



Georgina Parsons,
Research Data Manager, 2019.

Research Data Management 1: Introduction



- 
1. RDM definitions and importance.
 2. Data access and organisation.
 3. Data formats and backups.
 4. Data documentation.
 5. Data sharing and security.

[GP]



1. RDM definitions and importance.

2. Data access and organisation.

3. Data formats and backups.

4. Data documentation.

5. Data sharing and security.

[IS]



Definitions: “research data management”



[IS]

One possible definition. Data management practices cover the entire lifecycle of the data, from planning the research to long-term preservation of data after the research investigation has concluded.

[Creating: design research, plan data management, collect data (experiment, observe, measure, simulate) and metadata

Processing: data entry, transcription, validation, anonymisation, description, storage

Analysis: interpretation, derivation, preparation for publication

Preservation: format migration, documentation, archival storage

Sharing: distribution, access controls, copyright/licensing

Re-use: follow-up research, teaching and learning]



“Research data”: any data that underpins your findings

It could be: experimental, observational, simulation, or derived.

- textual data e.g. interview responses;
- numerical data e.g. experimental results;
- multimedia e.g. photos/videos of processes/interviews;
- models e.g. computational models;
- code e.g. scripts or software;
- physical items e.g. lab notebooks.

Poll: what research data will you be creating?

5

[IS]

RD could include survey responses, interview transcripts, financial data, experimental results, images, interview videos, CAD files, statistical models, 3D models, software written during the project, laboratory notebooks. If experimenting on samples, it's the results, not the samples.

Research Information Network classification:

Experimental: data from experimental results, e.g. from lab equipment, often reproducible, but can be expensive e.g. chromatograms, microassays

Observational: data captured in real time, usually unique and irreplaceable e.g. brain images, survey data

Simulation: data generated from test models where model and metadata may be more important than output data from the model e.g. economic or climate models

Derived or compiled: resulting from processing or combining 'raw' data, often reproducible but expensive e.g. compiled databases, text mining, aggregate census data

Reference or canonical: a (static or organic) conglomeration or collection of smaller (peer reviewed) datasets, most probably published and curated e.g. gene databanks, crystallographic databases



Why is good RDM important?



RDM covers a lot of aspects of handling data so is important for a number of reasons...

1. A large element of RDM is keeping data secure and backed up, minimising the risk of data loss, whether that's through natural disasters such as fire or flooding, equipment theft, or technical failures.
2. Much like good referencing, being organised and following RDM best practice principles from the outset is well worth it. 'There is a discipline you need to teach yourself, I think, to do it at the time because if you don't, you regret it later if you have to go back' [Professor Haywood in <https://www.youtube.com/watch?v=i2jcOJOFUZg>].
3. RDM also covers data sharing, so whilst your data had one purpose, it may be used for others, even across disciplines, with potentially important consequences.
 - The biomarkers for Alzheimer's were discovered after years of lack of progress, after researchers around the world agreed to "park their egos at the door" and immediately share their findings (http://www.nytimes.com/2010/08/13/health/research/13alzheimer.html?_r=1).
 - More recently data was combined from a variety of earlier studies on excavated dinosaur teeth, and led to a new discovery about the evolution of theropod dinosaurs "that significantly advances science as a whole." (<http://blogs.plos.org/paleo/2013/01/25/and-this-is-why-we-should-always-provide-our-data/>)
 - A model of e-resilience, to study the impact of online shopping on high street shops

and thus help them survive, was only possible because it drew on so many existing datasets to pull together to create this model and draw new conclusions, one project simply couldn't create all the data required (<https://www.ukdataservice.ac.uk/use-data/data-in-use/case-study/?id=202>).

4. Various studies have looked at whether sharing data affects citations of your papers – and all have found that it does increase citations. The effects were from 69% to 9% (independently of journal impact factor, date of publication, and author country of origin), all positive.* You also do not want a retraction to your name such as <http://dx.doi.org/10.1016/j.scijus.2015.04.005> !

5. RDM has become a priority in universities recently due to a “reproducibility crisis” in research, and a resultant push towards open science, i.e. transparency of methods and data as well as just results. Some would say that not sharing data is misconduct, it's certainly becoming questionable behaviour, because it's an obvious ‘safeguard’. If someone accuses you of having falsified data underpinning papers, how do you prove you'd done the work as described?

Also shift in attitudes due to increasing number of scandals: “What type of academic culture allowed Stapel to continue his misconduct for so long?” Multiple major scandals in the Netherlands have led the Dutch to fund an investigation into research integrity, costing 8million euros. With data routinely made available, it can still be fabricated but there's the acknowledgement that it will be checked for reproducibility, and with spreadsheets online, automated checks can be done (eg there are usually patterns in falsified data – people who falsify are lazy!).

Open data in particular makes people think twice before committing fraud, and think twice about the robustness of their research.

6. UKRI and other funding bodies have policies on how we must manage and share data. Notably, UKRI bodies are government-funded, so this is taxpayer money paying for the research: the principle is that publicly funded research data should be made publicly available in a timely and responsible manner (without damaging the research process). Funders are mandating good RDM because of the previous five reasons, not because they like giving extra work to researchers!

So RDM has a lot of benefits, achieved by adhering to best practice in data management.

Credits:

**Sharing Detailed Research Data Is Associated with Increased Citation Rate (DOI: 10.1371/journal.pone.0000308). See also: The enduring value of social science research (<http://hdl.handle.net/2027.42/78307>); On the citation advantage of linking to data (<https://hal-hprints.archives-ouvertes.fr/hprints-00714715v2>).*

Image: Labeling, CC-BY-NC by Paul Istoan

<https://www.flickr.com/photos/vamapaul/5752351132/>

Image: Scientific integrity,

<http://blogs.lse.ac.uk/impactofsocialsciences/2015/07/03/data-secrecy-bad-science-or-scientific-misconduct>

- 
1. RDM definitions and importance.
 - 2. Data access and organisation.**
 3. Data formats and backups.
 4. Data documentation.
 5. Data sharing and security.

7

Cranfield University

[GP]

The next four sections cover practical elements of RDM. It is likely you will soon have to write a DMP (all doctoral students must write one before starting data collection, and staff should also write one for each project), and your DMP is where you explain how you'll address each of the elements we are about to discuss.

[IS]

Introduce first section.



Data access and organisation



[NYU Health Sciences Library \(2012\) Data Sharing, Part 1 of 3](#)

[IS]

While you watch it, think about the mistakes you can spot in the researcher's behaviour. We'll then go through the RDM best practice elements that would have prevented these problems.

...

We noted: "it was published in Science, which requires that you share your data" ... "everything you need to know is in the article" ... "it is on a USB drive" ... "I forgot to label the boxes"



Data access statements or data citations

Every publication should include a data access statement saying how the underlying data can be accessed (or why it can't):

- "Data is available at <https://doi.org/10.17862/cranfield.rd.5519725>."
- "Due to the politically sensitive nature of the research, no participants consented to their data being retained or shared."
- "Data will be available at [10.17862/cranfield.rd.3507755.v1](https://doi.org/10.17862/cranfield.rd.3507755.v1) after a five-year embargo by agreement with the commercial partner."

Use a normal citation for other data you reused:

- M Partridge (2014) Spectra evolution during coating. Figshare. DOI: [10.6084/m9.figshare.1004612](https://doi.org/10.6084/m9.figshare.1004612)

[IS]

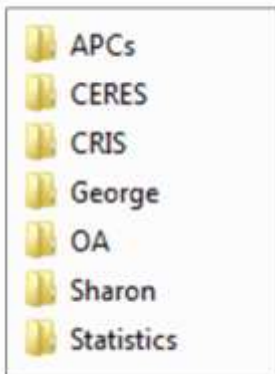
"Everything you need to know is in the article": it wasn't, as he was non-compliant with publisher requirements to share the data, and didn't include a data statement or citation.

Funders such as EPSRC are now compliance checking to ensure that a data access statement is present.

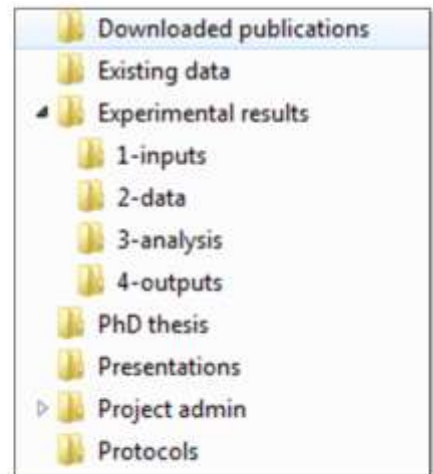


File and folder organisation.

Before:



After:



10

[IS]

“on a USB drive in unlabelled boxes” – badly labelled folders are the same principle as unlabelled boxes.

Most people simply don't consider folder structure as a whole, but it's really important and makes your life so much easier down the line. So before you start data collection, do consider how you will organise your data.

The slide shows three examples. The left two are a real-life example from my work, when we migrated content from our library network drive to a collaboration site. The 'before' image shows that even librarians can easily get in a mess! Why are APCs (OA payments) separate from OA? Where would you look for OA statistics – in the OA folder or the Statistics folder? If you want procedures for adding OA articles to CRIS and CERES (this is done in one go), do you look in OA or in CRIS or in CERES? If you were looking for some project work, would you remember whether it was George or Sharon who worked on it and thus know which folder to check? At the time, it all made perfect sense, but that's the point:

During your project, you'll just know where things are as it's fresh, but when you come back to the files in the future (or when other researchers do), it can be very

confusing if you haven't thought about a logical structure for your work. And it will waste your time trying to find things.

The example on the right is for a PhD project rather than library work, but it's not "the right way", just an example. You might find this suitable, or if you're doing lots of experiments, you might decide you want one folder for each key experimental condition. Whatever you choose, keep a clear structure as you go along.



File and folder names should:



- describe the contents/subject;
- be short and concise;
- avoid special characters/spaces;
- use ISO-8601 date format: YYYY-MM-DD;
- append with v01, v02... for version control.

E.g.:  wind.dat
 tdf11-taux-july.csv

[File name example discussion: Samantha Joel/tweet](#)

[IS]

Similarly, your file and folder naming should be clear - you should never have to open a file or folder to know what's in it. If you came back to a file called "experiment" or "paper 1", would you remember what they actually are? We always end up with more files than expected so it's really important to name them clearly, often using a date or version number to avoid things like ".final" and ".finalfinal"! (If you use Twitter you might have seen the "by the age of 35" advice going round – this one from a university assistant professor caught our eye!)

Again, it's easy to be lazy and assume we can go back and tidy the files up later, but there are always better things to do with your time.

The example is from a video of a researcher lamenting his earlier poor naming practice, when he realised he could reuse data from an earlier project, but ended up spending a lot of time going through all the different files trying to figure out what they were. He had to rename them as he went, so they would be reusable in future. In the new name: 'tape data family 11' is the data source, 'tau-x' is the variable (wind stress), 'july' is the time period of the observation, csv is an open format. What is arguably still wrong with his renamed file? [The year is unclear; also, writing

the month as a word means he can't sort chronologically, which he could do if files were tdf11_taux_2010-01, tdf11_taux_2010-02, etc.]

- 
1. RDM definitions and importance.
 2. Data access and organisation.
 - 3. Data formats and backups.**
 4. Data documentation.
 5. Data sharing and security.

[GP]



Data formats and backups



[NYU Health Sciences Library \(2012\) Data Sharing, Part 2 of 3](#)

[GP]

While you watch it, think about the mistakes you can spot in the researcher's behaviour. We'll then go through the RDM best practice elements that would have prevented these problems.

...

We noted: "I am not able to read hexadecimal" ... "maybe you can buy the program on eBay" ... "that is my only copy".



Ensuring sufficient backups

Ideally, use CU network drives:

- Personal or group drive;
- Easily accessible on- and off-site;
- Backed up daily by IT.

If keeping a local copy:

- Ensure it is equally secure;
- Ensure you're working on the right copy/version!

How much data would you lose if:

- your laptop got stolen,
- your lab burnt down,
- you lost your USB stick,
- your portable hard drive got damaged,
- data from your Dropbox/Google Drive account disappeared?

14

[GP]

When thinking about where to store data *during* your project, we recommend the University filestore as it is fully secure and backed up. Follow the link for guidance from IT on remote access, recovering deleted files (up to 7 days after deletion on Windows), etc. You can request extra storage space via the IT Service Desk servicedesk@cranfield.ac.uk.

Also, remember to make full backups/take snapshots before any “transformations” such as data processing, cleaning, recoding, deriving.

Sadly we still hear of laptop thefts and USB losses here, from distraught students who have lost months of work due to only having copies on that one device – backups are essential!



Cloud storage – good or bad?



- Read the terms (you may be granting them permissions).
- Check where data is stored (European Economic Area required by the Data Protection Act).
- Remember they don't guarantee data restoration.



15

Public domain images from pixabay.com

[GP]

Terms: with Drive or Gmail, you retain IP over the content, but grant Google and those they work with the rights to use/modify/publish your content to develop and promote Google services.

Storage: University storage is UK-based. For surveys, Qualtrics is on secure European servers, we also have a UK-based option in Blackboard.

iCloud terms: Apple does not guarantee or warrant that any content you may store or access through the service will not be subject to inadvertent damage, corruption, loss, or removal... It is your responsibility to maintain appropriate alternate backup of your information and data.



Preserving/sharing your finalised dataset

1. **Funder repository:** no compliance worries.
2. **Subject repository ([re3data](#)):** best visibility of your data.
3. **Institutional repository ([CORD](#)):** DOI and preservation.



[GP]

Data shouldn't remain on a USB drive, or even on your network drive, as it may not be accessible.

If you're a student, talk to your supervisor. For anyone wanting to use CORD, there's a separate hands-on training session (register your interest via DATES). Data will probably go on CORD, but you or supervisor might upload it. Data may be in your thesis/article but it's submerged and uneditable, whereas in eg csv in a repository, visibility and reusability is there, and you'll get clear (alt)metrics.

Don't forget – share data when you're publishing, not before you've finished exploiting it. Funders expect an online record of your data at point of publication or within 12 months if you're not publishing.



File formats: choose open

- Textual data: rtf, txt, xml, pdf/a.
- Tabular data: csv, tab, por (SPSS).
- Databases: xml, csv.
- Geospatial: shp, shx, dbf, geotiff.
- CAD: dwg.
- Video: mp4, mj2.
- Audio: wav, flac.
- Images: tif, png, svg, jpg.



Image by N. Hussein on Flickr

[GP]

“I am not able to read hexadecimal” ... “maybe you can buy the program on eBay”

Data must be accessible long term, which means that files need to be opened in different software/versions. An open format is one where the spec is published, so new software can be designed to open it. Companies such as Microsoft use proprietary formats (eg docx, xlsx) and don't publish their spec, so other programs have difficulty opening them – so that people have to buy Microsoft products. Long-term these are not reliable, so you should avoid them as far as possible.

E.g. a Word file from pre Word 97 will not be openable in 2010.

- 
1. RDM definitions and importance.
 2. Data access and organisation.
 3. Data formats and backups.
 - 4. Data documentation.**
 5. Data sharing and security.

[IS]



Data documentation



[NYU Health Sciences Library \(2012\) Data Sharing, Part 3 of 3](#)

[IS]

While you watch it, think about the mistakes you can spot in the researcher's behaviour. We'll then go through the RDM best practice elements that would have prevented these problems.

...

We noted: "I do not understand the data" ... "my co-author knows what the field names mean" ... "he is in China, his name is Sam Lee"



Metadata (vs documentation)

- Title;
- Authors;
- Categories;
- File type;
- Description;
- References;
- Funding;
- Licence.

See also: [RDA Directory of Metadata Standards.](#)

The screenshot shows a metadata form with the following fields:

- Title:** A text input field with the placeholder text "Please make it more descriptive".
- Authors:** A text input field with the placeholder text "Add co-authors by name or full email".
- Categories:** A dropdown menu with the text "Select categories".
- File type:** A dropdown menu with the text "Dataset".
- Keyword(s):** A text input field with the placeholder text "Add keywords for easy discovery".
- Description:** A text input field with the placeholder text "Describe your data as well as you can".

20

[IS]

When your data is stored or made available, it usually needs some extra information to make it usable – ie understandable without needing to track down a person. This extra information is generally considered in two parts: metadata (structured fields for computer use) and documentation (free text for human use).

CORD metadata fields on screen – very simple, just filling in a form. Metadata is structured i.e. it uses a schema (a standardised set of choices) for integration and use in search/filter/browse. Certain domains have specific schema, eg requiring coordinates for geographic data, start/end dates for longitudinal studies, sampling methods or universe for social science studies, etc. Know what will be expected in your domain and be prepared to add it when sharing your data.

[Metadata standard example: the ONS (Office for National Statistics) has the Standard Occupational Classification 2010: 9 types of occupation that you would use as your categories (and a mega spreadsheet of job titles saying which category they fall into).]



Documentation (vs metadata)



- **Dataset information:** file names, acronyms, variables, units, codes, date/location of data collection, software needed.
- **Methodology information:** methodology, data processing, instruments used, precision, calibration, quality controls.

21

Public domain image from unsplash.com

[IS]

Documentation is often just one text file and provides any information needed to ensure that someone unfamiliar with your data could understand and use it without needing to contact you with questions. This will also help you if you come back to your data months or years later, when you may not remember these details as clearly. It can be useful to put this information into a file as you go, throughout the project, rather than spending a day at the end putting it together. You'll probably want to submit a readme file with your dataset at project end.

You might find that not all this information is needed if you look after your data well – eg if your spreadsheet has very clear column and row headings, many of these fields are redundant. Keep the data files tidy to make this easier (e.g. in spreadsheets, have one worksheet for the raw data, start a new worksheet for each function you work on).



Readme file example (see [readme.txt template](#)).

1. Directory/file naming conventions: YYYY-MM-DD-MATERIAL-TEST
YYYY-MM-DD is the date the experiment was initiated
TEST is a code where: 01=stress test, 02=hardness test, ...
MATERIAL is the aluminium alloy tested eg 7068=7068 aluminium alloy
2. Definitions of acronyms, abbreviations, or other project-specific terms
AGL = above ground level
AIRC = Aerospace Integration Research Centre, Cranfield University, UK
3. Variables: units of measurement, note any special formats used
Variable "st1" = shear stress, measured in pascals (Pa)
Variable "st2" = tensile stress, measured in pascals (Pa)
4. Variables: codes for missing data
X = participant declined to answer
Y = answer obtained but illegible/inaudible

22

[IS]

Example of the start of our readme template which you could use to create your documentation. It's particularly useful for experimental work.

Also, if you've used SPSS, you can export the data dictionary to get a file of variable labels, missing value codes, etc.

- 
1. RDM definitions and importance.
 2. Data access and organisation.
 3. Data formats and backups.
 4. Data documentation.
 - 5. Data sharing and security.**

[GP]



Barriers to data sharing: ethical, commercial, legal.

My data contains personal information. ✓?	My data is too complicated. ✗	People may misinterpret my data. ✗	My data isn't very interesting. ✗
My commercial partner won't want it shared. ✓?	We might want to use it in another paper. ✓?	People might contact me to ask about stuff. ✗	Data protection/national security. ✓
It's too big. ✗	People might spot a mistake or see that my data's not very good. ✗	I want to patent my work. ✓?	It's not a priority and I'm busy. ✗
I don't know how. ✗	I'm not sure I own the data. ✓?	Someone might steal or plagiarise it. ✗	My funder doesn't mandate it. ✗

[GP]

Wherever possible, you should be planning to make your data openly accessible, for all those reasons we looked at earlier (scientific integrity, enabling innovations, aiding your own reuse, improving your citation rate, compliance with requirements).

However, data sharing must be responsible – it's important to consider any ethical, commercial, or legal aspects. The bingo card on screen shows a variety of responses people have when asked to share their data. Which are valid reasons not to share, and which are excuses that would not be accepted by funders/publishers?

Tick – if a threat to national security or DPA says you legally can't share your data, you can't share.

Tick/orange Q – this is a valid reason to delay access to your data until you've finished exploiting it. You should use an embargo, rather than withholding your data entirely.

Tick/red Q – this might be a valid reason, but you might be able to overcome it. If you're collecting personal information, you should plan to anonymise it. If you have a commercial partner, you should check whether parts of the data could be shared. If you don't know if you own the data, can you find out?

Cross – these are excuses.



Passwords and encryption

Passwords (see [Network Password Policy \(pdf\)](#))

- Use a strong password (avoid dictionary words).
- Don't let others see you type it in.
- Don't enter it on untrusted computers/networks.



Encryption (see [Encryption Guidelines \(pdf\)](#))

- Institutional storage encrypts data by default.
- Also: MS Office: File > Protect > Encrypt with password.
- For stricter requirements, ask IT Service Desk about our approved encryption software that they can support/install.

25

[GP]

Your network password is only for University accounts, for security reasons. Most security requirements are met by using network storage (network drives and collaboration sites) which has encryption at-rest and in-transit. If you want additional encryption using simple password protection, MS Office has this built in – but don't forget the encryption password!

Remember to take the same precautions if needing to transfer data – secure files before sending, and send files separately from the password/s needed to access them (Cranfield's Dropoff service may be preferable to email <http://dropoff.cranfield.ac.uk/>).



Anonymisation and data destruction

Anonymisation (see more detail on the [anonymisation intranet pages](#)):

- Data Protection Act: as soon as they're no longer required, the identifiable portions of data must be removed.
- NB. Anonymised means people can't be identified from this data **or by combining it with any other available dataset**.

Destruction (see [ICO data deletion guidelines](#)):

- Deleted files can be retrieved with common tools.
- Contact the IT Service Desk for data deletion or device destruction.

26

[GP]

Anonymisation is all about managing risk – you can't make guarantees as do not know what datasets/computational techniques will be available in the future.

Anonymisation techniques: remove identifiers e.g. names, generalise a variable e.g. change a postcode to a town or county, remove outliers e.g. 99yr old will be easy to identify.

If not confident that your data is sufficiently anonymised, make it restricted access.



How will I remember all this?!

Write a data management plan (DMP)! They:

- walk you through all the elements we've discussed;
- help save you time throughout your project;
- are mandatory for doctoral students and often required when applying for funding.



Sign up for the “RDM 2: Writing a DMP” session via the [DATES system](#).

[GP]

The next dates are...



Further help and information

RDM intranet site: <http://bit.ly/RDM-home>
(Research, Learning & Teaching > Research Data Management)

Personal support: researchdata@cranfield.ac.uk
(Georgina Parsons, 01234 754548 (x4548), g.l.parsons@cranfield.ac.uk)

Cranfield training:

- Workshops/webinars: <https://webapps3.cranfield.ac.uk/DATES/Application/>
- RDM module on VLE: <https://moodle.cranfield.ac.uk/RDM>