THE UNIVERSITY of York
Information

Managing your research data:
an introduction for Computer Science PGRs

Jenny Mitcham (Digital Archivist)
Lindsey Myers (Research Support Librarian)

16 June 2015

Overview

- What is research data management?
- Why manage research data?
- How to manage research data (best practice)
- Preserving and sharing research data
- Data management planning (this is key)
- Help & advice

What is research data management?

What is data?

“A reinterpretable representation of information in a formalized manner suitable for communication, interpretation, or processing.”

Digital Curation Centre

What is data?

2.1 Recorded material, irrespective of format or media, commonly retained and accepted in the academic community as being necessary to validate research findings. Created in the course of the research process, research data will be the recorded facts, observations, measurements, computations, statistics and results that underpin the research paper and grant or project outcomes.

What is data management?

Data management is a general term covering how you organize, structure, store, and care for the information used or generated during a research project

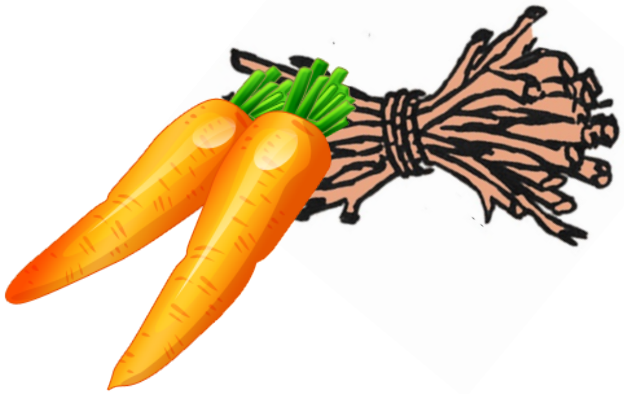
- It includes:
  - How you deal with information on a day-to-day basis over the lifetime of a project
  - What happens to data in the longer term - what you do with it after the project concludes

Why manage your research data?

Carrots and sticks

- Work efficiently and with minimum hassle over the lifetime of the project
- Save time and avoid problems in the future
- Make it easy to share your data

- University of York Research Data Management Policy
- Funding body requirements



University Policy

"4. Responsibilities

4.1 All staff and research students involved in research under the University's auspices have a responsibility to manage data they create and maintain it effectively and in line with University policy, regulations, codes of practice and associated training and guidance."

University Research Data Management Policy:
www.york.ac.uk/research-data-management

The Policy: what you need to know

Research data must be:

- accurate, complete, authentic and reliable
- identifiable, retrievable and available when needed
- kept safe and secure, avoiding data loss
- kept in a manner that is compliant with legal and ethical obligations, and (if applicable) funder requirements
- disposed of securely.

The Policy: What you need to know

Research data that are considered to have long-term value and potential re-use (with no legal, ethical or commercial constraints) must be:

- kept for 10 years from the date of last requested access
- preserved and deposited, with appropriate documentation.

Funder requirements

RESEARCH COUNCILS UK

Accessibility | Media Enquiries

Google™ Custom Search Search

Home | Funding | Research | Innovation | Skills | Public Engagement | News, Events and Publications | About Us

Home / Research / RCUK Common Principles on Data Policy

RCUK Common Principles on Data Policy

Making research data available to users is a core part of the Research Councils' remit and is undertaken in a variety of ways. We are committed to transparency and to a coherent approach across the research base. These RCUK common principles on data policy provide an overarching framework for individual Research Council policies on data policy.

Principles

- Publicly funded research data are a public good, produced in the public interest, which should be made openly available with as few restrictions as possible in a timely and responsible manner that does not harm intellectual property.
- Institutional and project specific data management policies and plans should be in accordance with relevant standards and community best practice. Data with acknowledged long-term value should be preserved and remain accessible and usable for future research.
- To enable research data to be discoverable and effectively re-used by others, sufficient metadata should be recorded and made openly available to enable other researchers to understand the research and re-use potential of the data. Published results should always include information on how to access the supporting data.
- RCUK recognises that there are legal, ethical and commercial constraints on release of research data. To ensure that the research process is not damaged by inappropriate release of data, research organisation policies and practices should ensure that these are considered at all stages in the research process.
- To ensure that research teams get appropriate recognition for the effort involved in collecting and analysing data, those who undertake Research Council funded work may be entitled to a limited period of privileged use of the data they have collected to enable them to publish the results of their research. The length of this period varies by research discipline and, where appropriate, is discussed further in the published policies of individual Research Councils.
- In order to recognise the intellectual contributions of researchers who generate, preserve and share key research datasets, all users of research data should acknowledge the sources of their data and abide by the terms and conditions under which they are accessed.
- It is appropriate to use public funds to support the management and sharing of publicly-funded research data. To maximise the research benefit which can be gained from limited budgets, the mechanisms for these activities should be both efficient and cost-effective in the use of public funds.

Funder requirements

Publicly funded research data are a public good, produced in the public interest, which should be made openly available with as few restrictions as possible in a timely and responsible manner that does not harm intellectual property.

Data with acknowledged long term value should be preserved and remain accessible and usable for future research.

EPSRC expectations

EPSRC
Pioneering research and skills

From **1 May 2015** both the University and its EPSRC-funded researchers are required to adhere to the EPSRC set of expectations

Required to: Manage your research data in accordance with the University RDM Policy.

For example:

- Research data must be kept safe and secure, avoiding data loss

EPSRC expectations

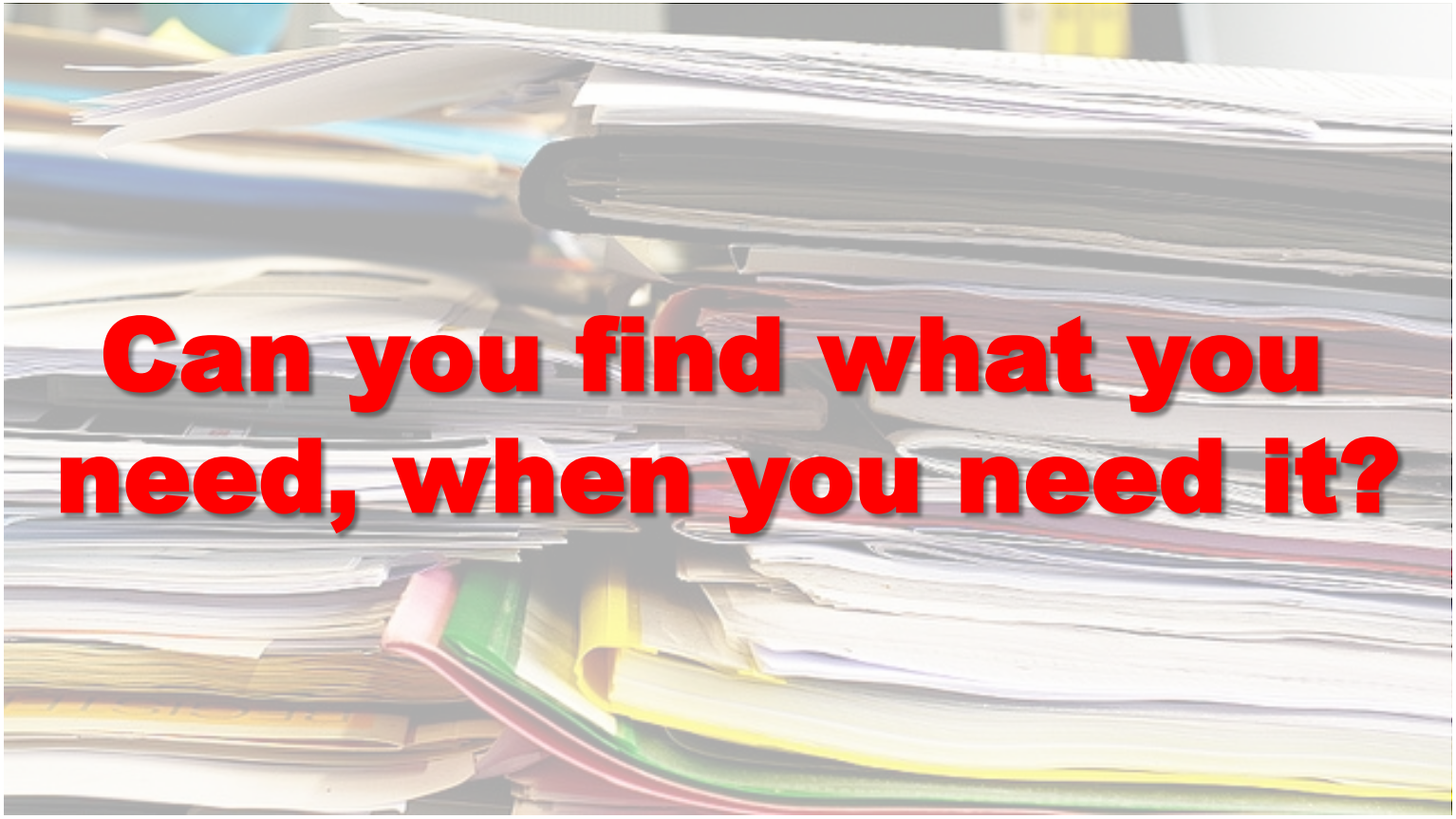
In addition, for papers (incl. theses) which acknowledge EPSRC funding, with a publication date after **1 May 2015,** you will need to:

EPSRC expectations

1. State in your published papers how the underlying data can be accessed
2. Store your data in a format which would facilitate sharing and use by others
3. Ensure your data are securely preserved for at least ten years
4. Describe your data using appropriate metadata to enable others to find your data, understand it and how to access it (PURE)
5. Anticipate the costs and resources needed to manage your data

Day-to-day data management

a. organising your data (good file management)



‘What a mess’ by .pst, via Flickr: www.flickr.com/photos/psteichen/3915657914

The Policy: what you need to know

Research data must be:

- accurate, complete, authentic and reliable
- identifiable, retrievable and available when needed
- kept safe and secure, avoiding data loss
- kept in a manner that is compliant with legal and ethical obligations, and (if applicable) funder requirements
- disposed of securely.

In practice

Are you using helpful, consistent file naming conventions? Is your file structure clear?

Develop plans for:

- file structures - *where to put data so you won't lose it*
- file (and folder) naming - *what to call data so you know what it is*
- version control - *keeping track of data*

Develop a system/convention that works for your project (and document it) - be consistent

File (and folder) naming

Decide on a file naming convention at the start of your project. Useful file names:

- are consistent
- are concise but informative
- classify broad file types
- are meaningful to you (and your colleagues)
- allow you to find the file easily
- do not contain special characters or spaces
- should not conflict when moved from one location to another.

File naming strategies

Think about the ordering of elements within a filename, e.g. YYYY-MM-DD dates allow chronological sorting

Order by date:

2013-04-12_interview-recording_THD.mp3
2013-04-12_interview-transcript_THD.docx
2012-12-15_interview-recording_MBD.mp3
2012-12-15_interview-transcript_MBD.docx

Order by type:

Interview-recording_MBD_2012-12-15.mp3
Interview-recording_THD_2013-04-12.mp3
Interview-transcript_MBD_2012-12-15.docx
Interview-transcript_THD_2013-04-12.docx

Order by subject:

MBD_interview-recording_2012-12-15.mp3
MBD_interview-transcript_2012-12-15.docx
THD_interview-recording_2013-04-12.mp3
THD_interview-transcript_2013-04-12.docx

Forced order with numbering:

01_THD_interview-recording_2013-04-12.mp3
02_THD_interview-transcript_2013-04-12.docx
03_MBD_interview-recording_2012-12-15.mp3
04_MBD_interview-transcript_2012-12-15.docx

Version control

- A common form for expressing data file versions is to use ordinal numbers (1,2,3 etc.) for major version changes and the decimal for minor changes (e.g. v1, v1.1, v2.6)
- Beware of using confusing labels (e.g. revision, final, final2, definitive_copy) these can accumulate
- Record every change irrespective of how minor that change may be
- Discard or delete obsolete versions (whilst retaining the original 'raw' copy)
- May create a version control table or file history w/in or alongside data file

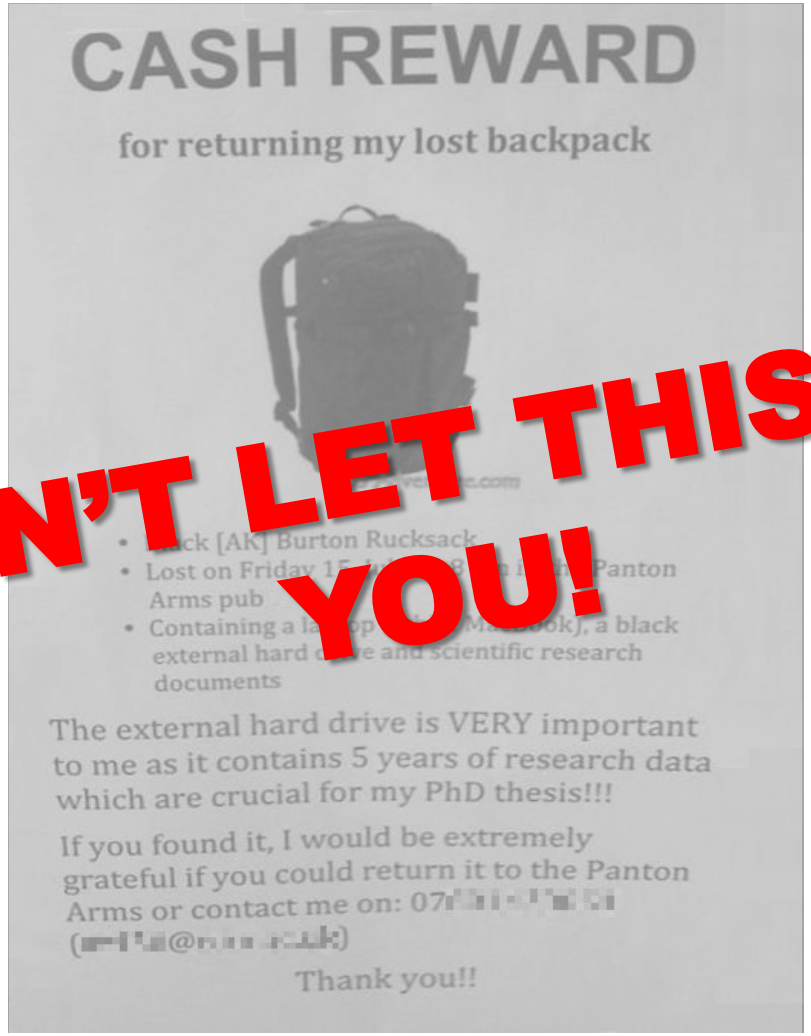
Day-to-day data management

b. storing your data (keeping your data safe)

The Policy: what you need to know

Research data must be:

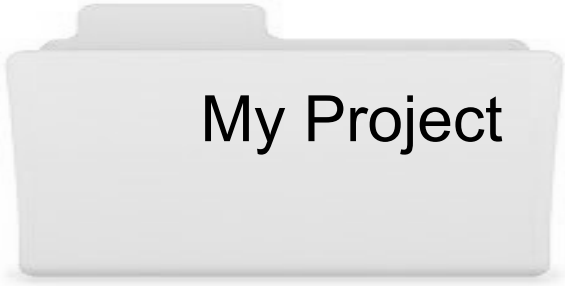
- accurate, complete, authentic and reliable
- identifiable, retrievable and available when needed
- kept safe and secure, avoiding data loss
- kept in a manner that is compliant with legal and ethical obligations, and (if applicable) funder requirements
- disposed of securely.



DON'T LET THIS BE YOU!

http://blogs.ch.cam.ac.uk/pmr/2011/08/01/why-you-need-a-data-management-plan/

Storage – Do's



The University offers a range of facilities to securely store your data, helping it live a long and useful life. The University recommends:

- University filestore/s – individual or shared
  - required for guaranteed UK storage
  - required for "must be kept on site"
- University (york.ac.uk) Google Drive
  - does not guarantee UK location, but does guarantee compliance with UK & EU data protection legislation

To keep your data safe and accessible to you and your collaborators within and without the University.

Storage: Don’ts

Don’t keep your data just on your working machine (laptop or a desktop) – it’s the perfect way to lose your data easily and permanently.

Don’t upload personal/sensitive data to services the University does not have a contract with

- never store personal data in services like Dropbox

Don’t use USB memory sticks

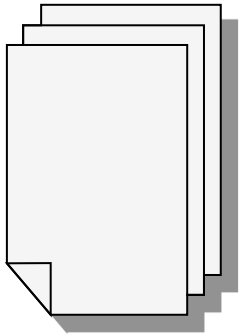
- encrypt them if you have to use them
  - use strong passphrases with all encryption products or the encryption has little value
- never keep the only copy of your data on one.

Backing up

What is the risk of losing your data?
(likelihood / consequence)

Don't have only 1 copy of your data or use only 1 type of data storage (LOCKSS)

multiple copies



keep in different places

automate the process

Day-to-day data management

c. documentation and metadata (describing your data)

Documentation and metadata

- **Documentation** is the contextual information required to make data intelligible and aid interpretation
  - A users' guide to your data
- **Metadata** is similar, but usually more structured
  - Can conform to set standards
  - Sometimes machine readable

Make material understandable

What's obvious now might not be in a few months, years, decades…

Adapted from 'Clay Tablets with Linear B Script' by Dennis, via Flickr: http://www.flickr.com/photos/archer10/5692813531/

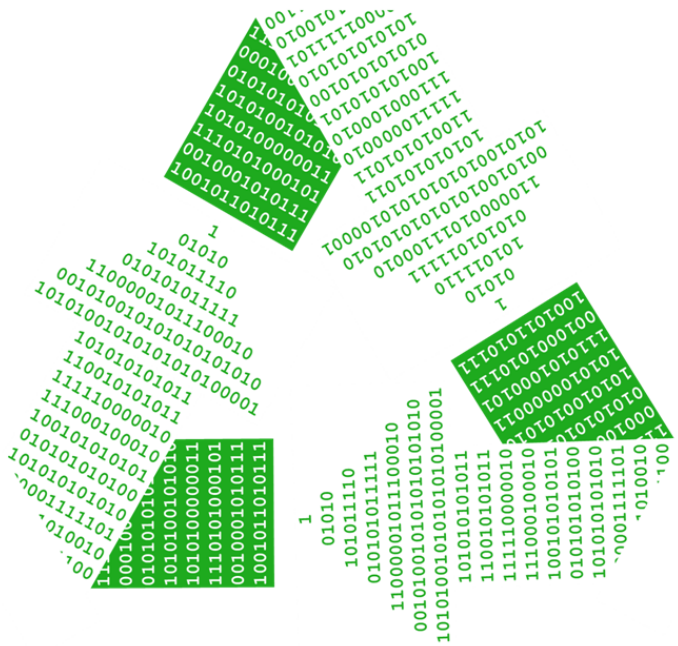
Make material verifiable

Image by woodleywonderworks , via Flickr:
http://www.flickr.com/photos/wwworks/4588700881/

- Detailing your methods helps people understand what you did
- And helps make your work reproducible
- Conclusions can be verified

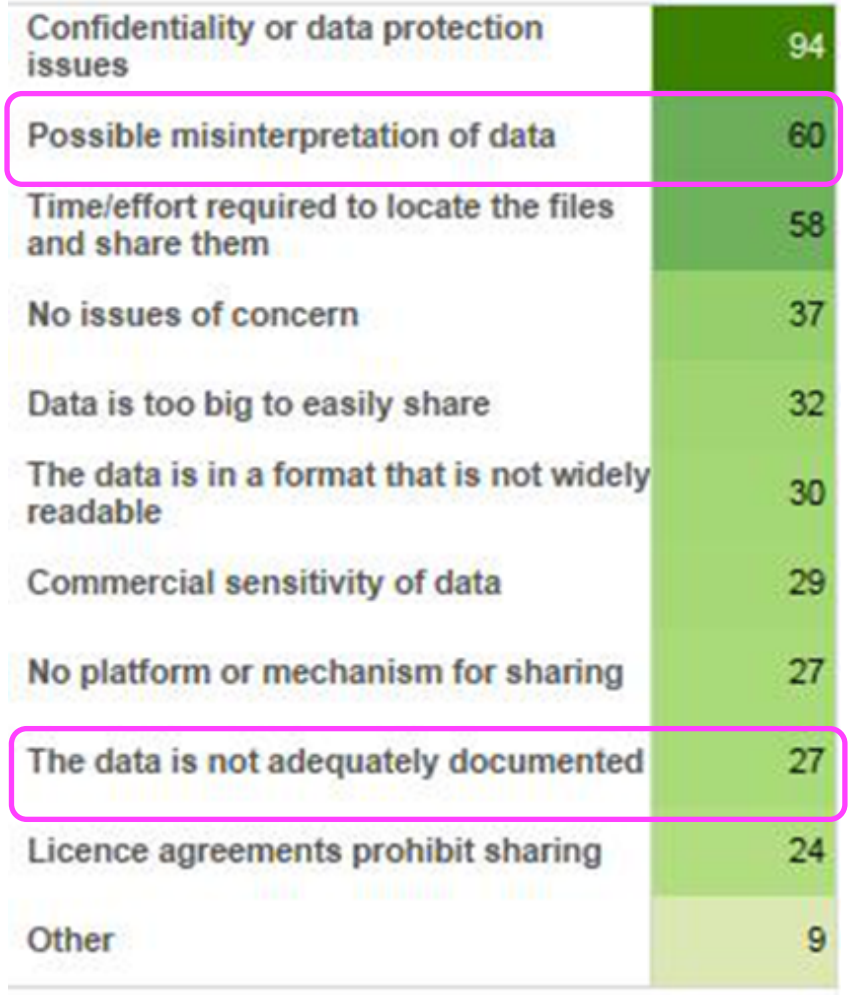
Make material reusable

- You may wish to re-use your own data later on
- Or you may wish to make it available for others to use
- Provide **context** to minimize the risk of misunderstanding or misuse
- Good metadata makes it easier to locate relevant data



Make material reusable

What access issues are of concern to you?



York Research Data Management Questionnaire results –2013

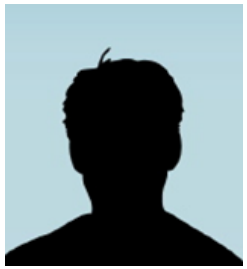
Will someone else understand your data if it isn't documented?

"*The single most useful thing you can do to ensure the long-term preservation of your data is to plan for it to be re-used.* ***Imagining it being reused by someone else who has never met you and who never will meet you****, will cause you to approach the creation and design of your data in a new light. ..... In short, always plan for re-use*"

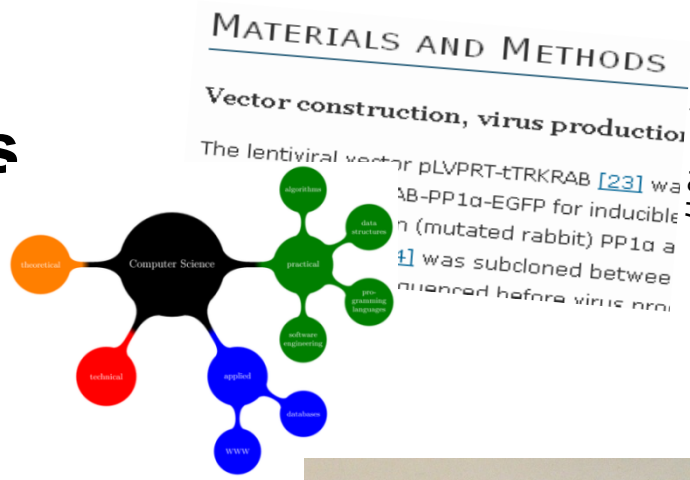
Prof. Julian D. Richards, Director, Archaeology Data Service, University of York

Documentation – what to include

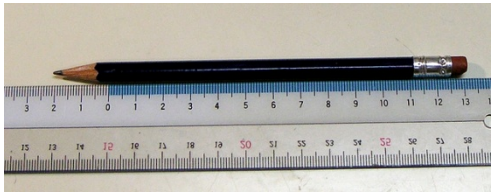
- **Who** created it, **when** and **why**



- **Description** of the item
- **Methodology** and **methods**
- **Units** of measurement
- **Definitions** of jargon, acronyms and code
- **References** to related data



www.texample.net

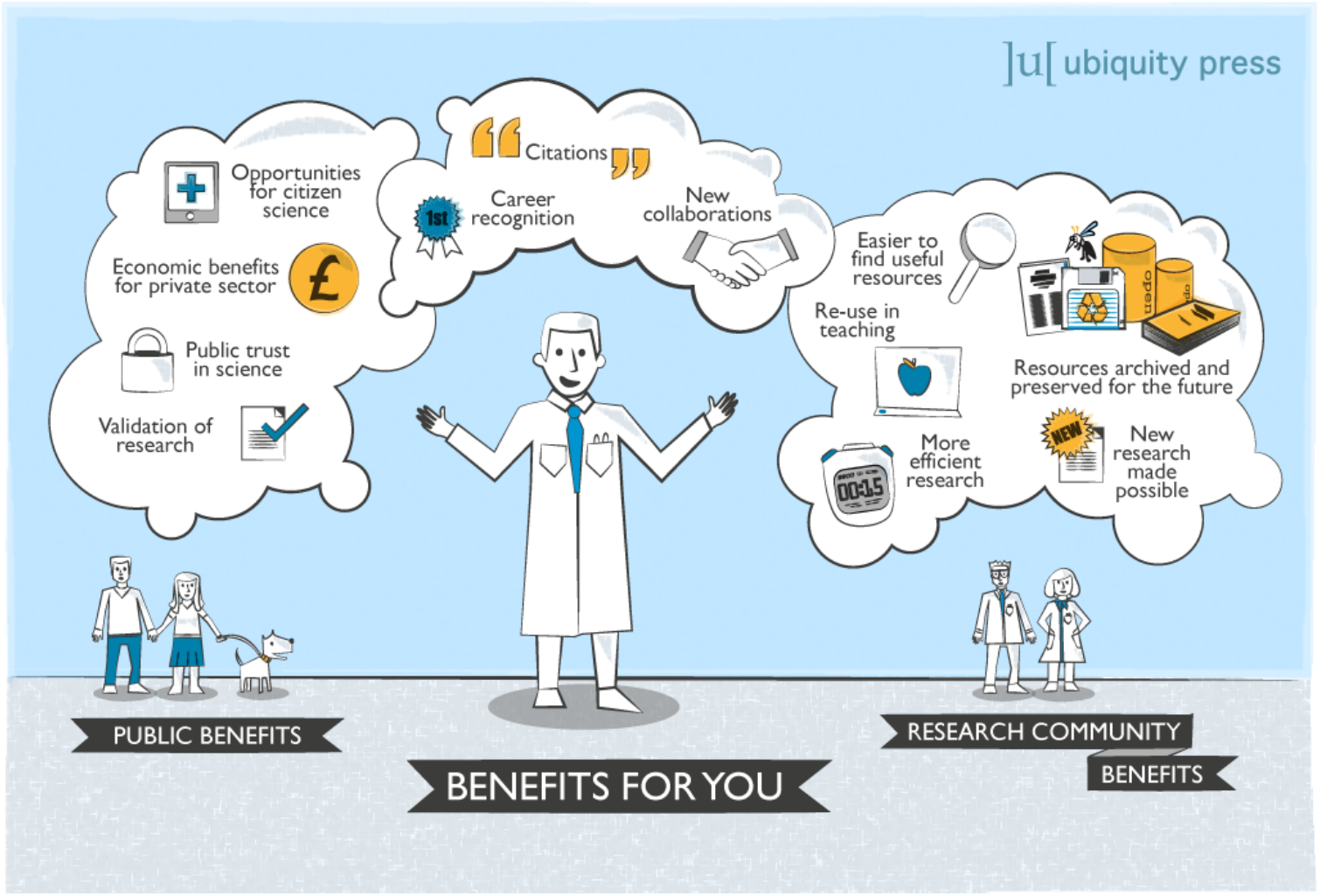


What happens at the end of the project?

Why share data? Be a trailblazer!

- A paradigm shift in how research outputs are viewed is occurring
- Data outputs are of increasing importance – and are likely to become even more so
  - Major journals are increasingly looking to publish datasets alongside articles
- Be at the forefront of an important shift in the academic world

Why share data?



Video by NYU Health Sciences Libraries: http://www.youtube.com/watch?v=N2zK3sAtr-4

Data sharing – concerns

- Ethical concerns
  - Confidential or sensitive data
- Legal concerns
  - Third party data
- Professional concerns
  - Intended publication
  - Commercial issues (e.g. patent protection)

Data sharing – concerns

- Redact or embargo if there is good reason
- Planning ahead can reduce difficulties

What data should you keep?

- ❑ Does the University or your funder stipulate a **retention period** for this material?
- ❑ Are there **legal reasons** to keep it, e.g. health & safety, financial regulation?
- ❑ Are you responsible for keeping the **master copy** (as its creator or owner)?
- ❑ Is the material **fundamental** to your project (e.g. scientific or historical value)?
- ❑ Does the material record one-off events that **cannot be recreated**?
- ❑ Does the record (e.g. email) provide **evidence** that you did something and why?
- ❑ Would the material be **useful in further research** (by you or others)?

Checklist can be found at:
www.lib.cam.ac.uk/dataman/resources/PrePARe_selection_retention_checklist.pdf

What data should you bin?

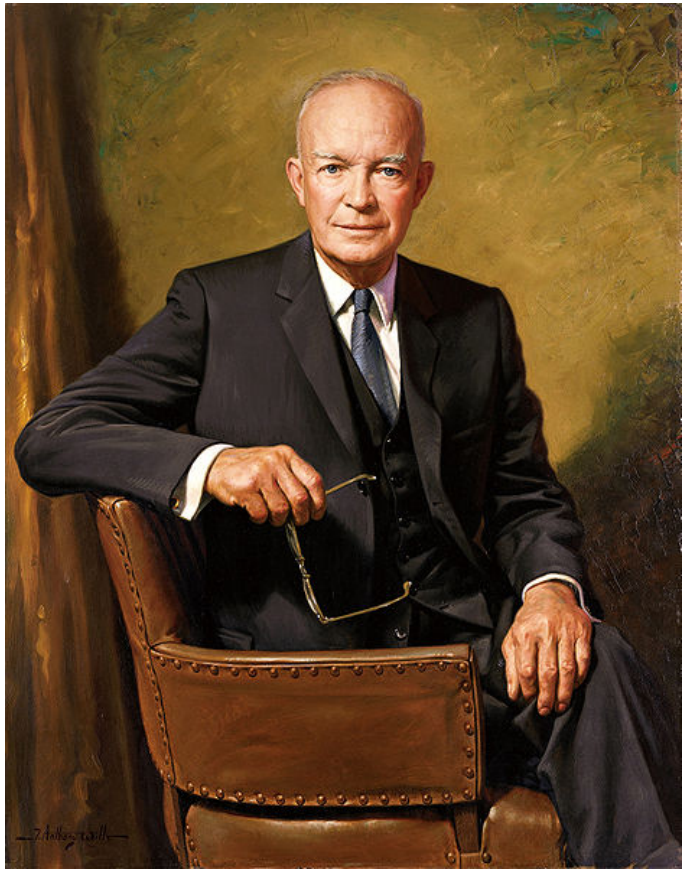
- ❑ Is **someone else responsible** for the master copy?
- ❑ Is it a **duplicate** of a master held elsewhere, e.g. an email attachment?
- ❑ Is the file a **draft** that was subsequently revised?
- ❑ Do **restrictions on reuse** of the material limit the justification for keeping it?
  - ❑ Does **copyright** prevent sharing or reuse of the material?
  - ❑ Are you prevented from archiving/reusing material identifying **living individuals**?
- ❑ Would it be easier / cheaper to **recreate or replicate** the material than to store it?

Checklist can be found at:
www.lib.cam.ac.uk/dataman/resources/PrePARe_selection_retention_checklist.pdf

Data management planning

‘In preparing for battle, I have always found that plans are useless but planning is indispensable.’

Dwight D. Eisenhower



Data Management Plans (DMPs)

A document which outlines how data will be managed over the course of a project.

Should include:

- Description of the data to be collected / created
- Standards / methodologies for data collection and management
- Ethics and Intellectual Property concerns or restrictions
- Plans for data sharing and access
- Strategy for long-term preservation.

Data Management Plans (DMPs)

Many funders require a data management plan as part of grant applications.

DMP*online*

The DCC Data Management Planning Tool

https://dmponline.dcc.ac.uk

Create a data management plan using the DMP*online* tool

My plans

The table below lists the plans that you have created, and any that have been shared with you by others.
These can be edited, shared, exported or deleted at anytime.

Filter plans | Filter

Name	Owner	Shared?	Last edited	Select an action
No funder - Edinburgh template	Me	No	12-09-2014	Edit Share Export Delete
No funder (DCC Template) - Glasgow	Me	No	12-09-2014	Edit Share Export Delete
No funder (DCC Template) - Bath	Me	No	11-09-2014	Edit Share Export Delete
No funder (DCC template) St Andrews	Me	No	26-09-2014	Edit Share Export Delete
DCC Template (with DCC guidance)	Me	No	15-10-2014	Edit Share Export Delete
MRC Template (Uni of Glasgow)	Me	No	17-10-2014	Edit Share Export Delete

Create plan

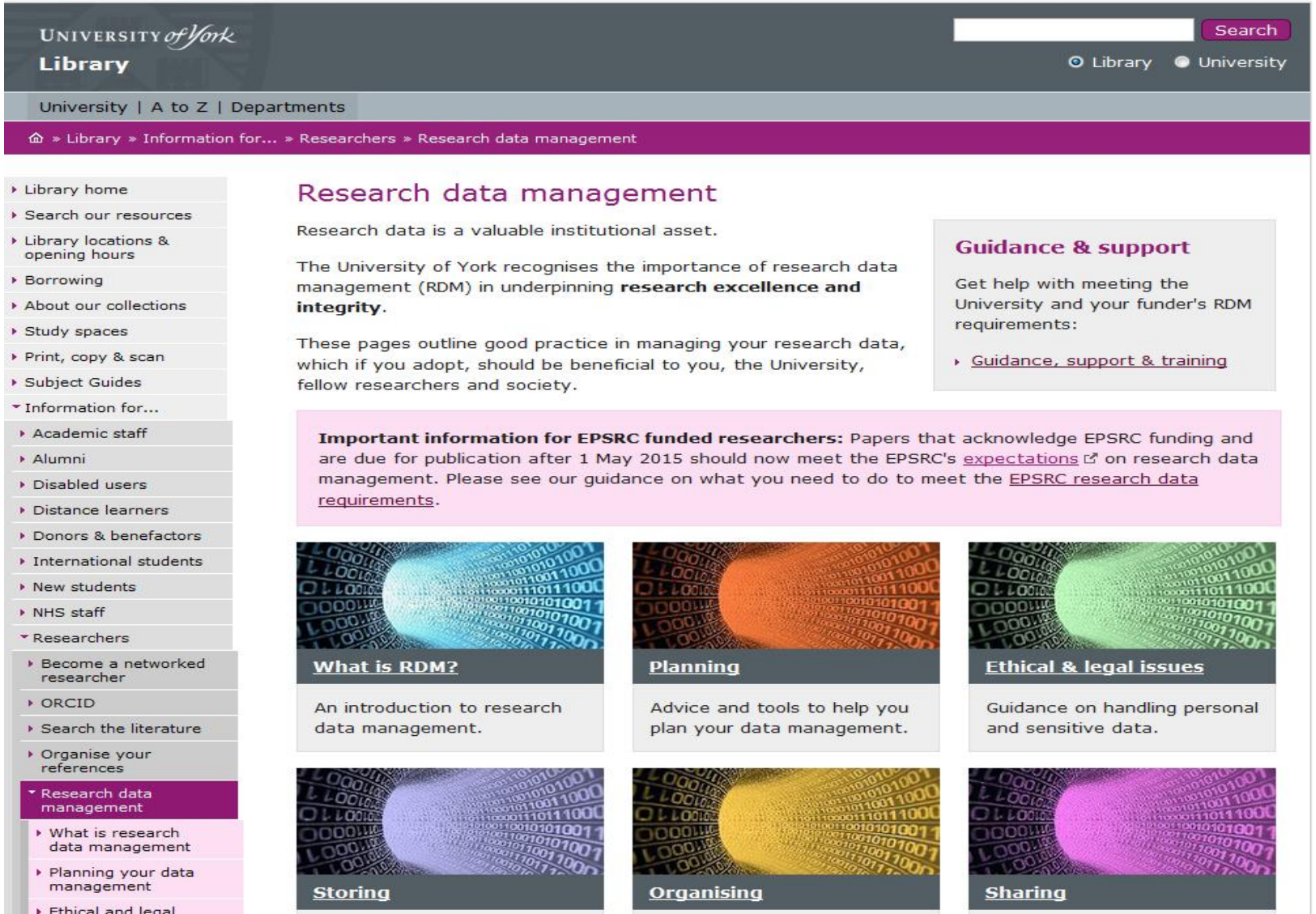
Contact us | Terms of use | DMPonline previous version

DCC Jisc

Further information and resources

RDM web pages

www.york.ac.uk/rdm



Research Data MANTRA

Free online interactive RDM training modules

http://datalib.edina.ac.uk/mantra

We are here to help

- Research Support Team: lib-research-support@york.ac.uk

- IT Support Office: itsupport@york.ac.uk

Rights and re-use

This presentation is an adapted version of the slideshow prepared by the DaMaRO Project at the University of Oxford. The University of York is re-using the slideshow within the terms of the Creative Commons Attribution Non-Commercial Share-Alike License.

Parts of this slideshow draw on teaching materials produced by the PrePARe Project , DATUM for Health and DataTrain Archaeology. All materials are shared under the same license.

Within the terms of this license, we actively encourage sharing, adaptation, and re-use of this material.

Any questions?