DCC because good research needs good data

Introduction to Research Data Management
Incentives and Barriers

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

Additional material from
Sarah Jones, DCC
Marjan Grootveld, DANS

My home – the DCC

- Miss
  capa
  for re
  services in UK
  institutions
- Not just a UK problem – an international one
- Training, shared services, guidance, policy, standards, futures

because good research needs good d

Repository Fringe 2014
Edinburgh, 30 - 31 July 2014

Latest news | Next events

RDAP14 Presentations now available
8 April, 2014 | in DCC News

Helping you with RDM
3 April, 2014 | in DCC News

RDM in South Africa - UCT Research Data Management Policy and Strategy Workshop
1 April, 2014 | in DCC News

How can the DCC help you?
About us
We are a world-leading centre of expertise in digital information curation...
What is digital curation?

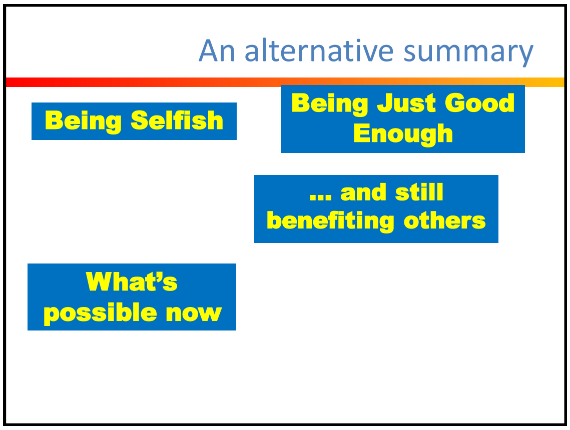
Editor's choice
Where are they now? An RDM update from UEL
A progress update from UEL following our institutional engagement...
Data Citation Principles

Recent blog posts
From RDM strategy to action -a glass half full
Where are they now? An RDM update from Queen's University Belfast

events/repository-fringe-2014

What will we cover?

1. What's research data management?
2. How does it relate to open science?
3. Why does it matter:
   - To you
   - To all of us
4. What gets in the way of good practice?



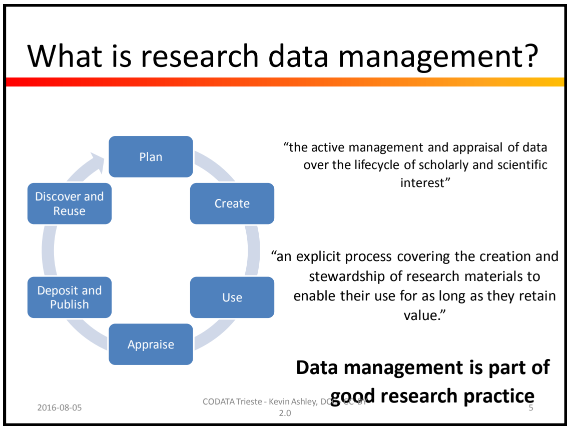
An alternative summary

Being Selfish

Being Just Good Enough

... and still benefiting others

What's possible now



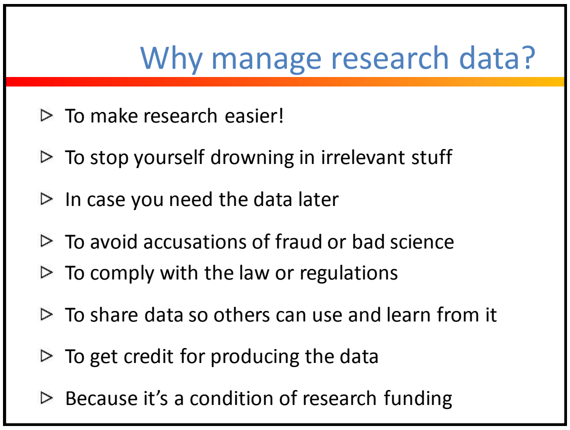
What is research data management?

"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

2016-08-05 CODATA Trieste - Kevin Ashley, DCC - CC-BY 2.0 5



Why manage research data?

- 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

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

EUDAT OpenAIRE

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

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

Definitions of research data?

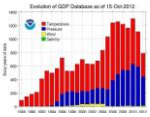
- "Research data, unlike other types of information is collected, observed, or created, for purposes of analysis to produce original research results."
- "Research data is defined as recorded factual material commonly retained by and accepted in the scientific community as necessary to validate research findings; although the majority of such data is created in digital format, all research data is included irrespective of the format in which it is created."
- "Evidence which is used or created to generate new knowledge and interpretations. 'Evidence' may be intersubjective or subjective; physical or emotional; persistent or ephemeral; personal or public; explicit or tacit; and is consciously or unconsciously referenced by the researcher at some point during the course of their research."

THE UNIVERSITY OF EDINBURGH

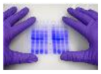
EPSRC

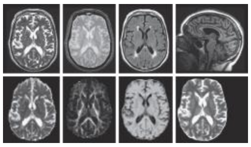
KAPTUR managing visual arts research data

So, what might this include?

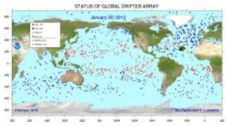


http://www.aoml.noaa.gov/phod/dac/array_growth.html





http://www.sbirc.ed.ac.uk/documents/lbc_protocol.pdf



http://www.aoml.noaa.gov/phod/graphics/dacdata/globpop.gif

Anything & everything produced in the course of research

Why make data available?

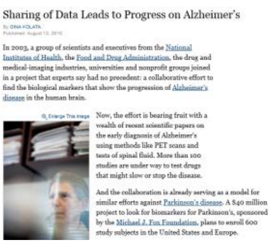
"It was *never* acceptable to publish papers without making data available."

- Ewan Birney

#OpenData
#OpenScience

Original image via doi:10.1038/461145a. "Research cannot flourish if data are not preserved and made accessible. Data management should be woven into every course in science." - *Nature* 461, 145

Sharing leads to breakthroughs



Sharing of Data Leads to Progress on Alzheimer's

By GINA KOLATA
Published: August 12, 2010

In 2003, a group of scientists and executives from the National Institutes of Health, the Food and Drug Administration, the drug and medical-imaging industries, universities and nonprofit groups joined in a project that experts say had no precedent: a collaborative effort to find the biological markers that show the progression of Alzheimer's disease in the human brain.

Enlarge This Image

Now, the effort is bearing fruit with a wealth of recent scientific papers on the early diagnosis of Alzheimer's using methods like PET scans and tests of spinal fluid. More than 100 studies are under way to test drugs that might slow or stop the disease.

And the collaboration is already serving as a model for similar efforts against Parkinson's disease. A $40 million project to look for biomarkers for Parkinson's, sponsored by the Michael J. Fox Foundation, plans to enroll 600 study subjects in the United States and Europe.

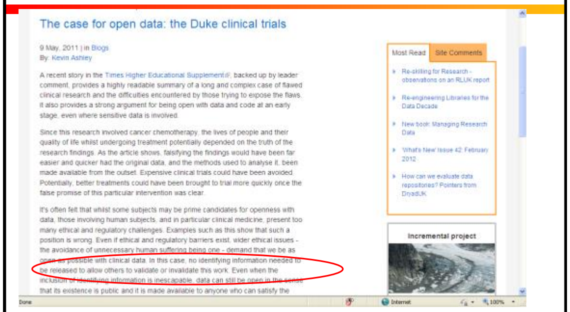
www.nytimes.com/2010/08/13/health/research/13alzheimer.html?pagewanted=all&_r=0

"It was unbelievable. Its not science the way most of us have practiced in our careers. But we all realised that we would never get biomarkers unless all of us parked our egos and intellectual property noses outside the door and agreed that all of our data would be public immediately."

Dr John Trojanowski, University of Pennsylvania

...and increases the speed of discovery

Sharing demonstrates integrity


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%

The case for open data: the Duke clinical trials

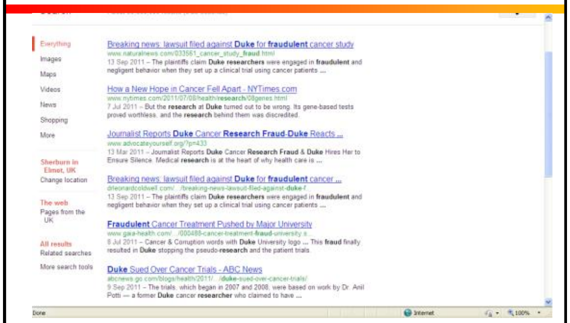
9 May, 2011 | in Blogs
By: Kevin Ashley

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

Not doing so has consequences…



Breaking news: lawsuit filed against Duke for fraudulent cancer study
www.naturalnews.com/033561_cancer_study_fraud.html
13 Sep 2011 – The plaintiffs claim Duke researchers were engaged in fraudulent and negligent behavior when they set up a clinical trial using cancer patients ...

How a New Hope in Cancer Fell Apart - NYTimes.com
www.nytimes.com/2011/07/08/health/research/08genes.html
7 Jul 2011 – But the research at Duke turned out to be wrong. Its gene-based tests proved worthless, and the research behind them was discredited.

Journalist Reports Duke Cancer Research Fraud-Duke Reacts ...
www.advocateyourself.org/?p=433
13 Mar 2011 – Journalist Reports Duke Cancer Research Fraud & Duke Hires Her to Ensure Silence. Medical research is at the heart of why health care is ...

Breaking news: lawsuit filed against Duke for fraudulent cancer ...
drleonardcoldwell.com/.../breaking-news-lawsuit-filed-against-duke-f...
13 Sep 2011 – The plaintiffs claim Duke researchers were engaged in fraudulent and negligent behavior when they set up a clinical trial using cancer patients ...

Fraudulent Cancer Treatment Pushed by Major University
www.gaia-health.com/.../000488-cancer-treatment-fraud-university.s...
8 Jul 2011 – Cancer & Corruption words with Duke University logo ... This fraud finally resulted in Duke stopping the pseudo-research and the patient trials.

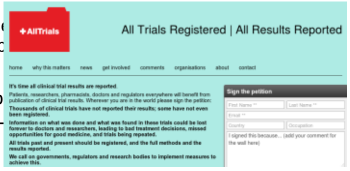
Duke Sued Over Cancer Trials - ABC News
abcnews.go.com/blogs/health/2011/.../duke-sued-over-cancer-trials/
9 Sep 2011 – The trials, which began in 2007 and 2008, were based on work by Dr. Anil Potti — a former Duke cancer researcher who claimed to have ...

Integrity – not without data

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

dies on intelligence.
ed 1976; now discredited

ng l
bab
es
Staper – 55 p
» Poldermans -



All Trials Registered | All Results Reported

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:

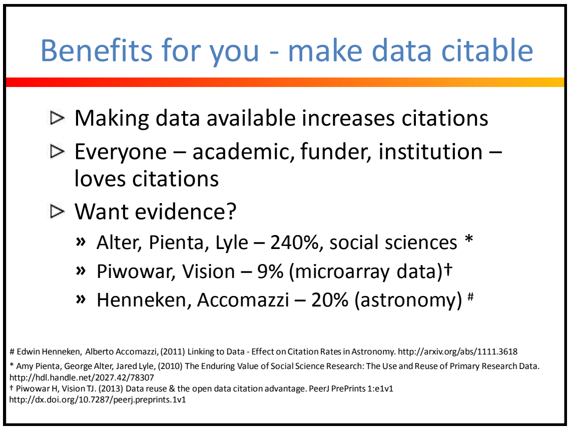
Thousands of clinical trials have not reported their results; some have not even been registered.

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.

"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



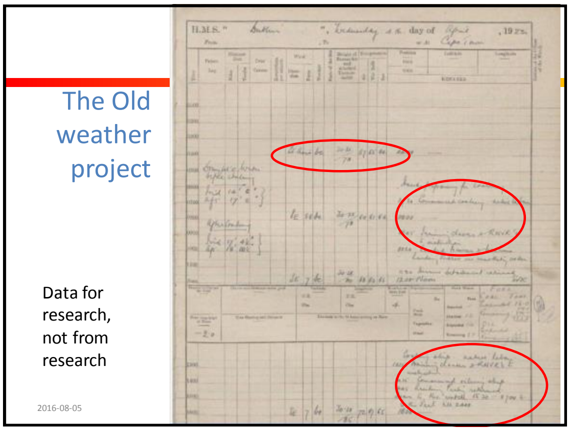
Benefits for you - make data citable

- ▷ 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

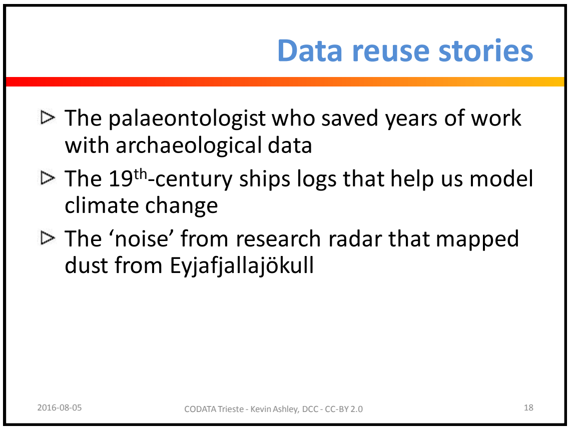
† Piwowar H, Vision TJ. (2013) Data reuse & the open data citation advantage. PeerJ PrePrints 1:e1v1 http://dx.doi.org/10.7287/peerj.preprints.1v1



The Old weather project

Data for research, not from research

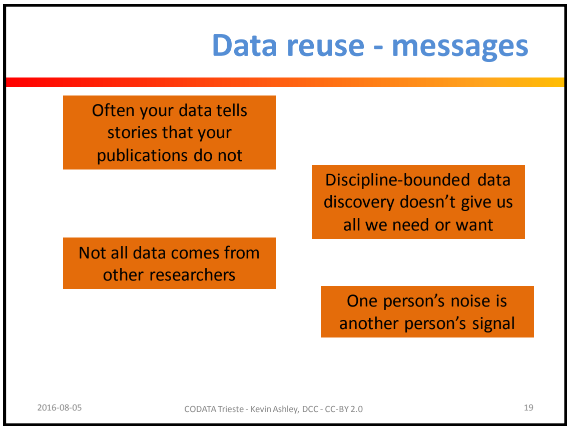
2016-08-05



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

2016-08-05 CODATA Trieste - Kevin Ashley, DCC - CC-BY 2.0 18



Data reuse - messages

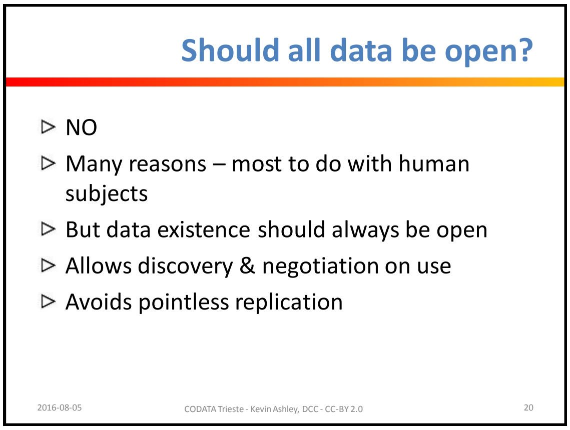
Often your data tells stories that your publications do not

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

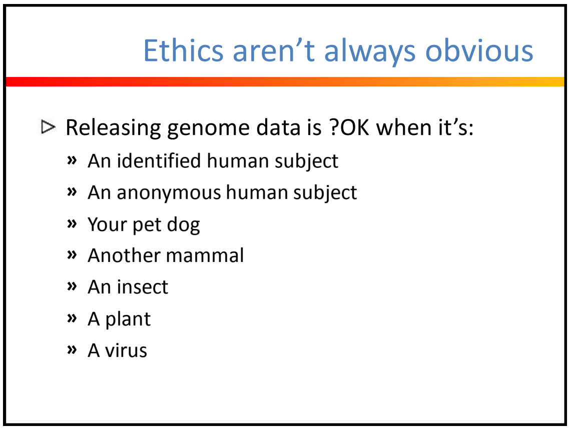
2016-08-05 CODATA Trieste - Kevin Ashley, DCC - CC-BY 2.0 19



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

2016-08-05 CODATA Trieste - Kevin Ashley, DCC - CC-BY 2.0 20

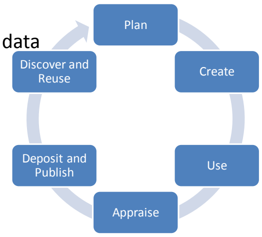


Ethics aren't always obvious

- Releasing genome data is ?OK when it's:
  - An identified human subject
  - An anonymous human subject
  - Your pet dog
  - Another mammal
  - An insect
  - A plant
  - A virus

What is involved in RDM?

- ▷ Data Management Planning
- ▷ Data creation
- ▷ Annotating / documenting data
- ▷ Analysis, use, versioning
- ▷ Storage and backup
- ▷ Publishing papers and data
- ▷ Preparing for deposit
- ▷ Archiving and sharing
- ▷ Licensing
- ▷ Citing…



Data management planning

- ▷ Many funders expect a DMP to be produced as part of project development, most expect it to be submitted with the grant application
- ▷ Examples of good DMPs are available here:
  - » http://www.dcc.ac.uk/resources/data-management-plans/guidance-examples
- ▷ The DCC provides an online tool to guide you through the process of developing a funder-specific DMP

DMP ONLINE

https://dmponline.dcc.ac.uk/

What data to keep

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

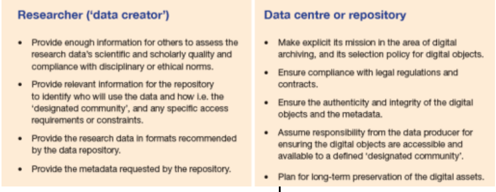
DCC

ands

How to Appraise & Select Research Data for Curation

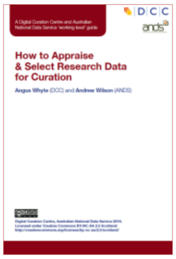
Angus Whyte (DCC) and Andrew Wilson (ANDS)

Roles and Responsibilities



Researcher ('data creator')	Data centre or repository
• 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.	• 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.

Appraisal and deposit



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

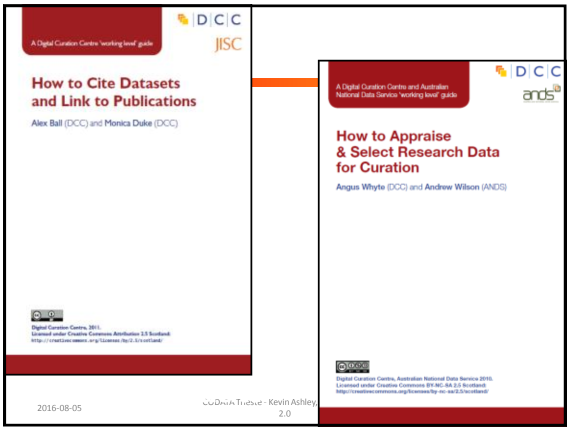
DCC ands

How to Appraise & Select Research Data for Curation

Angus Whyte (DCC) and Andrew Wilson (ANDS)

How to Appraise & Select Research Data for Curation
Angus Whyte, Digital Curation Centre, and Andrew Wilson, Australian National Data Service (2010)

1. **Relevance to Mission** – including any legal/funder requirement to retain the data beyond its immediate use.
2. **Scientific or Historical Value** – significance and relationship to publications etc.
3. **Uniqueness** – can it be found elsewhere / if we don't preserve it, who will?
4. **Potential for Redistribution** – quality / IP / ethical concerns are addressed.
5. **Non-Replicability** – either impossible to replicate (e.g. atmospheric or social science data) or not financially viable.
6. **Economic Case** – costs of managing and preserving the resource stack up well against potential future benefits.
7. **Full Documentation** – surrounding / contextual information necessary to facilitate future discovery, access, and reuse is adequate.



A Digital Curation Centre 'working level' guide

DCC JISC

How to Cite Datasets and Link to Publications

Alex Ball (DCC) and Monica Duke (DCC)

Digital Curation Centre, 2011.
Licensed under Creative Commons Attribution 2.5 Scotland:
http://creativecommons.org/licenses/by/2.5/scotland/

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

DCC ands

How to Appraise & Select Research Data for Curation

Angus Whyte (DCC) and Andrew Wilson (ANDS)

Digital Curation Centre, Australian National Data Service 2010.
Licensed under Creative Commons BY-NC-SA 2.5 Scotland:
http://creativecommons.org/licenses/by-nc-sa/2.5/scotland/

2016-08-05

CODATA Trieste - Kevin Ashley,
2.0

RDM and sharing : a best practice guide

MANAGING AND SHARING DATA

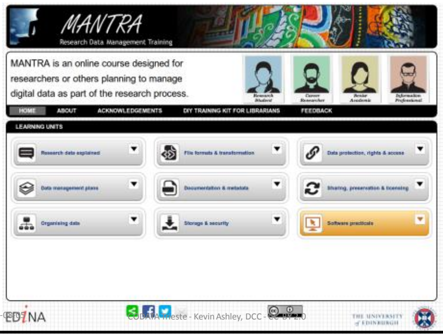
UK DATA ARCHIVE

BEST PRACTICE FOR RESEARCHERS

- Planning for sharing
- Consent and ethics
- Copyright
- Documenting your data
- Formatting your data
- Storing your data
- Strategies for centres

http://data-archive.ac.uk/media/2894/managingsharing.pdf

Acquire research data skills

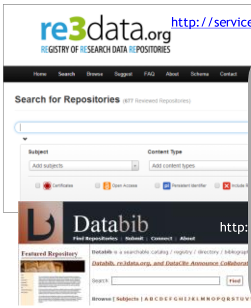


MANTRA
Research Data Management Training

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

2016-08-05 CODATA Trieste - Kevin Ashley, DCC - CC-BY 2.0 28

Data repositories



http://service.re3data.org/search

re3data.org
REGISTRY OF RESEARCH DATA REPOSITORIES

Search for Repositories (877 Reviewed Repositories)

Databib is a searchable catalog / registry / directory / bibliograp

Databib, re3data.org, and DataCite Announce Collaborat

http:

Zenodo

- OpenAIRE-CERN joint effort
- Multidisciplinary repository
- Multiple data types
  - Publications
  - Long tail of research data
- Citable data (DOI)
- Links to funder, publications, data & software

www.zenodo.org

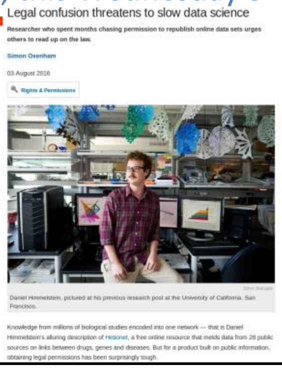
Why hand data over for preservation?

- To preserve the data themselves “Data rot”
  - Bitwise preservation
  - Format migration
- To preserve contextual information
  - Often held in a researcher’s head
  - Notes often aren’t detailed enough
- Protecting digital objects requires specialist skills and particular information to be captured
- The aim is to enable the reuse of data

Not everything can, or should be preserved!

LEGAL ISSUES

Nature, this Wednesday 3rd August



Legal confusion threatens to slow data science

Researcher who spent months chasing permission to republish online data sets urges others to read up on the law.

Simon Oxenham

03 August 2016

Rights & Permissions

Daniel Himmelstein, pictured at his previous research post at the University of California, San Francisco.

Knowledge from millions of biological studies encoded into one network — that is Daniel Himmelstein's alluring description of Hetionet, a free online resource that melds data from 28 public sources on links between drugs, genes and diseases. But for a product built on public information, obtaining legal permissions has been surprisingly tough.

Two types of issue

- Things that the law requires you to consider
- Things that the law allows you to do
- Can you think of any of these?

Requirements

- Data protection
  - If human subjects are involved
  - Common European framework
  - Informed consent essential
  - Make consent broad to allow reuse
  - Protect data
  - Provide subject access
  - Right of correction
  - Beware – law under review in Europe

FOI & EIR

- FOI = Freedom of Information
- EIR = Environmental Information Regulations
- First is nation-state specific; second from European regulation
- Both have similar effects, but differ in detail

Consequences for researchers

- Your organisation must know what data it possesses
- It must know whether exceptions to access may apply
- It must know if some of the data belongs to others
- It must know what data once existed, but has now been deleted – and why
- These are difficult questions for most of us!

What the law allows - licensing

- Licences allow you to constrain how others use your data
- They range from very open to very restrictive
- You MUST own the data in order to be able to licence it

License your data for reuse

A Digital Curation Centre and JISC Legal 'working level' guide

How to License Research Data

Alex Ball (DCC)

www.dcc.ac.uk/resources/how-guides/license-research-data

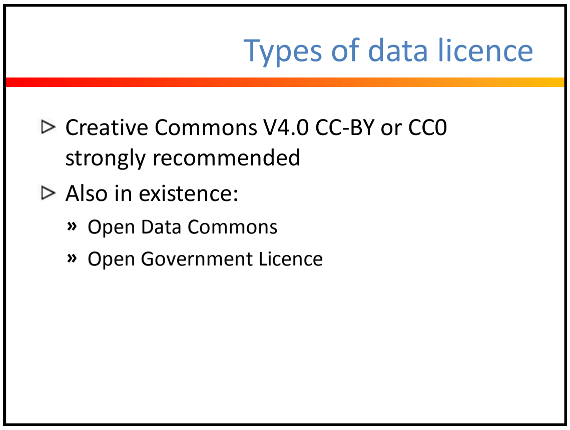
Outlines pros and cons of each approach and gives practical advice on how to implement your licence

CREATIVE COMMONS LIMITATIONS

NC	Non-Commercial What counts as commercial?
SA	Share Alike Reduces interoperability
ND	No Derivatives Severely restricts use

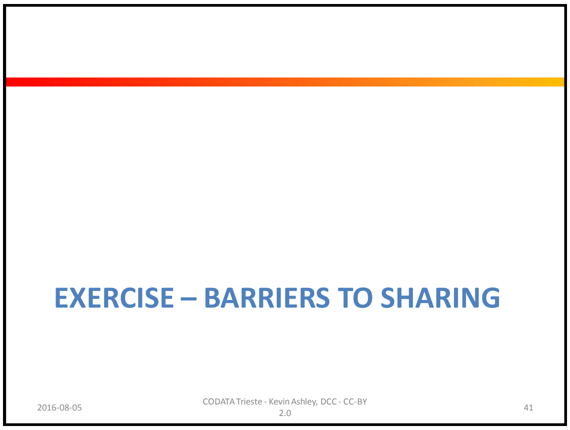
Data and copyright

- Ability to copyright data varies throughout the world
- Europe also offers ‘database right’ – applies even if data cannot be copyrighted.
- International licences help avoid this legal minefield
- Standard licences strongly recommended – we are not all legal experts

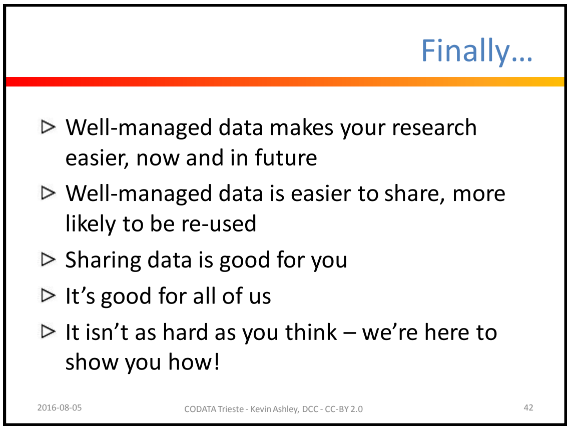


Types of data licence

- Creative Commons V4.0 CC-BY or CC0 strongly recommended
- Also in existence:
  - Open Data Commons
  - Open Government Licence

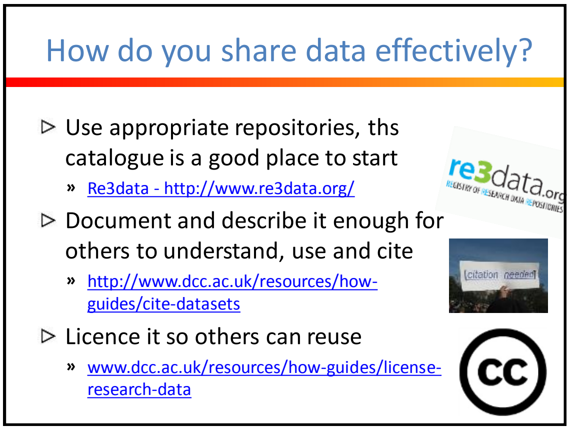


EXERCISE – BARRIERS TO SHARING



Finally…

- Well-managed data makes your research easier, now and in future
- Well-managed data is easier to share, more likely to be re-used
- Sharing data is good for you
- It's good for all of us
- It isn't as hard as you think – we're here to show you how!



How do you share data effectively?

- Use appropriate repositories, ths catalogue is a good place to start
  - Re3data - http://www.re3data.org/
- Document and describe it enough for others to understand, use and cite
  - http://www.dcc.ac.uk/resources/how-guides/cite-datasets
- Licence it so others can reuse
  - www.dcc.ac.uk/resources/how-guides/license-research-data

re3data.org
REGISTRY OF RESEARCH DATA REPOSITORIES
[citation needed]
cc