DCC because good research needs good data

# From My Data to Our Data

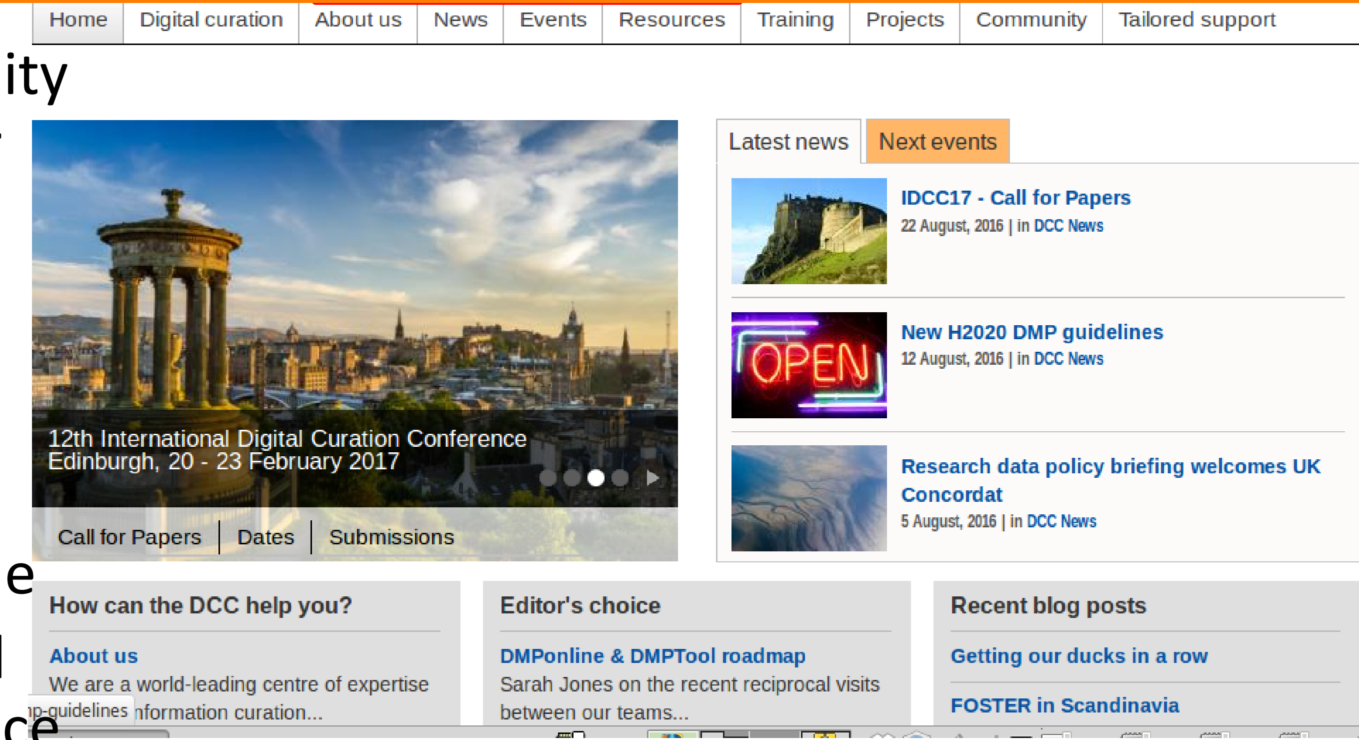
## Creating a culture of data reuse

**Kevin Ashley**
**Digital Curation Centre**
**www.dcc.ac.uk**
**@kevingashley**
**Kevin.ashley@ed.ac.uk**

# My home – the DCC

because good research needs good data

- 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



12th International Digital Curation Conference
Edinburgh, 20 - 23 February 2017

**IDCC17 - Call for Papers**
22 August, 2016 | in DCC News

**New H2020 DMP guidelines**
12 August, 2016 | in DCC News

**Research data policy briefing welcomes UK Concordat**
5 August, 2016 | in DCC News

**How can the DCC help you?**

**About us**
We are a world-leading centre of expertise ...nformation curation...

**Editor's choice**

**DMPonline & DMPTool roadmap**
Sarah Jones on the recent reciprocal visits between our teams...

**Recent blog posts**

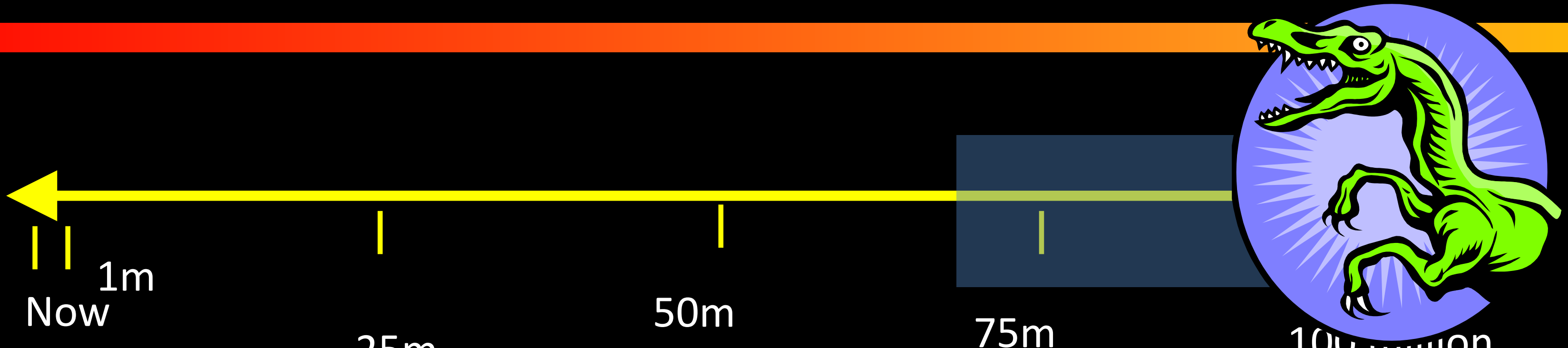
**Getting our ducks in a row**

**FOSTER in Scandinavia**

# Data reuse stories

- The palaeontologist who saved years of work with archaeological data

# What a paleontologist looks at



Now

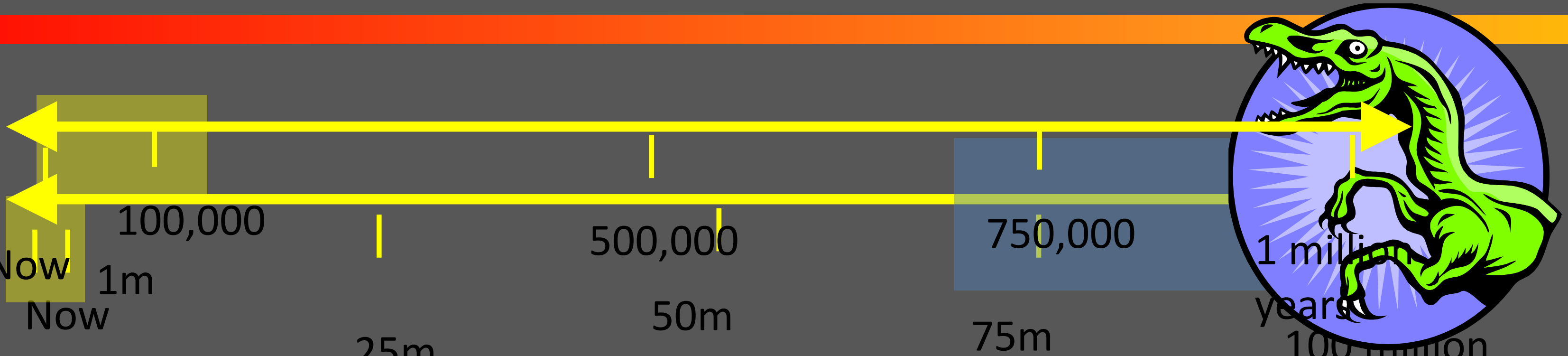
1m

25m

50m

75m

100 million years ago



Now

Now

100,000

1m

25m

500,000

50m

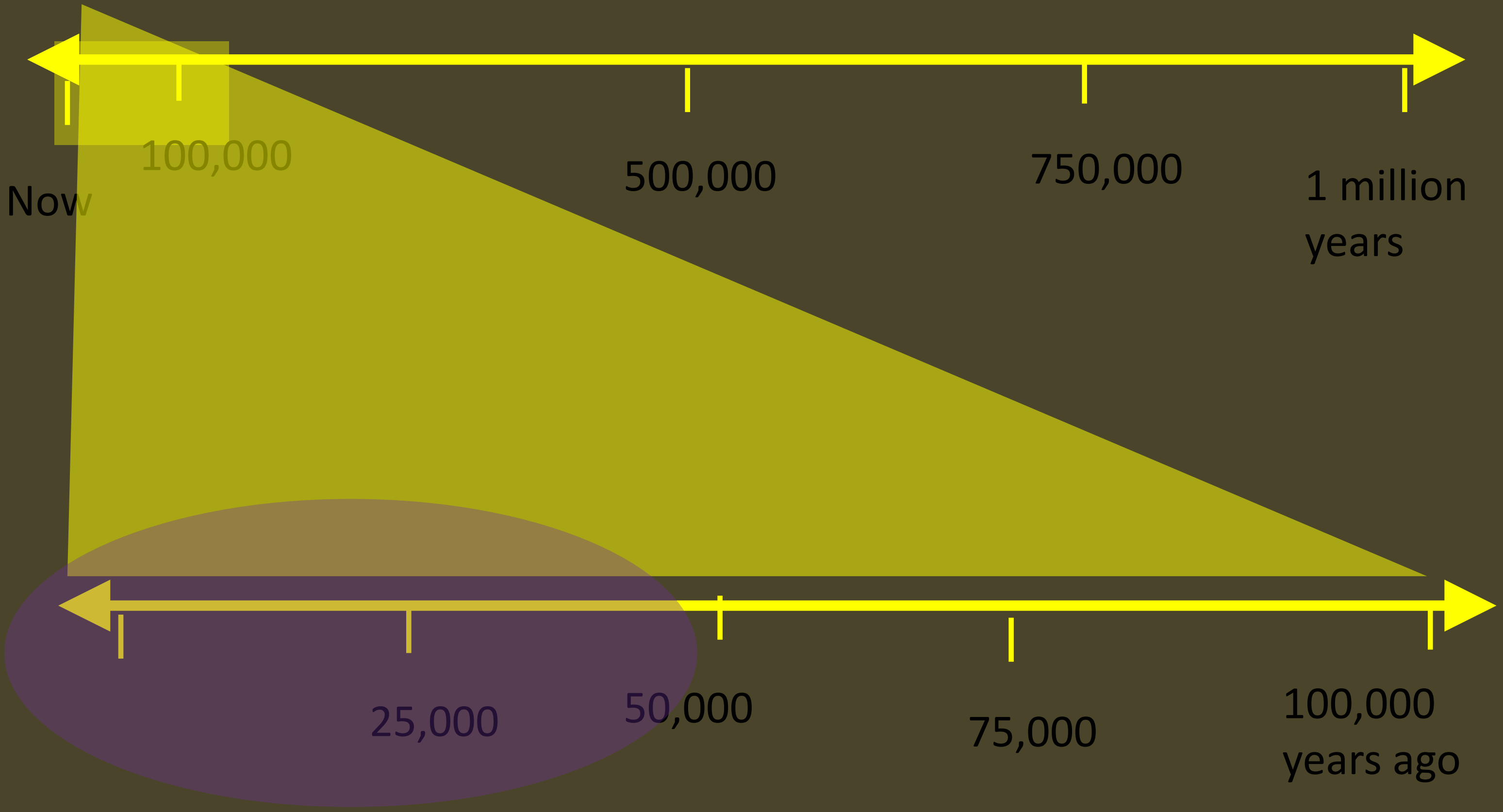
750,000

75m

1 million years

100 million years ago

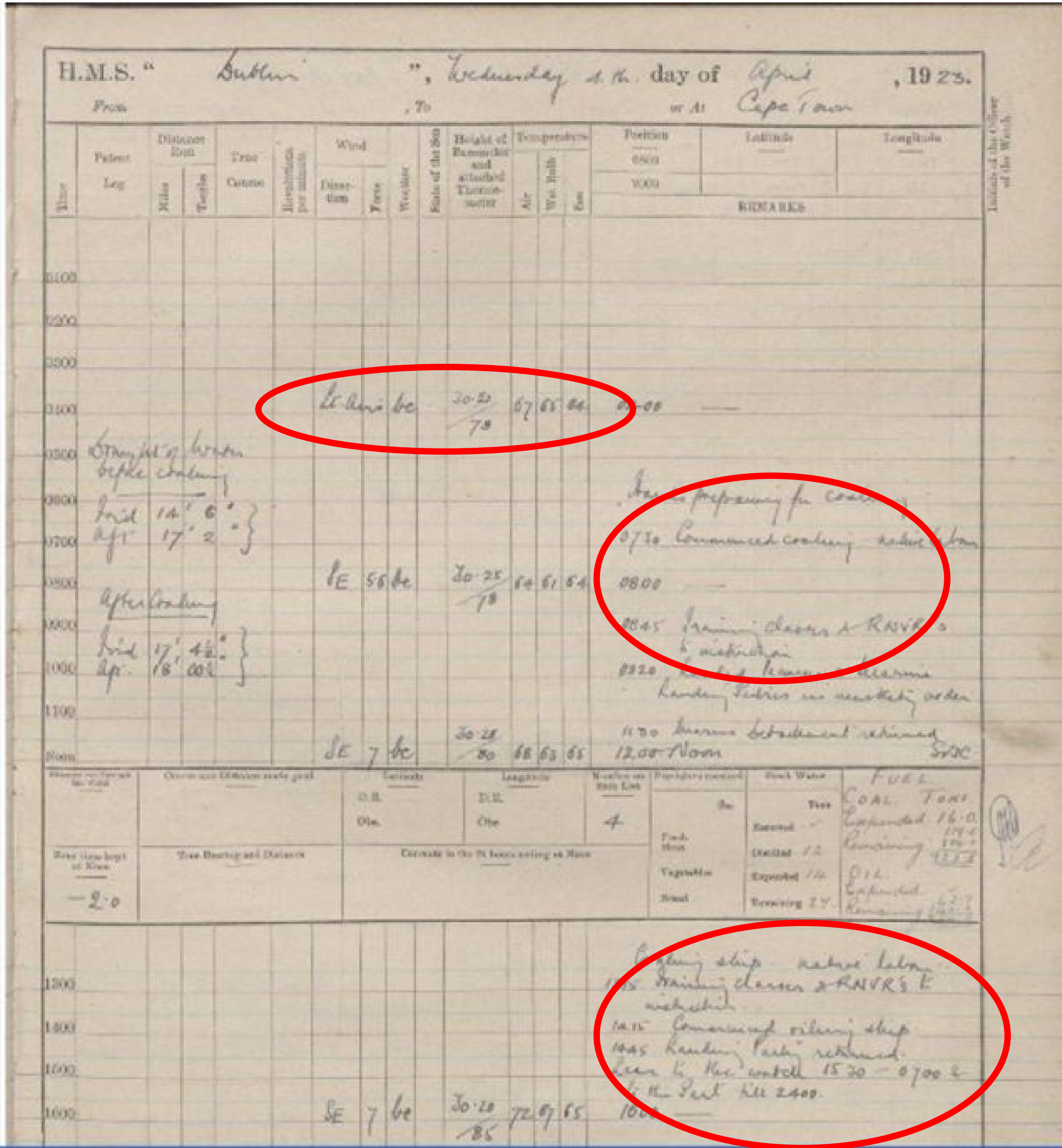
# What an archaeologist looks at



# Data reuse stories

- The palaeontologist who saved years of work with archaeological data
- The 19th-century ships logs that help us model climate change

# The Old weather project



H.M.S. " Dublin ", Wednesday 4th day of April , 1923.
From , To or At Cape Town

| Hour | Patent Log | Distance Run Miles | Distance Run Tenths | True Course | Revolutions per minute | Wind Direction | Wind Force | Weather | State of the Sea | Height of Barometer and attached Thermometer | Temperature Air | Temperature Wet Bulb | Temperature Sea | Remarks |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0100 | | | | | | | | | | | | | | |
| 0200 | | | | | | | | | | | | | | |
| 0300 | | | | | | | | | | | | | | |
| 0400 | | | | | | Lt. airs | | bc | | 30·21 / 73 | 67 | 65 | 64 | 0400 — |
| 0500 | Draught of water before coaling | | | | | | | | | | | | | |
| 0600 | Fwd 14' 6" | | | | | | | | | | | | | Hands preparing for coaling |
| 0700 | Aft 17' 2" | | | | | | | | | | | | | 0730 Commenced coaling native labour |
| 0800 | | | | | | E | 5 | bc | | 30·25 / 73 | 64 | 61 | 54 | 0800 — |
| 0900 | After Coaling | | | | | | | | | | | | | 0845 Training classes & RNVR's to instruction |
| 1000 | Fwd 17' 4½"<br>Aft 18' 00½" | | | | | | | | | | | | | 0920 Landed [illegible] Marine Landing Parties in marching order |
| 1100 | | | | | | | | | | | | | | 1130 Marine detachment returned |
| Noon | | | | | | SE | 7 | bc | | 30·28 / 80 | 68 | 63 | 65 | 12.00 Noon SRC |

| | | | | | | |
|---|---|---|---|---|---|---|
| Distance run through the Water | Course and Distance made good | Latitude<br>D.R.<br>Obs. | Longitude<br>D.R.<br>Obs. | Number on Sick List<br>4 | Provisions received<br>lbs.<br>Fresh Meat<br>Vegetables<br>Bread | Fresh Water<br>Tons<br>Received —<br>Distilled 12<br>Expended 14<br>Remaining 27 |
| Zone time kept at Noon<br>−2·0 | True Bearing and Distance | Currents in the 24 hours ending at Noon | | | | |

FUEL
COAL Tons
Expended 16·0
[illegible]
Remaining [illegible]
OIL
Expended
Remaining [illegible]

| Hour | Patent Log | Distance Run Miles | Distance Run Tenths | True Course | Revolutions per minute | Wind Direction | Wind Force | Weather | State of the Sea | Height of Barometer and attached Thermometer | Temperature Air | Temperature Wet Bulb | Temperature Sea | Remarks |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1300 | | | | | | | | | | | | | | Coaling ship native labour<br>1315 Training classes & RNVR's to instruction |
| 1400 | | | | | | | | | | | | | | 1415 Commenced oiling ship<br>1445 Landing Party returned |
| 1500 | | | | | | | | | | | | | | Leave to the watch 1530 – 0700 & to the Port till 2400. |
| 1600 | | | | | | SE | 7 | bc | | 30·20 / 85 | 72 | 67 | 65 | 1600 — |

Data for research, not from research

2017-03-10

# Data reuse stories

- The palaeontologist who saved years of work with archaeological data
- The 19th-century ships logs that help us model climate change
- The ‘noise’ from research radar that mapped dust from Eyjafjallajökull

# Data reuse - messages

Often your data tells stories that your publications do not

Discipline-bounded data discovery doesn’t give us all we need or want

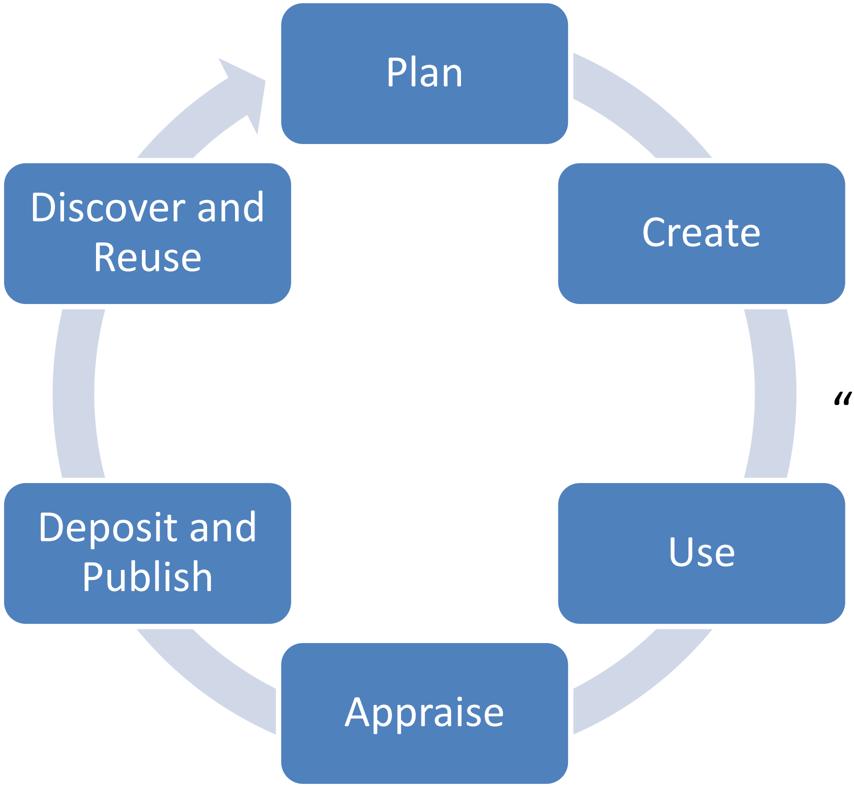
Not all data comes from other researchers

One person’s noise is another person’s signal

# What is data curation ?

- ▷ “Maintaining, preserving and adding value to research data throughout its lifecycle”
- ▷ More than preservation:
  - » Active management – dealing with change
- ▷ Less than preservation:
  - » Lifecycle sometimes involves destruction
- ▷ Sometimes, not always, about publication or citation
- ▷ Always about sharing in some way

# What is research data management?



"the active management and appraisal of data over the lifecycle of scholarly and scientific interest"

"an explicit process covering the creation and stewardship of research materials to enable their use for as long as they retain value."

**Data management is part of good research practice**

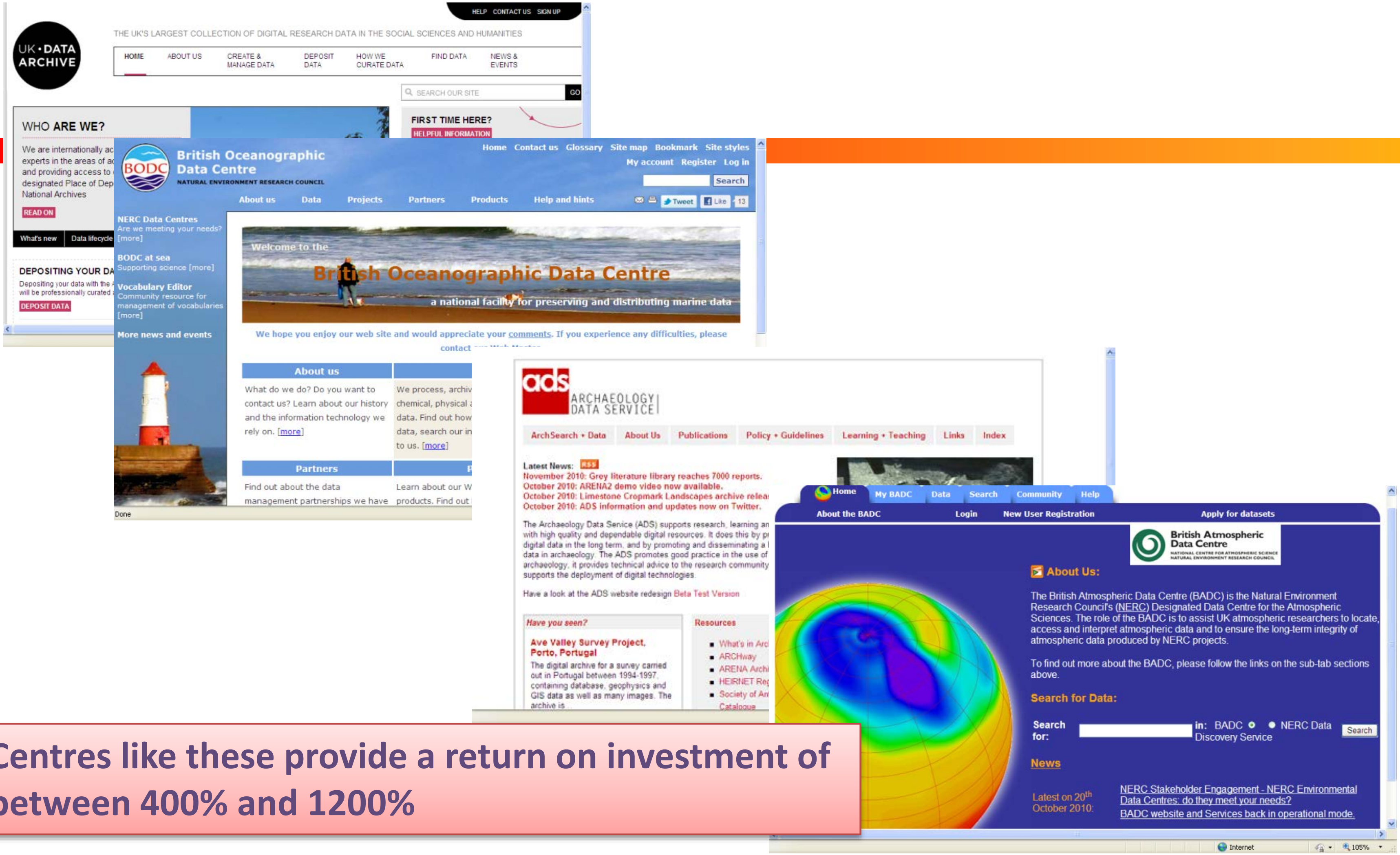
# Why care?

- Data is expensive – an investment
- Reuse:
  - More research
  - Teaching & Learning
  - Planning
- Impact – with or without publication
- Accountability
- Legal & regulatory requirements

# Why does this matter?

- 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



HELP CONTACT US SIGN UP

UK • DATA ARCHIVE

THE UK'S LARGEST COLLECTION OF DIGITAL RESEARCH DATA IN THE SOCIAL SCIENCES AND HUMANITIES

HOME ABOUT US CREATE & MANAGE DATA DEPOSIT DATA HOW WE CURATE DATA FIND DATA NEWS & EVENTS

SEARCH OUR SITE GO

WHO ARE WE?

We are internationally ac experts in the areas of ac and providing access to designated Place of Dep National Archives

READ ON

FIRST TIME HERE? HELPFUL INFORMATION

What's new Data lifecycle

DEPOSITING YOUR DA

Depositing your data with the will be professionally curated

DEPOSIT DATA

British Oceanographic Data Centre

NATURAL ENVIRONMENT RESEARCH COUNCIL

Home Contact us Glossary Site map Bookmark Site styles

My account Register Log in

Search

About us Data Projects Partners Products Help and hints

Tweet Like 13

NERC Data Centres
Are we meeting your needs? [more]

BODC at sea
Supporting science [more]

Vocabulary Editor
Community resource for management of vocabularies [more]

More news and events

Welcome to the

British Oceanographic Data Centre

a national facility for preserving and distributing marine data

We hope you enjoy our web site and would appreciate your comments. If you experience any difficulties, please contact

About us

What do we do? Do you want to contact us? Learn about our history and the information technology we rely on. [more]

We process, archiv chemical, physical data. Find out how data, search our in data to us. [more]

Partners

Find out about the data management partnerships we have

Learn about our W products. Find out

Done

ads ARCHAEOLOGY DATA SERVICE

ArchSearch • Data About Us Publications Policy • Guidelines Learning • Teaching Links Index

Latest News: RSS
November 2010: Grey literature library reaches 7000 reports.
October 2010: ARENA2 demo video now available.
October 2010: Limestone Cropmark Landscapes archive releas
October 2010: ADS information and updates now on Twitter.

The Archaeology Data Service (ADS) supports research, learning an with high quality and dependable digital resources. It does this by pr digital data in the long term, and by promoting and disseminating a data in archaeology. The ADS promotes good practice in the use of archaeology, it provides technical advice to the research community supports the deployment of digital technologies.

Have a look at the ADS website redesign Beta Test Version

Have you seen?

Ave Valley Survey Project, Porto, Portugal

The digital archive for a survey carried out in Portugal between 1994-1997, containing database, geophysics and GIS data as well as many images. The archive is ...

Resources

- What's in Arc
- ARCHway
- ARENA Archi
- HEIRNET Reg
- Society of An Catalogue

Home My BADC Data Search Community Help

About the BADC Login New User Registration Apply for datasets

British Atmospheric Data Centre

NATIONAL CENTRE FOR ATMOSPHERIC SCIENCE NATURAL ENVIRONMENT RESEARCH COUNCIL

About Us:

The British Atmospheric Data Centre (BADC) is the Natural Environment Research Council's (NERC) Designated Data Centre for the Atmospheric Sciences. The role of the BADC is to assist UK atmospheric researchers to locate, access and interpret atmospheric data and to ensure the long-term integrity of atmospheric data produced by NERC projects.

To find out more about the BADC, please follow the links on the sub-tab sections above.

Search for Data:

Search for: in: BADC ○ ● NERC Data Discovery Service Search

News

Latest on 20th October 2010: NERC Stakeholder Engagement - NERC Environmental Data Centres: do they meet your needs?
BADC website and Services back in operational mode.

Internet 105%

**Centres like these provide a return on investment of between 400% and 1200%**

http://www.jisc.ac.uk/whatwedo/programmes/di_directions/strategicdirections/badc.aspx

# Integrity – not without data

Bad Pharma
Ben Goldacre
Bestselling author of Bad Science
How drug companies mislead doctors and harm patients
384 pages
4th

dies on inte

ed 1976; n

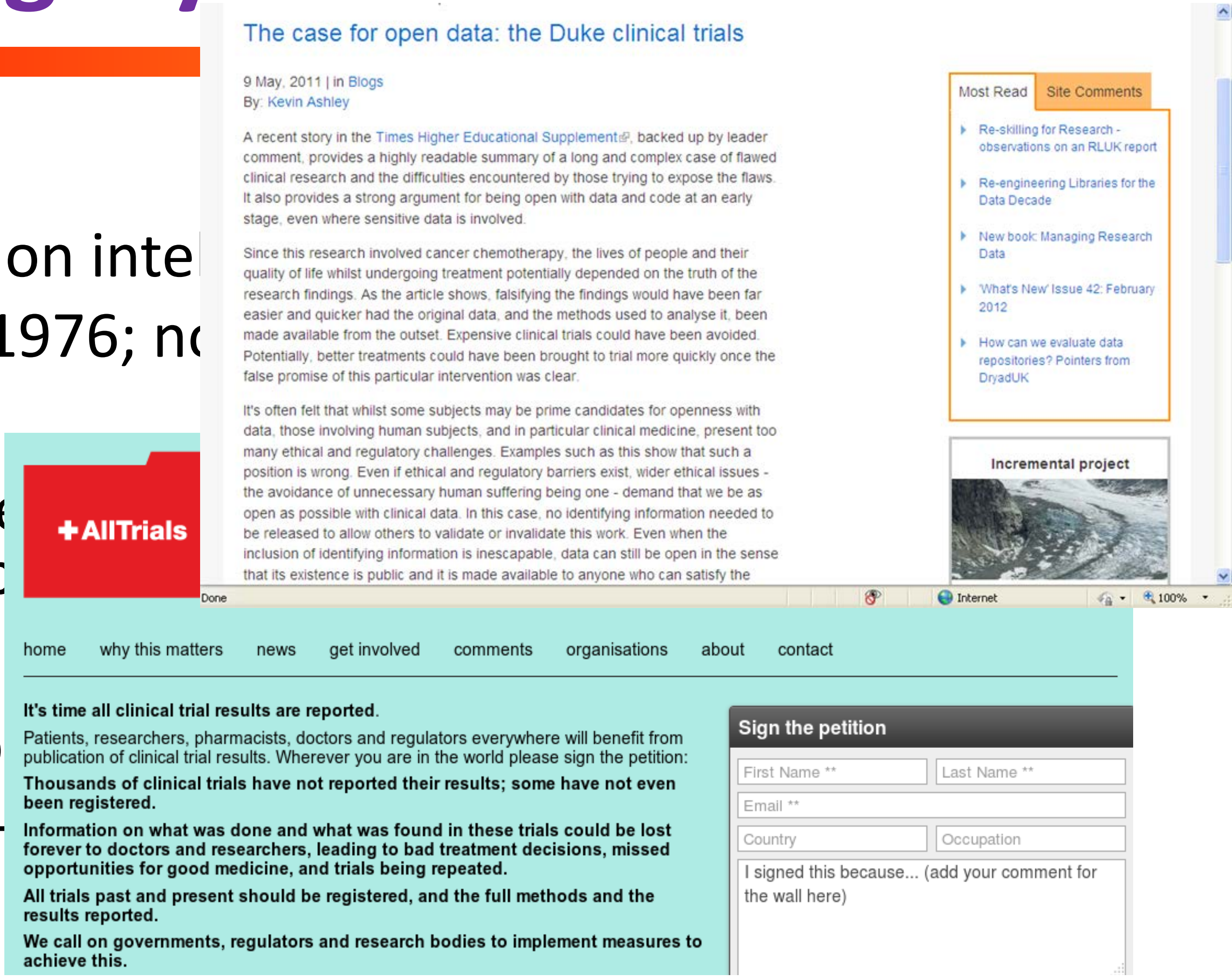
ng le

obab

es

Stapel – 55 p

» Poldermans -



## The case for open data: the Duke clinical trials

9 May, 2011 | in Blogs

By: Kevin Ashley

A recent story in the Times Higher Educational Supplement, backed up by leader comment, provides a highly readable summary of a long and complex case of flawed clinical research and the difficulties encountered by those trying to expose the flaws. It also provides a strong argument for being open with data and code at an early stage, even where sensitive data is involved.

Since this research involved cancer chemotherapy, the lives of people and their quality of life whilst undergoing treatment potentially depended on the truth of the research findings. As the article shows, falsifying the findings would have been far easier and quicker had the original data, and the methods used to analyse it, been made available from the outset. Expensive clinical trials could have been avoided. Potentially, better treatments could have been brought to trial more quickly once the false promise of this particular intervention was clear.

It's often felt that whilst some subjects may be prime candidates for openness with data, those involving human subjects, and in particular clinical medicine, present too many ethical and regulatory challenges. Examples such as this show that such a position is wrong. Even if ethical and regulatory barriers exist, wider ethical issues - the avoidance of unnecessary human suffering being one - demand that we be as open as possible with clinical data. In this case, no identifying information needed to be released to allow others to validate or invalidate this work. Even when the inclusion of identifying information is inescapable, data can still be open in the sense that its existence is public and it is made available to anyone who can satisfy the

Most Read
Site Comments
Re-skilling for Research - observations on an RLUK report
Re-engineering Libraries for the Data Decade
New book: Managing Research Data
'What's New' Issue 42: February 2012
How can we evaluate data repositories? Pointers from DryadUK
Incremental project
Done
Internet
100%
+AllTrials
home
why this matters
news
get involved
comments
organisations
about
contact

**It's time all clinical trial results are reported.**

Patients, researchers, pharmacists, doctors and regulators everywhere will benefit from publication of clinical trial results. Wherever you are in the world please sign the petition:

**Thousands of clinical trials have not reported their results; some have not even been registered.**

**Information on what was done and what was found in these trials could be lost forever to doctors and researchers, leading to bad treatment decisions, missed opportunities for good medicine, and trials being repeated.**

**All trials past and present should be registered, and the full methods and the results reported.**

**We call on governments, regulators and research bodies to implement measures to achieve this.**

**Sign the petition**

First Name **

Last Name **

Email **

Country

Occupation

I signed this because... (add your comment for the wall here)

"The case for open data: the Duke Clinical Trials "– blog post, Kevin Ashley, http://www.dcc.ac.uk/news/case-open-data-duke-clinical-trials

"Lies, Damned Lies and Research Data: Can Data Sharing Prevent Data Fraud?" – Doorn, Dillo, van Horik, IJDC 8(1); doi:10.2218/ijdc.v8i1.256

# Why manage research data –The selfish view

- To make research easier!
- To stop yourself drowning in irrelevant stuff
- In case you need the data later
- To avoid accusations of fraud or bad science
- To comply with the law or regulations
- To share data so others can use and learn from it
- To get credit for producing the data
- Because it’s a condition of research funding

# Data loss

Digital data are fragile and susceptible to loss for a wide variety of reasons

- Natural disaster
- Facilities infrastructure failure
- Storage failure
- Server hardware/software failure
- Application software failure
- Format obsolescence
- Legal encumbrance
- Human error
- Malicious attack
- Loss of staffing competencies
- Loss of institutional commitment
- Loss of financial stability
- Changes in user expectations

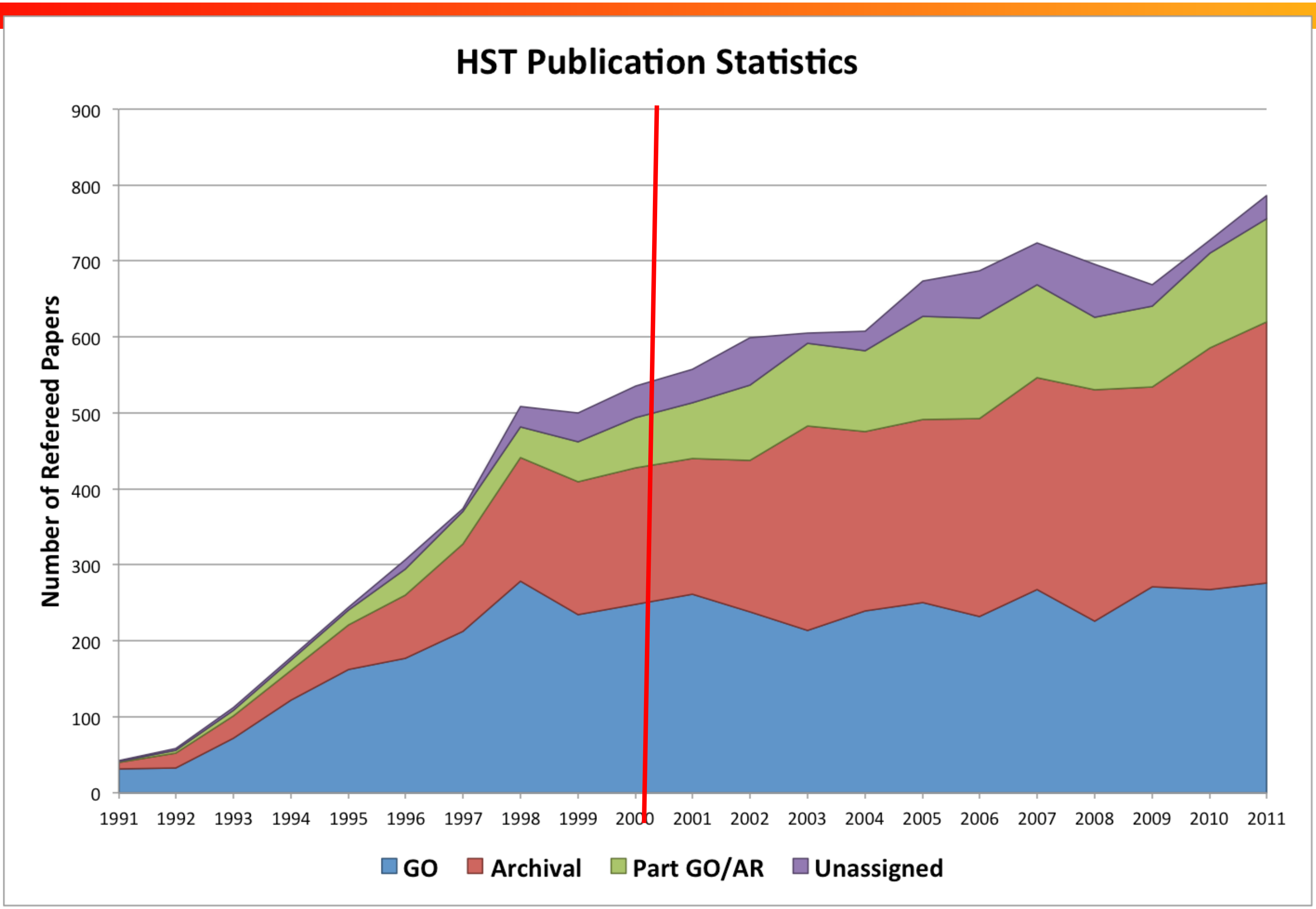
DRUCE
www.druce.com
3771
Very important!
LAPTOP LOST
in the bus 345
(silver macbook pro, lost in the bus, line 345, at South Kensington station on friday 16th at 7am, within a black bag also containing my ID card)
CRUCIAL scientific data
+ many YEARS of
research work inside!
Please bring it back to Imperial College (Prince consort road, between the museums and Hyde park) or CALL me to see how we can proceed:
020-75-94-11-55
or
0033-6-07-67-84-04
Very important!

Image CC BY-NC-SA 2.0 by Dave Hill https://www.flickr.com/photos/dmh650/4031607067

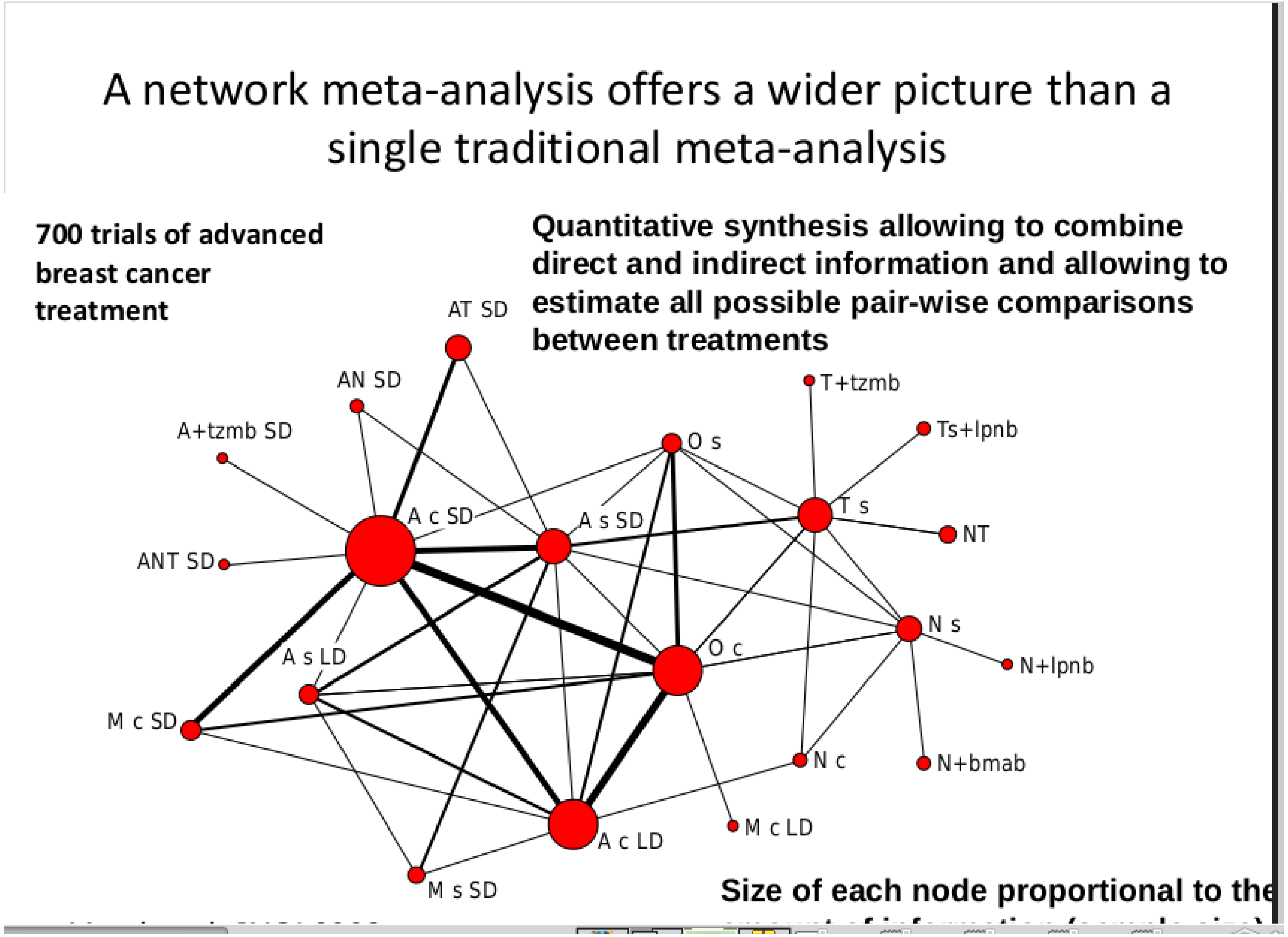
# Data is variable

- Not always textual
- Not always tabular
- Not always fixed – continual change
- Not always clearly authored – think of archival provenance
- Not always associated with publication
- Often with indistinct boundaries
- Multi-dimensional and non-linear

# Data reuse from Hubble



# New research with old data



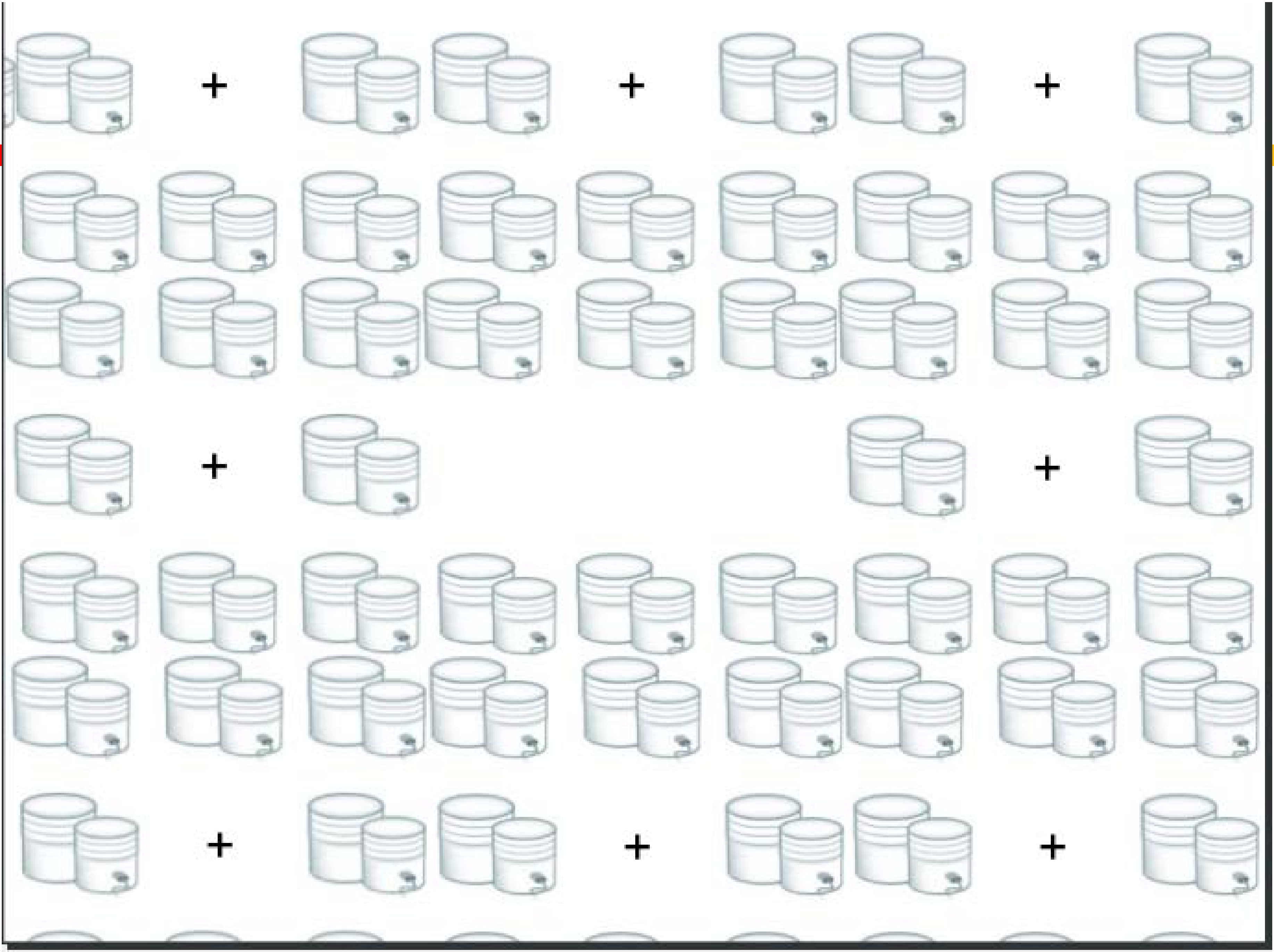
A network meta-analysis offers a wider picture than a single traditional meta-analysis

**700 trials of advanced breast cancer treatment**

**Quantitative synthesis allowing to combine direct and indirect information and allowing to estimate all possible pair-wise comparisons between treatments**

**Size of each node proportional to the amount of information (sample size)**

- ▷ Synthesis allows new analyses
- ▷ Research that cannot be done with any one of these datasets



# Make data citable

- Making data available increases citations
- Everyone – academic, funder, institution – loves citations
- Want evidence?
  - Alter, Pienta, Lyle – 240%, social sciences *
  - Piwowar, Vision – 9% (microarray data)†
  - Henneken, Accomazzi – 20% (astronomy) #

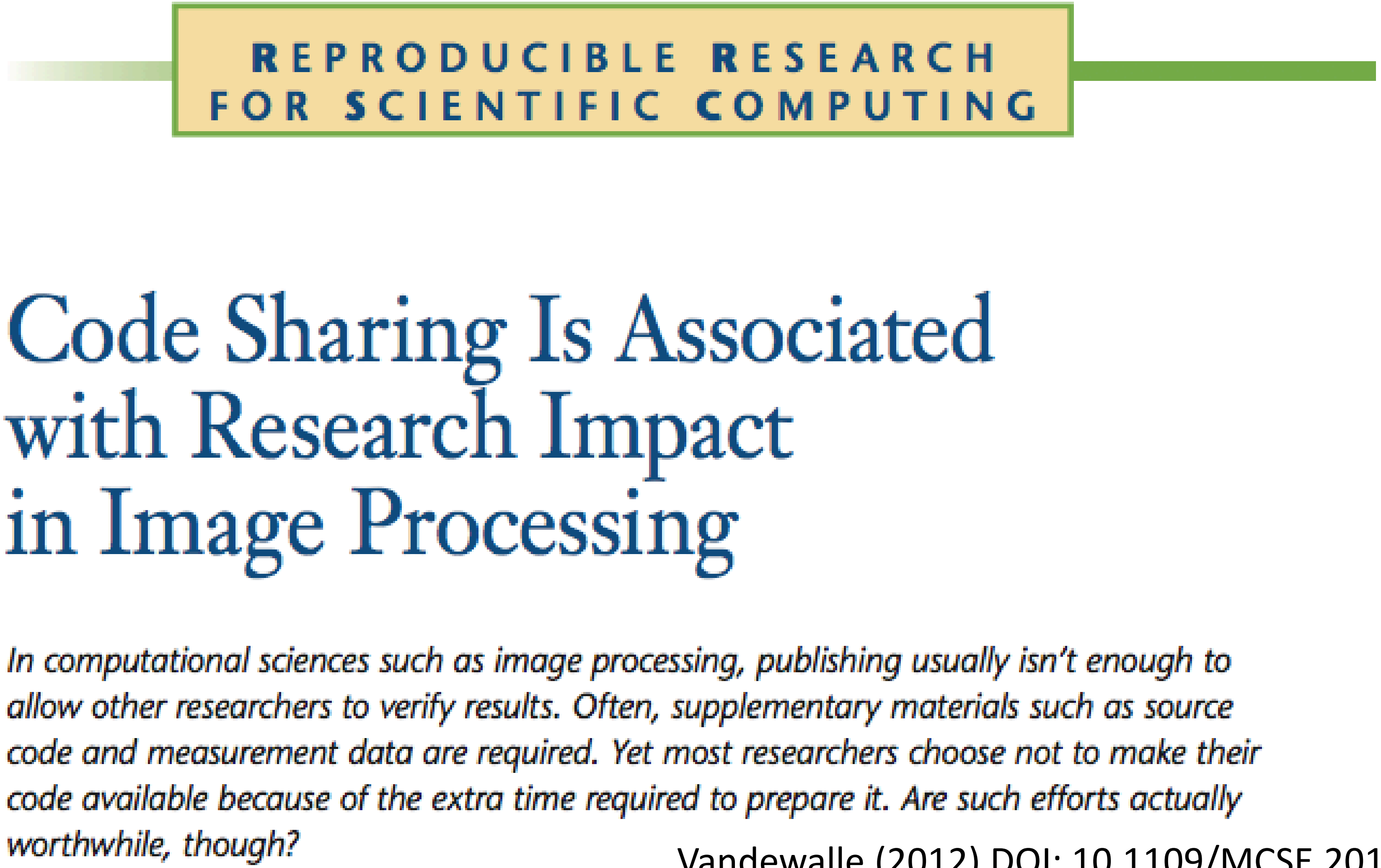
# Edwin Henneken, Alberto Accomazzi, (2011) Linking to Data - Effect on Citation Rates in Astronomy. http://arxiv.org/abs/1111.3618

* Amy Pienta, George Alter, Jared Lyle, (2010) The Enduring Value of Social Science Research: The Use and Reuse of Primary Research Data. http://hdl.handle.net/2027.42/78307

† Piwowar H, Vision TJ. (2013) Data reuse & the open data citation advantage. PeerJ PrePrints 1:e1v1 http://dx.doi.org/10.7287/peerj.preprints.1v1

Slide: Neil Chue Hong

# Improve your research impact

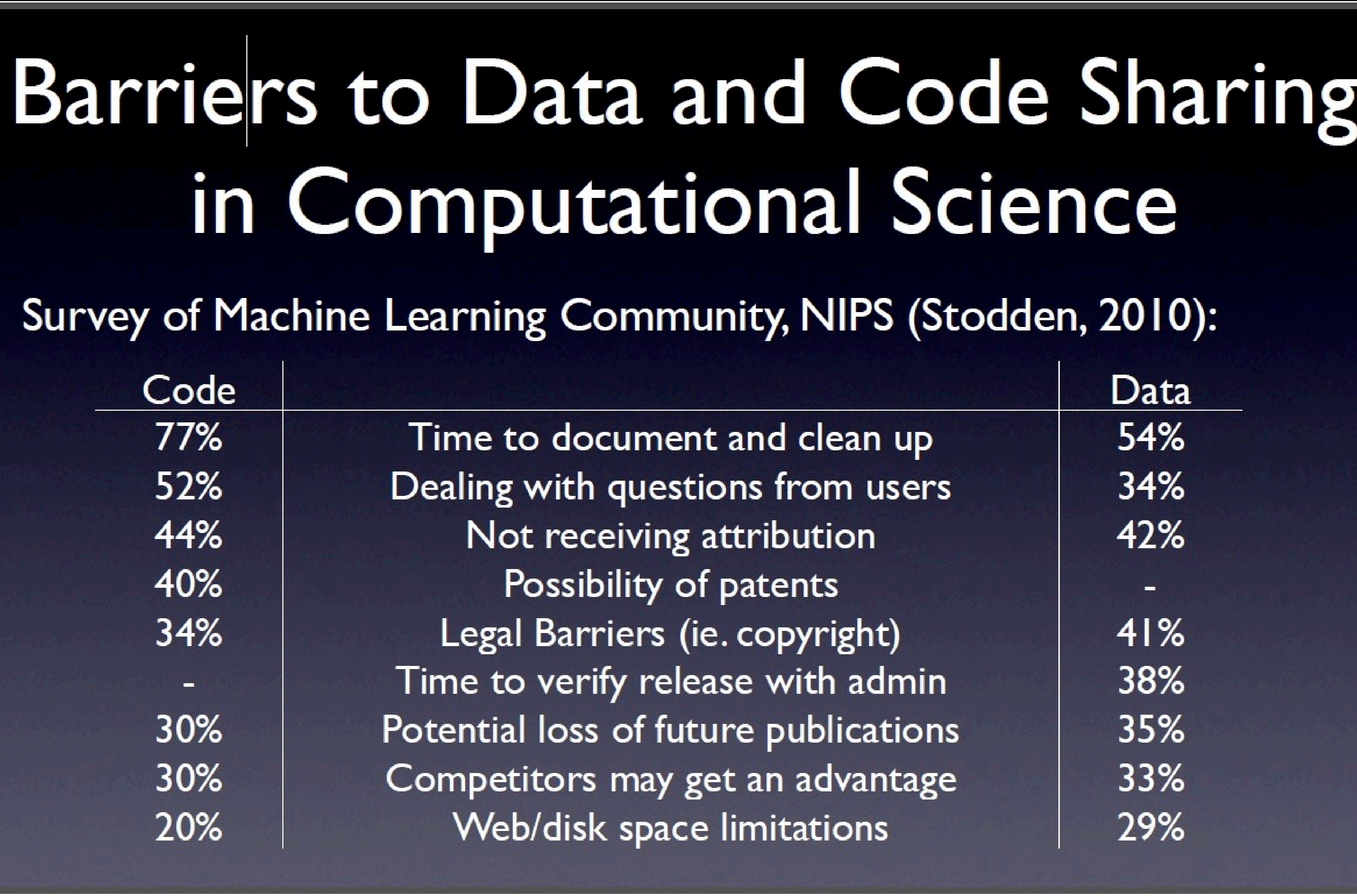


REPRODUCIBLE RESEARCH FOR SCIENTIFIC COMPUTING

## Code Sharing Is Associated with Research Impact in Image Processing

*In computational sciences such as image processing, publishing usually isn't enough to allow other researchers to verify results. Often, supplementary materials such as source code and measurement data are required. Yet most researchers choose not to make their code available because of the extra time required to prepare it. Are such efforts actually worthwhile, though?*

Vandewalle (2012) DOI: 10.1109/MCSE.2012.63



# Barriers to Data and Code Sharing in Computational Science

Survey of Machine Learning Community, NIPS (Stodden, 2010):

| Code | | Data |
|---|---|---|
| 77% | Time to document and clean up | 54% |
| 52% | Dealing with questions from users | 34% |
| 44% | Not receiving attribution | 42% |
| 40% | Possibility of patents | - |
| 34% | Legal Barriers (ie. copyright) | 41% |
| - | Time to verify release with admin | 38% |
| 30% | Potential loss of future publications | 35% |
| 30% | Competitors may get an advantage | 33% |
| 20% | Web/disk space limitations | 29% |

Victoria Stodden, AMP 2011 http://www.stodden.net/AMP2011/,

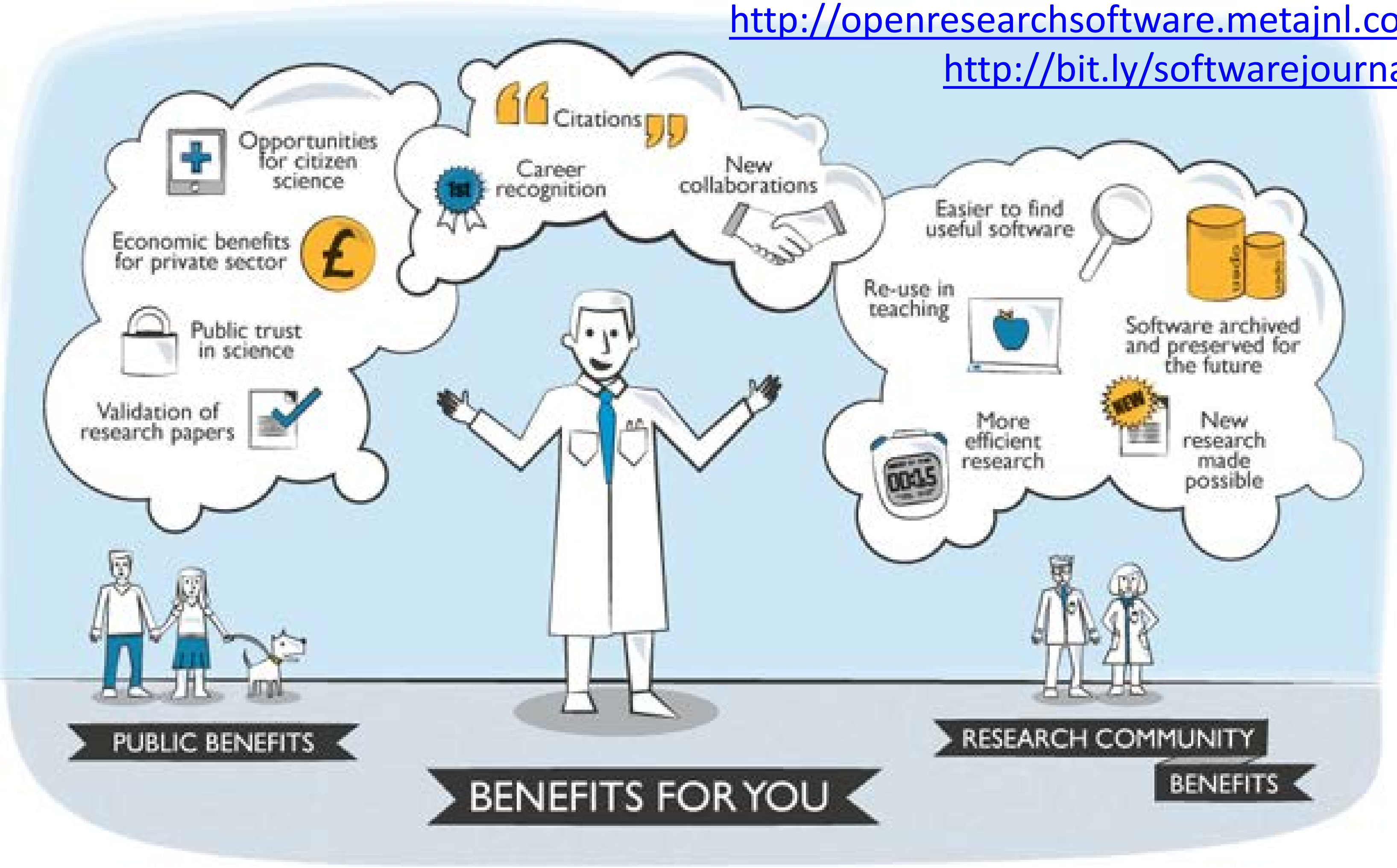
Special Issue Reproducible Research Computing in Science and Engineering July/August 2012, 14(4)

Howison and Herbsleb (2013) "Incentives and Integration In Scientific Software Production" CSCW 2013.

# Publishing data & software papers is easy

http://openresearchsoftware.metajnl.com

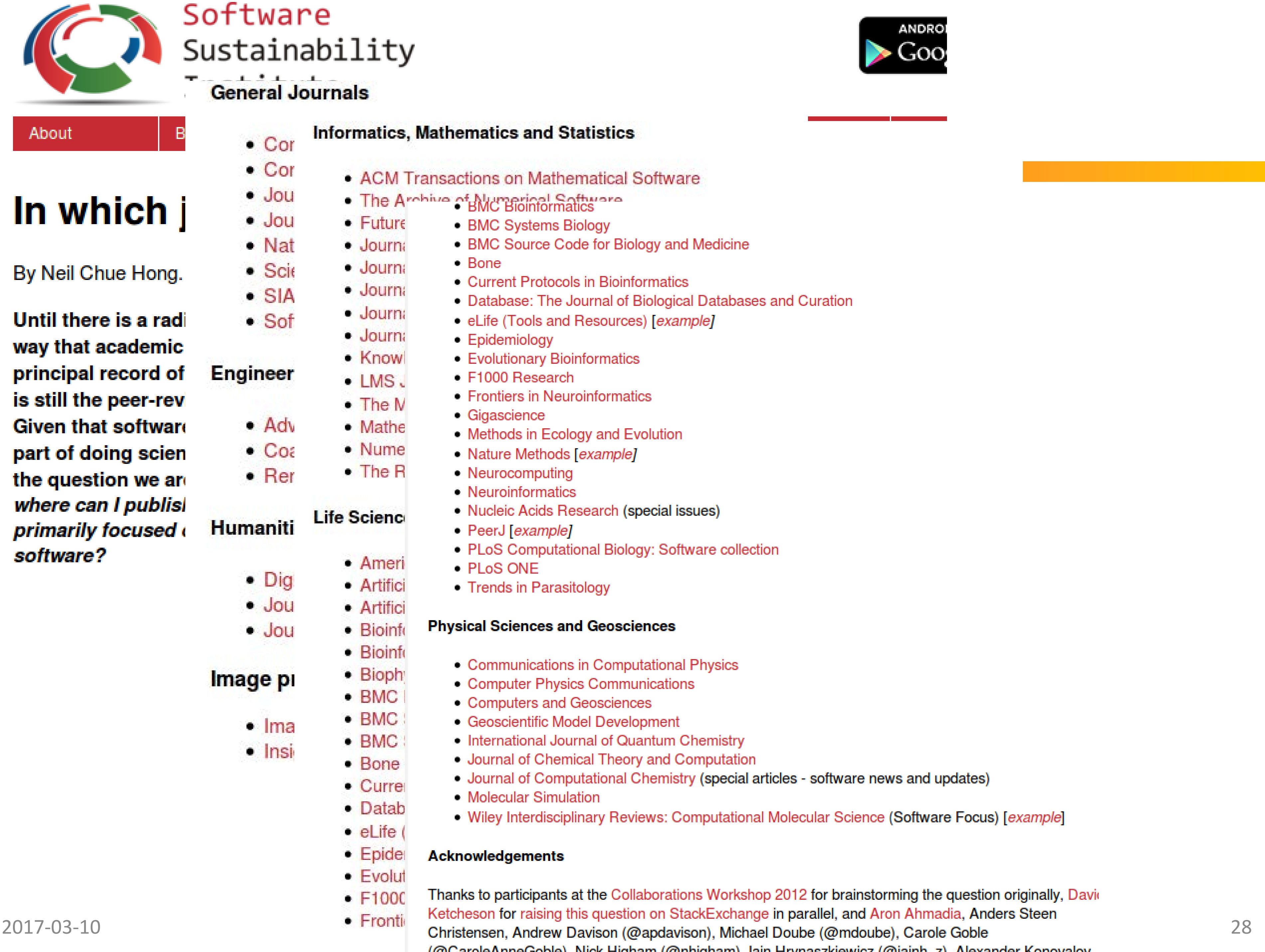
http://bit.ly/softwarejournals



Opportunities for citizen science
Economic benefits for private sector
£
Public trust in science
Validation of research papers
Citations
1st
Career recognition
New collaborations
Easier to find useful software
Re-use in teaching
Software archived and preserved for the future
NEW
More efficient research
00:15
New research made possible
PUBLIC BENEFITS
BENEFITS FOR YOU
RESEARCH COMMUNITY
BENEFITS

http://dx.doi.org/10.6084/m9.figshare.942289

Slide: Neil Chue Hong



# In which j

By Neil Chue Hong.

**Until there is a radi way that academic principal record of is still the peer-rev Given that software part of doing scien the question we ar *where can I publis primarily focused software?***

## General Journals

- Cor
- Cor
- Jou
- Jou
- Nat
- Scie
- SIA
- Sof

## Engineer

- Adv
- Coa
- Ren

## Humaniti

- Dig
- Jou
- Jou

## Image pr

- Ima
- Insi

### Informatics, Mathematics and Statistics

- ACM Transactions on Mathematical Software
- The Archive of Numerical Software
- Future
- Journ
- Journ
- Journ
- Journ
- Journ
- Knowl
- LMS J
- The M
- Mathe
- Nume
- The R

### Life Scienc

- Ameri
- Artific
- Artific
- Bioinf
- Bioinf
- Bioph
- BMC
- BMC
- BMC
- Bone
- Curre
- Datab
- eLife (
- Epide
- Evolut
- F1000
- Fronti

---

- BMC Bioinformatics
- BMC Systems Biology
- BMC Source Code for Biology and Medicine
- Bone
- Current Protocols in Bioinformatics
- Database: The Journal of Biological Databases and Curation
- eLife (Tools and Resources) [*example*]
- Epidemiology
- Evolutionary Bioinformatics
- F1000 Research
- Frontiers in Neuroinformatics
- Gigascience
- Methods in Ecology and Evolution
- Nature Methods [*example*]
- Neurocomputing
- Neuroinformatics
- Nucleic Acids Research (special issues)
- PeerJ [*example*]
- PLoS Computational Biology: Software collection
- PLoS ONE
- Trends in Parasitology

### Physical Sciences and Geosciences

- Communications in Computational Physics
- Computer Physics Communications
- Computers and Geosciences
- Geoscientific Model Development
- International Journal of Quantum Chemistry
- Journal of Chemical Theory and Computation
- Journal of Computational Chemistry (special articles - software news and updates)
- Molecular Simulation
- Wiley Interdisciplinary Reviews: Computational Molecular Science (Software Focus) [*example*]

### Acknowledgements

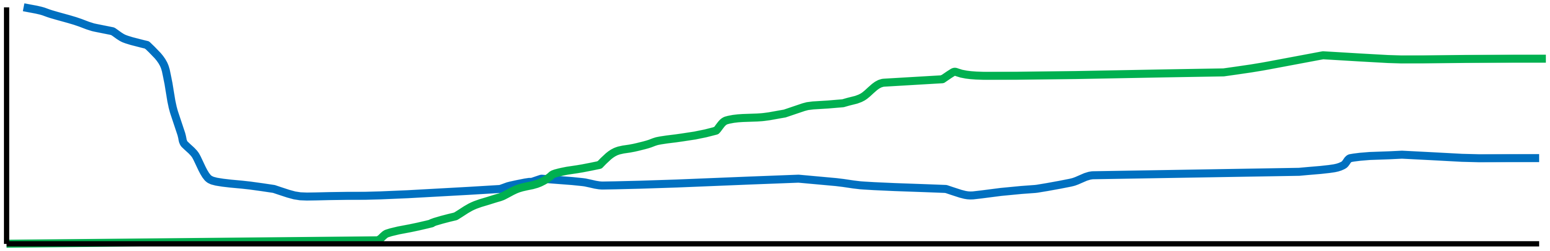
Thanks to participants at the Collaborations Workshop 2012 for brainstorming the question originally, Davi Ketcheson for raising this question on StackExchange in parallel, and Aron Ahmadia, Anders Steen Christensen, Andrew Davison (@apdavison), Michael Doube (@mdoube), Carole Goble (@CaroleAnneGoble), Nick Higham (@nhigham), Iain Hrynaszkiewicz (@iainh_z), Alexander Konovalov

# PRACTICE AROUND THE WORLD – AND COSTS

# UKRDS – vision and business case

- Assume need for data preservation within each university
- Linked by common national services:
  - Discovery
  - Data Management Planning
  - Permanent Identifiers
- Working with international data infrastructure
- £5m (5.57m Euro) over 5 years – investment then repaid by increased efficiency

# A simplified business case



# Where should data go and when?

- Where a national or international subject repository exists – use it
- Where there is no repository – the university is responsible
- It takes responsibility for data once active use is finished
- Not all data is kept for ever

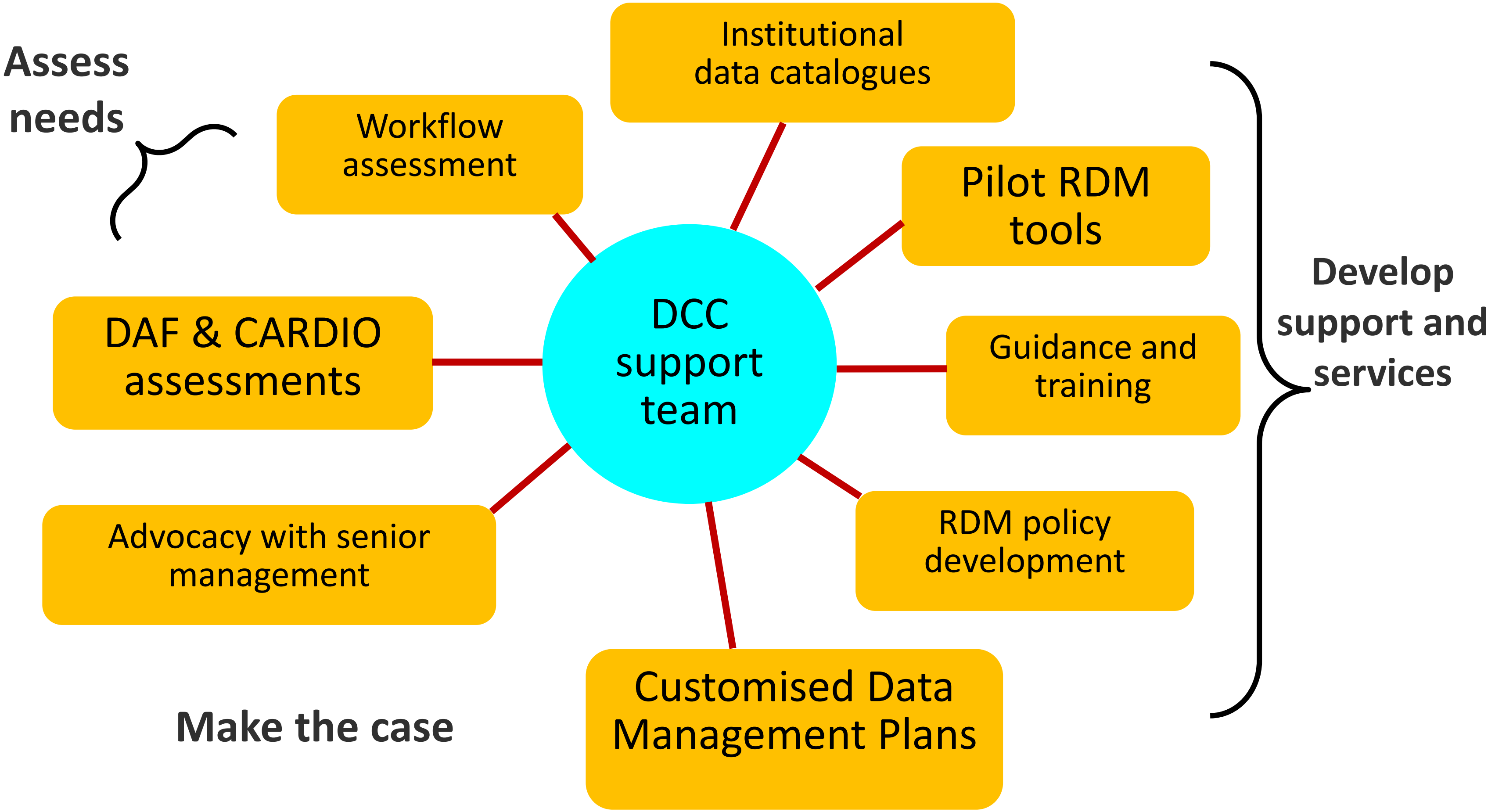
# Funder policy helps the change

- NERC (environment) – 2010
- ESRC (social science) – 1990s
- EPSRC (engineering, physical sciences) – 2013
- European Commission – 2012
- Most require data management plans
- Policies influence researchers and their universities

# The source of my lessons

- In 2011, DCC began working closely with 20 UK universities to develop research data management (RDM) services
- Putting guidance, learning into practice
- Since expanded to > 60 universities and other organisations around the world

# DCC 'institutional engagement'

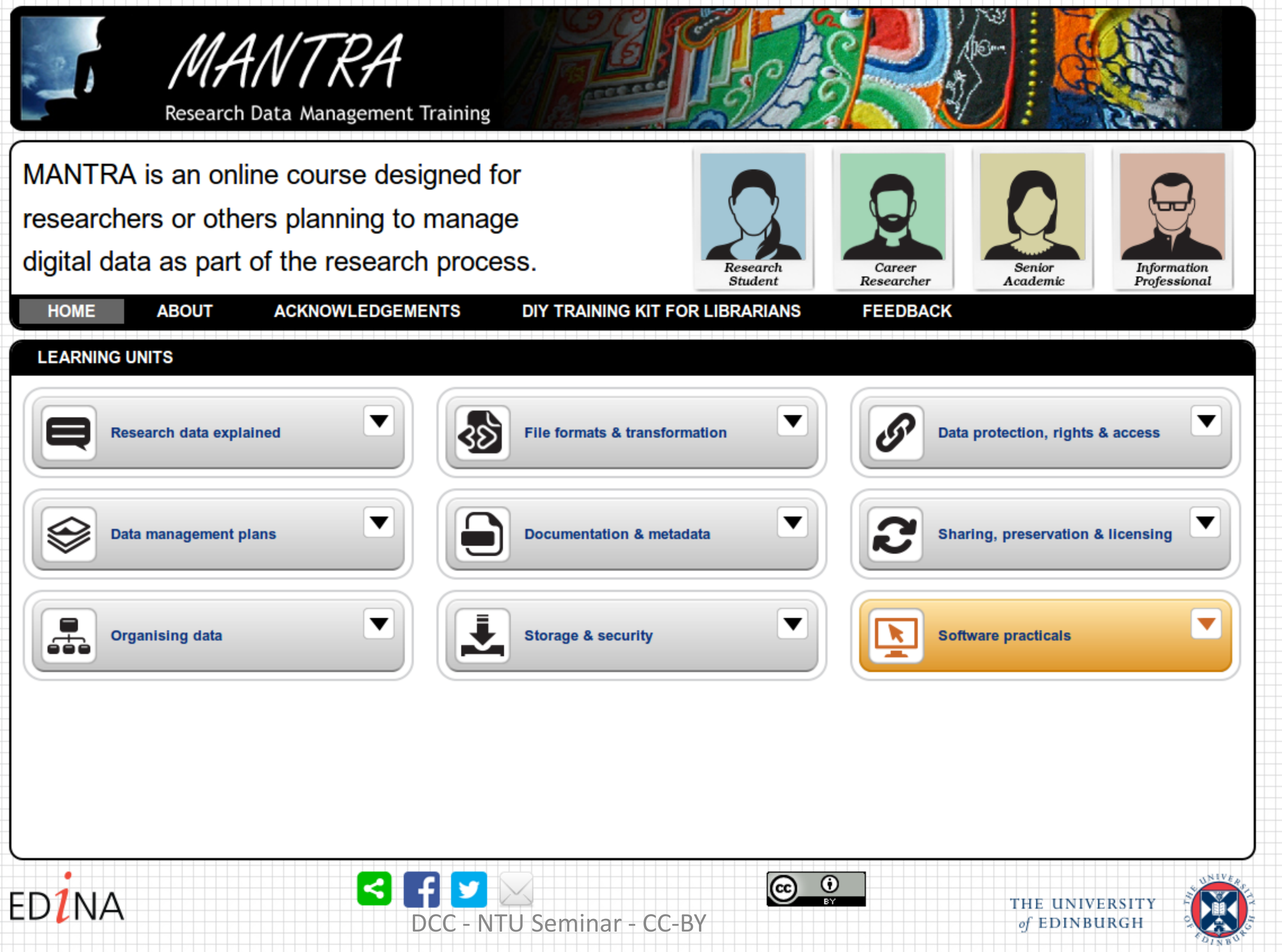


…and support policy implementation

# Some institutional roles

- Leadership – coordinate action
- Audit – who has what, where does it go?
- Advice on access – data, wherever it is
- Preservation – permanence
- Citability
- Data/publication linking
- Promoting data in teaching
- Selection
- Education – early career researchers

# Acquire research data skills



MANTRA

Research Data Management Training

MANTRA is an online course designed for researchers or others planning to manage digital data as part of the research process.

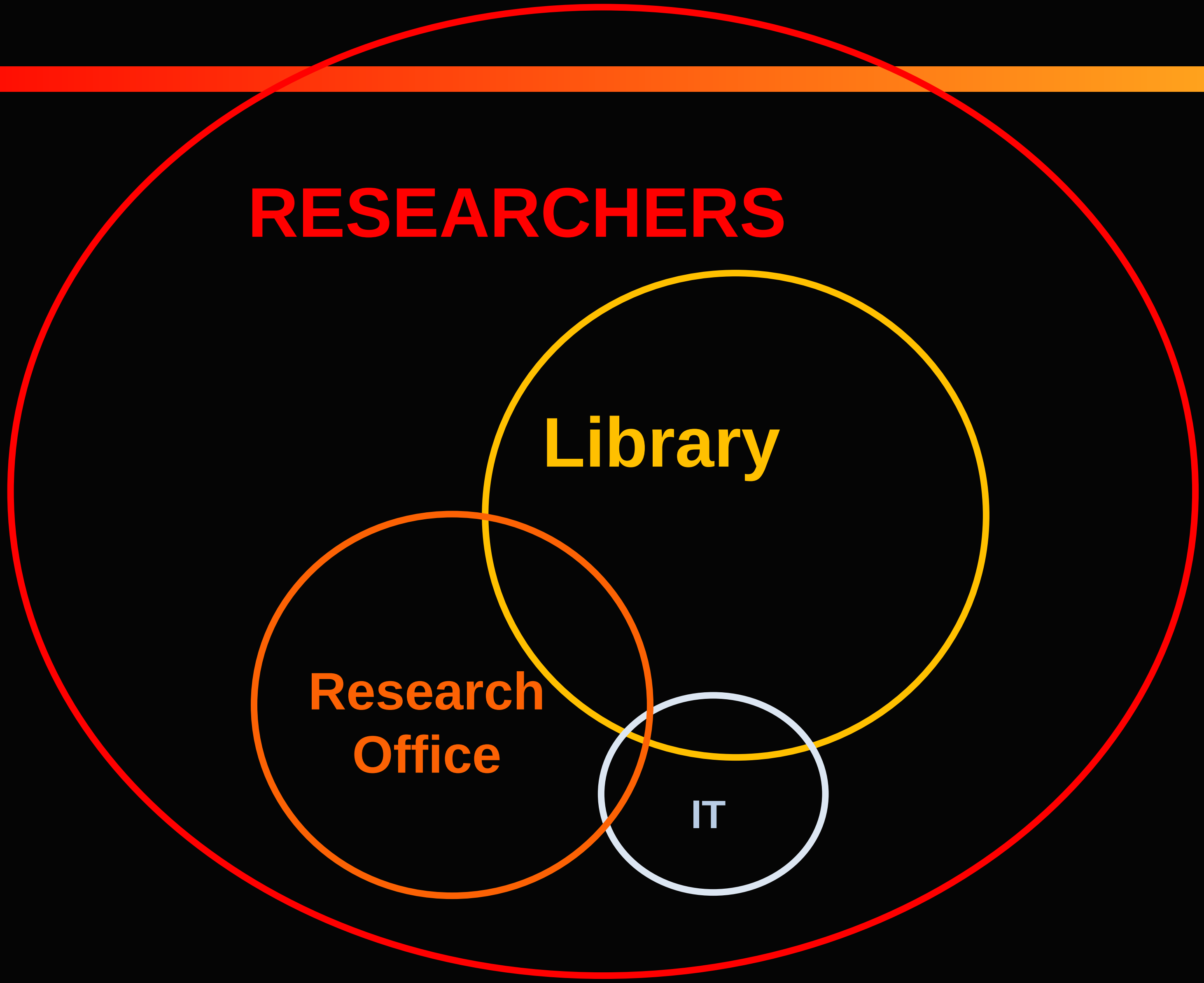
LEARNING UNITS

- Research data explained
- File formats & transformation
- Data protection, rights & access
- Data management plans
- Documentation & metadata
- Sharing, preservation & licensing
- Organising data
- Storage & security
- Software practicals

# LESSON 1

Without senior management attention and researcher involvement, your initiative will fail

# Who (in the UK) is leading RDM work?



# LESSON 2

Research data management services cannot involve the library alone

*"I just back everything up onto data sticks. I didn't even know you could back-up to servers".*

http://www.flickr.com/photos/mattimattila/3003324844/

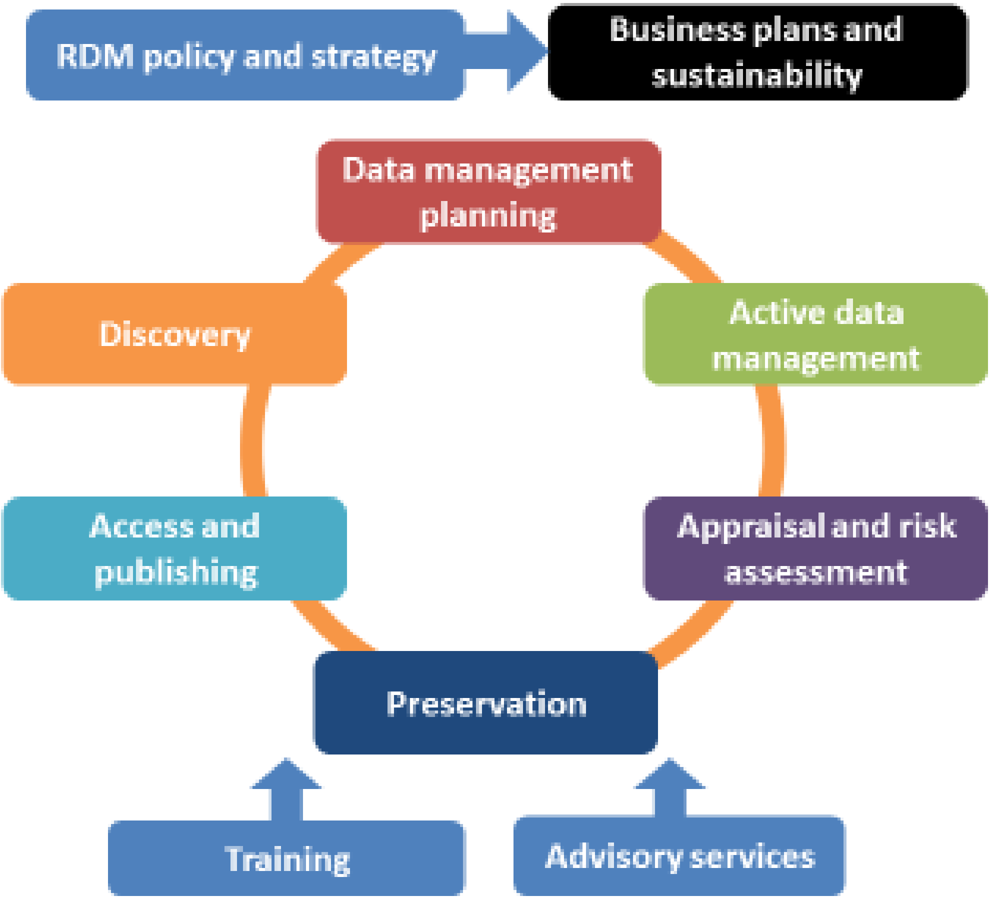
*"Departments don't have guidelines or norms for personal back-up and researcher procedure, knowledge and diligence varies tremendously. Many have experienced moderate to catastrophic data loss"*

Incremental Project Report, June 2010

# LESSON 6

Researchers need to know your services exist

# Institutional support



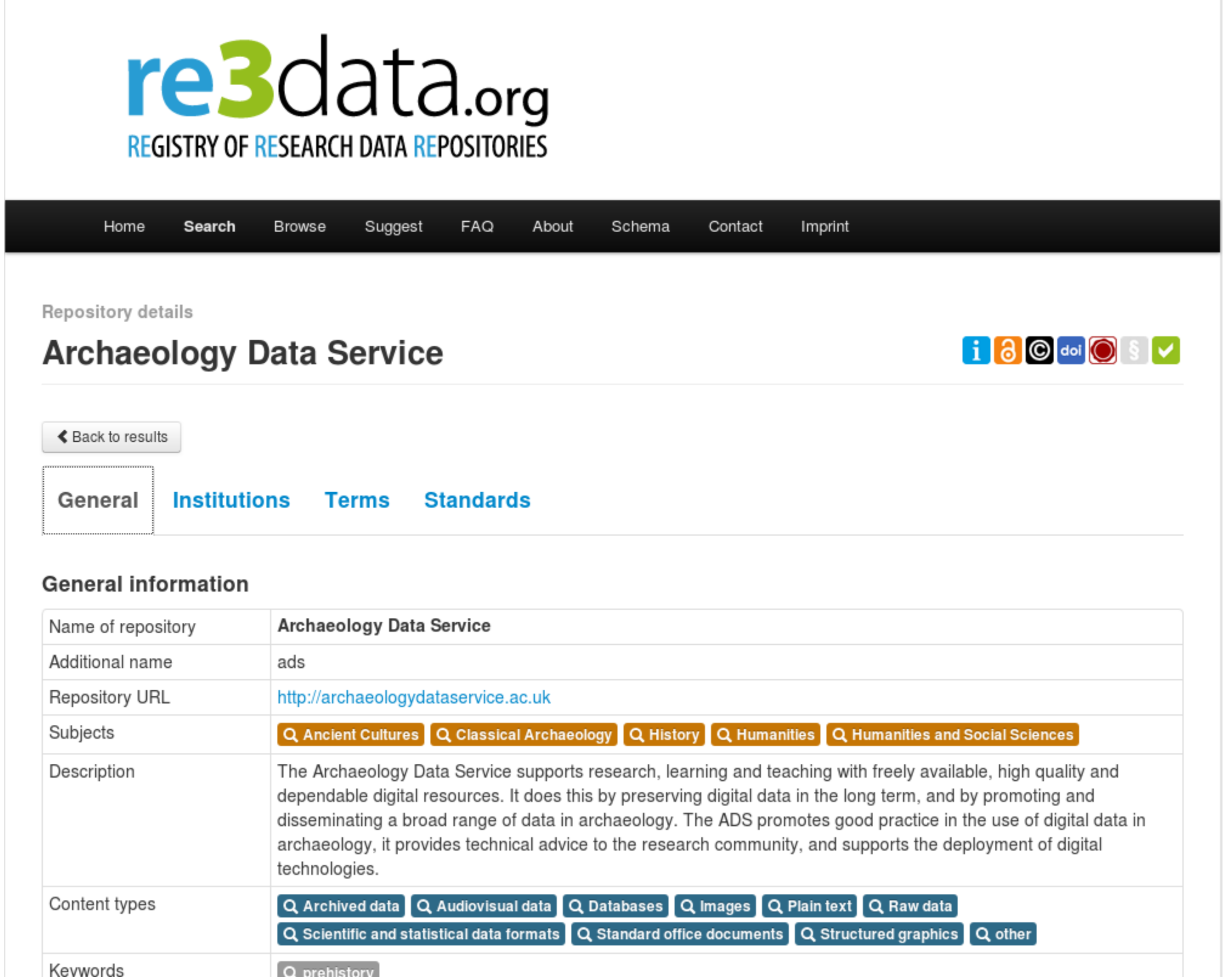
http://www.dcc.ac.uk/resources/how-guides/how-develop-rdm-services

# Goals for the university

- Each university can provide a safe home for data that is discoverable
- Each university has RDM skills in library, IT, research support
- Each university is training new researchers in RDM skills
- Where sensible, universities work together to provide services
- Each university uses national, international services where appropriate

# Is there a better home for my data?

An international service – building on work by Purdue, DCC, Biomedcentral and others



Repository details

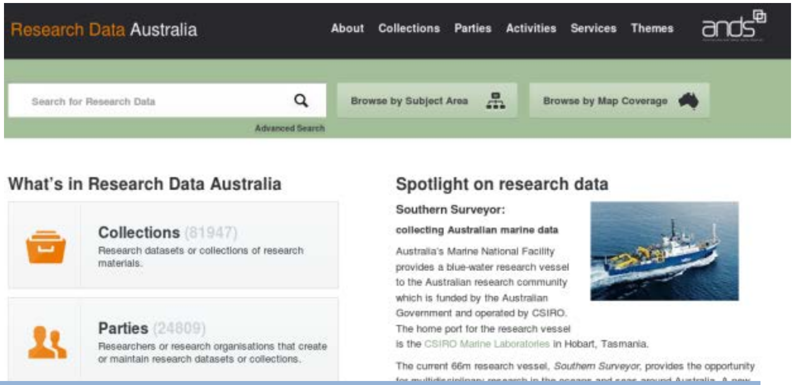
## Archaeology Data Service

### General information

| | |
|---|---|
| Name of repository | **Archaeology Data Service** |
| Additional name | ads |
| Repository URL | http://archaeologydataservice.ac.uk |
| Subjects | Ancient Cultures<br>Classical Archaeology<br>History<br>Humanities<br>Humanities and Social Sciences |
| Description | The Archaeology Data Service supports research, learning and teaching with freely available, high quality and dependable digital resources. It does this by preserving digital data in the long term, and by promoting and disseminating a broad range of data in archaeology. The ADS promotes good practice in the use of digital data in archaeology, it provides technical advice to the research community, and supports the deployment of digital technologies. |
| Content types | Archived data<br>Audiovisual data<br>Databases<br>Images<br>Plain text<br>Raw data<br>Scientific and statistical data formats<br>Standard office documents<br>Structured graphics<br>other |
| Keywords | prehistory |

# Data discovery around the world

- Research Data Australia
- UK data registry pilot & Gateway2Research
- Research Data Netherlands
- World Data System



Research Data Australia

About Collections Parties Activities Services Themes

Search for Research Data

Browse by Subject Area

Browse by Map Coverage

Advanced Search

What's in Research Data Australia

Collections (81947)
Research datasets or collections of research materials.

Parties (24809)
Researchers or research organisations that create or maintain research datasets or collections.

Spotlight on research data

Southern Surveyor:
collecting Australian marine data

Australia's Marine National Facility provides a blue-water research vessel to the Australian research community which is funded by the Australian Government and operated by CSIRO. The home port for the research vessel is the CSIRO Marine Laboratories in Hobart, Tasmania.

The current 66m research vessel, *Southern Surveyor*, provides the opportunity

Still focussed on discovery of whole datasets – we need to move to discovery of what's inside them

# Australia

- Significant long-term funding for national services & support
- ANDS – data discovery, software, skills, coordination, advice
- National storage & HPC infrastructure
- Financial incentives for universities to use common services in standard ways

# Canada

- Like Australia, national research funding but province-level university funding
- Two initiatives – one from universities (PROTAGE), one from federal level (RDC)
- Data management planning, discovery, skills, repositories
- Common tools for local deployment

# Netherlands

- Strong national data repository – DANS
- One cooperative service – 3TU (now 4TU)
- Combined to produce RDNL – back office tech services, front office liaison
- National Dataverse instance
- Some shared services now being proposed by SURF-SARA

# Portugal

- ▷ Existing national publication repository
- ▷ Extending to cover research data
- ▷ One provider, university-branded front-ends
- ▷ Copying other aspects of UK model – e.g. regular meetings for professional staff, funders, other stakeholders

# USA

- Scale means many initiatives
- Much is NSF project-funded
- Some existing university collaborations
- Similar spread to UK – 3 or 4 tiers from Ivy League to small institutions
- Very complex funding model
- Technology is useful – models less so

# What about the cost?

- A rough guide – 5% of total project cost on data curation
- May not all fall to original research group
- How to pay depends on funder and university costing models
- Benefits to society & industry are proven
- Automation and simplification of many processes is helping

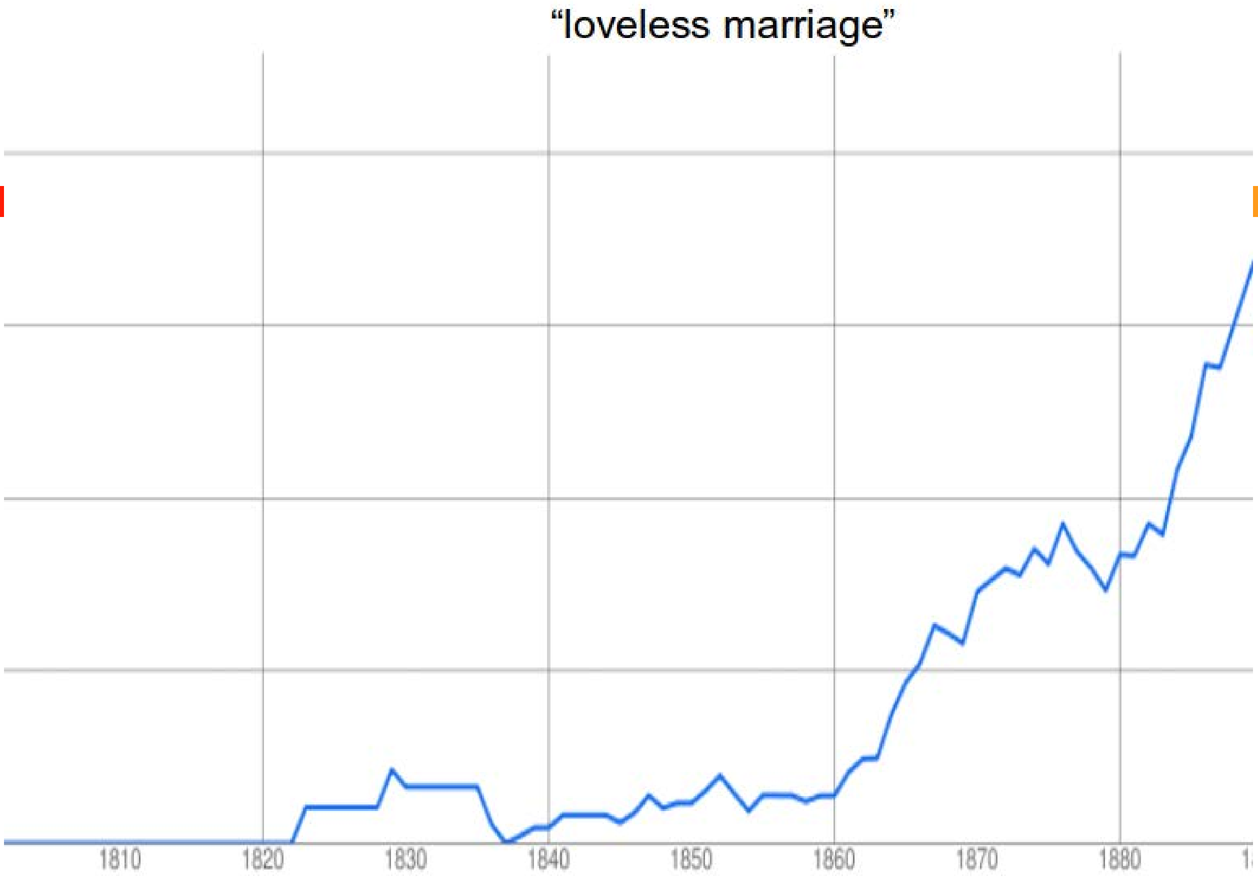
# SOME FINAL MESSAGES

# Should all data be open?

- NO
- Many reasons – most to do with human subjects
- But data existence should always be open
- Allows discovery & negotiation on use
- Avoids pointless replication

# Data isn’t all about numbers

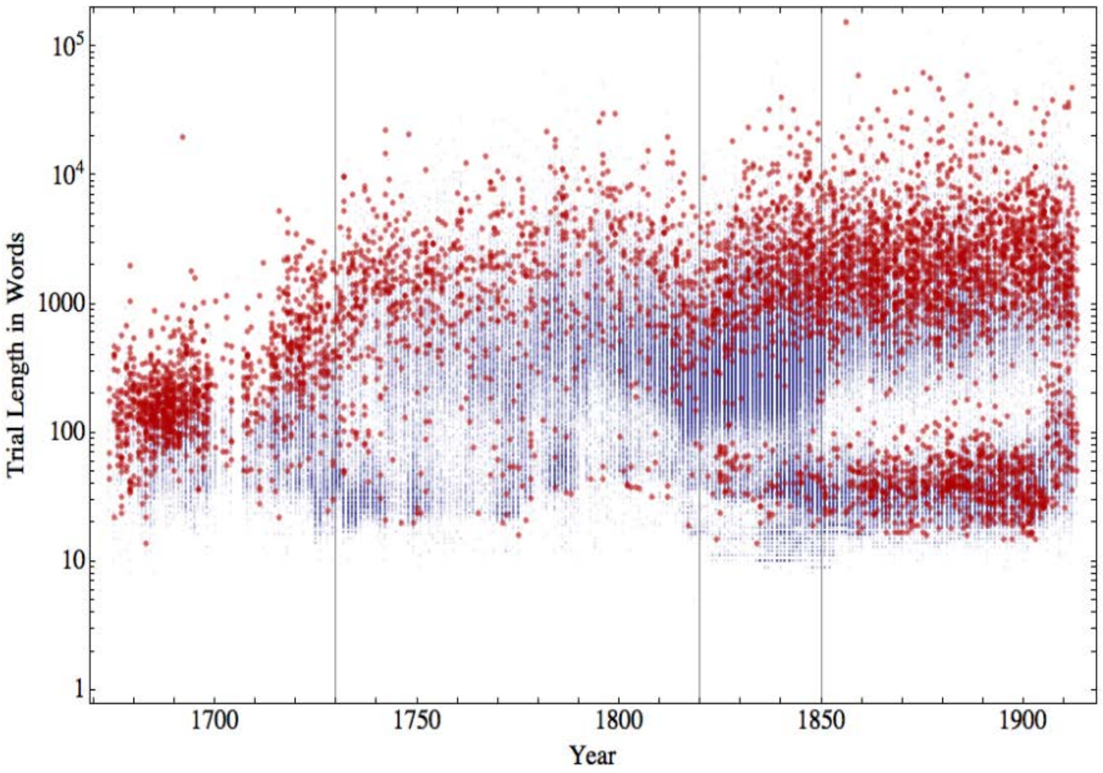
- Data can be words, images, sound, video…
- Anything which can be analysed to provide insight
- Some examples from Old Bailey online – 300 years of English court records



Credit: “Criminal Intent” – Cyril Briquet, Dan Cohen, Frederick Gibbs, Tim Hitchcock, Jamie McLaughlin, Geoffrey Rockwell, Joerg Sander, Robert Shoemaker, John Simpson, Stefan Sinclair, Sean Takats, William J. Turkel

http://criminalintent.org/

Distribution of trial lengths in words for ‘killing’ displayed in red; all other trials in grey. ‘Killing’ includes all trials tagged as including the offences of, ‘Infanticide’, ‘murder’, ‘petty treason’, ‘manslaughter’, and ‘killing: other’, by the Old Bailey online.



# Length of trial – killing – from Old Bailey Online

Credit: “Criminal Intent” – Cyril Briquet, Dan Cohen, Frederick Gibbs, Tim Hitchcock, Jamie McLaughlin, Geoffrey Rockwell, Joerg Sander, Robert Shoemaker, John Simpson, Stefan Sinclair, Sean Takats, William J. Turkel
http://criminalintent.org/

# Traditional skills can win

Data made available before paper was published – result was immediate impact

ets it wr

s us why

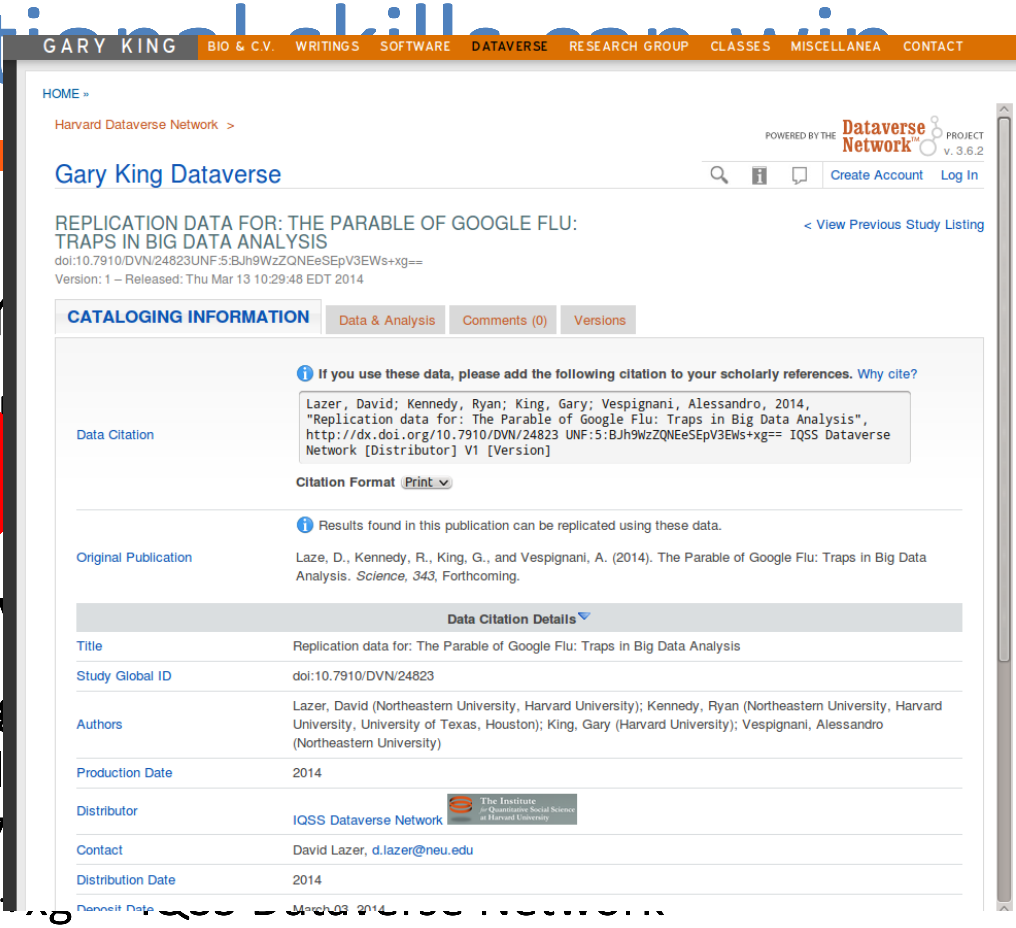
y, Ryan; King

: The Parabl

Analysis", http://dx.doi.org/10.7

UNF:5:BJh9WzZQNEeSEpV3EWs+xg== IQSS Dataverse Network

[Distributor] V1 [Version]



Gary King Dataverse

REPLICATION DATA FOR: THE PARABLE OF GOOGLE FLU: TRAPS IN BIG DATA ANALYSIS

doi:10.7910/DVN/24823UNF:5:BJh9WzZQNEeSEpV3EWs+xg==

Version: 1 – Released: Thu Mar 13 10:29:48 EDT 2014

< View Previous Study Listing

| | |
|---|---|
| Data Citation | If you use these data, please add the following citation to your scholarly references. Why cite?<br>Lazer, David; Kennedy, Ryan; King, Gary; Vespignani, Alessandro, 2014, "Replication data for: The Parable of Google Flu: Traps in Big Data Analysis", http://dx.doi.org/10.7910/DVN/24823 UNF:5:BJh9WzZQNEeSEpV3EWs+xg== IQSS Dataverse Network [Distributor] V1 [Version]<br>Citation Format Print |
| Original Publication | Results found in this publication can be replicated using these data.<br>Laze, D., Kennedy, R., King, G., and Vespignani, A. (2014). The Parable of Google Flu: Traps in Big Data Analysis. *Science*, *343*, Forthcoming. |

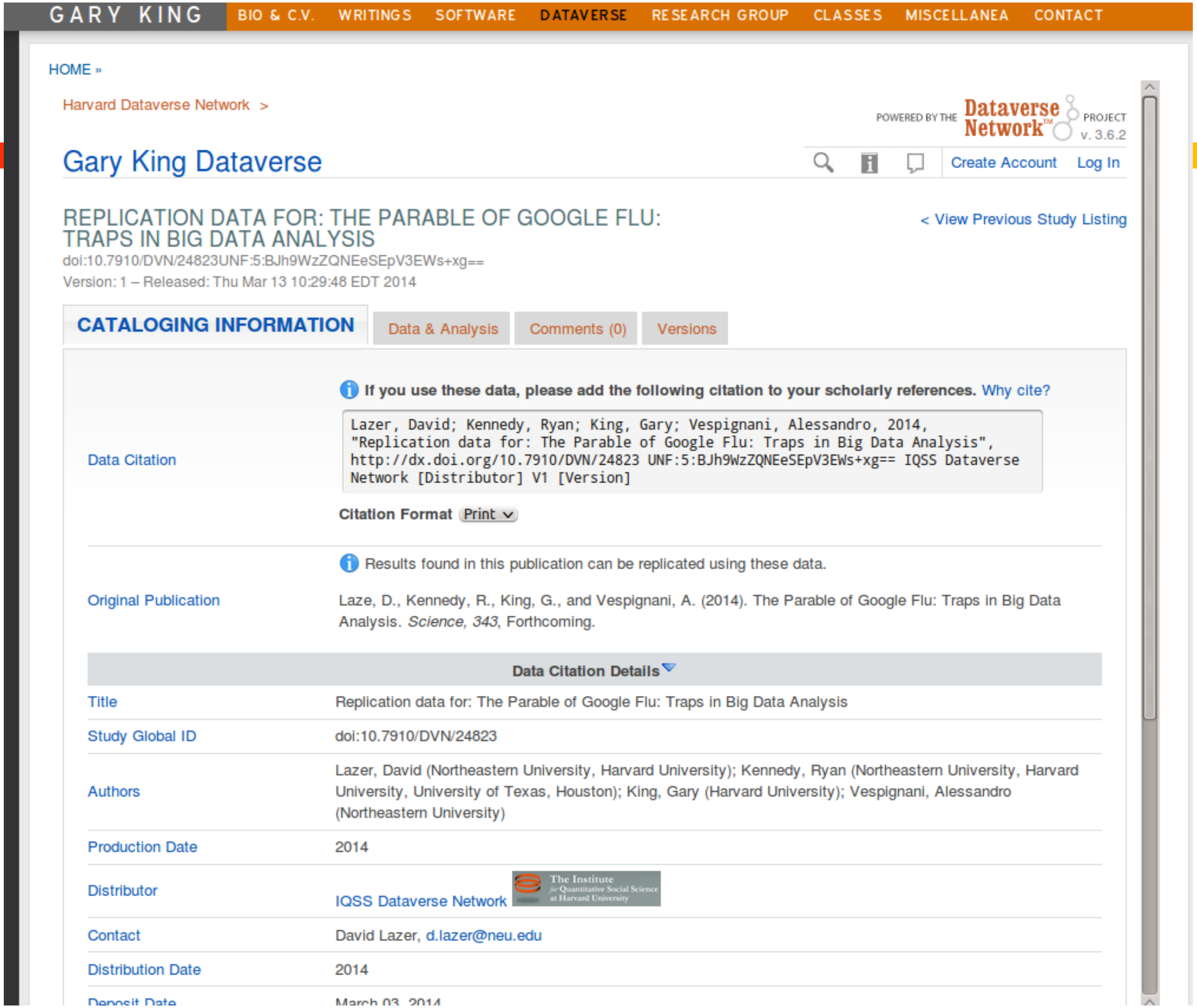
Data Citation Details

| | |
|---|---|
| Title | Replication data for: The Parable of Google Flu: Traps in Big Data Analysis |
| Study Global ID | doi:10.7910/DVN/24823 |
| Authors | Lazer, David (Northeastern University, Harvard University); Kennedy, Ryan (Northeastern University, Harvard University, University of Texas, Houston); King, Gary (Harvard University); Vespignani, Alessandro (Northeastern University) |
| Production Date | 2014 |
| Distributor | IQSS Dataverse Network |
| Contact | David Lazer, d.lazer@neu.edu |
| Distribution Date | 2014 |
| Deposit Date | March 03, 2014 |

# Gking.harvard.edu/data

Tools to make data findable & reusable

Researcher-friendly: incremental approach to metadata



Gary King Dataverse

REPLICATION DATA FOR: THE PARABLE OF GOOGLE FLU: TRAPS IN BIG DATA ANALYSIS
doi:10.7910/DVN/24823UNF:5:BJh9WzZQNEeSEpV3EWs+xg==
Version: 1 – Released: Thu Mar 13 10:29:48 EDT 2014

**If you use these data, please add the following citation to your scholarly references.** Why cite?

Data Citation

```
Lazer, David; Kennedy, Ryan; King, Gary; Vespignani, Alessandro, 2014,
"Replication data for: The Parable of Google Flu: Traps in Big Data Analysis",
http://dx.doi.org/10.7910/DVN/24823 UNF:5:BJh9WzZQNEeSEpV3EWs+xg== IQSS Dataverse
Network [Distributor] V1 [Version]
```

Results found in this publication can be replicated using these data.

Original Publication: Laze, D., Kennedy, R., King, G., and Vespignani, A. (2014). The Parable of Google Flu: Traps in Big Data Analysis. *Science, 343*, Forthcoming.

| | |
|---|---|
| Title | Replication data for: The Parable of Google Flu: Traps in Big Data Analysis |
| Study Global ID | doi:10.7910/DVN/24823 |
| Authors | Lazer, David (Northeastern University, Harvard University); Kennedy, Ryan (Northeastern University, Harvard University, University of Texas, Houston); King, Gary (Harvard University); Vespignani, Alessandro (Northeastern University) |
| Production Date | 2014 |
| Distributor | IQSS Dataverse Network The Institute for Quantitative Social Science at Harvard University |
| Contact | David Lazer, d.lazer@neu.edu |
| Distribution Date | 2014 |
| Deposit Date | March 03, 2014 |

# Some messages for you

- Some things we need to know about data:
  - When/where/what is it about?
  - Who owns it
  - What rights apply
  - What it is derived from & how
  - What software may be associated
  - What data management plan applies
  - How do I gain access ?
  - Where is it ?
  - When was/will it be destroyed?

# My messages to researchers

- ▷ Sharing is difficult
- ▷ Reusing is difficult
- ▷ Both are key to advancing science, and advancing your own career
- ▷ Your data can live longer than your findings
- ▷ All this can be easier than you think

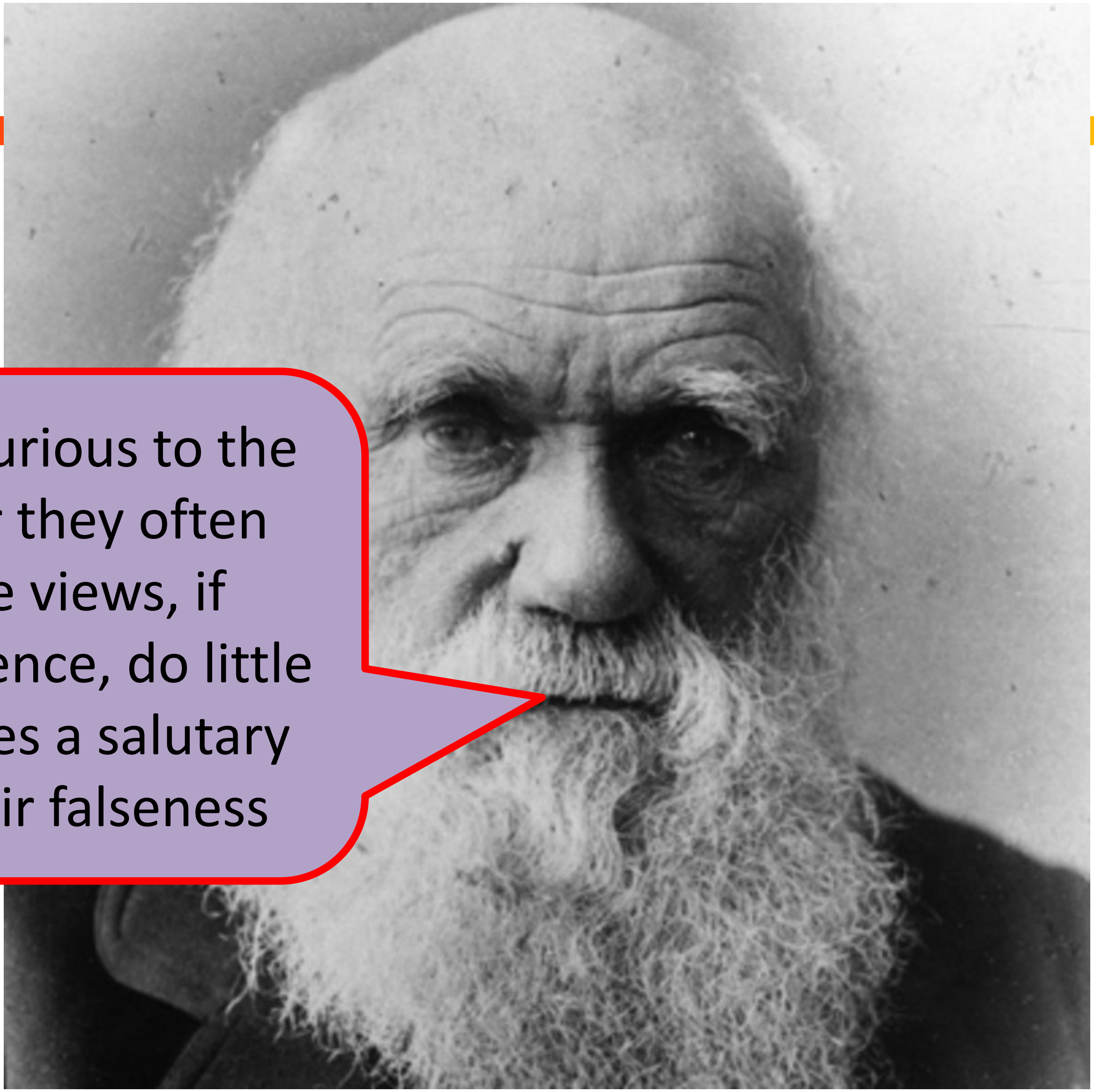
# The value of data in astronomy

- Zij star catalogues – 8th century onwards
- Abd Al-Rahman Al-Sufi – book of fixed stars
- Abu-Mahmud al-Khujandi
  - meridian transits of sun
  - calculate earth’s angle of tilt
  - Different to earlier Indian/Greek calculations (but data lost)
- Lots of work refinin motion
- Kepler – data from
- Modern day – chine us measure change

**The old theories are discredited**

**The old data has value**

# Darwin had something to say about this



False facts are highly injurious to the progress of science, for they often endure long; but false views, if supported by some evidence, do little harm, for every one takes a salutary pleasure in proving their falseness

# Our message to researchers

- The credit belongs to you
- The data belongs to all of us
- Share, and we all reap the benefits

# Excuses – and responses

- “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/