: building research relationships



Corpus Protocols Project Team

- Seth Cayley: Head of Research Solutions, Cengage Learning EMEA
- Mike Gardner: Analyst and Web Developer in Web Technologies, University of Nottingham
- Kat Gupta: Corpus Protocols Researcher, University of Nottingham
- Michaela Mahlberg: Professor of English Language and Linguistics, University of Nottingham
- Neil Smyth: Faculty Team Leader – Arts, University of Nottingham
- Stella Wisdom: Digital Curator at the British Library

Context

- Focusing on textual data (corpus) we review relationships between institutional infra-structure, external business partners and research questions and methods that deal with textual data
- Exploring opportunities to future developments (protocols)

Problem (practical and research)

- Some University of Nottingham owned data sets are not available for research for several reasons, including:
  - licence agreements restrict copying for data analytics;
  - additional charges for additional licenses;
  - technical issues: copying storage, accessibility of data;
  - and, a lack of shared understanding of research trends in applied linguistics, corpus linguistics and corpus stylistics.

Aims

- To articulate the relationship between data and tools for text and data mining
- Prototype a methodology for using locally stored data using existing corpus linguistic tools
- Use an existing corpus linguist tool to analyse a University owned data set that will be made available on the University network.

Corpus Protocols: didn’t start with discourse research questions

- Declassified Documents as opportunistic data

- The Declassified Documents corpus
  - 197,837,328 words in total
  - divided into 30 yearly subcorpora (between 3.3 to 8.3 million words each)

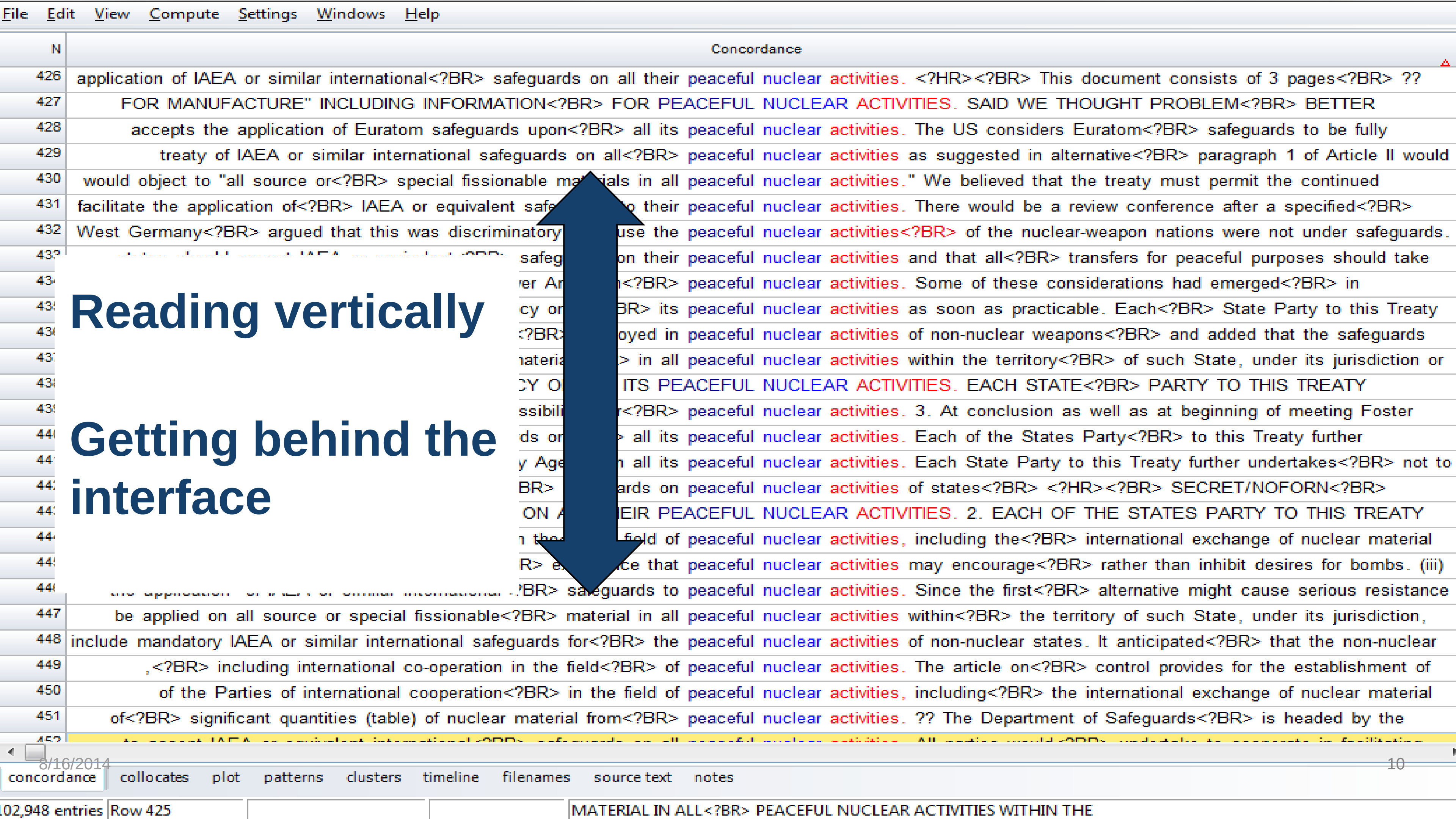
UK Copyright legislation: 2014

"the making of a copy of a work by a person who has lawful access to the work does not infringe copyright in the work provided that (a) the copy is made in order that a person who has lawful access to the work may carry out a computational analysis of anything recorded in the work for the sole purpose of research for a non-commercial purpose, and (b) the copy is accompanied by a sufficient acknowledgement."

(The Copyright and Rights in Performances (Research, Education, Libraries and Archives) Regulations 2014.

Corpus methods:

- Discourse as representation: labeling, naming, creating reality (e.g. *weapons of mass destruction).*
- Lexis in context

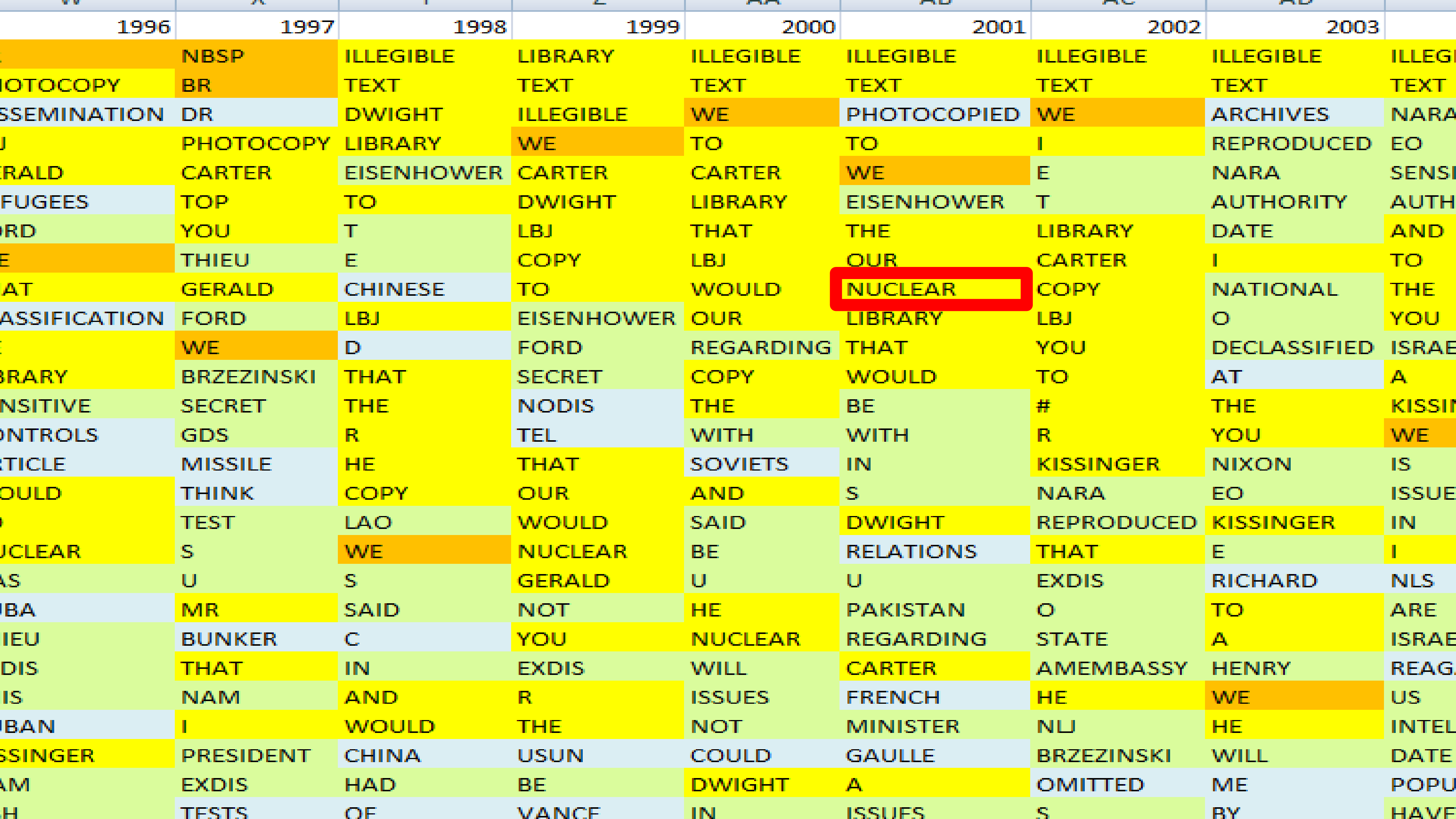


Reading vertically

Getting behind the interface

```
N    Concordance
426  application of IAEA or similar international<?BR> safeguards on all their peaceful nuclear activities. <?HR><?BR> This document consists of 3 pages<?BR> ??
427  FOR MANUFACTURE" INCLUDING INFORMATION<?BR> FOR PEACEFUL NUCLEAR ACTIVITIES. SAID WE THOUGHT PROBLEM<?BR> BETTER
428  accepts the application of Euratom safeguards upon<?BR> all its peaceful nuclear activities. The US considers Euratom<?BR> safeguards to be fully
429  treaty of IAEA or similar international safeguards on all<?BR> peaceful nuclear activities as suggested in alternative<?BR> paragraph 1 of Article II would
430  would object to "all source or<?BR> special fissionable ma    ials in all peaceful nuclear activities." We believed that the treaty must permit the continued
431  facilitate the application of<?BR> IAEA or equivalent safe    o their peaceful nuclear activities. There would be a review conference after a specified<?BR>
432  West Germany<?BR> argued that this was discriminatory    use the peaceful nuclear activities<?BR> of the nuclear-weapon nations were not under safeguards.
433  states should accept IAEA or equivalent<?BR> safeg    on their peaceful nuclear activities and that all<?BR> transfers for peaceful purposes should take
43   er Ar    n<?BR> peaceful nuclear activities. Some of these considerations had emerged<?BR> in
43   cy or    BR> its peaceful nuclear activities as soon as practicable. Each<?BR> State Party to this Treaty
43   <?BR>    oyed in peaceful nuclear activities of non-nuclear weapons<?BR> and added that the safeguards
43   materia    > in all peaceful nuclear activities within the territory<?BR> of such State, under its jurisdiction or
43   CY O    ITS PEACEFUL NUCLEAR ACTIVITIES. EACH STATE<?BR> PARTY TO THIS TREATY
43   ssibili    r<?BR> peaceful nuclear activities. 3. At conclusion as well as at beginning of meeting Foster
44   ds or    > all its peaceful nuclear activities. Each of the States Party<?BR> to this Treaty further
44   y Age    n all its peaceful nuclear activities. Each State Party to this Treaty further undertakes<?BR> not to
44   BR>    ards on peaceful nuclear activities of states<?BR> <?HR><?BR> SECRET/NOFORN<?BR>
44   ON A    EIR PEACEFUL NUCLEAR ACTIVITIES. 2. EACH OF THE STATES PARTY TO THIS TREATY
44   n the    field of peaceful nuclear activities, including the<?BR> international exchange of nuclear material
44   R> e    ce that peaceful nuclear activities may encourage<?BR> rather than inhibit desires for bombs. (iii)
44   the application of IAEA or similar international ?BR> safeguards to peaceful nuclear activities. Since the first<?BR> alternative might cause serious resistance
447  be applied on all source or special fissionable<?BR> material in all peaceful nuclear activities within<?BR> the territory of such State, under its jurisdiction,
448  include mandatory IAEA or similar international safeguards for<?BR> the peaceful nuclear activities of non-nuclear states. It anticipated<?BR> that the non-nuclear
449  ,<?BR> including international co-operation in the field<?BR> of peaceful nuclear activities. The article on<?BR> control provides for the establishment of
450  of the Parties of international cooperation<?BR> in the field of peaceful nuclear activities, including<?BR> the international exchange of nuclear material
451  of<?BR> significant quantities (table) of nuclear material from<?BR> peaceful nuclear activities. ?? The Department of Safeguards<?BR> is headed by the
452  to accept IAEA or equivalent international<?BR> safeguards on all peaceful nuclear activities. All parties would<?BR> undertake to cooperate in facilitating
```

```
102,948 entries | Row 425 |  |  | MATERIAL IN ALL<?BR> PEACEFUL NUCLEAR ACTIVITIES WITHIN THE
```



1996	1997	1998	1999	2000	2001	2002	2003	
	NBSP	ILLEGIBLE	LIBRARY	ILLEGIBLE	ILLEGIBLE	ILLEGIBLE	ILLEGIBLE	ILLEGI
IOTOCOPY	BR	TEXT	TEXT	TEXT	TEXT	TEXT	TEXT	TEXT
SSEMINATION	DR	DWIGHT	ILLEGIBLE	WE	PHOTOCOPIED	WE	ARCHIVES	NARA
J	PHOTOCOPY	LIBRARY	WE	TO	TO	I	REPRODUCED	EO
RALD	CARTER	EISENHOWER	CARTER	CARTER	WE	E	NARA	SENSI
FUGEES	TOP	TO	DWIGHT	LIBRARY	EISENHOWER	T	AUTHORITY	AUTH
RD	YOU	T	LBJ	THAT	THE	LIBRARY	DATE	AND
E	THIEU	E	COPY	LBJ	OUR	CARTER	I	TO
AT	GERALD	CHINESE	TO	WOULD	NUCLEAR	COPY	NATIONAL	THE
ASSIFICATION	FORD	LBJ	EISENHOWER	OUR	LIBRARY	LBJ	O	YOU
	WE	D	FORD	REGARDING	THAT	YOU	DECLASSIFIED	ISRAE
BRARY	BRZEZINSKI	THAT	SECRET	COPY	WOULD	TO	AT	A
NSITIVE	SECRET	THE	NODIS	THE	BE	#	THE	KISSIN
ONTROLS	GDS	R	TEL	WITH	WITH	R	YOU	WE
TICLE	MISSILE	HE	THAT	SOVIETS	IN	KISSINGER	NIXON	IS
OULD	THINK	COPY	OUR	AND	S	NARA	EO	ISSUE
)	TEST	LAO	WOULD	SAID	DWIGHT	REPRODUCED	KISSINGER	IN
UCLEAR	S	WE	NUCLEAR	BE	RELATIONS	THAT	E	I
AS	U	S	GERALD	U	U	EXDIS	RICHARD	NLS
JBA	MR	SAID	NOT	HE	PAKISTAN	O	TO	ARE
IEU	BUNKER	C	YOU	NUCLEAR	REGARDING	STATE	A	ISRAE
DIS	THAT	IN	EXDIS	WILL	CARTER	AMEMBASSY	HENRY	REAG.
IS	NAM	AND	R	ISSUES	FRENCH	HE	WE	US
JBAN	I	WOULD	THE	NOT	MINISTER	NLJ	HE	INTEL
SSINGER	PRESIDENT	CHINA	USUN	COULD	GAULLE	BRZEZINSKI	WILL	DATE
AM	EXDIS	HAD	BE	DWIGHT	A	OMITTED	ME	POPU
H	TESTS	OF	VANCE	IN	ISSUES	S	BY	HAVE

TIDAL: TImes Data Archive Lab project

- Continuing partnership with Cengage
- Times Digital Archive
- University of Michigan
- Investigating social reality in 19th century based on news discourse (to complement picture of Dickensian London)

British Library News Data: Future Plans

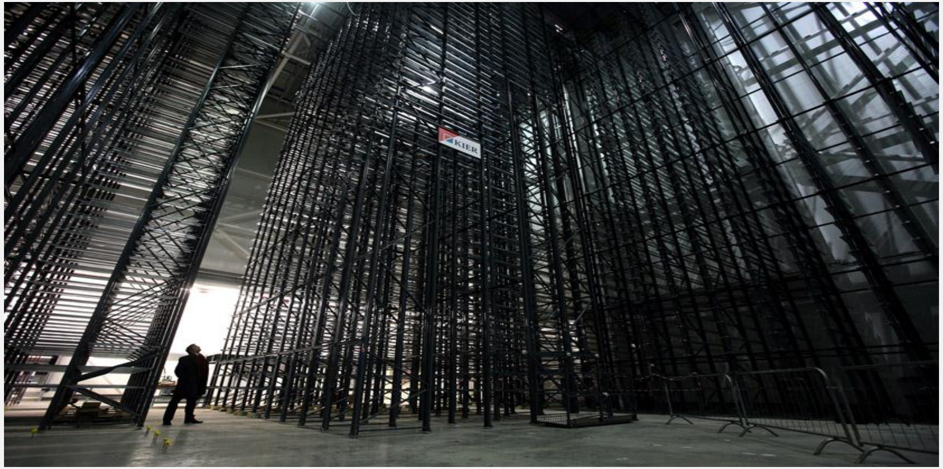
The British Library has one of the world's greatest news archives. Our collection of UK, Irish and world newspapers numbers over 60 million issues, from the 17th century to the present day, and we have growing collections of television, radio and web news.

http://britishlibrary.typepad.co.uk/thenewsroom/

https://twitter.com/BL_newsroom

British Library News Data: Future Plans

Print newspapers from Colindale moved to a new newspaper storage building in Boston Spa in Yorkshire



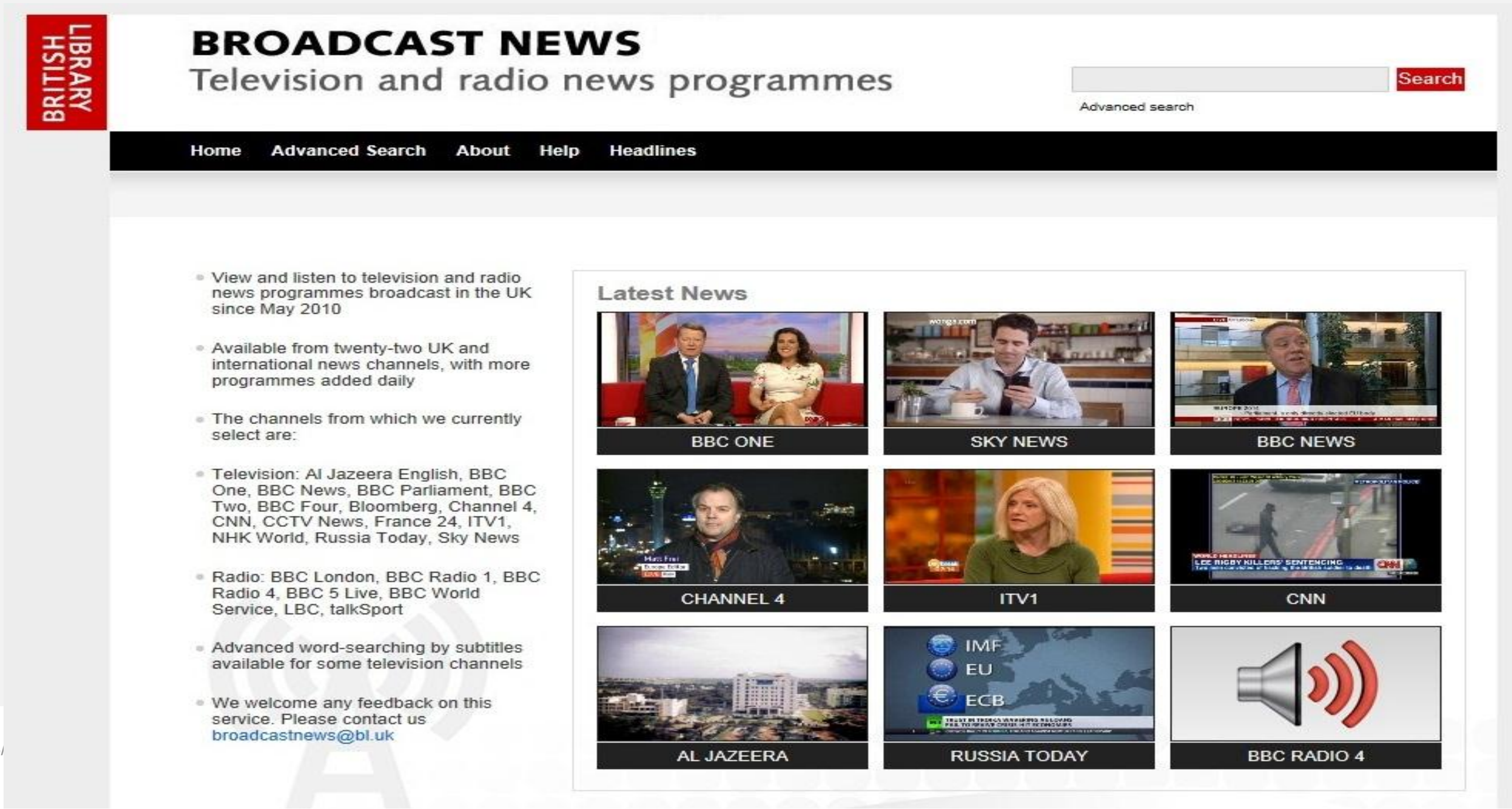
British Library: Newsroom Opened 7th April 2014

British Library News Data: Future Plans

Broadcast News:

Daily television and radio news programmes broadcast in the UK since May 2010 available though an instant access service in the Reading Rooms, new programmes available within hours of broadcast.

Over 60 hours of news are recorded every day from 22 channels including BBC, ITV, Channel 4, Sky News, Al Jazeera English and CNN.



BROADCAST NEWS

Television and radio news programmes

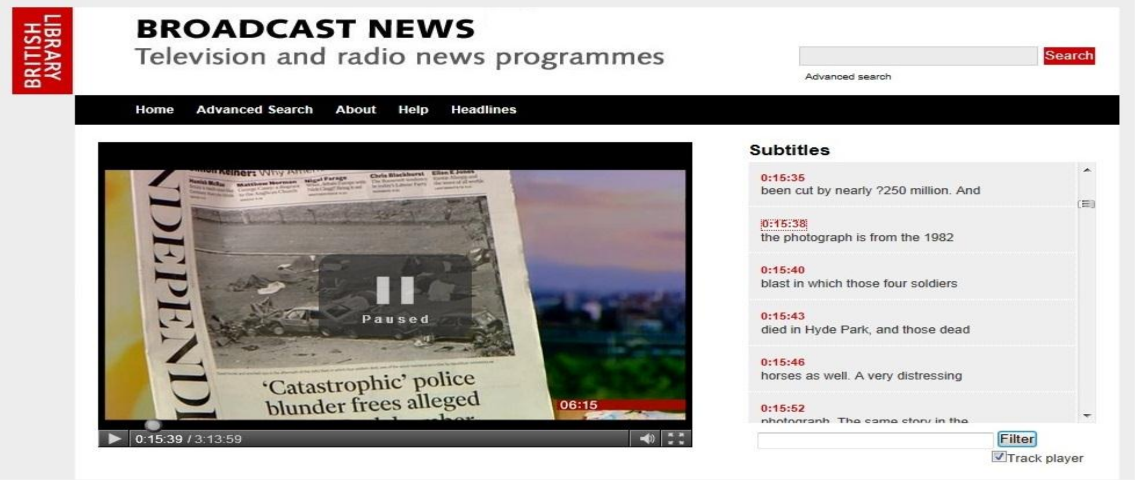
- View and listen to television and radio news programmes broadcast in the UK since May 2010
- Available from twenty-two UK and international news channels, with more programmes added daily
- The channels from which we currently select are:
- Television: Al Jazeera English, BBC One, BBC News, BBC Parliament, BBC Two, BBC Four, Bloomberg, Channel 4, CNN, CCTV News, France 24, ITV1, NHK World, Russia Today, Sky News
- Radio: BBC London, BBC Radio 1, BBC Radio 4, BBC 5 Live, BBC World Service, LBC, talkSport
- Advanced word-searching by subtitles available for some television channels
- We welcome any feedback on this service. Please contact us broadcastnews@bl.uk

Latest News

British Library News Data: Future Plans

Many recorded news programmes come with subtitles

A corpus of data and metadata with potential for innovative research



BROADCAST NEWS
Television and radio news programmes

Subtitles

0:15:35
been cut by nearly ?250 million. And

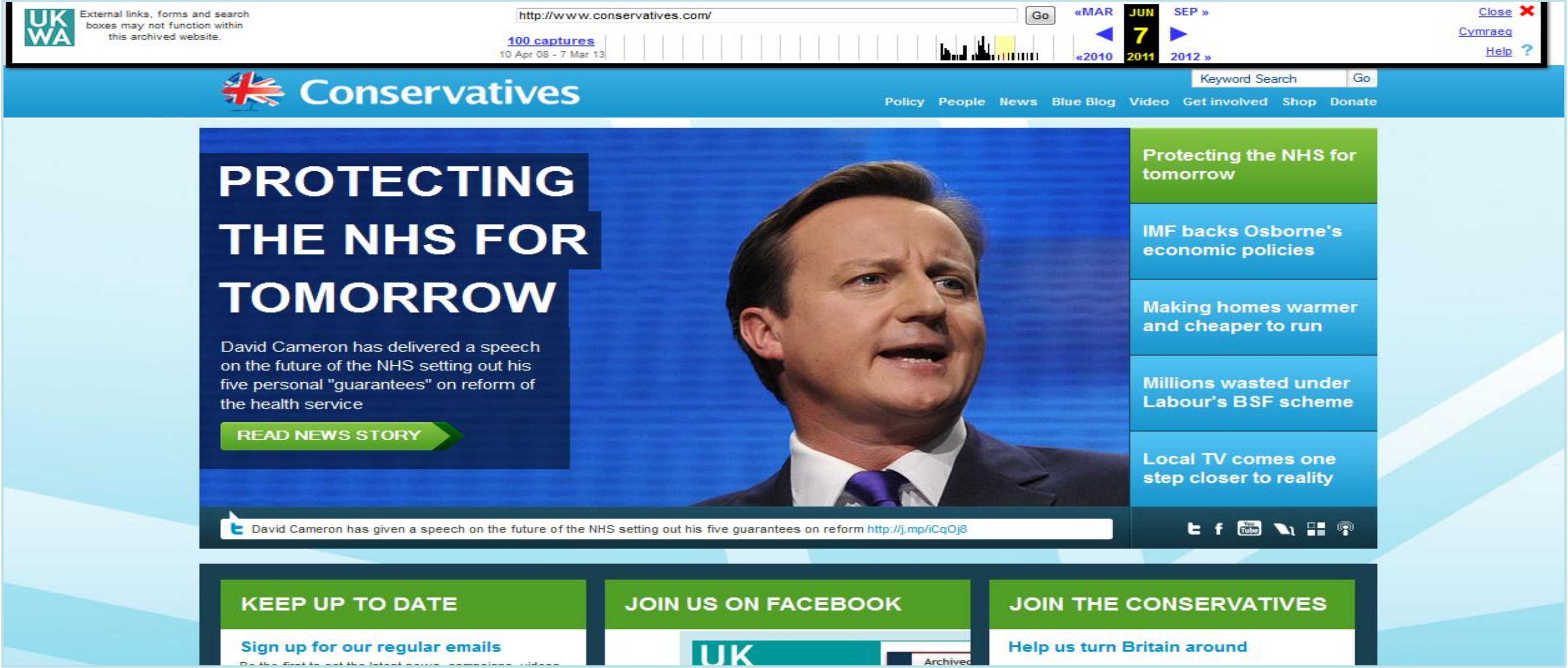
0:15:38
the photograph is from the 1982

0:15:40
blast in which those four soldiers

0:15:43
died in Hyde Park, and those dead

0:15:46
horses as well. A very distressing

0:15:52
photograph. The same story in the



UK WA External links, forms and search boxes may not function within this archived website.

http://www.conservatives.com/ Go

100 captures
10 Apr 08 - 7 Mar 13

«MAR JUN SEP »
7
«2010 2011 2012 »

Close ✖
Cymraeg
Help ?

Conservatives

Keyword Search Go

Policy People News Blue Blog Video Get involved Shop Donate

PROTECTING THE NHS FOR TOMORROW

David Cameron has delivered a speech on the future of the NHS setting out his five personal "guarantees" on reform of the health service

READ NEWS STORY

Protecting the NHS for tomorrow

IMF backs Osborne's economic policies

Making homes warmer and cheaper to run

Millions wasted under Labour's BSF scheme

Local TV comes one step closer to reality

David Cameron has given a speech on the future of the NHS setting out his five guarantees on reform http://j.mp/iCqOj8

KEEP UP TO DATE

Sign up for our regular emails

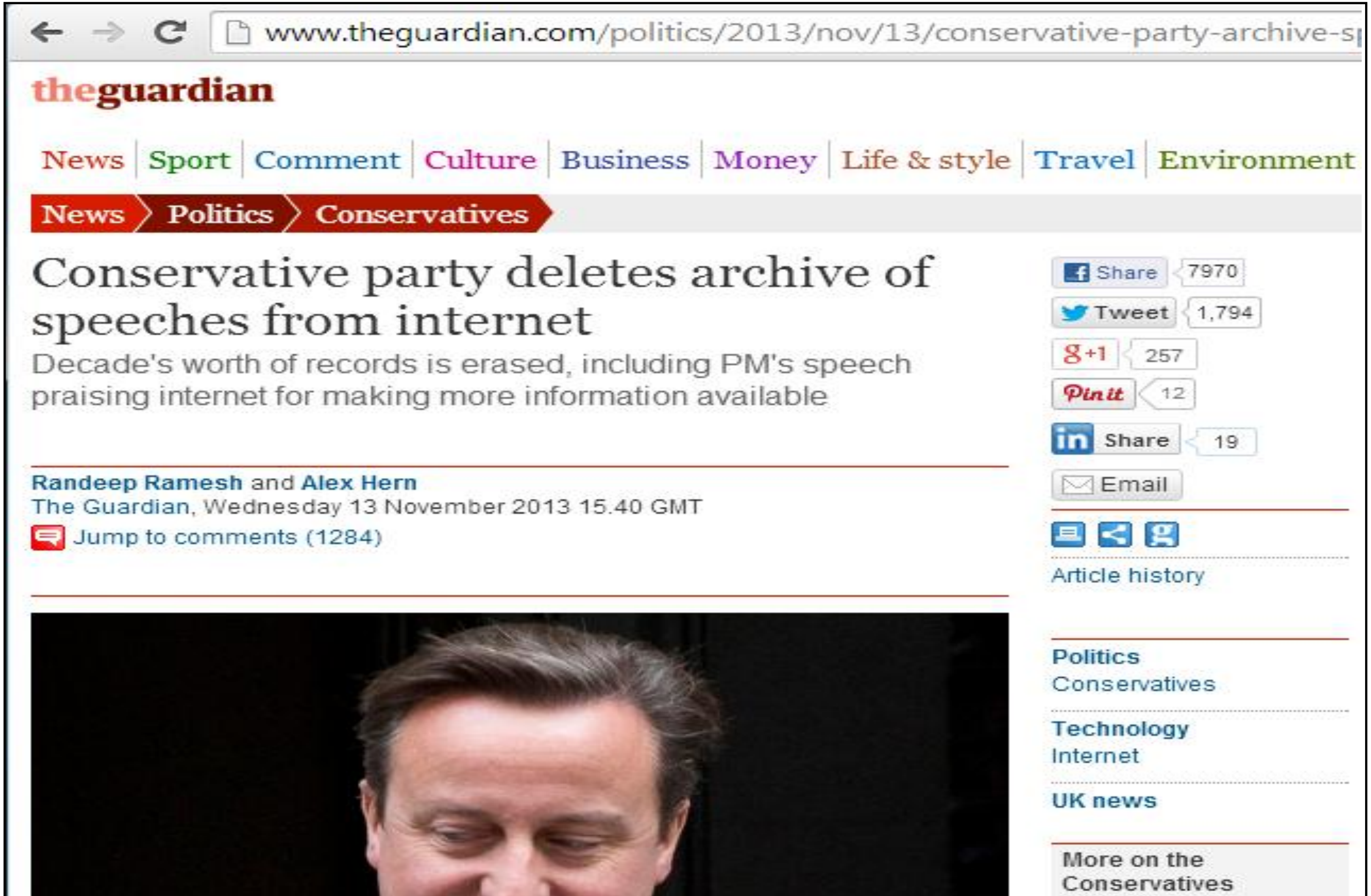
JOIN US ON FACEBOOK

UK

Archived

JOIN THE CONSERVATIVES

Help us turn Britain around



www.theguardian.com/politics/2013/nov/13/conservative-party-archive-s

theguardian

News | Sport | Comment | Culture | Business | Money | Life & style | Travel | Environment
News › Politics › Conservatives

Conservative party deletes archive of speeches from internet

Decade's worth of records is erased, including PM's speech praising internet for making more information available

Randeep Ramesh and **Alex Hern**

The Guardian, Wednesday 13 November 2013 15.40 GMT

Jump to comments (1284)

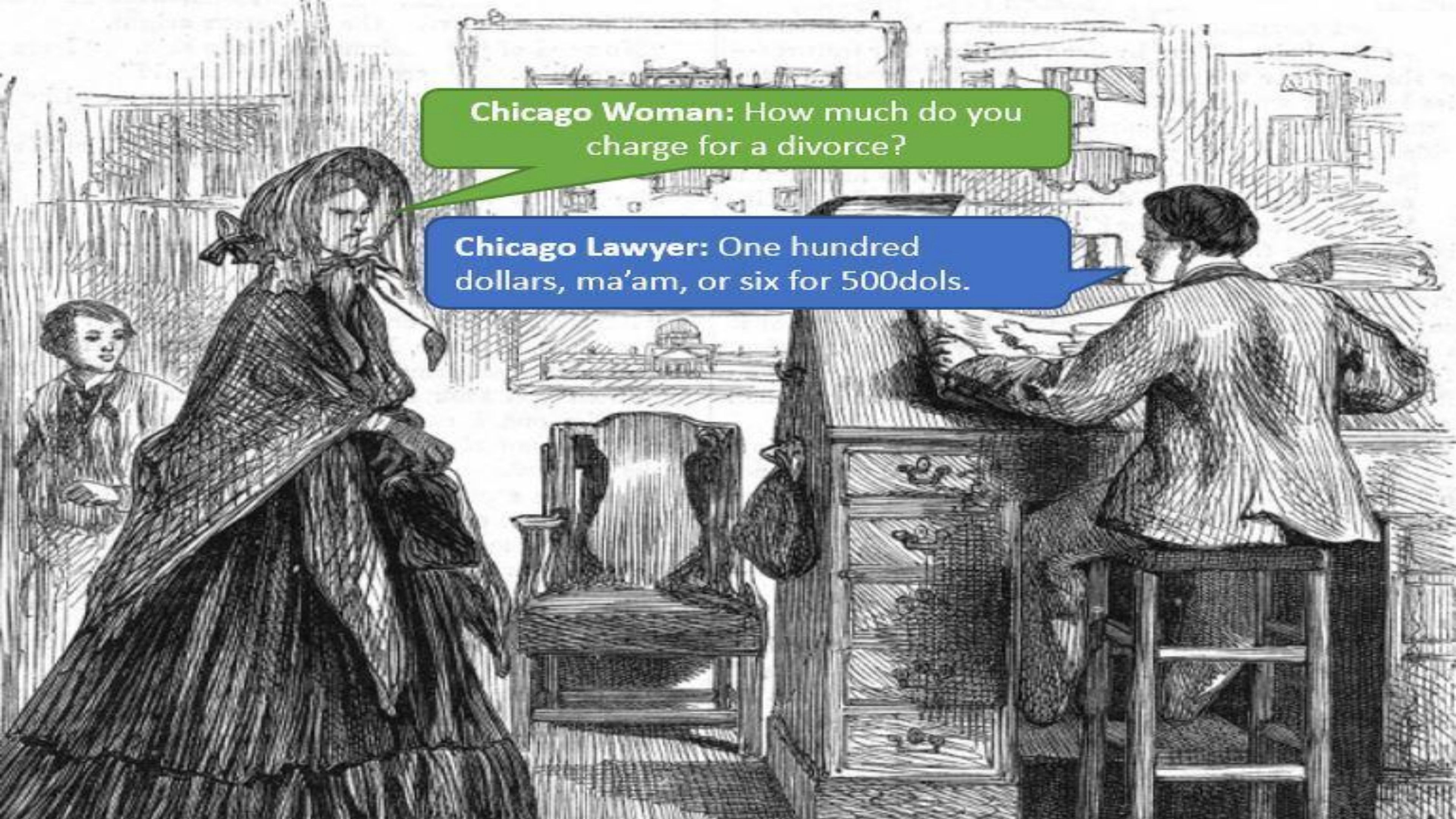
Share 7970
Tweet 1,794
+1 257
Pin it 12
in Share 19
Email
Article history
Politics
Conservatives
Technology
Internet
UK news
More on the Conservatives

BRITISH LIBRARY

British Library Labs

Experiment with our digital collections

http://labs.bl.uk


Chicago Woman: How much do you charge for a divorce?
Chicago Lawyer: One hundred dollars, ma'am, or six for 500dols.

Corpus Protocols:

digital transformations of commercial newspaper collections for text and data mining to support academic research

IFLA 2014 Pre-Conference: Digital Transformation and the Changing Role of News Media in the 21st Century

August 13-14, 2014

International Telecommunication Union, Geneva, Switzerland

Arts & Humanities Research Council

horizon
DIGITAL ECONOMY RESEARCH

The University of Nottingham
UNITED KINGDOM · CHINA · MALAYSIA