

Data Reuse from Government & Research: observations from experience

Kevin Ashley

Digital Curation Centre, University of Edinburgh

www.dcc.ac.uk

@kevingashley

Kevin.ashley@ed.ac.uk



Reusable with attribution: CC-BY

The DCC is supported by Jisc & FP7

Jisc



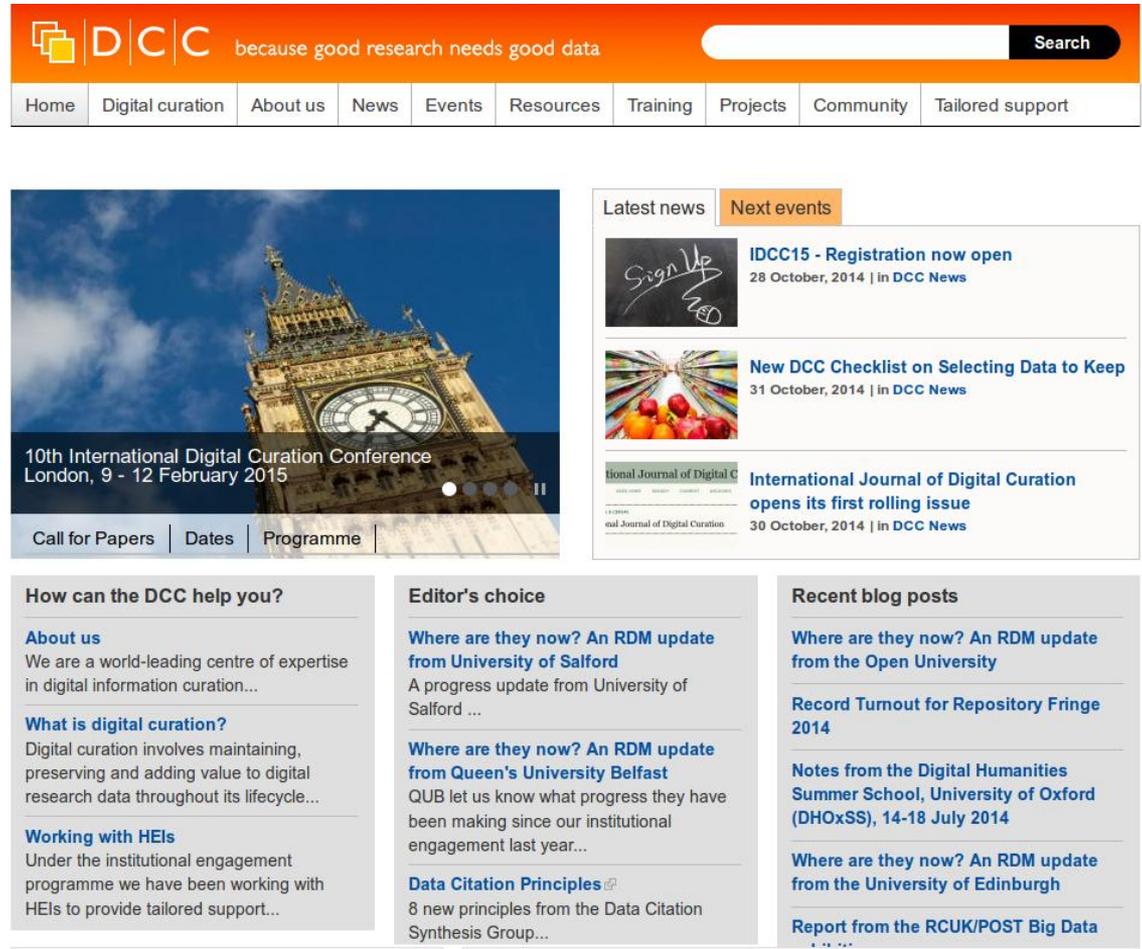
because good research needs good data

Me – some background

- Currently – director of Digital Curation Centre
 - Focus on digital curation of research materials

My home – the DCC

- Mission – to increase capability and capacity for research data services in UK institutions
- Not just a UK problem – an international one
- Training, shared services, guidance, policy, standards, futures

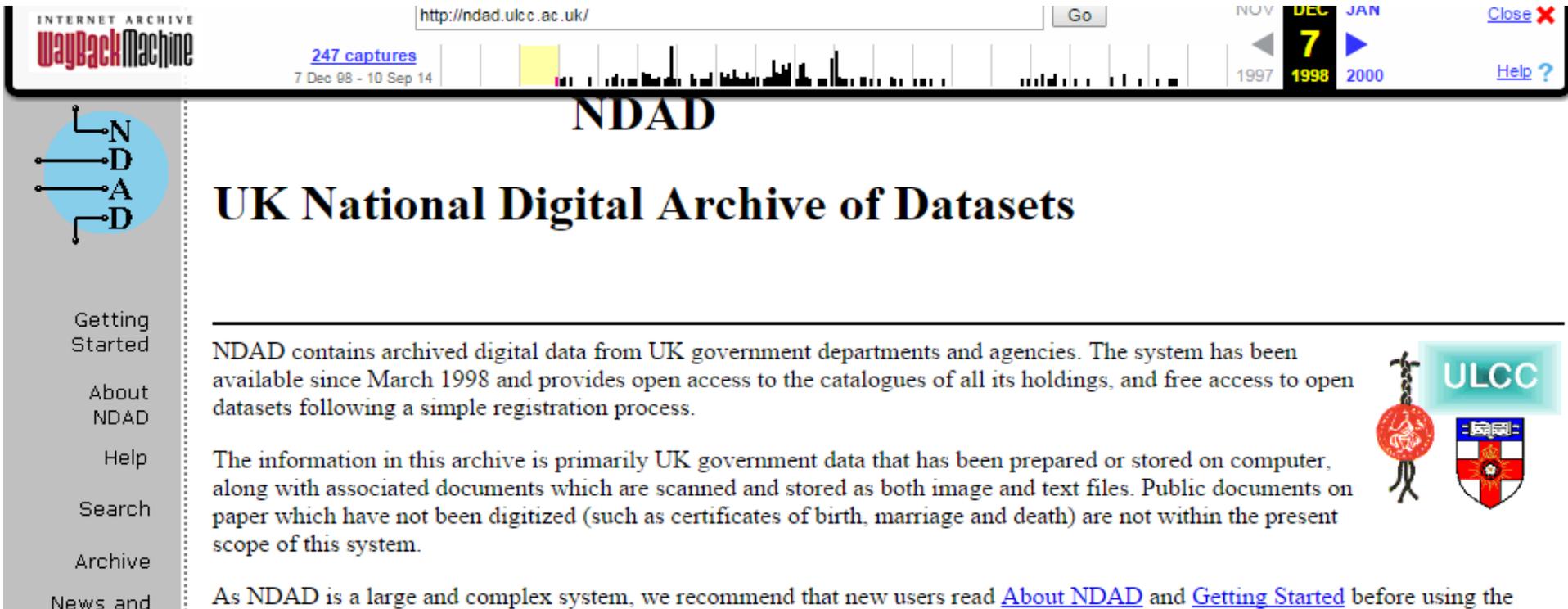


The screenshot shows the DCC website homepage. At the top is the DCC logo with the tagline "because good research needs good data" and a search bar. Below the logo is a navigation menu with links: Home, Digital curation, About us, News, Events, Resources, Training, Projects, Community, and Tailored support. The main content area features a large banner for the "10th International Digital Curation Conference" in London, with a call to action for papers, dates, and programme. To the right of the banner are three news items: "IDCC15 - Registration now open", "New DCC Checklist on Selecting Data to Keep", and "International Journal of Digital Curation opens its first rolling issue". Below the banner are three columns of content: "How can the DCC help you?" with sub-sections for "About us", "What is digital curation?", and "Working with HEIs"; "Editor's choice" with sub-sections for "Where are they now? An RDM update from University of Salford" and "Where are they now? An RDM update from Queen's University Belfast"; and "Recent blog posts" with sub-sections for "Where are they now? An RDM update from the Open University", "Record Turnout for Repository Fringe 2014", "Notes from the Digital Humanities Summer School, University of Oxford (DHOxSS), 14-18 July 2014", "Where are they now? An RDM update from the University of Edinburgh", and "Report from the RCUK/POST Big Data".

Me – some background

- Currently – director of Digital Curation Centre
 - Focus on digital curation of research materials
- Initially – software/systems in clinical research
 - Frequently rescuing old data
- 1997-2010 Head of ULCC Digital Archives Group
 - First online access to archived government datasets by a national archive

NDAD – National Digital Archive of Datasets



The screenshot shows the NDAD website interface. At the top, there is a WayBack Machine search bar with the URL <http://ndad.ulcc.ac.uk/> and a 'Go' button. Below the search bar is a calendar navigation showing the date 7 Dec 1998. The main content area features the NDAD logo and the title 'UK National Digital Archive of Datasets'. A navigation menu on the left includes links for 'Getting Started', 'About NDAD', 'Help', 'Search', 'Archive', and 'News and'. The main text describes the archive's contents and provides a recommendation for new users.

INTERNET ARCHIVE
WayBackMachine

<http://ndad.ulcc.ac.uk/> Go

247 captures
7 Dec 98 - 10 Sep 14

NOV DEC JAN
1997 1998 2000

Close X
Help ?

NDAD

UK National Digital Archive of Datasets

Getting Started
About NDAD
Help
Search
Archive
News and

NDAD contains archived digital data from UK government departments and agencies. The system has been available since March 1998 and provides open access to the catalogues of all its holdings, and free access to open datasets following a simple registration process.

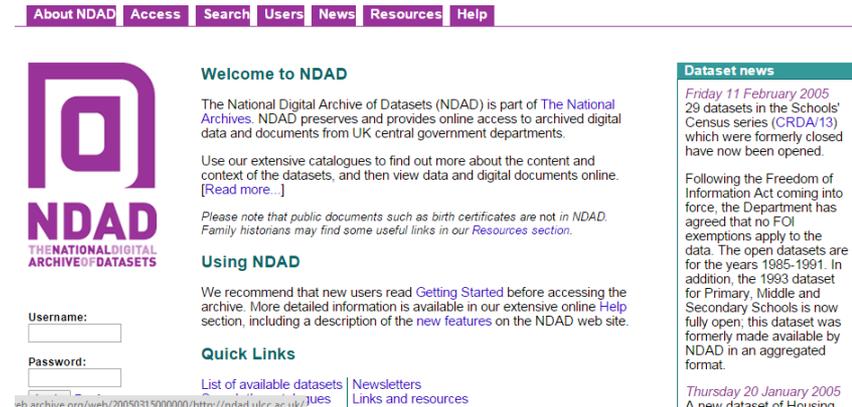
The information in this archive is primarily UK government data that has been prepared or stored on computer, along with associated documents which are scanned and stored as both image and text files. Public documents on paper which have not been digitized (such as certificates of birth, marriage and death) are not within the present scope of this system.

As NDAD is a large and complex system, we recommend that new users read [About NDAD](#) and [Getting Started](#) before using the



NDAD – National Digital Archive of Datasets

- Archivists & data specialists, liaising directly with government departments & contractors
- Documentation, software critical
- Data as a record – not as ‘truth’



The screenshot shows the NDAD website homepage. At the top is a navigation menu with links for About NDAD, Access, Search, Users, News, Resources, and Help. The main content area features the NDAD logo (a stylized 'N' and 'D' in a square) and the text 'THE NATIONAL DIGITAL ARCHIVE OF DATASETS'. Below the logo is a login form with fields for Username and Password. To the right of the login form is a 'Welcome to NDAD' section with a paragraph of introductory text, a 'Use our extensive catalogues...' sentence, and a 'Please note that public documents such as birth certificates are not in NDAD.' notice. Below this is a 'Using NDAD' section with a paragraph of advice for new users. Further down is a 'Quick Links' section with links for 'List of available datasets', 'Newsletters', and 'Links and resources'. On the far right, there is a 'Dataset news' sidebar with a date 'Friday 11 February 2005' and text about '29 datasets in the Schools' Census series (CRDA/13) which were formerly closed have now been opened.' and another entry for 'Thursday 20 January 2005' about a 'new dataset of Housing'.

An aside – purpose, context of a data archive is important

- PSI – “this is the data we have now” (may be incomplete; subject to change)
- National Archive – “this is the data the government used” (may be inaccurate; will not change)
- Research publication – “this is the data we used for this paper” (may be wrong; should not change)
- Data bank – “This is the best data available now” (may be old; should change)

Full of interest!

B	C	D	E	F	G	H	I	J	K	L
1281	8203	2	HARTLEPOOLS WATER COMPANY	8104	1	60	1	1		4T
1281	8303	1	HARTLEPOOLS WATER COMPANY	8204	1	60	1	1		4T
1281	8403	1	HARTLEPOOLS WATER COMPANY	8304	1	60	1	1		4T
1281	8503	1	HARTLEPOOLS WATER COMPANY	8404	1	60	1	1		4T
1281	8603	1	HARTLEPOOLS WATER COMPANY	8504	1	60	1	1		4T
1521	8112	2	WREXHAM AND EAST DENBIGHSHIRE WATER COMP	8101	1	60	1	1		3T
1521	8212	1	WREXHAM AND EAST DENBIGHSHIRE WATER CO	8201	1	60	1	1		3T
1521	8312	1	WREXHAM AND EAST DENRIGHSHIRE WATER CO	8301	1	60	1	1		3T
1521	8412	1	WREXHAM AND EAST DENBIGHSHIRE WATER CO.	8401	1	60	1	1		3T
1521	8512	1	WREXHAM AND EAST DENBIGHSHIRE WATER CO	8501	1	60	1	1		3T
1581	8109	2	MID-SUSSEX WATER COMPANY	8010	1	60	1	1		2T
1581	8109	4	MID-SUSSEX WATER COMPANY	8010	1	60	1	1		2T
1581	8209	1	MID-SUSSEX WATER COMPANY	8110	1	60	1	1		2T
1581	8403	1	MID-SUSSEX WATER COMPANY	8210	1	60	1	1		4T
1581	8503	1	MID-SUSSEX WATER COMPANY	8404	1	60	1	1		4T
1581	8503	4	MID-SUSSEX WATER COMPANY	8404	1	60	1	1		4T
1581	8603	1	MID-SUSSEX WATER COMPANY	8504	1	60	1	1		4T
1781	8203	2	LEE VALLEY WATER COMPANY	8104	1	60	1	1		4T
1781	8303	1	LEE VALLEY WATER COMPANY	8204	1	60	1	1		4T
1781	8403	1	LEE VALLEY WATER COMPANY	8304	1	60	1	1		4T
1781	8503	1	LEE VALLEY WATER COMPANY	8404	1	60	1	1		4T
1781	8603	1	LEE VALLEY WATER COMPANY	8504	1	60	1	1		4T
1891	8112	2	MERSEY DOCKS & HARBOUR COMPANY	8101	1	70	1	1		3T
1891	8212	1	MERSEY DOCKS & HARBOUR COMPANY	8201	1	70	1	1		3T
1891	8312	1	MERSEY DOCKS AND HARBOUR COMPANY	8301	1	70	1	1		3T
1891	8412	1	MERSEY DOCKS AND HARBOUR COMPANY	8401	1	70	1	1		3T

The Company Accounts Database

Excuses not to share – much the same

- “People will ask questions”
 - So use a data centre or repository
- “It will be misinterpreted”
 - Stuff happens. Also, openness encourages correction
- “It’s not interesting”
 - Let others be the judge – your noise is my signal
- “I might get another paper out of it”
 - Up to a point. We might get more research out of it
- “I don’t have permission”
 - A real problem. But solvable at senior level
- “It’s too bad/complicated” –see above
- “It’s not a priority”
 - Unfortunately, funders are making it so. But if you looked at the evidence, it would be your priority as well

See e.g. Carly Strasser’s blog:

<http://datapub.cdlib.org/2013/04/24/closed-data-excuses-excuses/>



Research Data vs Government Data

- Collected by those who will use it (often)
- Often used just once (if at all)
- Expensive to create
- Hard to find
- Poorly documented
- Easily lost
- Often connected to publication
- Little regulatory framework

- Collected by those who will use it (often)
- Often used continuously
- Expensive to create
- Hard to find
- Poorly documented
- Easily lost
- Rarely connected to publication
- Significant regulatory framework

Types of government data

- Collected to support an administrative or regulatory function
- Collected to answer a specific research question
 - Collection, analysis often contracted out
- Created as a reference source
 - Administrative & research functions

What we got

Department	Series	NDAD ref	PRO ref
Agricultural Departments	Agricultural and Horticultural Census	CRDA/4	MAF 408 MAF 410
	Coast Protection Survey of England	CRDA/10	MAF 406
	Internal Drainage Board Database	CRDA/9	MAF 407
British Railways Board	British Rail Electronically Archived Documents	CRDA/37	AN 186
Countryside Agencies	Survey of Rural Services	CRDA/30	D 16 D 10
Department of Culture, Media and Sport	National Lottery Awards Database	CRDA/39	PF 1
Department of Trade and Industry	Oil and Gas Directorate, North Sea Geographical Information System	CRDA/26	EG 17
Education Departments	Grant Maintained Schools Database	CRDA/36	ED 278
	Learning Partnerships	CRDA/53	NV 4
	Learning and Training at Work	CRDA/52	NV 3
	Register of Educational Establishments	CRDA/47	NV 2
	Schools' Census (Form 7)	CRDA/13	ED 267
Environment Departments	Countryside Information System	CRDA/46	AT 73
Forestry Commission	National Inventory of woodland and Trees	CRDA/3	F 45
Health Departments	1994 General Household Survey: Follow-up Survey of the Health of People aged 65 and over	CRDA/28	JA 6
	AIDS Advertising Evaluation	CRDA/35	BN 97
	Anatomy Office - Anatomy dataset	CRDA/21	JA 3
	...ies on Starting Infant School	CRDA/33	BN 94

REFERENCE DATA

ADMIN DATA

ONE-OFF DATA

What we got...

	AIDS Advertising Evaluation	CRDA/35	BN 97
	Anatomy Office - Anatomy dataset	CRDA/21	JA 3
	Children's Difficulties on Starting Infant School	CRDA/33	BN 94
	Children in care	CRDA/38	BN 98
	Elderly and their Medicines	CRDA/34	BN 96
	Public Health Common Dataset	CRDA/24	JA 5
	Reasons for Retirement	CRDA/27	BN 93
	Survey of Abortion Patients for the Committee on the Working of the Abortion Act	CRDA/32	BN 95
HM Customs and Excise	Beer Duty	CRDA/19	CUST 134
Home Office	British Crime Survey	CRDA/2	HO 400
Lord Chancellor's Department	Judge Advocate General's Office Case Index System	CRDA/23	LCO60
	Judicial Statistics	CRDA/8	LE 1
	Presentations under the Public Records Act 1958, s.3(6)	CRDA/59	PRO 72
Metropolitan Police	Crime Statistics System (ME)	CRDA/1	MEPO 36 MEPO 37
Museum and Galleries Commission	Digest of Museum Statistics (DOMUS)	CRDA/12	EB 6
Nature Conservation Departments	Ancient Woodland Inventory	CRDA/43	FT 43
	British Bats	CRDA/17	FT 40
Social Security Departments	Earnings Top-up Scheme	CRDA/48	NB5
Statistical Departments	Historic Mortality Data Files	CRDA/20	RG 69
	Primary Births	CRDA/5	RG 71
Transport Departments	Survey of Heavy Goods Vehicles	CRDA/14	LM 1
Welsh Office	Coastal Survey - Wales	CRDA/6	BD 80
	Survey of Contaminated Land in Wales	CRDA/15	BD 83

Potential reuse

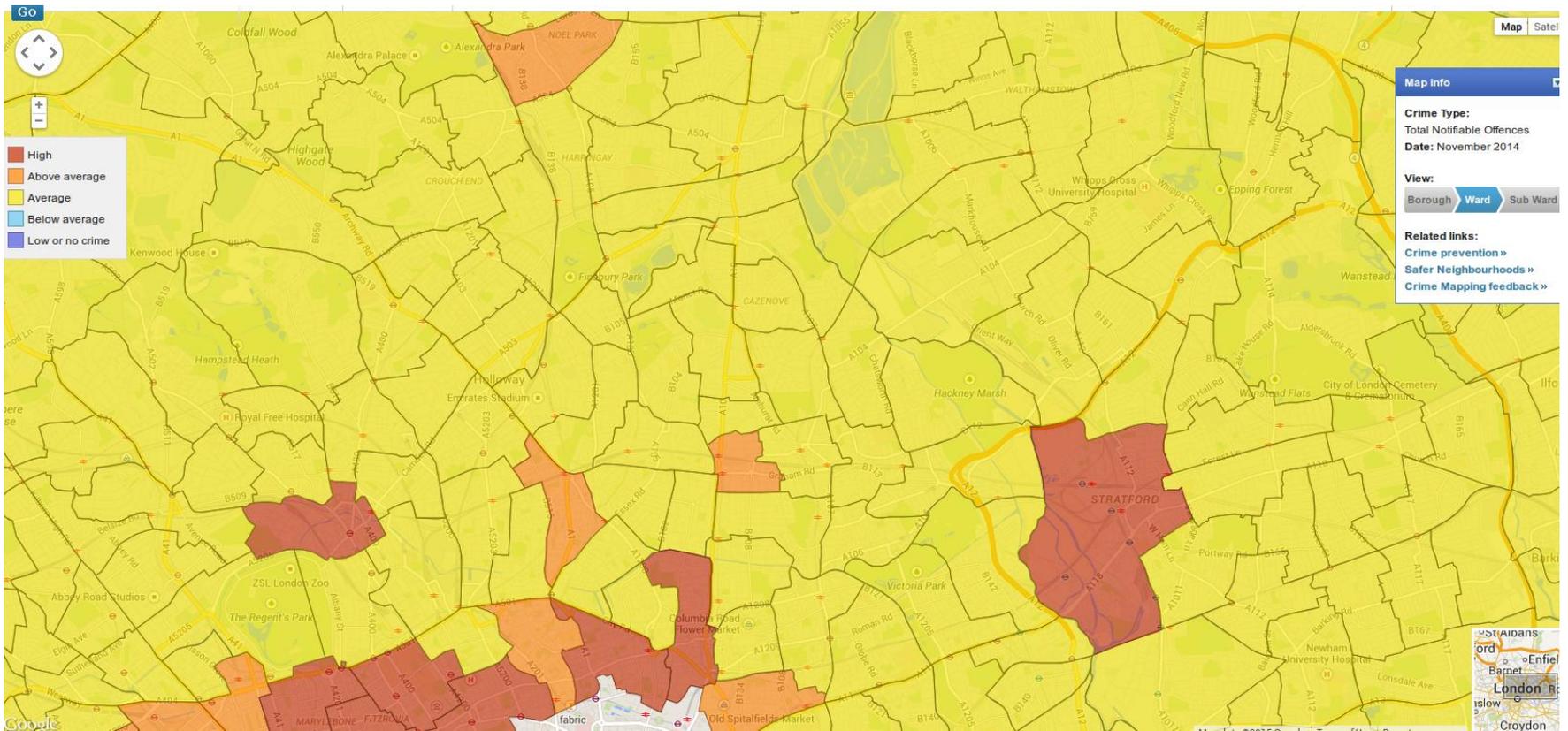
- By other parts of government – or other governments
- By journalists
- By the public
- By historians
- In academic and commercial research
- In education

Potential gains from research data sharing

- Research quality
 - How close can we get to the truth?
- Research speed
 - How quickly can we get to the truth?
- Research finance
 - How much does the truth cost?
- Improving one or more of these is of interest to all actors:
- Researchers as data creators
- Researchers as data reusers
- Research institutions
- Funders – hence government and society

The story of Metropolitan Police Data

- Crime information is accessible to all on the web – from a system called CRIS



If you don't like pictures, you can have numbers

		Violence Against The Person			Sexual Offences			Robbery			Burglary			Theft/Taking Of Motor Vehicle
		Murder	Wounding/GBH	Assault With Injury	Rape	Other Sexual	Total	Personal Property	Business Property	Total	Burglary in A Dwelling	Burglary in Other Buildings	Total	
					118	140	258	619	48	667	1835	705	2540	808
					100	200	300	1001	72	1073	3438	1368	4806	699
					43	135	178	205	37	242	997	686	1683	477
					122	257	379	1369	85	1454	2800	906	3706	580
					76	163	239	523	67	590	2064	1229	3293	762
					90	241	331	994	70	1064	1483	1445	2928	592
					178	296	474	1877	155	2032	2908	1214	4122	922
					112	263	375	1100	85	1185	2895	1031	3926	842
					109	197	306	946	137	1083	2688	1054	3742	805
					88	243	331	494	56	550	1466	802	2268	603
					136	241	377	1171	61	1232	1436	1118	2554	686
					85	177	262	656	40	696	1115	600	1715	529
					116	244	360	979	70	1049	2314	836	3150	871
					58	144	202	619	25	644	1877	614	2491	259
					57	126	183	350	51	401	2102	653	2755	873
					3	26	29	3	1	4	0	4	4	36
					89	187	276	741	47	788	2200	893	3093	522
					94	232	326	642	37	679	2116	846	2962	522
					88	205	293	1094	69	1163	1266	1056	2322	618
					63	142	205	618	21	639	1011	741	1752	472
					59	104	163	201	16	217	848	533	1381	131
					170	320	490	2431	191	2622	2577	1174	3751	926
					132	286	418	1226	108	1334	2460	922	3382	833
					54	126	180	501	43	544	1288	708	1996	361
					168	222	390	2151	109	2260	1983	1043	3026	1069
					105	140	245	840	84	923	2651	688	3339	1226

Now vs 1997

- CRDA/1 (now MEPO 36) – first dataset acquired by NDAD
- From retired system ME – predecessor to CRIS
- An administrative system – documenting workload
- E.g. records time of crime, time reported to police and date entered into system

Browse The National Archives' catalogue

Browse home

Browse by reference

You are currently viewing

MEPO - Records of the Metropolitan Police Office

↳ Division 2 within MEPO - Records of the Receiver

Inside you will find

◀◀ First ◀ Prev 30

 **MEPO 36** 1990-1997

Metropolitan Police: Crime Statistics Database

The datasets derived from the Metropolitan Police's Crime Statistics System (known within the Metropolitan Police by the code 'ME') contain data relating to crimes reported within the Metropolitan Police District which were input to the Crime Statistics System between 1990 and 1997.

The datasets derived from the ME Crime Statistics System fall into the following categories:

Datasets for the years 1990-1992, where the division between one year and the next is based on a calendar year. Each of these datasets comprises two flat files: a year end extract file (YTDEXTRACT), consisting of data relating to offences, clear-ups, arrests and victims input during that year; and a year end 'No Crime' extract file (YTDNOCREXT) holding data on allegations classed as 'No Crime', also input during the year in question.

Datasets for 1992-1993, 1993-1994 and 1994 where the division between one year and the next is based on a financial year. The 1994 dataset does not include a 'No Crime' file

-  [MEPO 36/1](#) 1990
Crime Statistics System (ME): 1990 dataset
[Details](#)
-  [MEPO 36/2](#) 1991
Crime Statistics System (ME): 1991 dataset
[Details](#)
-  [MEPO 36/3](#) 1992
Crime Statistics System (ME): 1992 dataset
[Details](#)
- 

Interpretation could be harder...

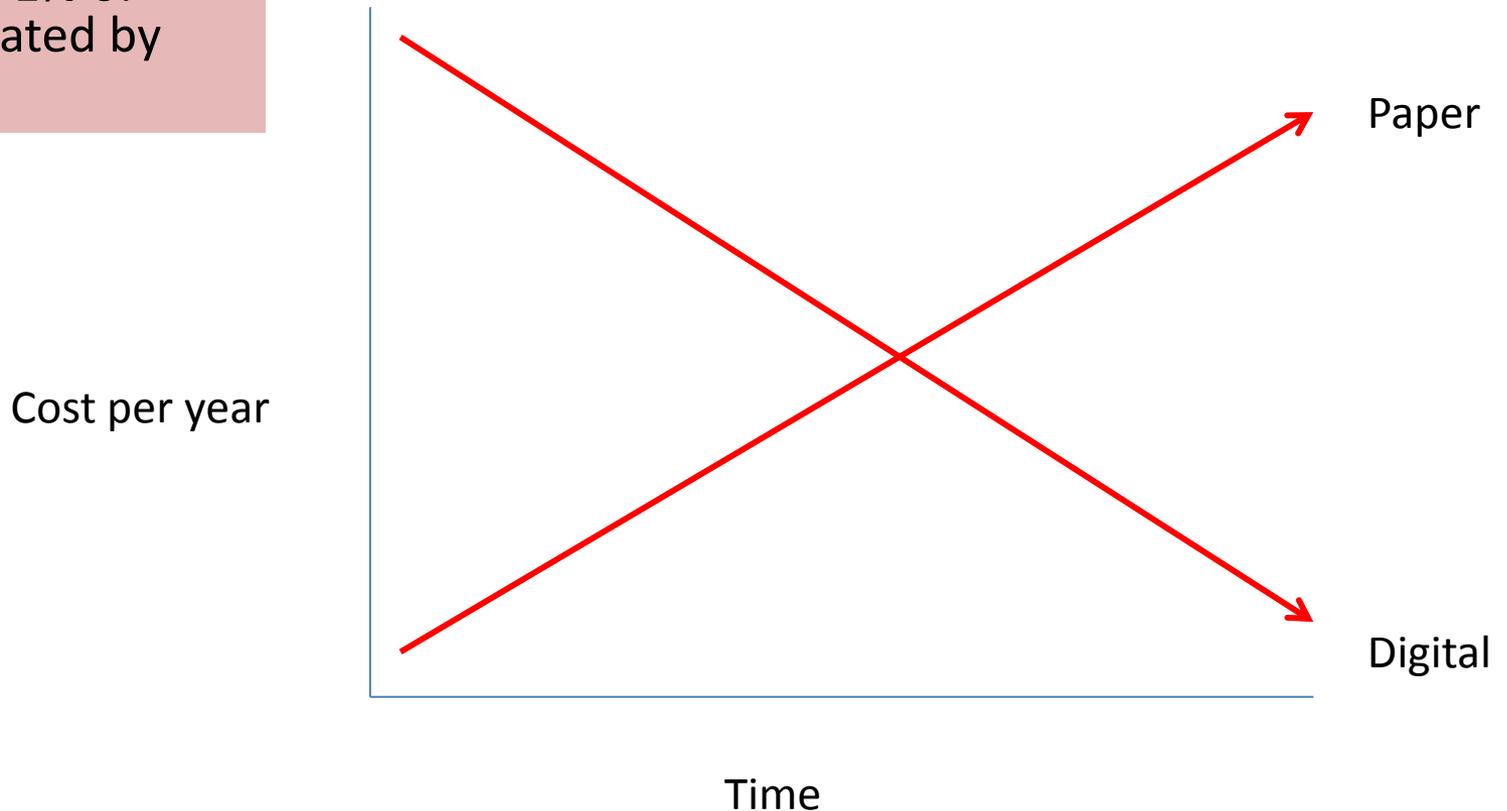
Field name:	Offence-Subdivision-Code
Data type:	Fixed Length String (length 2)
Original description:	<i>The identifying code for an MPD-SUBDIVISION</i>
Further descriptive information:	<p>The identifying code for the Metropolitan Police Subdivision (station) where the offence was reported. Station codes of 2 alpha chars; also 'C1' & 'C6' for CO branches (see Offence-Division-Code)</p> <p>The Subdivision is equivalent to an individual police station within a Division (except for the codes C1 and C6 which represent New Scotland Yard branches). The interpretation of the codes for values in this field is based on the codes for Metropolitan Police Stations given in the <i>Police and Constabulary Almanac 1990: Official Register</i> (Henley-on-Thames: R. Hazell and Company, 1990), pp. 34-36., and on the List of Station Codes (CRDA/1/DD/3/3 and CRDA/1/DD/3/4) and the Specification of Borough Tables (CRDA/1/DD/3/5) supplied by the Metropolitan Police (see the Dataset Documentation Catalogue).</p>
Missing value(s):	One or more spaces or zeroes
Attributes:	
Constraints:	[As specified by Encoding]
Encoding:	This field may only take one of the 160 pre-defined values listed

Legislative context

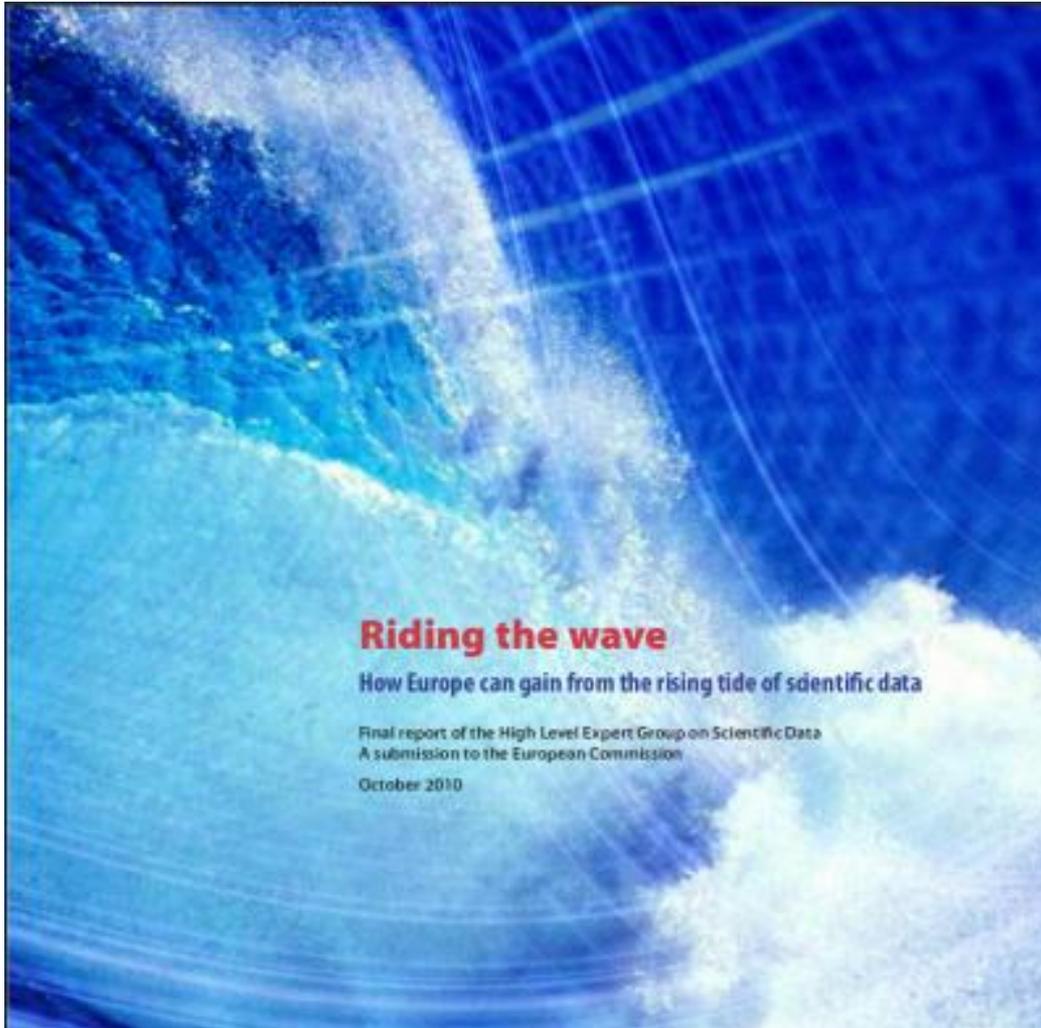
- No Freedom Of Information Act
- No PSI regulations
- Crown Copyright rigorously enforced
- Data Protection Act in force
- Yet – we got & made available the whole thing
- One record per reported crime – 8 years of detailed information for London

We thought selection was getting easier

National Archives keeps about 1% of paper generated by government



The Data Deluge is upon us



Sensor's ability
to produce data
outstrips IT's
ability to
process it

What data to keep

Roles and Responsibilities



A Digital Curation Centre and Australian National Data Service 'working level' guide

How to Appraise & Select Research Data for Curation

Angus Whyte (DCC) and Andrew Wilson (ANDS)

Researcher ('data creator')

- Provide enough information for others to assess the research data's scientific and scholarly quality and compliance with disciplinary or ethical norms.
- Provide relevant information for the repository to identify who will use the data and how i.e. the 'designated community', and any specific access requirements or constraints.
- Provide the research data in formats recommended by the data repository.
- Provide the metadata requested by the repository.

Data centre or repository

- Make explicit its mission in the area of digital archiving, and its selection policy for digital objects.
- Ensure compliance with legal regulations and contracts.
- Ensure the authenticity and integrity of the digital objects and the metadata.
- Assume responsibility from the data producer for ensuring the digital objects are accessible and available to a defined 'designated community'.
- Plan for long-term preservation of the digital assets.

IDCC15 – London, Feb 9-12 2015



The 10th
International
Digital
Curation
Conference

<http://www.dcc.ac.uk/events/idcc15>

Some conclusions

- There are many parallels between PSI & research data
- There are a few differences to remember also
- Some data can cross back and forth between the domains
- So – best not to keep the conversations separate

Thanks for your attention

kevin.ashley@ed.ac.uk
www.dcc.ac.uk
@kevingashley