Diamonds are forever.
What about research data?

Dr Adam Farquhar
Head of Digital Scholarship, The British Library

14 March 2016

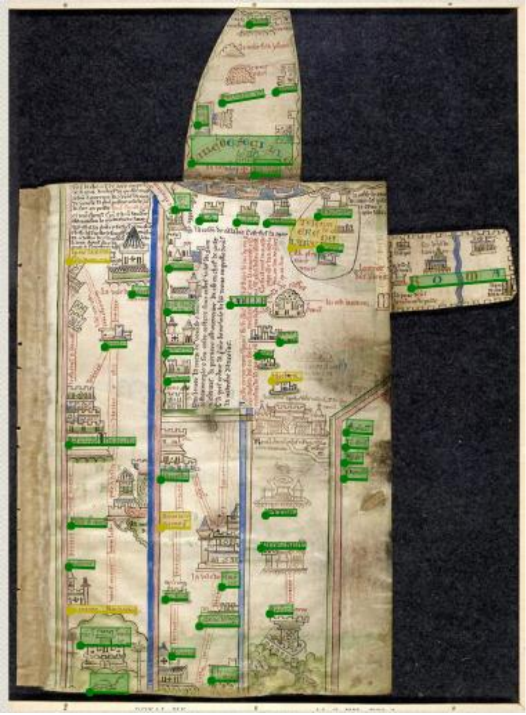
- Over 200 million items in most known languages stored in London and in Yorkshire
- 3 million new items added every year in every format/content type
- 2014 Web Archive crawl: 20m seeds, 56Tb data, 2.5b webpages incl 4.7Gb of viruses & 3.2Tb screenshots.
- Headquarters site for the Alan Turing Institute – the UK’s National Institute for Data Science
- Founding member of the Knowledge Quarter

Digital scholarship

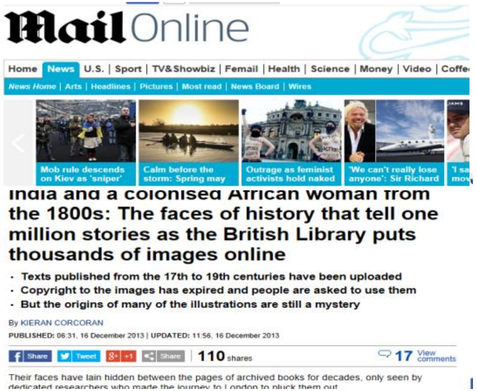
- At the intersection of academic research, cultural heritage and technology
- Driving research and innovation through
  - Getting content in digital form and online
  - Engaging with researchers across the UK
  - Collaborative research projects
  - Providing support and guidance
  - BL Labs

Getting content in digital form

- Non-print legal deposit
- Digitisation: manuscripts, maps, archives, printed books, newspapers, audio, video
- Annotation: transcription, speech-to-text, NLP, entity-extraction
- Crowd-sourcing
  - c750,000 Asian & African items listed in hard-copy
  - Geo-referencing



One million images on Flickr
Over 375,358,000 views & 515,000 tags



Mail Online

Home | News | U.S. | Sport | TV&Showbiz | Femail | Health | Science | Money | Video | Coffe

News Home | Arts | Headlines | Pictures | Most read | News Board | Wires

Mob rule descends on Kiev as 'sniper'
Calm before the storm: Spring may
Outrage as feminist activists hold naked
'We can't really lose anyone': Sir Richard
'I sa mov

India and a colonised African woman from the 1800s: The faces of history that tell one million stories as the British Library puts thousands of images online

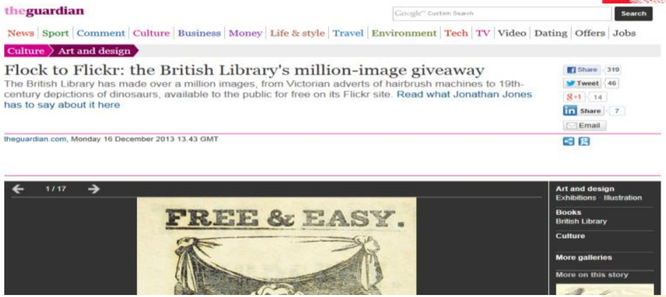
- Texts published from the 17th to 19th centuries have been uploaded
- Copyright to the images has expired and people are asked to use them
- But the origins of many of the illustrations are still a mystery

By KIERAN CORCORAN

PUBLISHED: 06:31, 16 December 2013 | UPDATED: 11:56, 16 December 2013

Share Tweet +1 Share 110 shares 17 View comments

Their faces have lain hidden between the pages of archived books for decades, only seen by dedicated researchers who made the journey to London to pluck them out



theguardian

News | Sport | Comment | Culture | Business | Money | Life & style | Travel | Environment | Tech | TV | Video | Dating | Offers | Jobs

Culture › Art and design

Flock to Flickr: the British Library's million-image giveaway

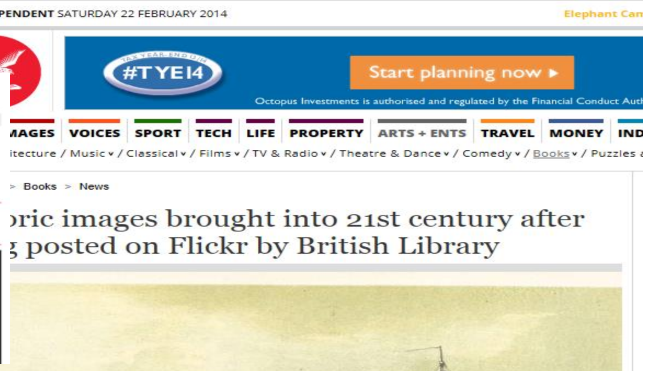
The British Library has made over a million images, from Victorian adverts of hairbrush machines to 19th-century depictions of dinosaurs, available to the public for free on its Flickr site. Read what Jonathan Jones has to say about it here

Share 319 Tweet 46 +1 14 Share 7 Email

theguardian.com, Monday 16 December 2013 13.43 GMT

1 / 17

Art and design
Exhibitions · Illustration
Books
British Library
Culture
More galleries
More on this story



THE INDEPENDENT SATURDAY 22 FEBRUARY 2014

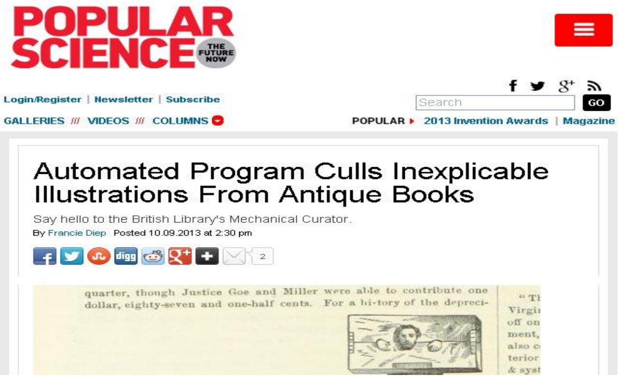
#TYE14
Start planning now ▸
Octopus Investments is authorised and regulated by the Financial Conduct Aut

MAGES | VOICES | SPORT | TECH | LIFE | PROPERTY | ARTS + ENTS | TRAVEL | MONEY | IND

itecture / Music / Classical / Films / TV & Radio / Theatre & Dance / Comedy / Books / Puzzles

› Books › News

oric images brought into 21st century after g posted on Flickr by British Library



POPULAR SCIENCE THE FUTURE NOW

Login/Register | Newsletter | Subscribe

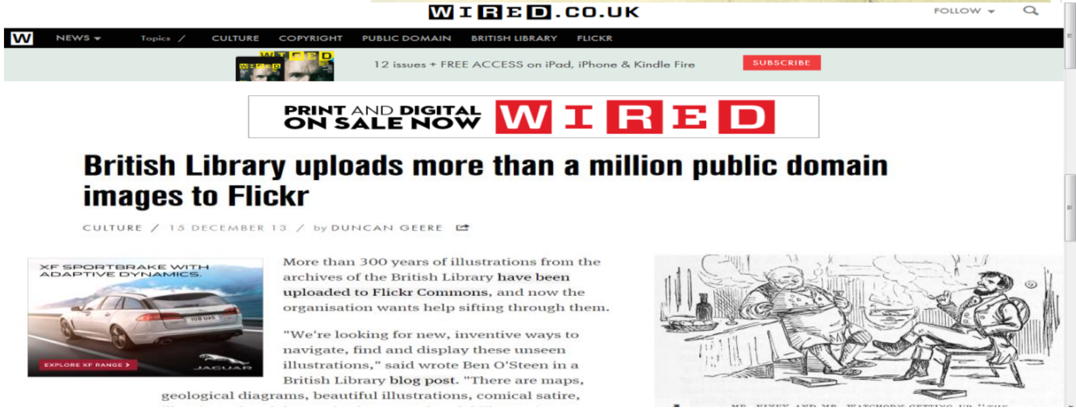
GALLERIES /// VIDEOS /// COLUMNS — POPULAR ▸ 2013 Invention Awards | Magazine

Automated Program Culls Inexplicable Illustrations From Antique Books

Say hello to the British Library's Mechanical Curator.

By Francie Diep Posted 10.09.2013 at 2:30 pm

quarter, though Justice Goe and Miller were able to contribute one dollar, eighty-seven and one-half cents. For a hi-tory of the depreci-



WIRED.CO.UK

NEWS Topics / CULTURE COPYRIGHT PUBLIC DOMAIN BRITISH LIBRARY FLICKR

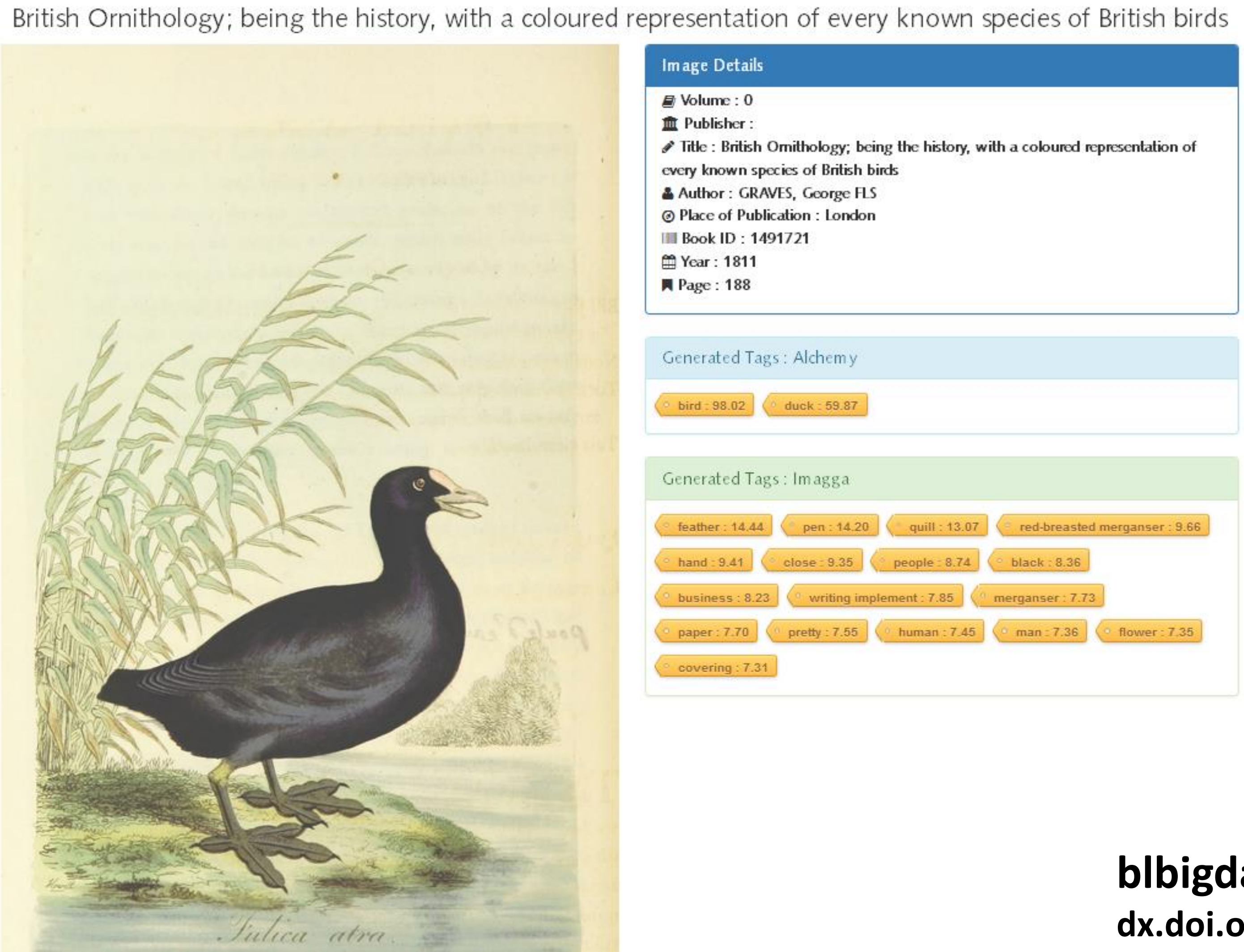
12 issues + FREE ACCESS on iPad, iPhone & Kindle Fire SUBSCRIBE

British Library uploads more than a million public domain images to Flickr

CULTURE / 15 DECEMBER 13 / by DUNCAN GEERE

More than 300 years of illustrations from the archives of the British Library have been uploaded to Flickr Commons, and now the organisation wants help sifting through them.

"We're looking for new, inventive ways to navigate, find and display these unseen illustrations," said wrote Ben O'Steen in a British Library blog post. "There are maps, geological diagrams, beautiful illustrations, comical satire,



British Ornithology; being the history, with a coloured representation of every known species of British birds

Image Details

Volume : 0
Publisher :
Title : British Ornithology; being the history, with a coloured representation of every known species of British birds
Author : GRAVES, George FLS
Place of Publication : London
Book ID : 1491721
Year : 1811
Page : 188

Generated Tags : Alchemy

bird : 98.02
duck : 59.87

Generated Tags : Imagga

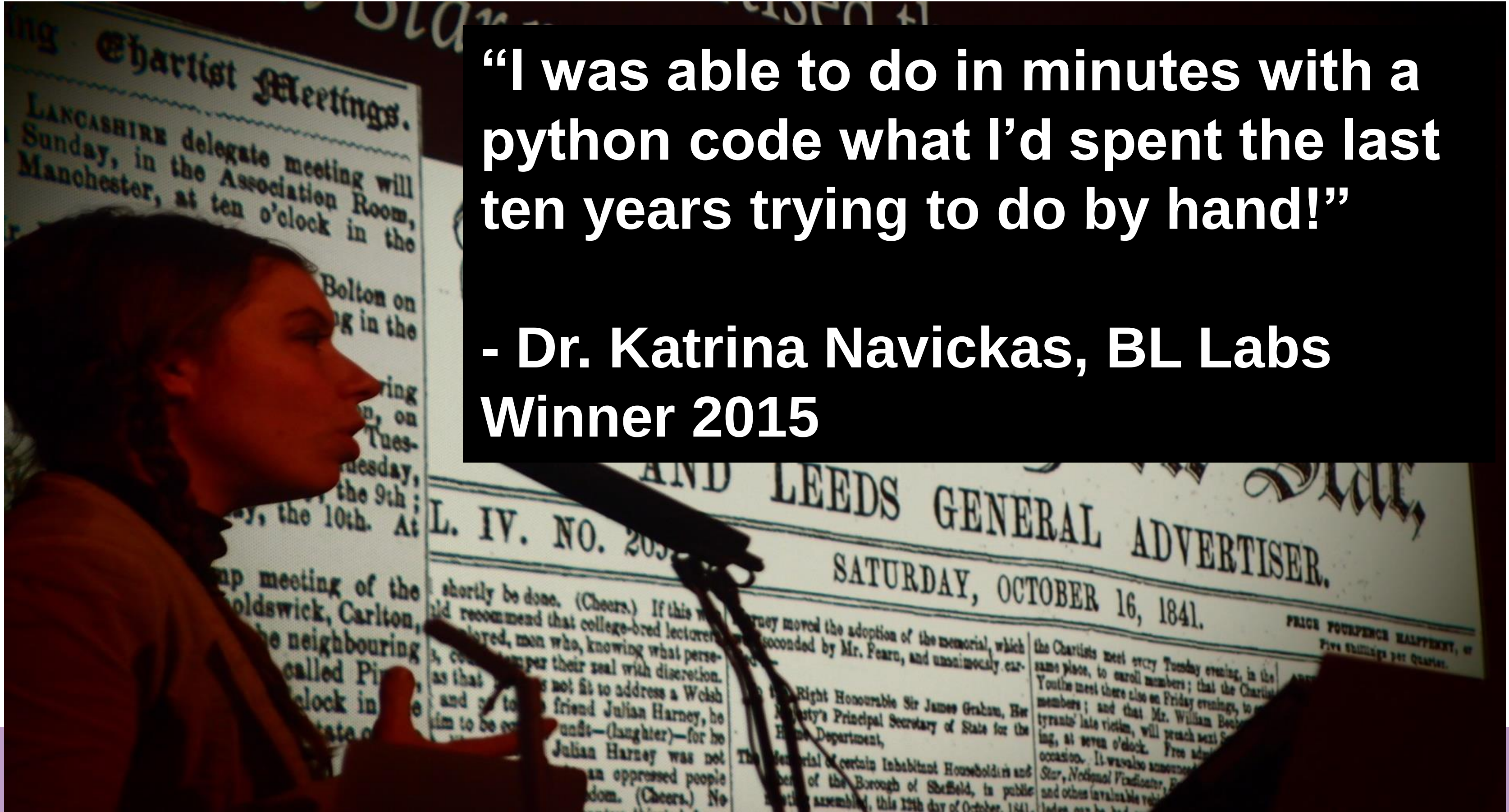
feather : 14.44
pen : 14.20
quill : 13.07
red-breasted merganser : 9.66
hand : 9.41
close : 9.35
people : 8.74
black : 8.36
business : 8.23
writing implement : 7.85
merganser : 7.73
paper : 7.70
pretty : 7.55
human : 7.45
man : 7.36
flower : 7.35
covering : 7.31

In partnership with

blbigdata.herokuapp.com

dx.doi.org/10.5281/zenodo.17168

Content is data: Political Meetings Mapper



“I was able to do in minutes with a python code what I’d spent the last ten years trying to do by hand!”

- Dr. Katrina Navickas, BL Labs Winner 2015

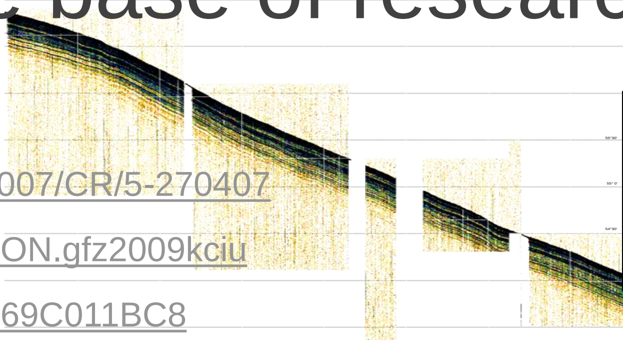
Modern science relies on good data

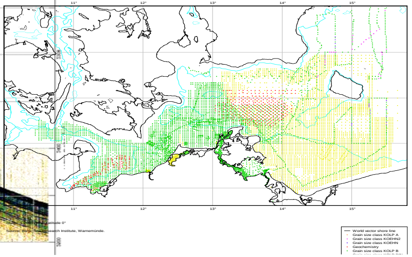
Data are the evidence base of research

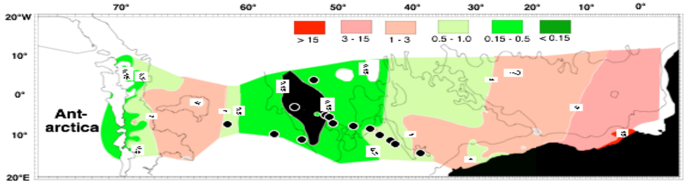
- Datasets are a crucial component of the scholarly record
- As the National Library, we have a responsibility to protect the nation’s scholarly record
- Datasets are essential to the British Library’s mission to advance the World’s knowledge

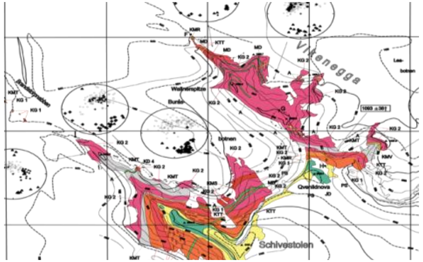
Data – the evidence base of research

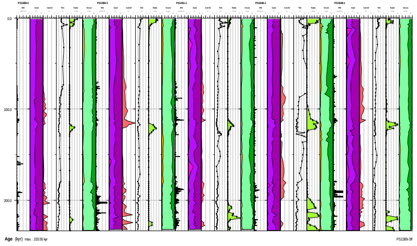
- Gravitational waves – doi:10.7914/SN/LI
- Medical case studies - doi:10.1594/eaacinet2007/CR/5-270407
- Earth quake events - doi:10.1594/GFZ.GEOFON.gfz2009kciu
- Computational models - doi:10.4225/02/4E9F69C011BC8
- Climate models - doi:10.1594/WDCC/dphase_mpeps
- Distributed samples - doi:10.1594/PANGAEA.51749
- Sea bed photos - doi:10.1594/PANGAEA.757741
- Audio records - doi:10.1594/PANGAEA.339110
- Grey Literature - doi:10.2314/GBV:489185967
- Videos - doi:10.3207/2959859860
- And much more

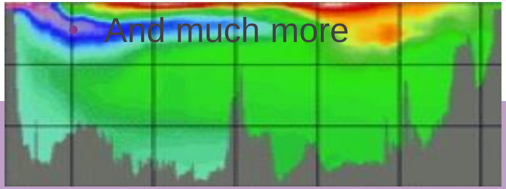


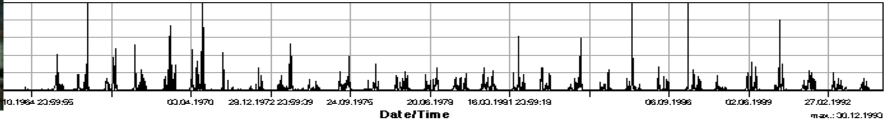












Mind the gaps between

- Articles and data
- Publishers and data
- People and data
- Reviews and data
- Repositories and data
- Today and tomorrow

Articles and data

In 2009:

- No effective way to link between articles and datasets
- No widely used method to identify datasets
- No widely used method to cite datasets

Data citation

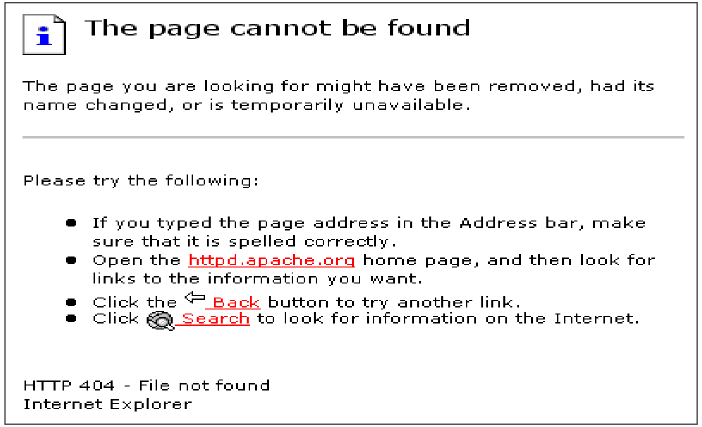
- The use of published digital data, like the use of digitally published literature, depends upon the ability to identify, authenticate, locate, access, and interpret them.
- Data citations provide necessary support for these functions, as well as other functions such as attribution of credit and establishment of provenance.

Data Science Journal, Sep 2013

Data citation

URLs are not persistent

(e.g. Wren JD: **URL decay in MEDLINE- a 4-year follow-up study**. Bioinformatics. 2008, Jun 1;24(11):1381-5).



The page cannot be found

The page you are looking for might have been removed, had its name changed, or is temporarily unavailable.

Please try the following:

- If you typed the page address in the Address bar, make sure that it is spelled correctly.
- Open the httpd.apache.org home page, and then look for links to the information you want.
- Click the Back button to try another link.
- Click Search to look for information on the Internet.

HTTP 404 - File not found
Internet Explorer

Digital Object Identifiers (DOIs) offer a solution

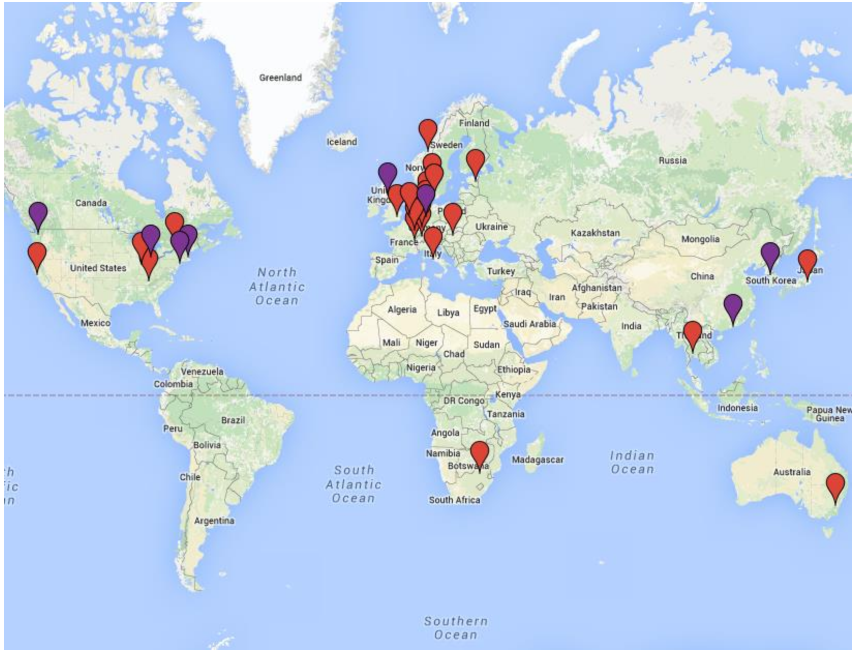
A DOI is a unique identifier, similar in concept to an ISBN.

- Mostly widely used identifier for scientific articles
- Researchers, authors, publishers know how to use them
- Put datasets on the same playing field as articles

Dataset
Yancheva et al (2007). Analyses on sediment of Lake Maar. PANGAEA. doi:10.1594/PANGAEA.587840

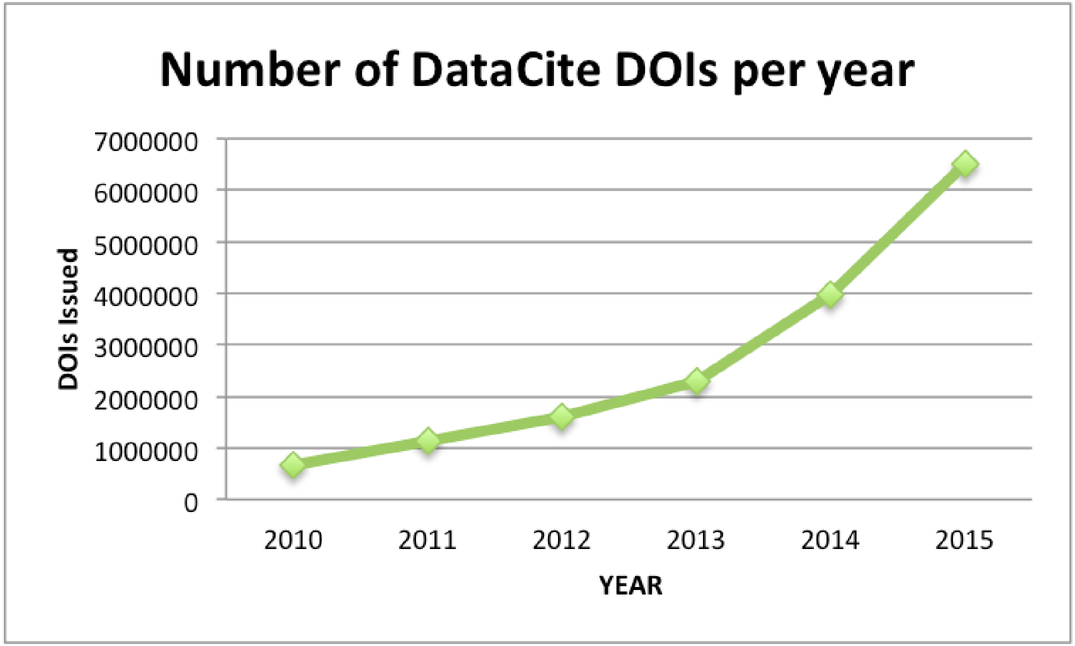
DataCite

- Makes research better by enabling people to find, share, use, and cite data
- A leading global membership organization offering reliable persistent data identification
- Engages researchers, scholars, data centres, libraries, publishers, and funders through advocacy, guidance and services



Use of DOIs for data is growing rapidly

- 27 DataCite Members
- Total data centers: 635 up 209 in 2015
- Total DOIs: 6,168,989 up 2,571,770 in 2015
- 208,000 DOIs with an ORCID iDs



Publishers and data

- Publishers vary in support for data
  - Policies
  - Services

- Laurie Goodman’s (GigaScience) call to action for journal publishers:
  - Accept data
  - Include data citations references
  - Track in citation indexes
  - Encourage and enable metrics by the community…

- Rice 3k – 13.4TB of data, linked to data paper with basic information; citability reassured authors

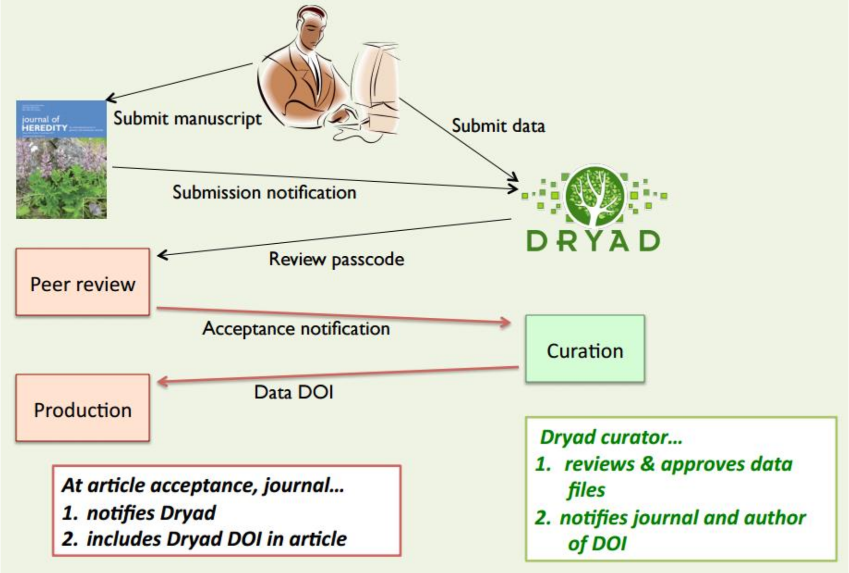
Reviewing and data

- As recently as 2000, some viewed data publication as unethical!
- Data review
  - Essential for validity
  - Viewed as important
  - Practice daunting and variable
- Publishers and systems starting to support and integrate

- Callaghan (doi:10.1045/january2015-callaghan)
  - Examined 7 datasets in earth sciences from DataCite MDs
  - Only 2 of 7 met basic checklist
  - Major challenges
    - Accessibility
    - Human readable metadata
    - Adequate metadata
    - Dead links to related information

Dryad Digital Repository

- A curated resource that makes the data underlying scientific publications discoverable, freely reusable, and citable (datadryad.org)
- Coordinates manuscript and data submission
- Provides access to data during review
- Allows choice of repositories
- Assigns DOIs to data and includes into article
- Includes manual curation process



People and data

- Gives credit where it is due
- Re-assures data creators
- Stimulates data sharing
- Connects researchers to all of their outputs
- Helps researchers build their profiles

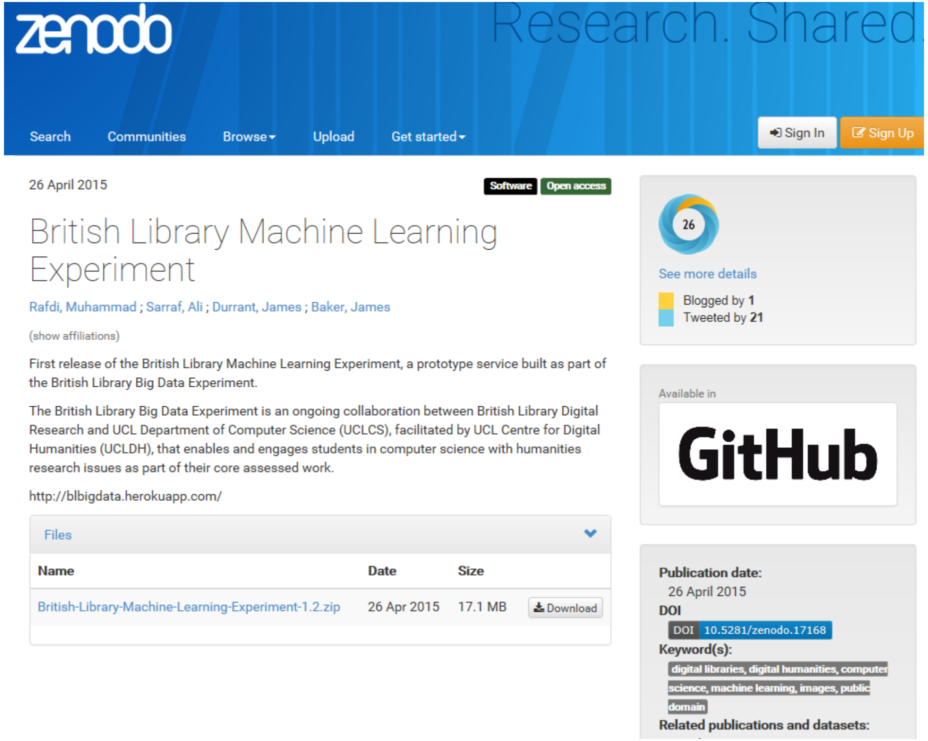
ORCID

- ORCID is an open, non-profit community driven effort to create and maintain a registry of unique researcher identifiers
- ORCID identifiers aim to disambiguate author names and to enable all relevant publications and outputs to be associated with individual researchers
- Over 208,000 DataCite DOIs connect to an ORCID identifier
- Automated integrations with publishers (e.g. Thomson Reuters and Elsevier) and research data registries (e.g. ANDS, DataCite)
- Over one million ORCID identifiers have been issued since its launch in October 2012
- Over 1000 journals are using ORCID in manuscript submission systems

Software and Data - Zenodo and github

- Zenodo features
  - Research. Shared. — all research outputs from across all fields of science are welcome!
  - Citeable. Discoverable. — uploads gets a Digital Object Identifier (DOI) to make them easily and uniquely citeable.
  - Community Collections — accept or reject uploads to your own community collections
  - Funding — integrated in reporting lines for research funded via OpenAIRE
  - Flexible licensing —not everything is Creative Commons
  - Safe — outputs stored in the same cloud infrastructure as data from CERN's Large Hadron Collider

- GitHub integration
  - Archives software releases and assigns a DOI
  - Enables you to cite GitHub repositories in academic literature



26 April 2015

British Library Machine Learning Experiment

Rafdi, Muhammad ; Sarraf, Ali ; Durrant, James ; Baker, James

(show affiliations)

First release of the British Library Machine Learning Experiment, a prototype service built as part of the British Library Big Data Experiment.

The British Library Big Data Experiment is an ongoing collaboration between British Library Digital Research and UCL Department of Computer Science (UCLCS), facilitated by UCL Centre for Digital Humanities (UCLDH), that enables and engages students in computer science with humanities research issues as part of their core assessed work.

http://blbigdata.herokuapp.com/

Files

Name	Date	Size
British-Library-Machine-Learning-Experiment-1.2.zip	26 Apr 2015	17.1 MB

Publication date:
26 April 2015

DOI
DOI 10.5281/zenodo.17168

Keyword(s):
digital libraries, digital humanities, computer science, machine learning, images, public domain

Related publications and datasets:

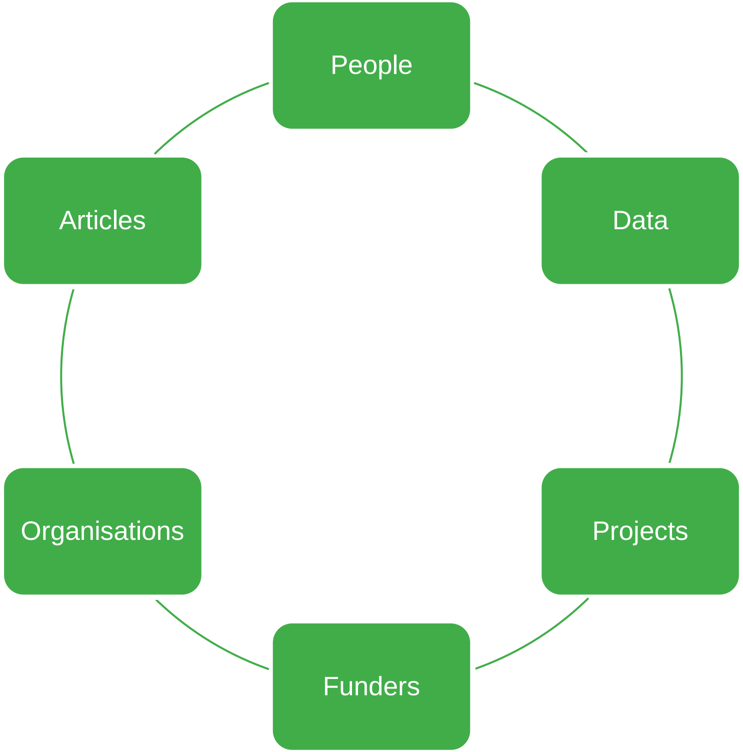
Software and data

- Data-driven research methods use complex software pipelines
  - Even the simplest rely on complex applications
- Software
  - Has complex dependencies
  - Has bugs
  - Can be configured incorrectly
  - Can be invoked incorrectly
- Researchers
  - Are often artisanal software creators
  - Rarely have formal training in software engineering or development
  - Are in a hurry to get results
  - Are happy if it works once – on their machine
  - Get no credit for creating scripts
  - Have tiny user communities for their code
  - Have complex workflows of which software may be a small part

Software and data

- Cultural and technical trends in software development are in our favour
  - Better dependency management for software
  - Better testing and integration
  - Better version control
  - More collaborative open development
  - Developers hate the ‘it worked on my machine’ bugs
- Computational cloud culture
  - Cheaper for research groups
  - Better for collaborative projects
  - Better for scaling to large data sets
  - Requires running code in different environments
    - Local, single machine, multi-machine
- Virtual machines and related technology work
  - Fit researcher workflow
  - Make it easier to develop and debug
  - Must manage dependencies for commercial environments
- Virtual machines let you bring computation to data
  - Data is too big to download
  - Data protection, licensing issues get in the way, too
- How can virtual machines last?
  - Is there a Universal Virtual Machine?
  - Is there a virtual machine designed for the longer term?
  - Can virtual machines be part of the scholarly record?

THOR: Making connections



- Persistent identifiers - link without doubt

- THOR project links
  - People (ORCID, ISNI)
  - Articles (DOI)
  - Data / research outputs (DOI)
  - Organisations (ISNI)
  - Funders (DOI)
  - Projects

- Basis for measuring impact of research, encourages data sharing and reuse, enhances trust

- Infrastructure that spans collaborative big-data to breakthroughs in the long tail of research

- Partners: BL, DataCite, ORCID, CERN, EBI, PANGAEA, ANDS, Dryad, Elsevier, PLoS

- Funded under H2020 http://project-thor.eu

Technical and Human Infrastructure for Open Research
http://project-thor.eu

Our goal is to ensure that every researcher, at any phase of their career, or at any institution, will have seamless access to Persistent Identifiers (PIDs) for their research artifacts and their work will be uniquely attributed to them

Oh, and the scholarly record is data

- Research library content is big data
  - Volume – hundreds of terabytes
  - Variety – every format, every language, every subject, every decade
  - Velocity – grows by billions (web included) per year
- Repositories hold
  - Documents, recordings, archives, manuscripts, images, audio, video, web
  - Articles, news, books, archives, sheet music, maps, …
  - Metadata including names, organisations, and references
  - Research data
- Opportunities and challenges
  - Feature extraction
  - Understanding collections
  - Analysing textual and multi-modal data
  - Abstraction, summarisation, comparison
- Improved automation
  - Can accelerate research progress
  - Drive new developments in data science
  - Measure impact of research funding
  - Understand interdisciplinary activity
  - Understand the advancement or development of research areas
  - …

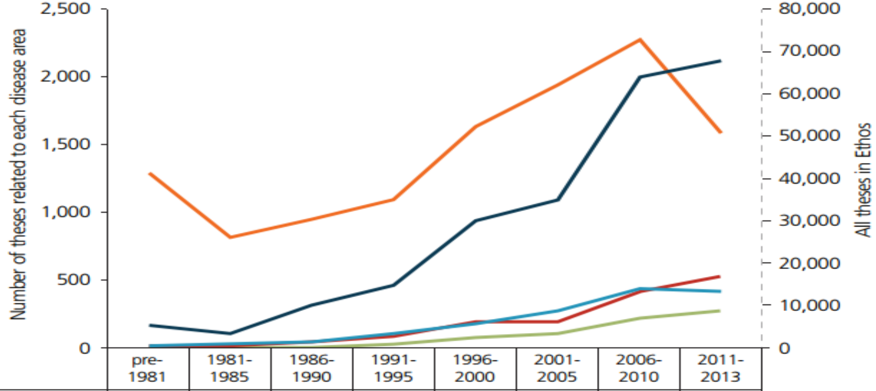
Data mining UK theses

- Aggregation of UK doctoral (PhD) theses
  - EThOS – http://ethos.bl.uk
  - 430,000 records – 90% of all PhDs
  - 170,000 full-text theses (40% of all records)
  - 20,000 awarded a year, avg 300 pages, 6m pages/year
- Good quality, accurate, consistent, de-duplicated metadata
- Metadata includes:
  - Author, title, year, university name
  - Abstract (for 90%)
  - Supervisor names, funder/sponsor body
  - A few DOI and ORCiD identifiers
  - Subject discipline
  - Via OAI-PMH or download

Alzheimer’s Society & RAND Europe

- Mapping the UK’s Dementia Research Landscape
- Workforce pipeline
- Tracked PhD to senior research
- 1/5 dementia PhD graduates remain in dementia research
- 70% leave dementia research within 4 years of completing PhD
- Used EThOS metadata to analyse trend

Figure 16. Number of theses published over time (and indexed in the EThOS database) overall, and in dementia, cancer, CHD and stroke.



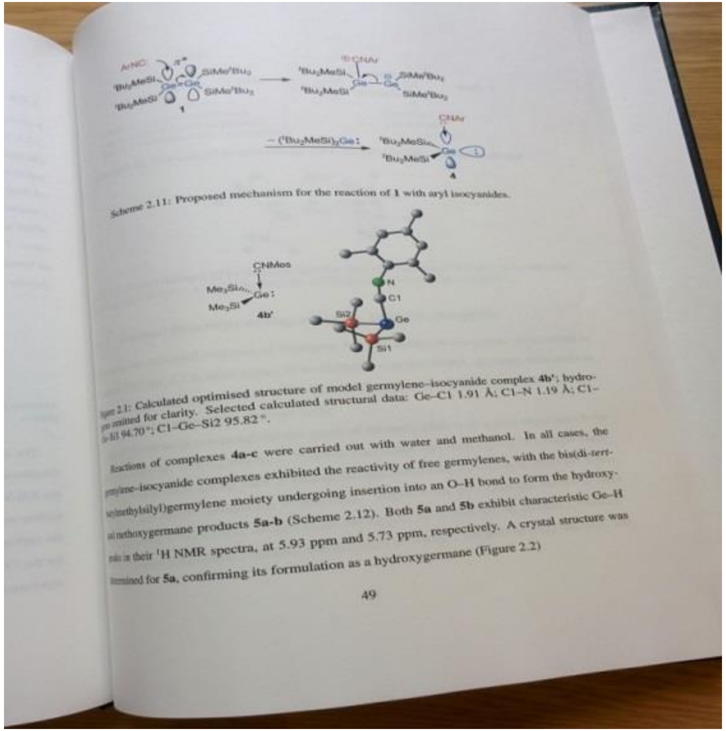
	pre-1981	1981-1985	1986-1990	1991-1995	1996-2000	2001-2005	2006-2010	2011-2013
Dementia	16	13	40	108	183	276	440	424
Cancer	175	114	316	457	946	1094	2015	2127
CHD	36	22	49	85	194	202	415	521
Stroke	7	14	10	31	73	98	227	276
All Ethos theses	41,481	26,217	30,354	34,883	52,428	62,002	72,593	50,902

Source: EThOS at the British Library.

http://britishlibrary.typepad.co.uk/science/2015/09/a-novel-use-of-phd-data.html

National Compound Collection

- Are there useful molecules in PhD theses?
- Extract the compounds; re-draw in ChemDraw; input into ChemSpider
- Bristol Uni & Royal Society Chemistry
- Manual pilot – could process be automated?
- Used full-text theses “likely to reveal new compounds”
- 50% were new compounds



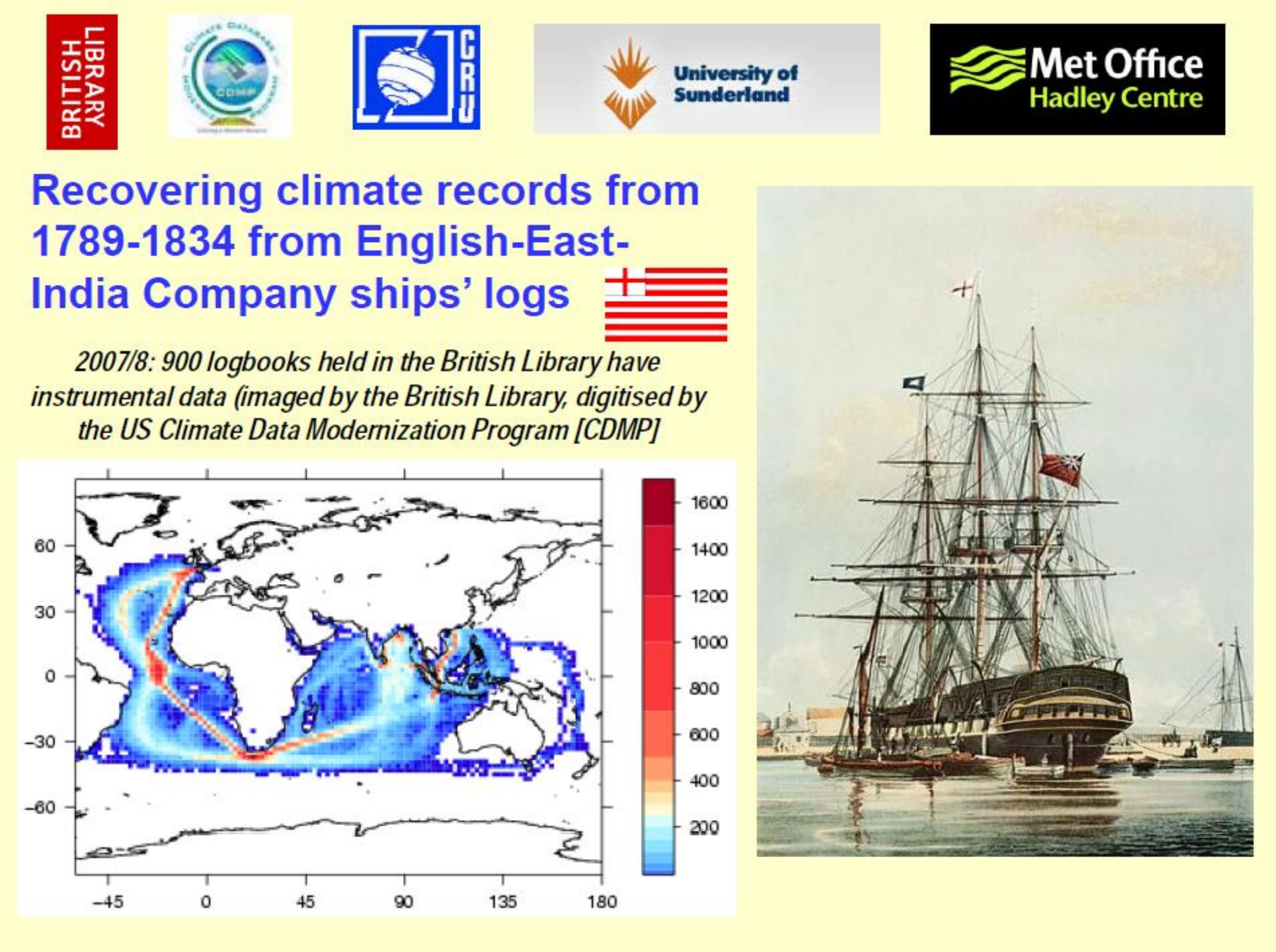
ArNC
$^tBu_2MeSi$
$^tBu_2MeSi$
Ge=Ge
$SiMe^tBu_2$
$SiMe^tBu_2$
1
CNAr
Ge
4
$-(^tBu_2MeSi)_2Ge:$
CNMes
Me_3Si
Me_3Si
Ge:
4b'
N
C1
Si2
Ge
Si1

Scheme 2.11: Proposed mechanism for the reaction of 1 with aryl isocyanides.

gure 2.1: Calculated optimised structure of model germylene–isocyanide complex **4b'**; hydro- omitted for clarity. Selected calculated structural data: Ge–C1 1.91 Å; C1–N 1.19 Å; C1–Si1 94.70°; C1–Ge–Si2 95.82°.

actions of complexes **4a-c** were carried out with water and methanol. In all cases, the ylene-isocyanide complexes exhibited the reactivity of free germylenes, with the bis(di-tert- thylsilyl)germylene moiety undergoing insertion into an O–H bond to form the hydroxy- methoxygermane products **5a-b** (Scheme 2.12). Both **5a** and **5b** exhibit characteristic Ge–H in their ^{1}H NMR spectra, at 5.93 ppm and 5.73 ppm, respectively. A crystal structure was mined for **5a**, confirming its formulation as a hydroxygermane (Figure 2.2)

49



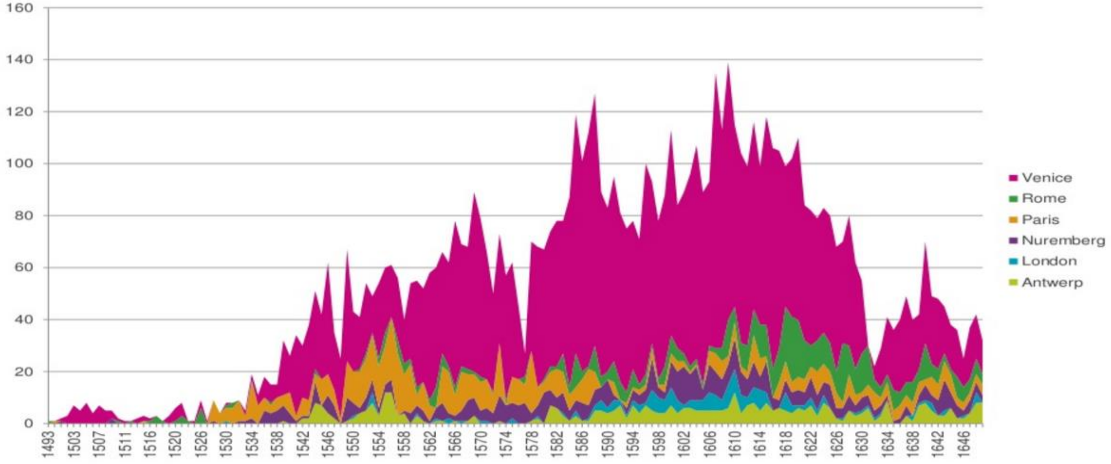
Recovering climate records from 1789-1834 from English-East-India Company ships' logs

2007/8: 900 logbooks held in the British Library have instrumental data (imaged by the British Library, digitised by the US Climate Data Modernization Program [CDMP]

Big Data History of Music

How can bibliographic data held by research libraries be unlocked for music researchers?

Can this data be interrogated in ways that challenge the traditional narratives of music history?



Analyses showed new patterns such as the rise and fall of music printing in 16th- and 17th-century Europe (huge dips in output in Venice were due to plague and war)

Perspective on forever

- At a National Library, forever is a long time – hundreds of years
- For a venerable research university?
- For a major facility, like LHC?
- For a data centre?
- For a professor?
- For a project?
- For an early career researcher?
- For a PhD student?

- Value is neither a simple function of time, nor easy to predict – like antiques
- Memory institutions, Libraries, Archives aren’t bad at guessing the future value of information
  - And they can be pretty conservative
- The diamonds of research data should last forever
  - Gravitational waves
  - Higgs Boson
  - John Snow’s cholera map
  - Florence Nightingale’s army mortality graph
  - Ship logs?

British Library Labs Awards and Competition 2016

- Competition: Transformative ideas to use our digital collections and data – by11 April
- Awards: The best work done using our digital collections and data – by 5 Sep
- Find out more:
  http://labs.bl.uk/
  @BL_Labs

Sheffield, Bristol, Sussex events to go

http://www.bl.uk/projects/british-library-labs

Thanks!

- Publishers and journal editors
  - Include data citations in articles
  - Include data in review process
- Researchers
  - Anticipate sharing and reuse
  - Accept and give credit
- Memory institutions
  - Research = narrative + data + software
- Infrastructure providers
  - Include data, plan for software
  - Consider the long term

- Today, research data is not forever
- Scholarly infrastructure for print articles is imperfect, but robust
- Data and software must be woven into the fabric of scholarly communication
- There is a lot to do, but great partners to do it with
- Let’s get to work!