

Introduction to Research Data Management

Presentation / Seminar | 19-Feb-2018

Michael Murphy

mmurphy@rvc.ac.uk

researchdata@rvc.ac.uk

Aims of course

- Introduction to world of 'research data' and concepts and procedures of RDM.
- Introduction to delegates' research.
- Understand RVC RDM policy, and funder requirements as applicable.
- Get you thinking about data, and data-related issues that might relate to your own research.
- Advertise the research drives.
- Identify potential avenues for sharing your research data, and highlight potential issues.
- Start a dialogue.

What is research data?

“representations of observations, objects, or other entities used as evidence of phenomena for the purposes of research or scholarship”includes:

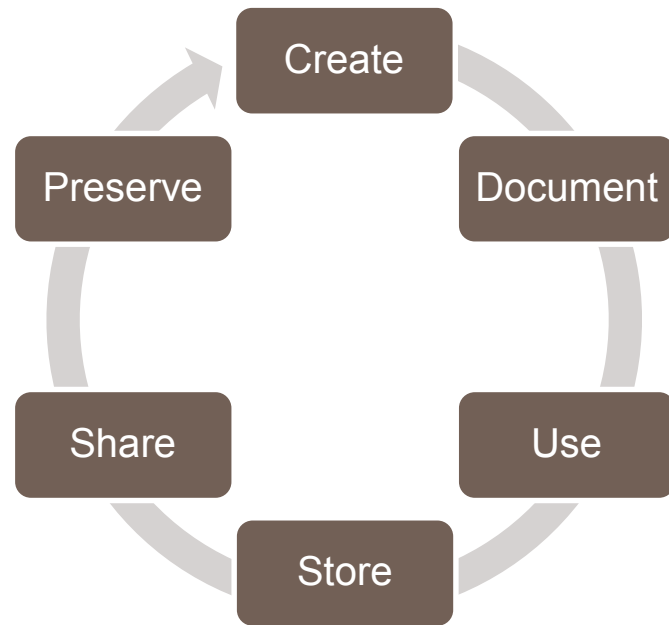
- source data - all data collected, created and used by the research, including data held elsewhere*
- assembled datasets - data extracted or derived from the above*
- referenced data - any subset of the above that has been used in analysis or to draw conclusions. Consistent with whatever is considered ‘supplementary material’ to research findings in your domain.”*

'Categories' of research data

- **Observational:** data captured in real time that is usually unique and irreplaceable (remote sensing data, survey data, field recordings, sample data)
- **Experimental:** data captured from lab equipment that is often reproducible (gene sequences, chromatograms, magnetic field data)
- **Models or simulation:** data generated from test models where model and metadata may be more important than output data from the model.
- **Derived or compiled:** resulting from processing or combining 'raw' data (text and data mining, compiled databases, 3D models)
- **Reference or canonical:** a static or organic conglomeration or collection of datasets, probably published and curated (gene sequence databanks)

What is research data management?

The “organisation of data, from its entry to the research cycle through to the dissemination and archiving of valuable results.”



Collection + Representation + Formatting +
Documentation + Monitoring + Access and
sharing + Updates + Security + Quality
control + Transformations + Destruction

Whyte, A., & Tedds, J. (2011). *Making the case for Research Data Management*. Edinburgh: Digital Curation Centre. Retrieved from http://www.dcc.ac.uk/webfm_send/487.

Image credit: Davidson, J. [Introduction to data management planning](#). ACDH Tool Gallery 2.1: Data Management Plan - Prepare your data for the long term. March 16 2016.

'Representation'

From: Georgakopoulos, Frison, Alvarez, Bertrand, Wells and Campanella (2017) [Reversible Keap1 inhibitors are preferential pharmacological tools to modulate cellular mitophagy](#). Scientific Reports (Nature), 7. p. 10303.

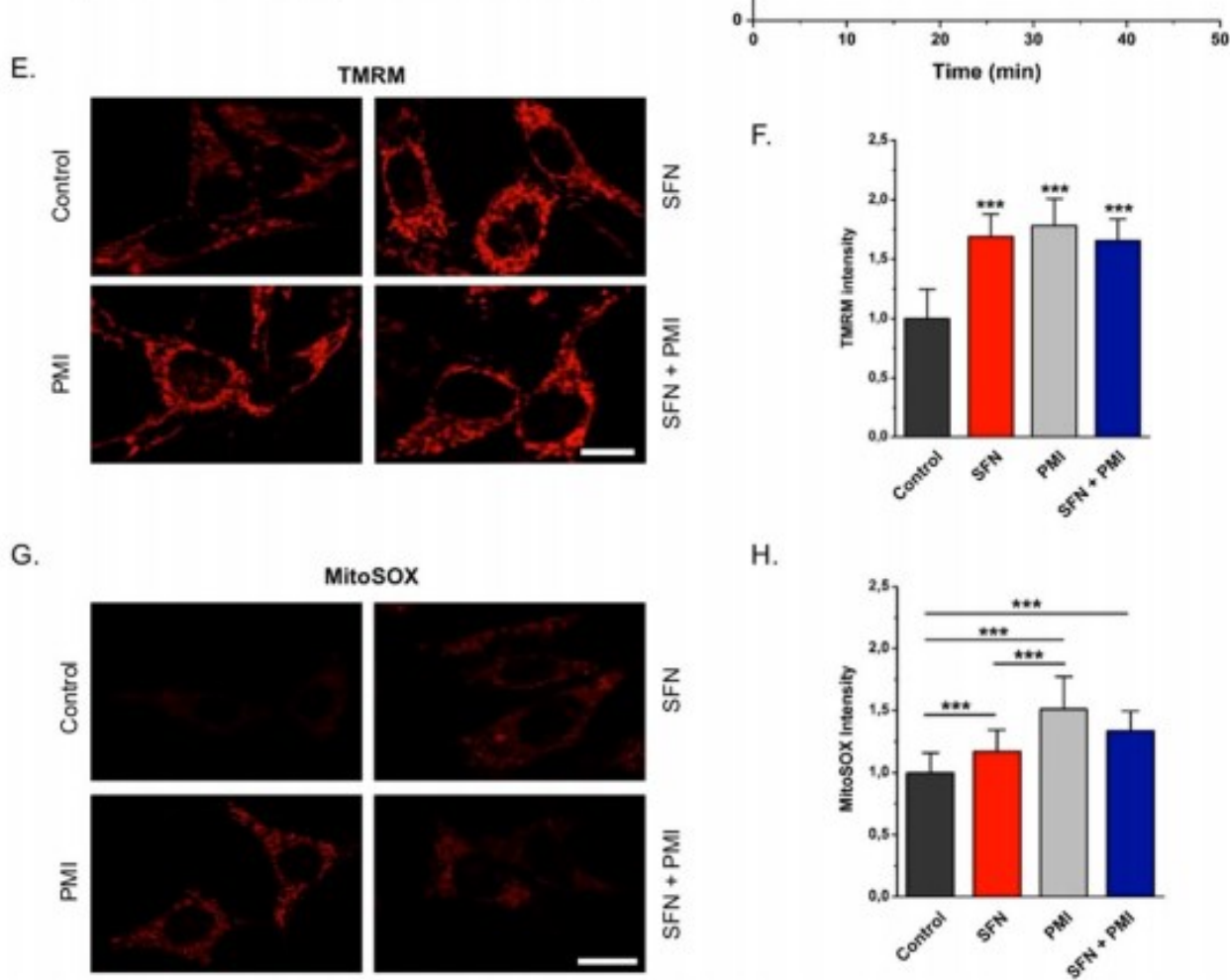


Figure 5. Oxidative metabolism within mitochondria is more enhanced in the presence of PMI compared to SFN. (A) Representative high-resolution confocal images of MEF cells transfected with mt-GFP and treated with DMSO vehicle control, 10 μ M PMI and/or 1 μ M SFN for 24 h. Scale bar represents 10 μ m. A magnification of the merge images is shown in areas demarcated by the white box. (B and C) Graphs showing no differences throughout conditions in (B) elongation and (C) branching of mitochondria ($n > 15$). (D) Graph showing OCR of cells treated with DMSO vehicle control, 10 μ M PMI and/or 1 μ M SFN for 24 h, $n \geq 2$. (E) Representative high-resolution confocal images showing differences in basal $\Delta\Psi_m$ of MEFs treated with DMSO vehicle control, 10 μ M PMI and/or 1 μ M SFN for 24 h and loaded with the potentiometric fluorescent probe TMRM (red) for 30 min. Scale bar represents 10 μ m. (F) Quantification of mean TMRM fluorescence intensity ($n > 30$, *** $p < 0.001$). (G) Representative confocal images of MEFs treated with DMSO vehicle control, 10 μ M PMI and/or 1 μ M SFN for 24 h and incubated with the mitochondrial superoxide sensitive probe mitoSOX (red) for 30 min. (H) Quantification of mean mitoSOX fluorescence intensity ($n > 40$, *** $p < 0.001$). All values are

Reproducibility of research

Most scientists 'can't replicate studies by their peers'

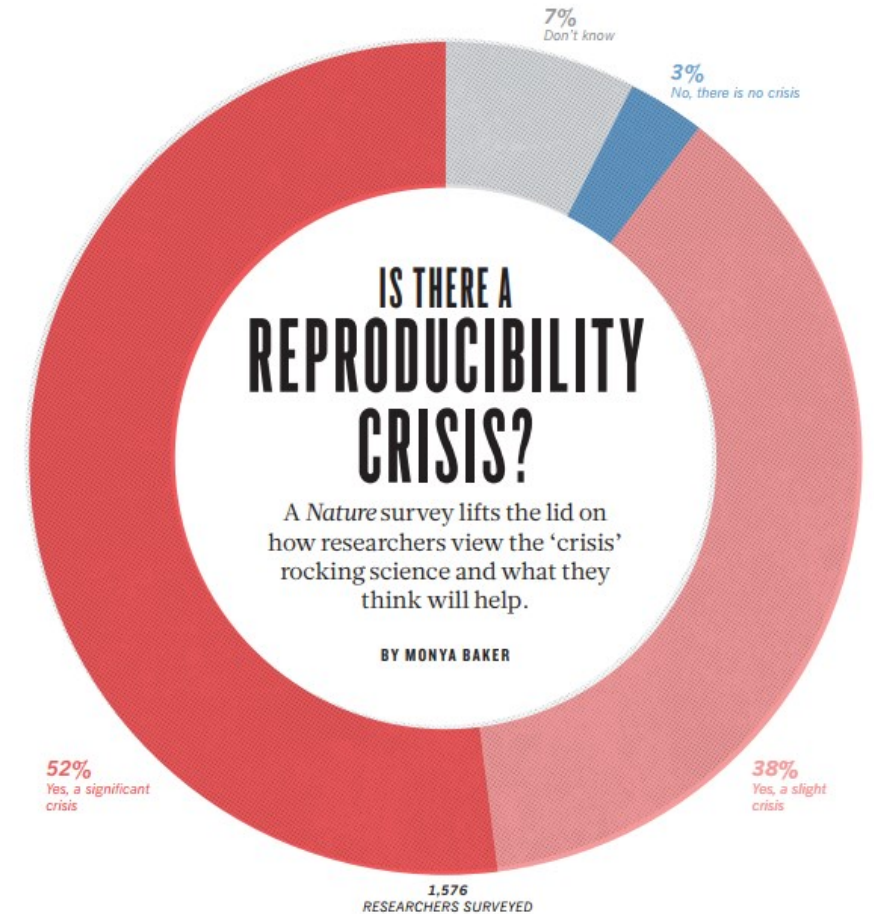
By Tom Feilden
Science correspondent, Today programme

22 February 2017 | Science & Environment



GETTY IMAGES

Scientists attempting to repeat findings reported in five landmark cancer studies confirmed only two



More than 70% of researchers have tried and failed to reproduce another scientist's experiments, and more than half have failed to reproduce their own experiments. Those are some of the telling figures that emerged from *Nature's* survey of 1,576 researchers who took a brief online questionnaire on reproducibility in research.

Failing to reproduce results is a rite of passage, says Marcu biological psychologist at the University of Bristol, UK, who has a standing interest in scientific reproducibility. When he says, "I tried to replicate what looked simple from the lit wasn't able to. Then I had a crisis of confidence, and then I my experience wasn't uncommon."

When contact changes minds: An experiment on transmission of support for gay equality

Michael J. LaCour¹, Donald P. Green²

+ See all authors and affiliations

Science 12 Dec 2014:
Vol. 346, Issue 6215, pp. 1366-1369
DOI: 10.1126/science.1256151

[Article](#) [Figures & Data](#) [Info & Metrics](#) [eLetters](#) [PDF](#)

You are currently viewing the abstract.

[View Full Text](#)

The New York Times

How Bright Promise in Cancer Testing Fell Apart

By GINA KOLATA JULY 7, 2011

This article has been retracted. Please see:
[Is retracted by - June 05, 2015](#)

THE
NEW YORKER

[News](#) [Culture](#) [Books](#) [Business & Tech](#) [Humor](#) [Cartoons](#) [Magazine](#) [Video](#) [Podcasts](#) [Archive](#) [Goings On](#) [Subscribe](#)

JOHN CASSIDY

THE REINHART AND ROGOFF CONTROVERSY: A SUMMING UP

By John Cassidy April 26, 2013

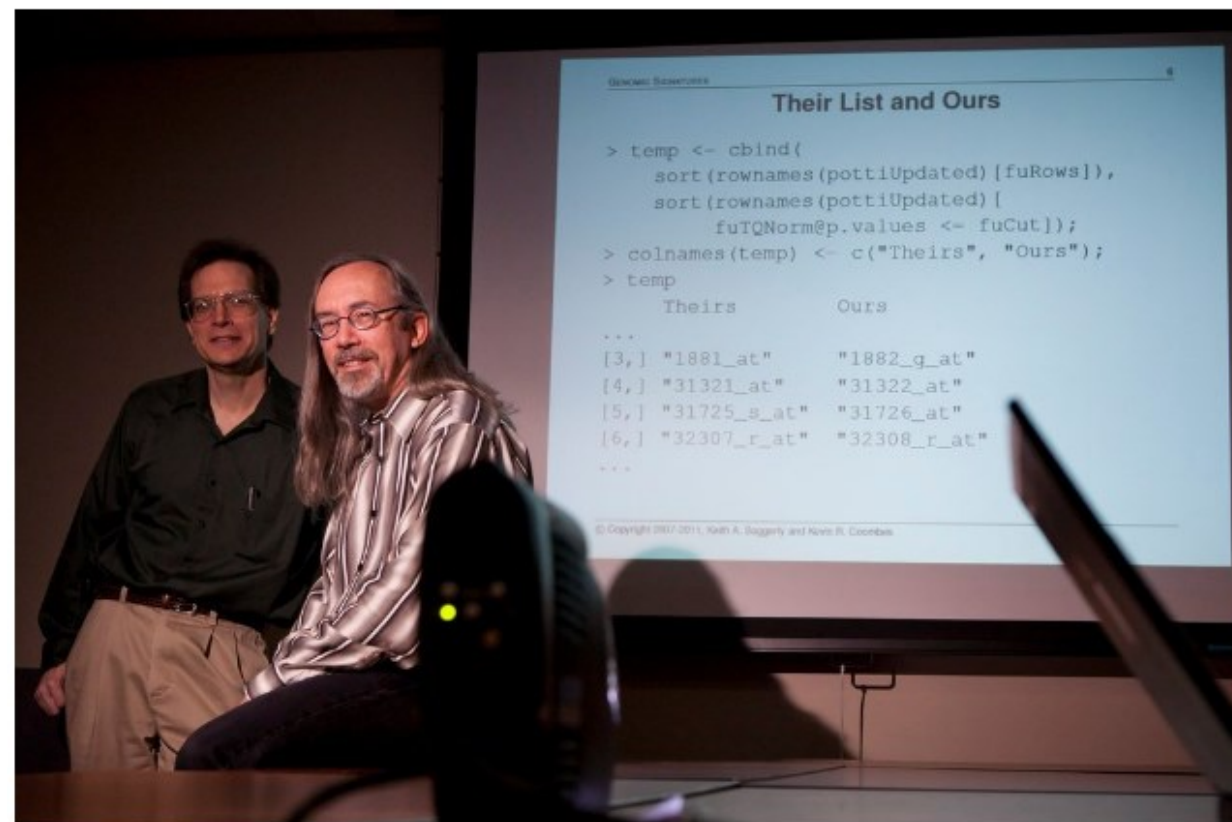


In one of life's little ironies, last Friday's disappointing G.D.P. figures, which reflected a sharp fall in government spending, appeared on the same day that the economists Carmen Reinhart and Kenneth Rogoff published an Op-Ed in the *Times* defending their famous (now infamous) research that conservative



THE
NEW YORKER

The best writing
anywhere, everywhere.
Subscribe for \$1 a week
and get a free tote.

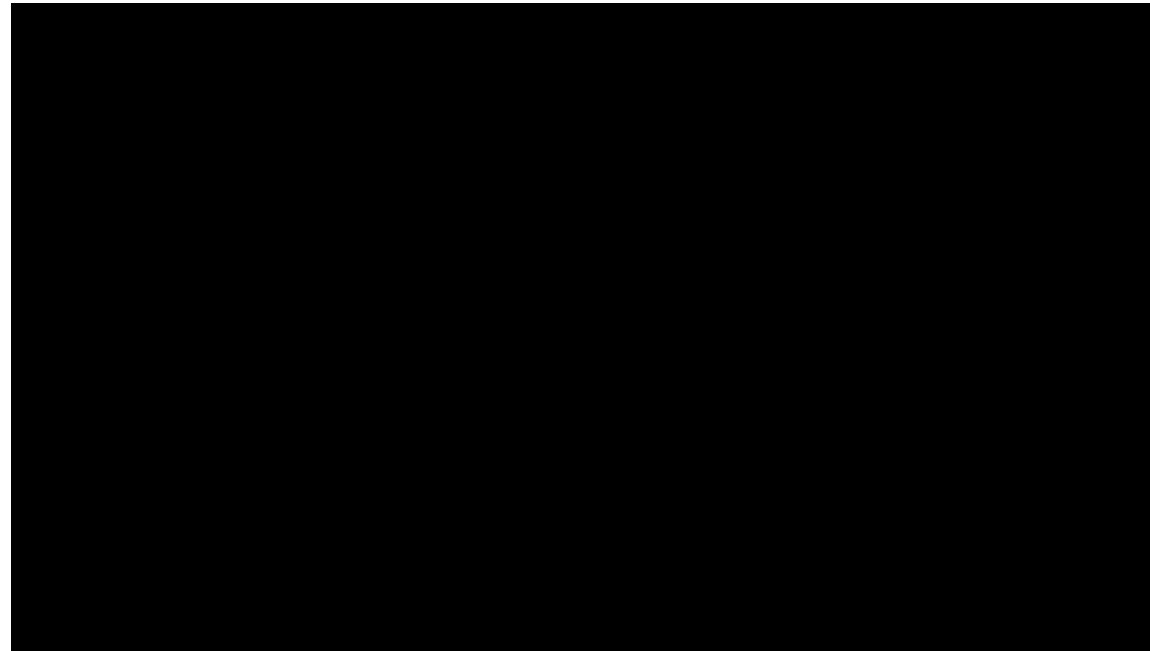


Keith Baggerly, left, and Kevin Coombes, statisticians at M. D. Anderson Cancer Center, found flaws in research on tumors. Michael Stravato for The New York Times

Genomic signatures to guide the use of chemotherapeutics

1. Original article (2006): <https://www.nature.com/articles/nm1491>
2. Reply from Coombes, Wang, and Baggerly: <https://www.nature.com/articles/nm1107-1276b> ('We report here our inability to reproduce their findings')
3. Reply to reply: <https://www.nature.com/articles/nm1107-1277> ('they have not followed our methods in several crucial contexts and have made unjustified conclusions in others, and as a result their interpretation of our process is flawed')
4. [Eventual] Retraction (2011): <https://www.nature.com/articles/nm0111-135>

Excerpt from Keith Baggerly's lecture



<https://youtu.be/7gYIs7uYbMo>

Why is RDM important?

- Helps avoid scandal and retraction of job offers and grant awards.....
- Helps focus your research:
 - What do you need to collect? How will it be collected? How much storage space do you think it will require? How will it be shared? Is the information sensitive? What might someone else need in order to find, evaluate, understand, and reuse the data?
- Vouches for the integrity of your research and research findings in the future (“enable validation”).
- Makes it easier for others to build on your research (increase impact) = more citations
- It has become an obligation for researchers.

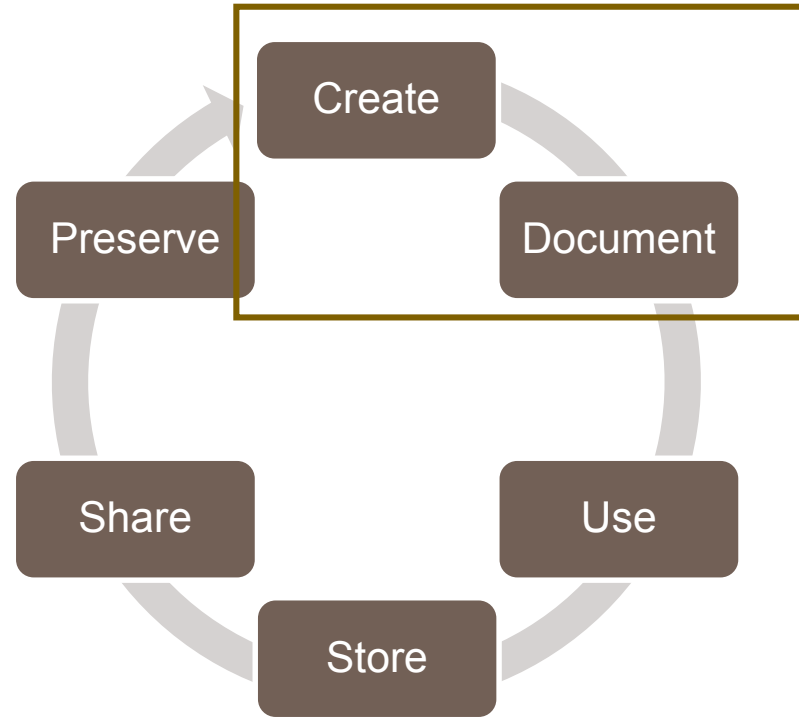


Image credit: Davidson, J. [Introduction to data management planning](#). ACDH Tool Gallery 2.1: Data Management Plan - Prepare your data for the long term. March 16 2016.

Thinking about data collection

Think about the best ‘type’ of measurement during collection phase:

- Nominal (e.g. race, sex)
- Ordinal (e.g. education level, cancer stage)
- Continuous (e.g. temperature, blood pressure)

Stay consistent with formatting (e.g. ‘f’ vs. ‘female’, 2017-11-16 vs 16-Nov-17). Think of how the variable name suggests the input value (e.g. ‘Smoker/Non-smoker’ vs. ‘YES/NO’)

Think about access to ‘physical’ data that you collect/analyse: how will you facilitate access if someone requests it?

Methodologies and standards

Terminology: “organized set of terms in a specific subject field whose meanings have been defined or are generally understood in the relevant field.” [International Standard [ISO 5127](#), *Information and documentation*]

“Broken leg” vs. “Fractured femur” – consistency in collection/recording of data makes it easier to store, to migrate, to analyse, to expose to machine learning, etc.

Standards exist to try to control for the ambiguities of language, e.g. [Clinical Data Interchange Standards Consortium \(CDISC\)](#).

Taxonomies & controlled vocabularies

How many ways can you say ‘female’? A list of values submitted in the ‘sex’ attribute for sequences and samples that all represent female [in European Nucleotide Archive]

18-day pregnant females	female (phenotype)	hexaploid female	dikaryon	female, 6–8 weeks old	pseudohermaphroditic female
2-yr-old female	female (pregnant)	individual female	dioecious female	female, other	remale
3 female	female (worker)	lgb*cc females	diploid female	female, pooled	semi-engorged female
400-yr-old female	female child	mare	f	female, spayed	sex: female
adult female	female mice	metafemale	famale	female, virgin	sexual oviparous female
asexual female	female parent	monosex female	femal	female, worker	sf
castrate female	female plant	normal female	femal	female(gynocious)	sterile female
cf.female	female with eggs	ovigerous female	female	femalen	sterile female worker
cystocarpic female	female worker	oviparous sexual females	female - worker	females	strictly female
dikaryon	female, 6–8 weeks old	pseudohermaphroditic female	female (alate sexual)	females only	tetraploid female
dioecious female	female, other	remale	female (calf)	femele	thelytoky
diploid female	female, pooled	semi-engorged female	female (f-o)	femlale	vitellogenic replete female
f	female, spayed	sex: female	female (gynocious)	gynocious	worker
famale	female, virgin	sexual oviparous female	female (lactating)	healthy female	worker bee

from: Silvester et al. (2015).
 Content discovery and
 retrieval services at the
 European Nucleotide Archive.
Nucleic Acids Research
[.https://doi.org/10.1093/nar/gku1129](https://doi.org/10.1093/nar/gku1129)

Over to you (DMP checklist)

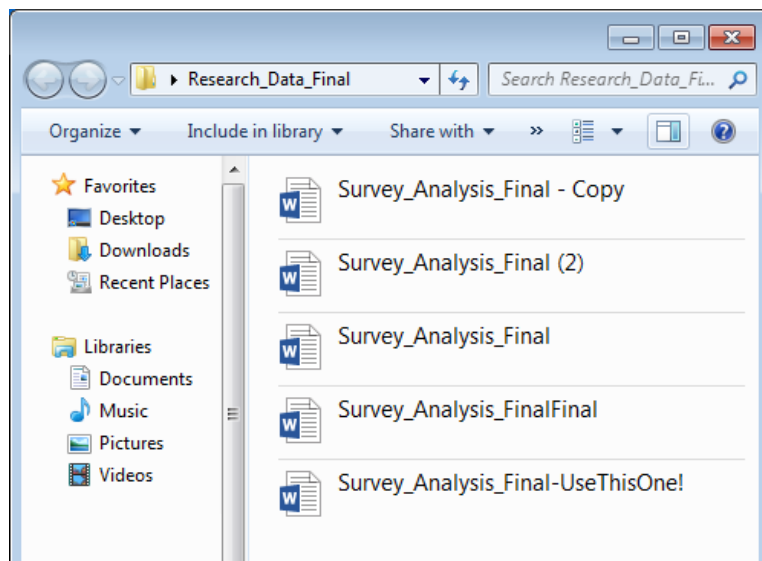
- Are you creating ‘new’ data in the course of your research? If so, how is it collected?
 - Observation, surveys, DNA, MRIs, measurements..
- What format does it take?
 - survey responses, case reports, spreadsheet data, “high density data” (MRI images, slo-mo video), biometric data, geospatial, etc...
- What type of analysis will be applied?
 - coding, statistical, modelling, quantitative, qualitative..
- What tools are used for analysis?
 - Excel, STATA, Matlab, R, Python...

Storing, describing, depositing, citing datasets

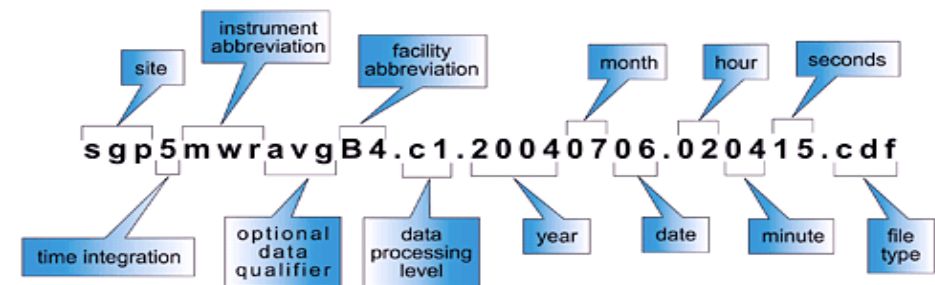
- Where will you store your data? Where else will you store it? How will you secure it?
 - Research drive storage (O: or R: drive) complies with RCUK “Common Principles” standards on retention, backups, security.
- How will you describe it?
 - Is there a metadata standard for research data in your discipline?
- How will you name and order your files?
 - Remember to link storage of datasets to specific publications where possible.
- What would someone need to know/do in order to make sense of your data? To reproduce your results?

Naming your files

- Adopt a convention with your research group.
- Include dates.
- Use hyphens or underscores not spaces e.g. day-sheet, day_sheet
- Dile naming conventions might be prescribed by the nature of the data



An example netCDF data file name is depicted below:



Example from ARM Climate Research Facility
<https://www.arm.gov/policies/datapolicies/formattin-g-and-file-naming-protocols>

File media, file types

Is your data on physical media? Does your laptop have a disc drive? Did your previous laptop have one?

Is a proprietary programme (e.g. MS Access) required to make use of the file? Is there a non-proprietary alternative?

.doc/.docx > .odf/.rtf

.ai/.psd > .tiff

.xls > .csv

<http://www.data-archive.ac.uk/create-manage/format/formats-table>

Metadata (concept)

A formal description of a dataset which conforms to a particular structure.

One typical use of metadata is to create a catalogue record for a dataset held in an archive.

Making data intelligible, verifiable, reusable, retrievable, machine readable (where possible) offers potential for greater impact.

dc.contributor.author	Tugume, Arthur	en
dc.contributor.author	Cuellar, Wilmer	en
dc.contributor.author	Mukasa, Settumba	en
dc.contributor.author	Valkonen, Jari	en
dc.date.accessioned	2010-05-14T15:41:40Z	
dc.date.available	2010-05-14T15:41:40Z	
dc.date.issued	2010-07-01	
dc.identifier	doi:10.5061/dryad.1573	
dc.identifier.citation	Tugume AK, Cuéllar WJ, Mukasa SB, Valkonen JPT (2010) Molecular genetic analysis of virus isolates from wild and cultivated plants demonstrates that East Africa is a hotspot for the evolution and diversification of Sweet potato feathery mottle virus. <i>Molecular Ecology</i> 19: 3139-3156.	
dc.identifier.uri	http://hdl.handle.net/10255/dryad.1573	
dc.description	Sweet potato feathery mottle virus (SPFMV, genus Potyvirus) is globally the most common pathogen of cultivated sweetpotatoes (<i>Ipomoea batatas</i> ; Convolvulaceae). Although >100 SPFMV isolates have been sequence-characterized from cultivated sweetpotatos across the world, little is known about SPFMV isolates from wild hosts and the evolutionary forces shaping SPFMV population structures. In this	en

Metadata (practice)

- DataCite Schema for the Publication and Citation of Research Data - <http://doi.org/10.5438/0013> : a list of core metadata properties chosen for an accurate and consistent identification of a resource for citation and retrieval purposes
- SDMX standard (<http://sdmx.org/>) for statistical data
- DDI metadata standard (<http://www.ddialliance.org/>) for social science data
- Discipline-specific metadata schema:
 - FAIRsharing.org (recommendations by funder, discipline, journal)
 - NIH Common Data Elements:
https://www.nlm.nih.gov/cde/summary_table_1.html
 - DCC list: <http://www.dcc.ac.uk/resources/metadata-standards>.

What metadata should be created?

- Title
- Grant Number (if applicable)
- Description: paragraph that describes the content
- Creator + Contributors
- Date (of publication/completion)
- Rights: Ownership and other rights associated with the data
- Access Restrictions
- Temporal coverage: period over which the data collection took place
- Spatial coverage: geographic region in which data collection took place

Description/documentation

Elements to capture:

- 1) Project information: research question/hypotheses, data capture and analysis methods, sampling frame used (e.g. geographic location and time period), instruments and measures used.
- 2) Description of individual files within the dataset: context in which it was created (e.g. geographic location and time period), audit trail of activities performed when capturing, processing, and analysing contained content, relationship between files and project as a whole, copyright claims and licencing info
- 3) Codebook: information necessary to understand variables contains within the data, including variable description.

Description/documentation

- **Project level:** what the study set out to do, how it contributes new knowledge to the field, what the research questions/hypotheses were, what methodologies were used, what sampling frames were used, what instruments and measures were used, etc. A complete academic thesis normally contains this information in detail, but a published article may not. If a dataset is shared, a detailed technical report will need to be included for the user to understand how the data were collected and processed. You should also provide a sample bibliographic citation to indicate how you would like secondary users of your data to cite it in any publications, etc.
- **File or database level:** how all the files (or tables in a database) that make up the dataset relate to each other; what format they are in; whether they supercede or are superceded by previous files. A readme.txt file is the classic way of accounting for all the files and folders in a project.
- **Variable or item level:** the key to understanding research results is knowing exactly how an object of analysis came about. Not just, for example, a variable name at the top of a spreadsheet file, but the full label explaining the meaning of that variable in terms of how it was operationalised.

An example from LSHTM

Nash, S, Mentzer, AJ, Lule, SA, Kizito, D, Smits, G, van der Klis, FR and Elliott, A. 2017. Entebbe Mother and Baby Study - Data at one year. [Online]. *London School of Hygiene & Tropical Medicine*, London, United Kingdom. Available from: [10.17037/DATA.128](https://doi.org/10.17037/DATA.128)

Note:

- Metadata elements (incl. grant number, locational coordinates, data capture method).
- File formats: STATA file (.dta) and tab delimited text file (.txt).
- User guide (information about dataset and accompanying documentation) + link to copy of data collection form, codebook, related publications.

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”

RCUK Common Principles on Data Policy



<http://www.dcc.ac.uk/resources/policy-and-legal/funders-data-policies>



- A 'Data Sharing Plan' is required at proposal stage. Compliance is checked during final assessment.
- Data to be released when findings published / within 3yrs of project completion.
- Data must be available and appropriate for secondary use for a minimum of 10 years after project end.
- Sharing via a repository is expected, but a specific repository is not mandated.
- Funding to support the management and sharing of research data can be budgeted.
- Research outputs must provide access to any underpinning research datasets.

RVC RDM policy (abridged)

- The College is committed to the ethos of open access to research data, within the parameters of intellectual property protection and contractual requirements.
- Responsibility for research data management and the creation of data management plan lies primarily with Principal Investigators (PIs) as part of their project management responsibilities.
- **All new research proposals must include research data management plans addressing data capture, management, integrity, confidentiality, retention, sharing and publication.**
- The College will provide mechanisms and services for storage, backup, registration, deposit and retention of research data assets, supporting future access after project completion.

RVC facilities and services


- Secure storage HH and Camden.
- Up to 2 hours of battery backup in each data centre.
- 500TB = 500,000GB Usable storage.
- 130TB allocated O:\Research_Storage and R:\Research_Storage (Access to authorised users ONLY).
- Hourly backups (0000, 0600, 1200, 1800) (5 backup copies retained).
- Daily backup 0010 (5 backup copies retained).
- Weekly – Sundays 0015 (16 backup copies retained).
- Long-term data retention on backup tape cartridges (10 years).
- Data is also mirrored between HH and CM storage systems, on hourly basis.
- 2 virus scanning servers.

<http://www.rvc.ac.uk/research/about/research-data-management>

Good alternative to



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!

Backing up

- Make use of the RVC Research Storage Drive.
- Make multiple copies and keep them in different places (LOCKSS)
- Consider automating the process.
- Create a Data backup plan.
- Ensure consideration of data sensitivity

Cloud storage

- Not a permanent solution!
- UK Data Protection Law & Personal Data: Don't store sensitive information



Publishing

DOIs: unique identifiers for datasets (be mindful of storing sensitive data)

- [Figshare.com](https://www.figshare.com)
- [Synapse.org](https://www.synapse.org)
- data.mendeley.com
- [UK Data Archive](https://www.ukdataservice.ac.uk)

[Creativecommons.com](https://creativecommons.org/licenses/) (licensing)

Depositing

Be familiar with the requirements or expectations of your funder:

Studies may share their data by archiving their data collection (or a subset) at a discipline based repository like the UK Data Archive (www.data-archive.ac.uk), or at an institutional repository that can preserve data and make them available to users. [MRC data policy]

If the data is associated with a journal article, be aware of the journal's data deposit policy (if applicable):

Sequences must be submitted to the EMBL Database Library or GenBank. Protein sequences that have been determined by direct sequencing of the protein must be submitted to SWISS-PROT at the EBI. All accession numbers should be included in the manuscript. [JEB data deposit policy]

We expect that all researchers submitting to PLOS submissions in which software is the central part of the manuscript will make all relevant software available without restrictions upon publication of the work. [PLoS]

<http://journals.plos.org/plosone/s/materials-and-software-sharing>

Repositories



[Dryad.org](http://dryad.org)



[UK Data Archive](http://ukdataservice.ac.uk)



[Figshare](http://figshare.com)



[ENA](http://ena.ebi.ac.uk)



[GenBank](http://ncbi.nlm.nih.gov)



[GE Omnibus](http://ncbi.nlm.nih.gov/geo)

<http://www.re3data.org/> : search or browse by subject

Tools and resources

ORCID: get yourself an ID @ orcid.org

Digital Curation Centre: online Data Management Plan tool, plus metadata resources, guides and checklists and FAQs @ dcc.ac.uk/resources/

RVC RDM site @ rvc.ac.uk/research/about/research-data-management

RCUK Common Principles on Data Policy @ rcuk.ac.uk/research/datapolicy/

Next steps

Book a DMP consultation (or: teach me about the data you create and how you make it work for you): researchdata@rvc.ac.uk

Stake a claim to some storage space on the O: or R: drive: have your PI/supervisor email researchdata@rvc.ac.uk