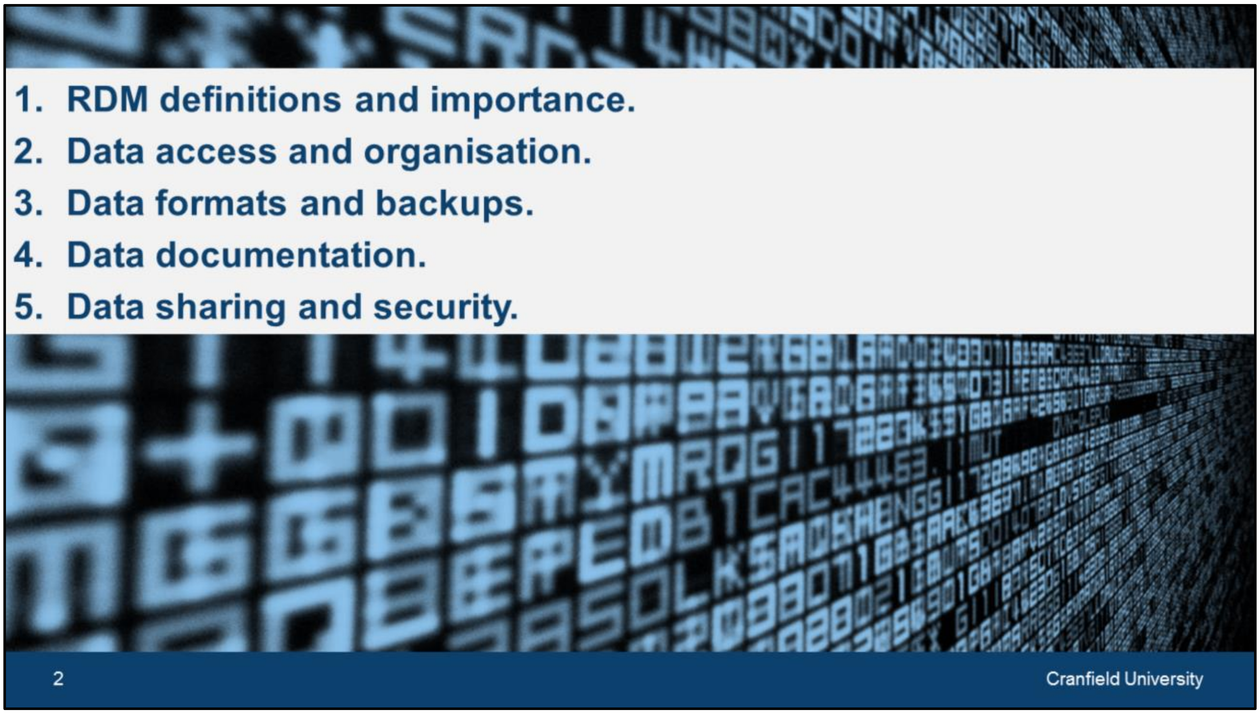
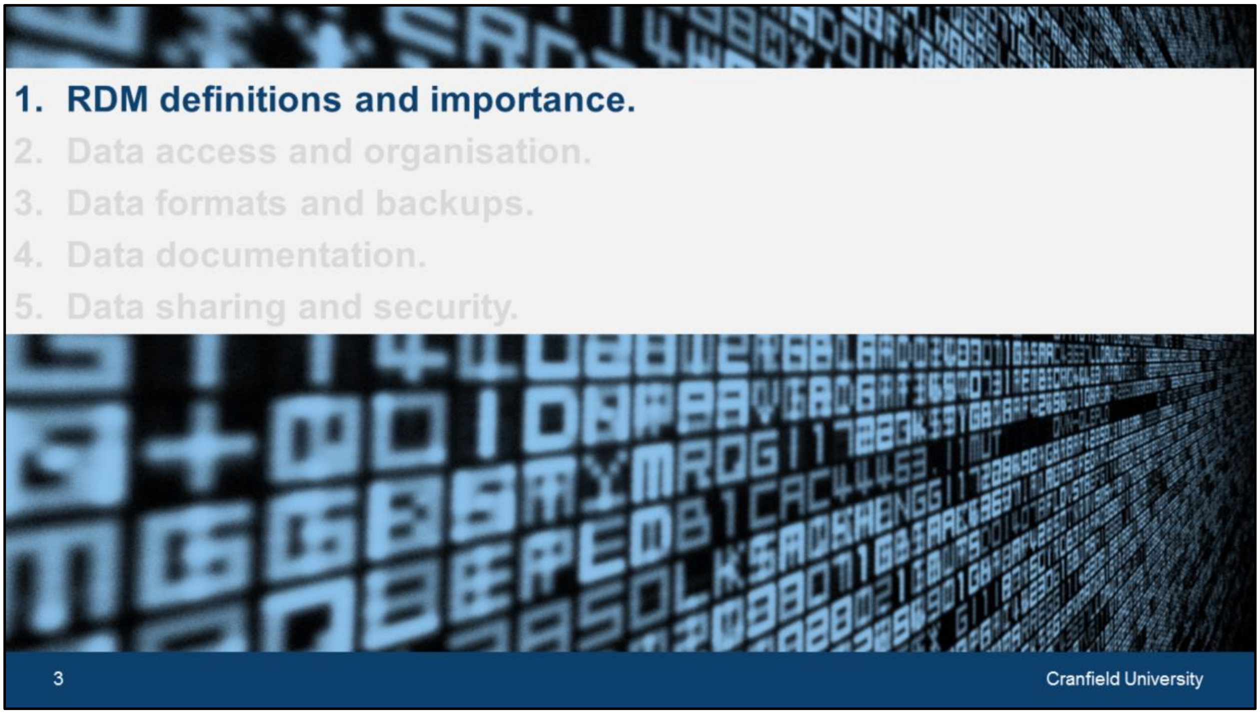
Research Data Management 1: Introduction

Georgina Parsons,
Research Data Manager, 2019.



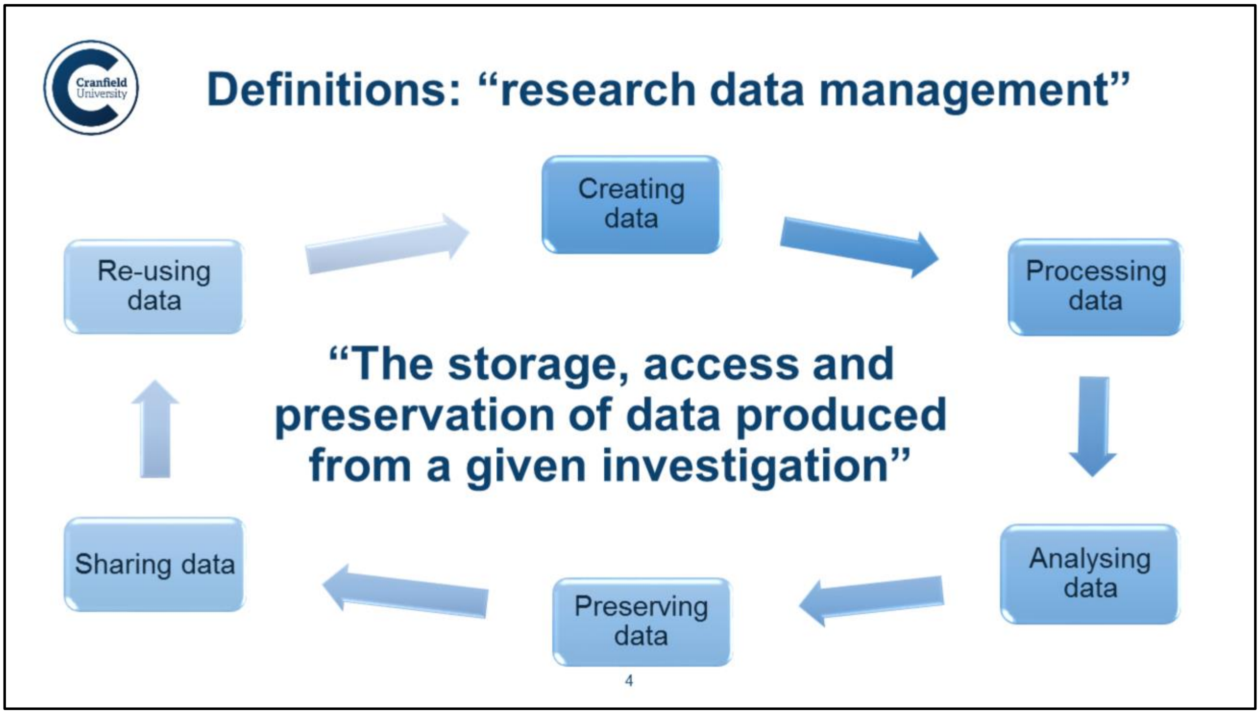
1. RDM definitions and importance.
2. Data access and organisation.
3. Data formats and backups.
4. Data documentation.
5. Data sharing and security.

[GP]
The first section will cover the theory: what exactly is RDM and why is it important. The next four sections cover a variety of practical aspects of RDM to help you through your research.



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”

Creating data

Processing data

Analysing data

Preserving data

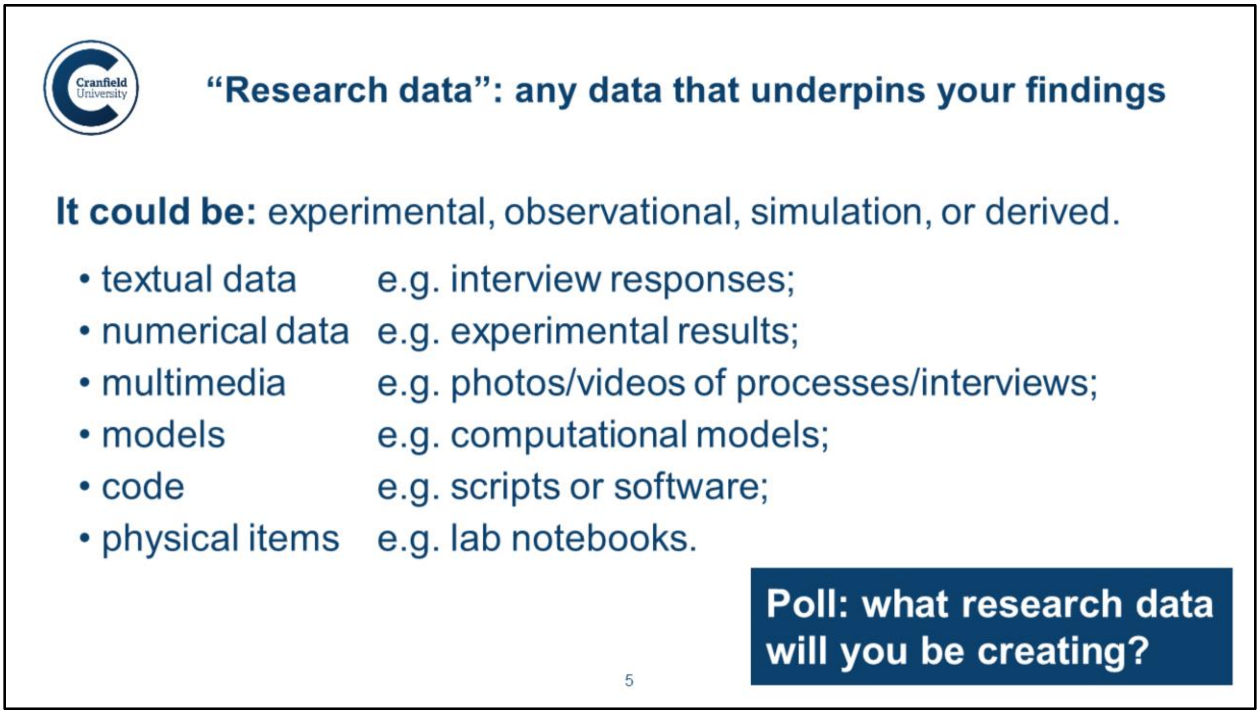
Sharing data

Re-using data

“The storage, access and preservation of data produced from a given investigation”

[IS]
So what is RDM? One possible definition is on screen [read].
It’s important also to note that data management practices cover the entire lifecycle of the data, from planning the research through to the long-term preservation of data after the project ends.

NB background ref:
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 = in follow-up research, teaching and learning, industry, etc.]



“Research data”: any data that underpins your findings

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

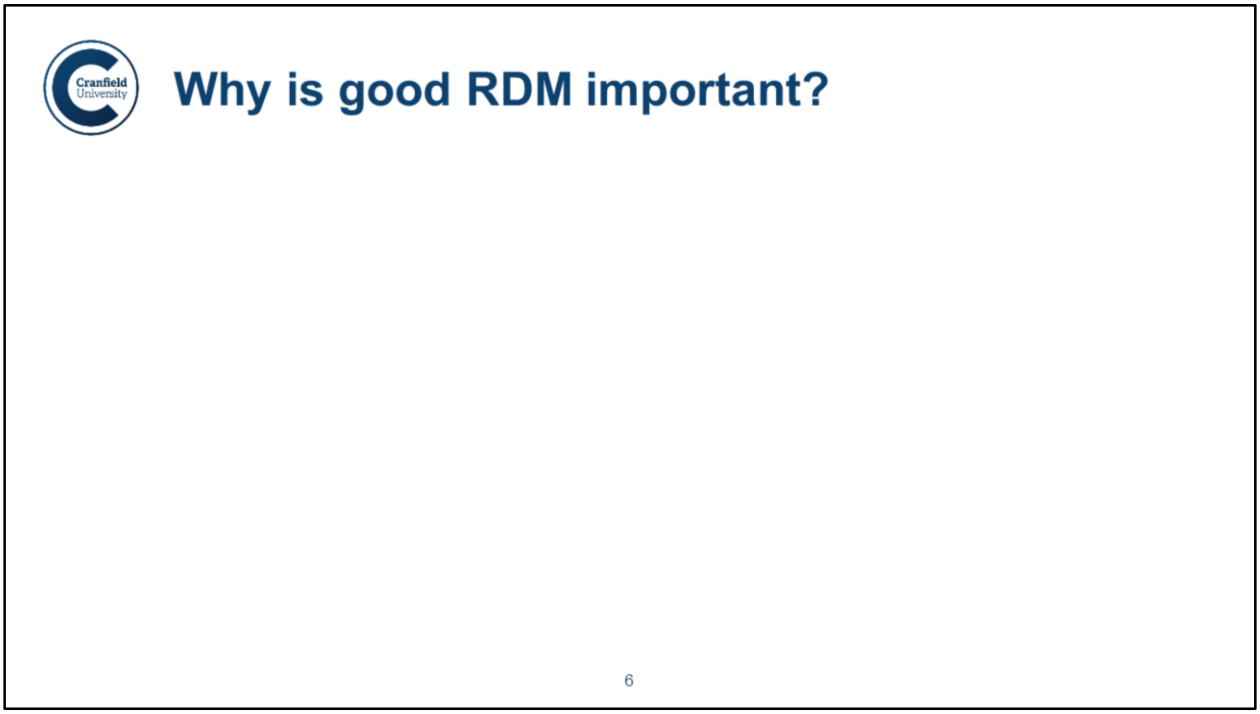
• textual data	e.g. interview responses;
• numerical data	e.g. experimental results;
• multimedia	e.g. photos/videos of processes/interviews;
• models	e.g. computational models;
• code	e.g. scripts or software;
• physical items	e.g. lab notebooks.

Poll: what research data will you be creating?

[IS]
And what is “research data”? We define it as the data that underpins your findings, i.e. the evidence to your results.
It can be a variety of formats and could include survey responses, interview transcripts, financial data, experimental results, images, interview videos, CAD files, statistical models, 3D models, software written during the project, and more.
When we say “physical items” we really mean laboratory notebooks. If you’re experimenting on samples, your research data is the experimental results, not the samples.

[NB background ref: Research Information Network classification is:
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?

[GP]

[NB animations don’t work in WebEx so need six separate slides.]

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

Why is good RDM important?

It keeps your research safe and secure.

[GP]
1. A large element of RDM is keeping data secure and backed up, minimising the risk of data loss, whether that’s through an accident, natural or manmade disasters such as fire or flooding, equipment theft, or technical failures.

Why is good RDM important?

It keeps your research safe and secure.

It aids your own data reuse in the future.

[GP]
2. Much like good referencing, being organised and following RDM best practice principles from the outset is well worth it as you will save yourself time in the future. ‘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].

Why is good RDM important?

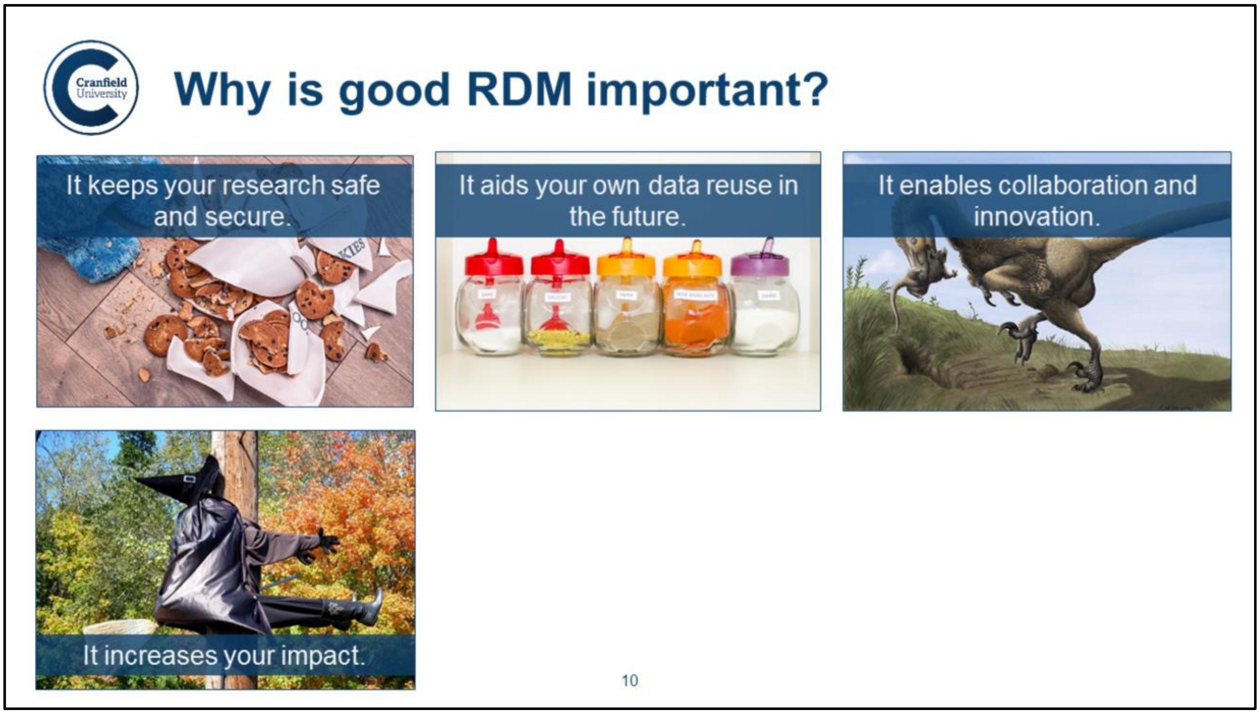
It keeps your research safe and secure.

It aids your own data reuse in the future.

It enables collaboration and innovation.

[GP]
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).



Why is good RDM important?

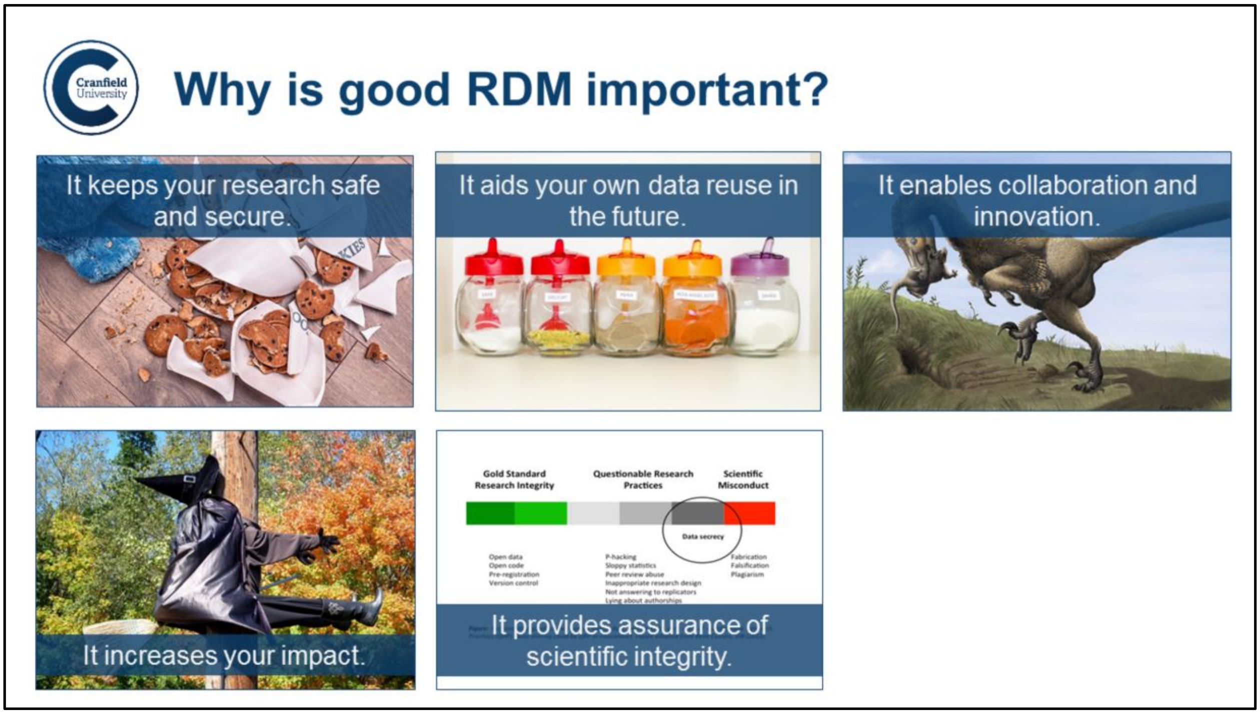
It keeps your research safe and secure.

It aids your own data reuse in the future.

It enables collaboration and innovation.

It increases your impact.

[GP]
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 !



Why is good RDM important?

It keeps your research safe and secure.

It aids your own data reuse in the future.

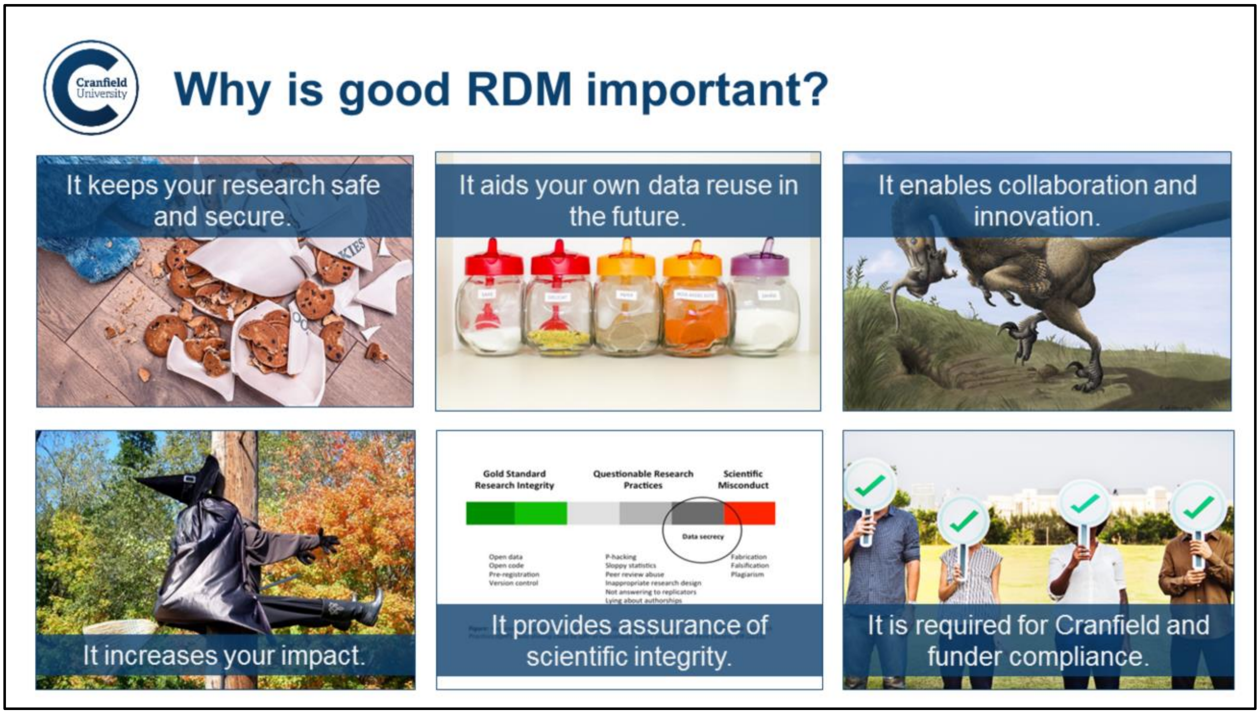
It enables collaboration and innovation.

It increases your impact.

It provides assurance of scientific integrity.

[GP]
5. RDM has also become a priority in universities recently due to a "reproducibility crisis" in research, where it was found that only a third of papers studied were reproducible, 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' to demonstrate robust research practices.

Attitudes have also shifted due to some high-profile scandals, such as Dr Stapel in the Netherlands who falsified data for over a decade. People asked "What type of academic culture allowed Stapel to continue his misconduct for so long?" and the Dutch funded 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.



Why is good RDM important?

It keeps your research safe and secure.

It aids your own data reuse in the future.

It enables collaboration and innovation.

It increases your impact.

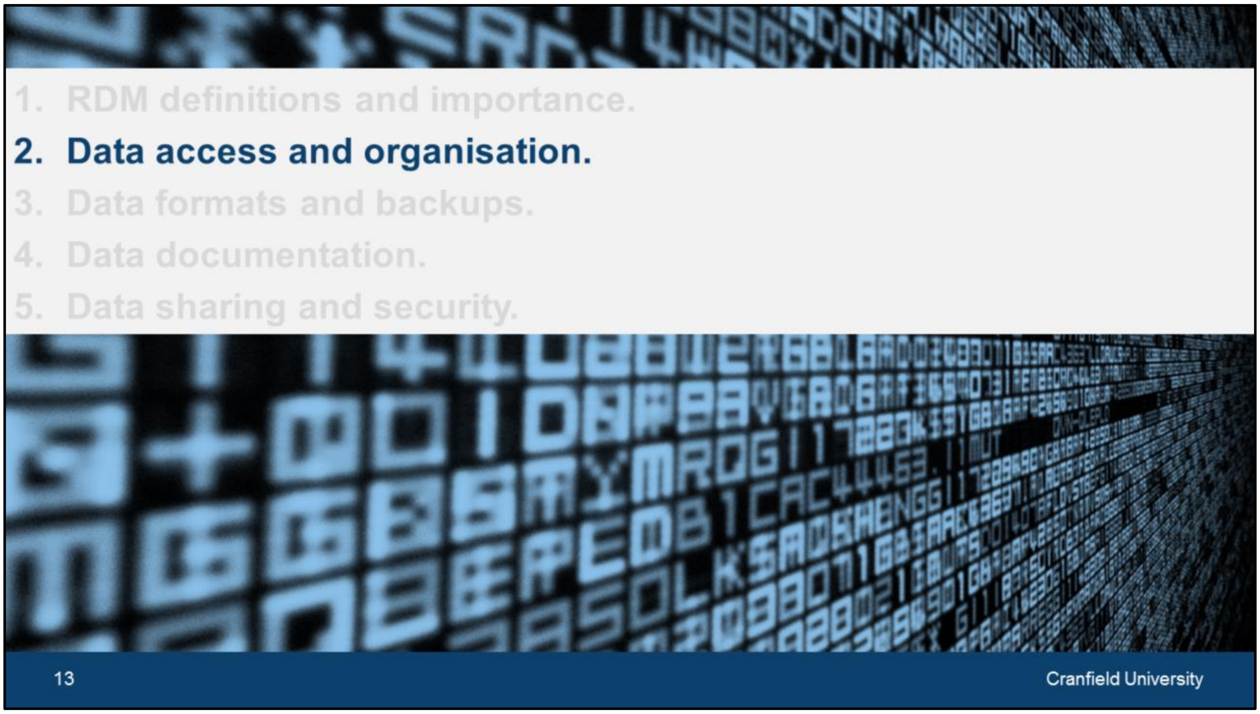
It provides assurance of scientific integrity.

It is required for Cranfield and funder compliance.

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.

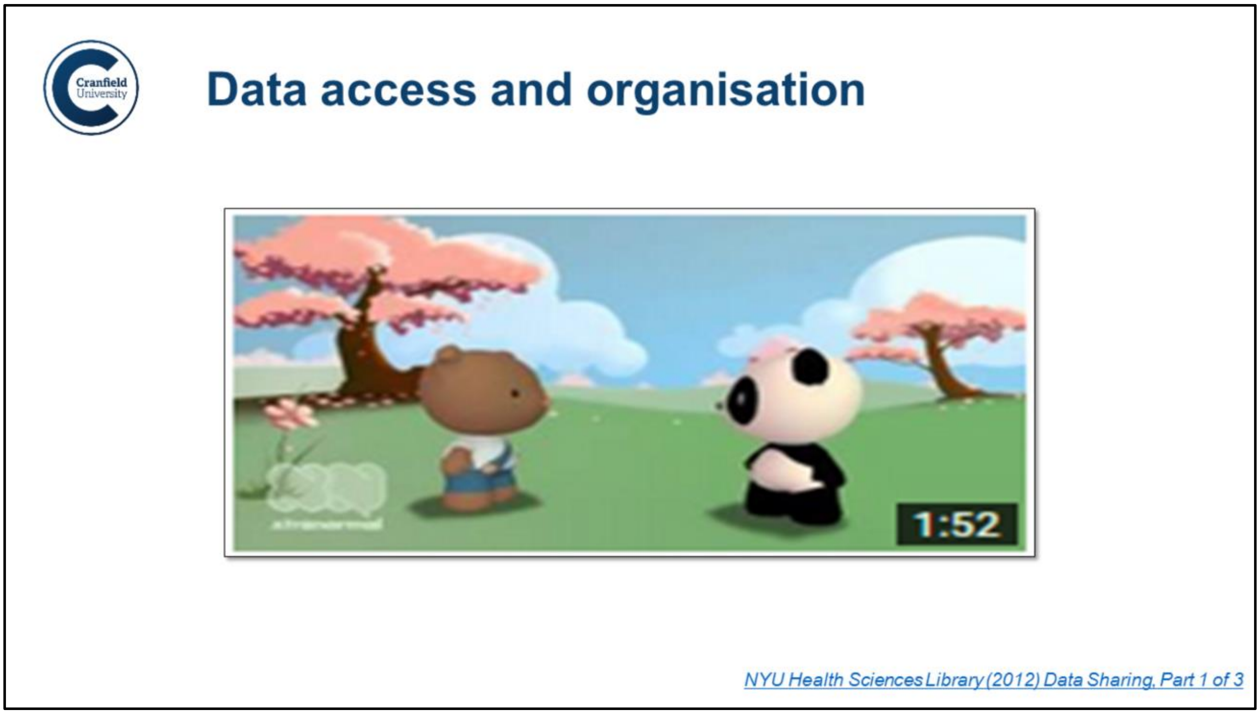
Credits:
**Sharing Detailed Research Data Is Associated with Increased Citation Rate (DOI: 10.1371\journal.pone.0000308). See also: The enduring value of social science research (http://hdl.handle.net/2027.42/78307); On the citation advantage of linking to data (https://hal-hprints.archives-ouvertes.fr/hprints-00714715v2).*
Image: Labeling, CC-BY-NC by Paul Istoan
https://www.flickr.com/photos/vamapaull/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.

[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.

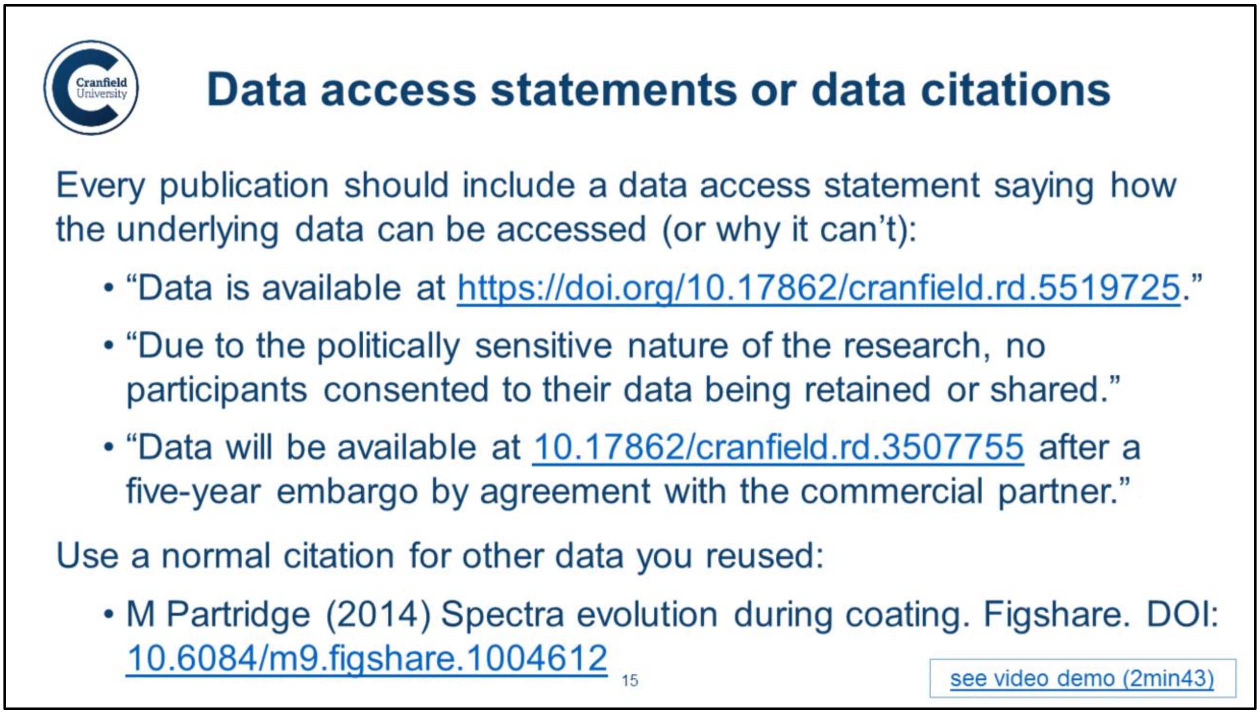


Data access and organisation

NYU Health Sciences Library (2012) Data Sharing, Part 1 of 3

[IS]
We’ll start the first three sections with a series of videos from New York University’s library on how *not* to do RDM.

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 after a five-year embargo by agreement with the commercial partner."

Use a normal citation for other data you reused:

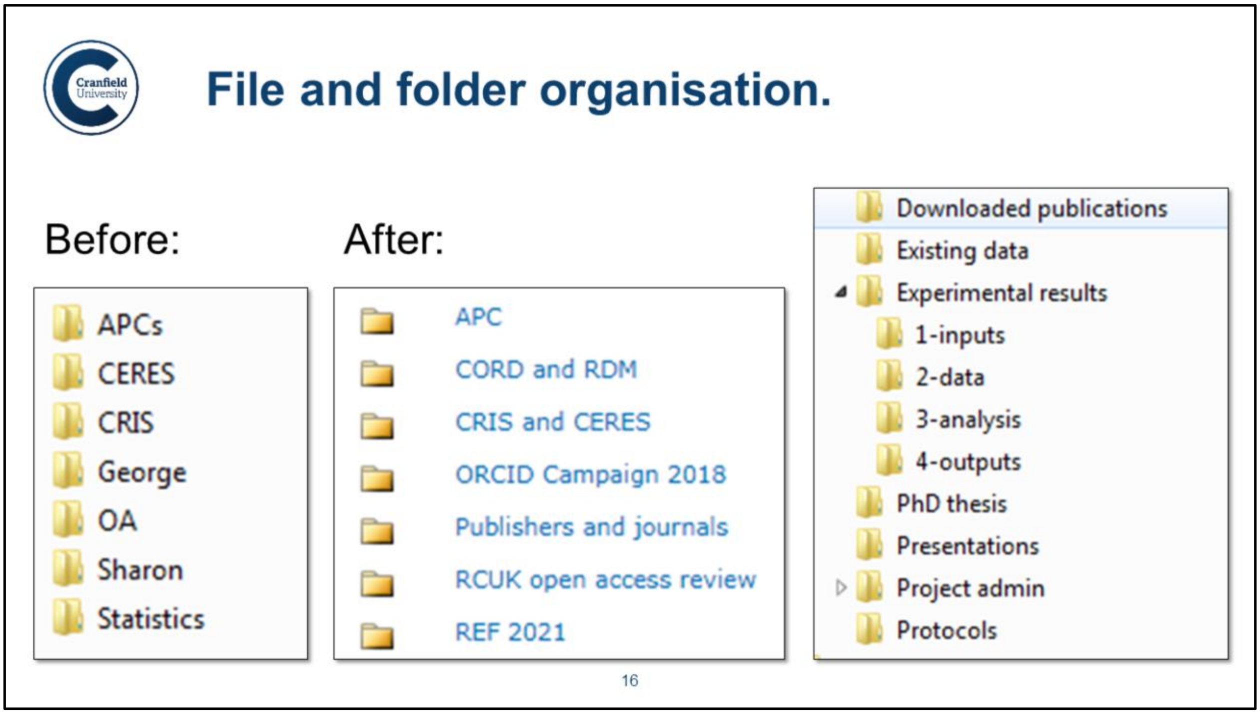
- M Partridge (2014) Spectra evolution during coating. Figshare. DOI: 10.6084/m9.figshare.1004612

see video demo (2min43)

[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.

Every publication needs to include a data access statement in the Acknowledgments section, either linking to the data, or explaining why you can't link to it.

Cranfield's RDM policy requests a data statement, and funders such as EPSRC are now compliance checking to ensure that one is present.



File and folder organisation.

Before:

- APCs
- CERES
- CRIS
- George
- OA
- Sharon
- Statistics

After:

- APC
- CORD and RDM
- CRIS and CERES
- ORCID Campaign 2018
- Publishers and journals
- RCUK open access review
- REF 2021

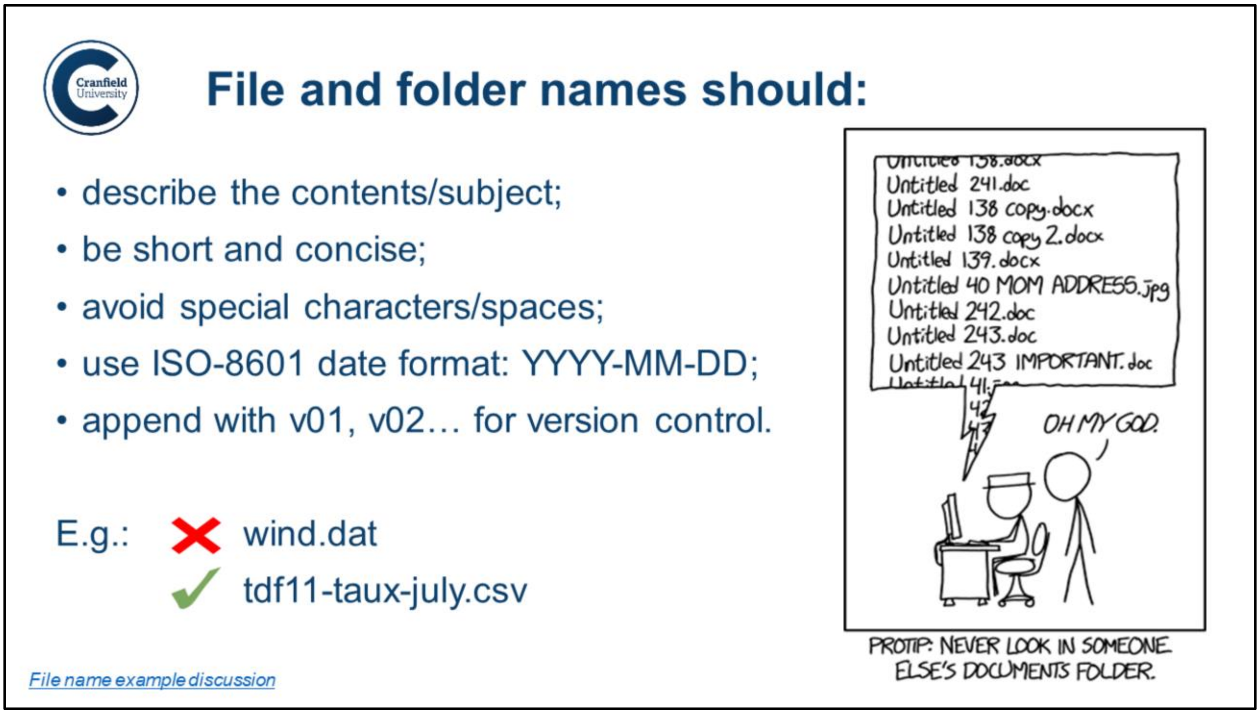
- Downloaded publications
- Existing data
- Experimental results
  - 1-inputs
  - 2-data
  - 3-analysis
  - 4-outputs
- PhD thesis
- Presentations
- Project admin
- Protocols

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

Organising your files logically sounds like basic common sense, but it’s important to stress because it’s so easy to overlook. This is the part that’s like referencing – it’s very easy just to dump all your files in one place because during your project you just know where things are, but when you come back to the work later, or someone else does, it is a huge waste of time if you can’t quickly locate everything you need. A little bit of discipline and planning a clear folder structure will make your life so much easier down the line.

On the slide, the two images on the left are a real-life example from my work when we had to migrate library content – it shows that even librarians can get in a mess! In the ‘before’ folder, where would I look for OA Statistics – in the OA folder or the Statistics folder? And when I wanted to look for project work, I had to look in both ‘George’ and ‘Sharon’ because I didn’t know which staff member had done the work. In the ‘after’ image, it’s really clear what is in each folder with no overlap, and anyone could find information really easily.

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.

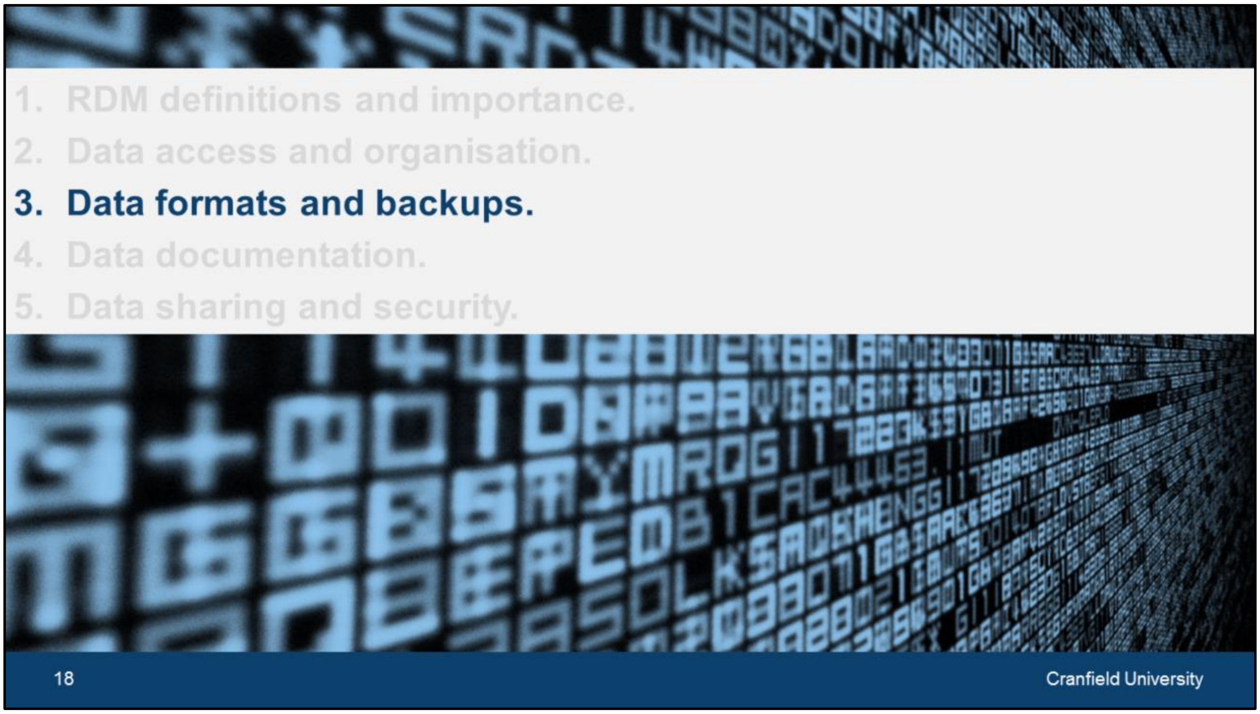
E.g.: ✗ wind.dat
✓ tdf11-taux-july.csv

File name example discussion

PROTIP: NEVER LOOK IN SOMEONE ELSE'S DOCUMENTS FOLDER.

[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 “experiment FINAL FINAL”, would you remember what it really was? You’ll inevitably accumulate more files than you think, but never give in to the lure of leaving files untitled or labelled ”new” etc – always use a date or version number if you’ve got multiple versions of the same base data. If you’re creating a lot of files, consider setting a naming convention, such as date_experiment_sample_variable, or participantnumber_type where ‘type’ is audio, transcripts, or anonymisedtranscript. Also think about how you’re likely to want to sort – for me it’s usually chronologically so I put my dates at the start of the file name, but in my ‘training’ folder, I put the session title at the start and the date at the end, so I can sort or group by session.

The example on the slide – wind.dat – is a real-life example from a video of a researcher lamenting his earlier poor naming practice. He realised that he could reuse data from an earlier project, but ended up spending a lot of time going through all his old files, opening them to figure out what they were then renaming them so they would be reusable in future. You will always have better things to do with your time than this, so learn from his mistakes! In the new filename: tdf11 is his data source (tape data family 11), tau-x is the variable (wind stress), and july is the time period of the observation. I would still say this he should have used ISO-8601, because when this file is taken on its own, you don’t know which year the observation was from, and he can’t sort his files chronologically because he used ‘july’ instead of a numerical date.



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

[GP]

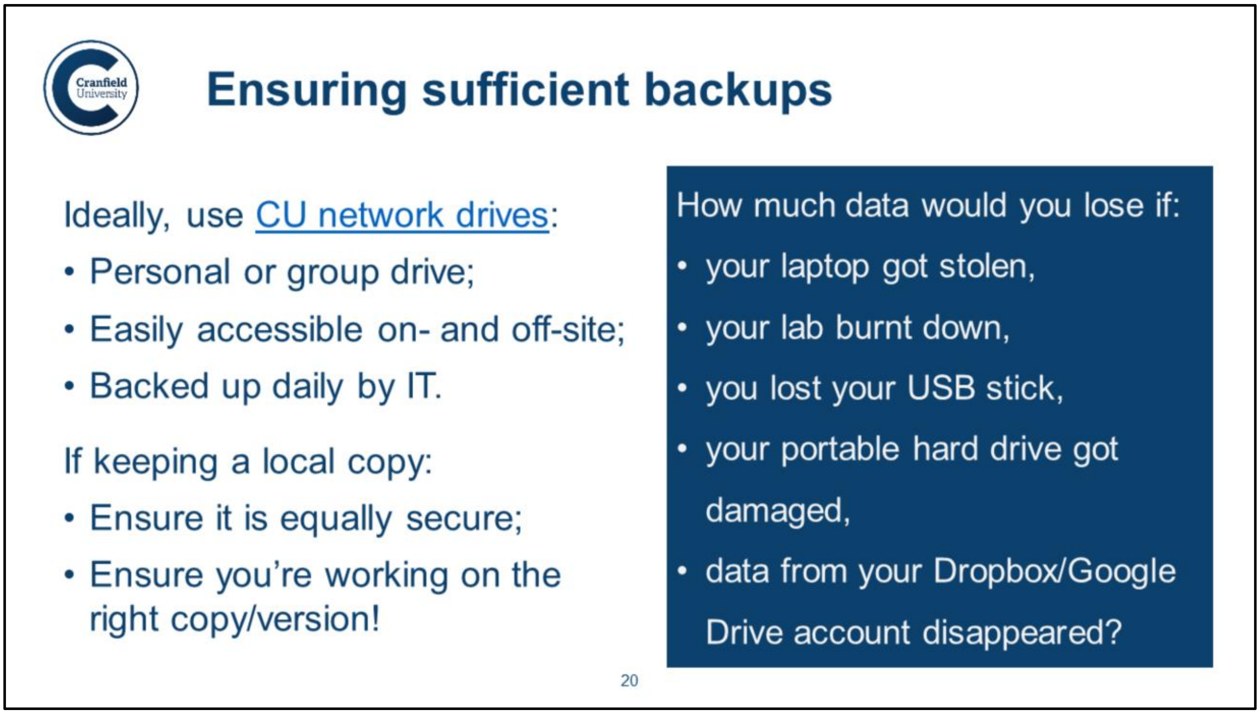
Data formats and backups

NYU Health Sciences Library (2012) Data Sharing, Part 2 of 3

[GP]
Here's the second in the trio of bad RDM videos. Again, 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:

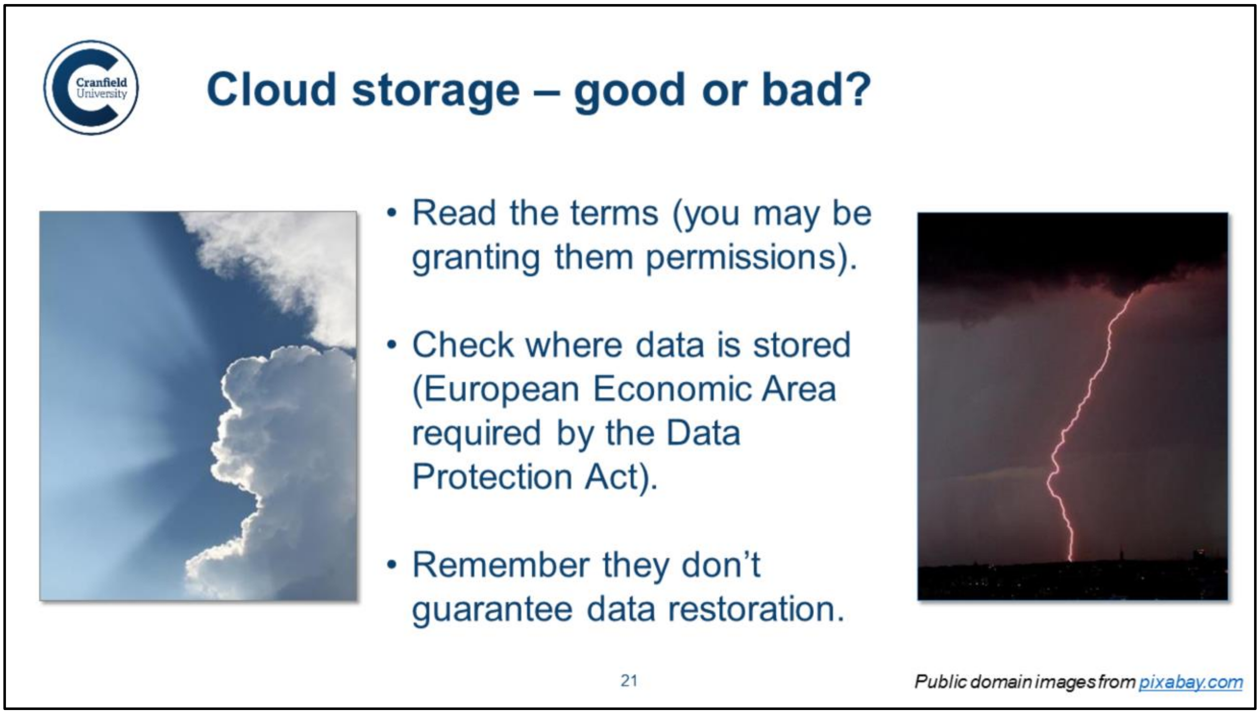
- your laptop got stolen,
- your lab burnt down,
- you lost your USB stick,
- your portable hard drive got damaged,
- data from your Dropbox/Google Drive account disappeared?

[GP]
“that is my only copy” – hopefully it’s obvious that you need to make sure your work is backed up, but it’s upsetting how many times someone has run into the library asking if anyone’s handed in a USB drive because it contains the later version of all their work.

During your project, you need to be using the Cranfield network drives. These are fully secure and IT back up content daily to two separate data centres for extra resilience. The link goes to the IT intranet page with guidance on remote access (you can map your drive so it’s constantly accessible from your laptop or a personal computer), and also on recovering deleted files (up to 7 days after deletion on Windows).

You’ll get a personal drive, accessible only to you, and you can request extra storage space via the IT Service Desk servicedesk@cranfield.ac.uk if you will be creating lots of data. You’ll also have access to a group drive, and you can request a new area accessible to a specific group of people, if that’s the most appropriate for your project.

Do also take a full copy of all your work before you do any major “transformations” such as data processing, cleaning, recoding, deriving.



Cloud storage – good or bad?

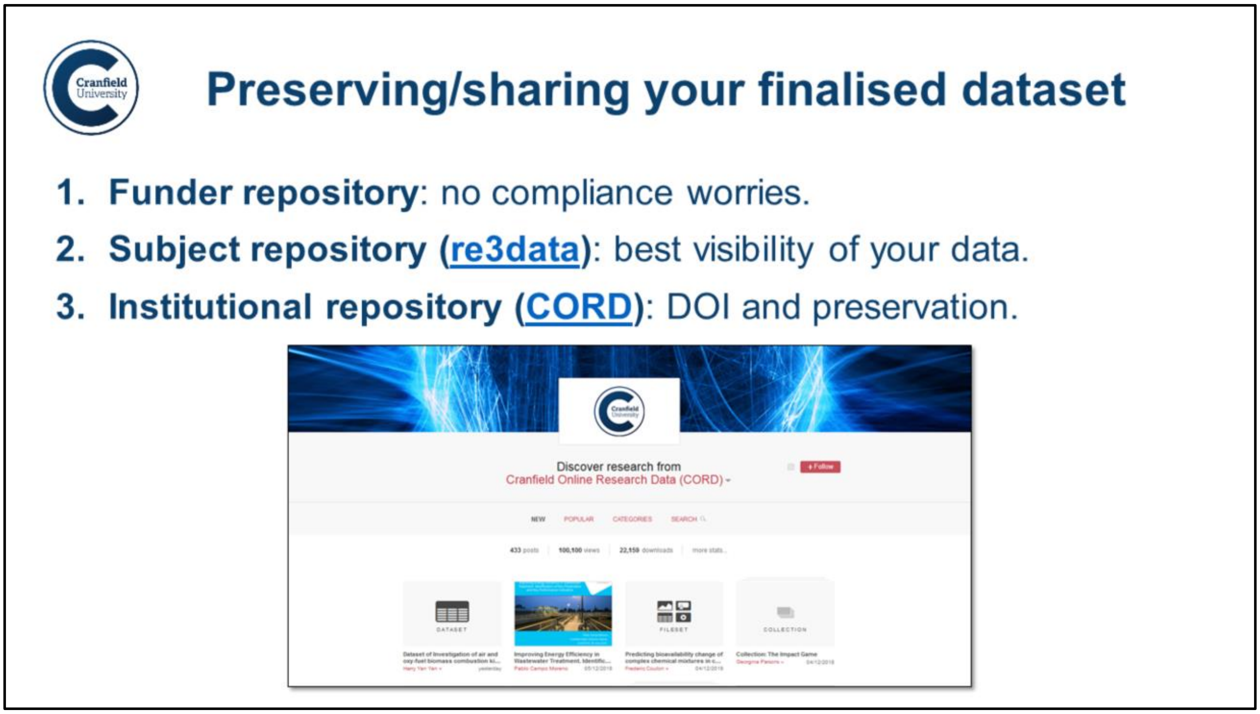
- Read the terms (you may be granting them permissions).
- Check where data is stored (European Economic Area required by the Data Protection Act).
- Remember they don't guarantee data restoration.

Public domain images from pixabay.com

[GP]
People are often tempted to use third party cloud storage such as Google Drive, iCloud, or Dropbox, because it's easier off-site or they're just more familiar with it than with the network drives. Cranfield's policy says not to use these solutions, and I would just stress three different aspects that you need to think about if you're tempted to use these:

1. The terms, because with Drive or Gmail, you are granting Google and partners the rights to use/modify/publish your content to develop and promote their services.
2. The storage location: personal data must legally be in the EEA or under equivalent security and commercial partners may also have requirements. University storage is UK-based but cloud services often aren't. As an aside, if you're doing surveys, Qualtrics is on secure European servers and we also have a UK-based option in Blackboard.
3. The lack of guarantee: for example, the iCloud terms specifically state that 'Apple does not guarantee that any content you may store will not be subject to inadvertent loss… It is your responsibility to maintain appropriate alternate backup of your data.'

Fundamentally, you're safe using Cranfield IT and if anything does go wrong, it won't be your fault and I'm sure you'll be given an extension or whatever necessary, but if you lost data because you used an unapproved third party solution…!



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.

Discover research from
Cranfield Online Research Data (CORD)

433 posts | 100,100 views | 22,159 downloads | more stats...

Dataset of Investigation of air and oxy-fuel biomass combustion ki... Harry Yan Yan yesterday

Improving Energy Efficiency in Wastewater Treatment. Identific... Pablo Campo Moreno 05/12/2018

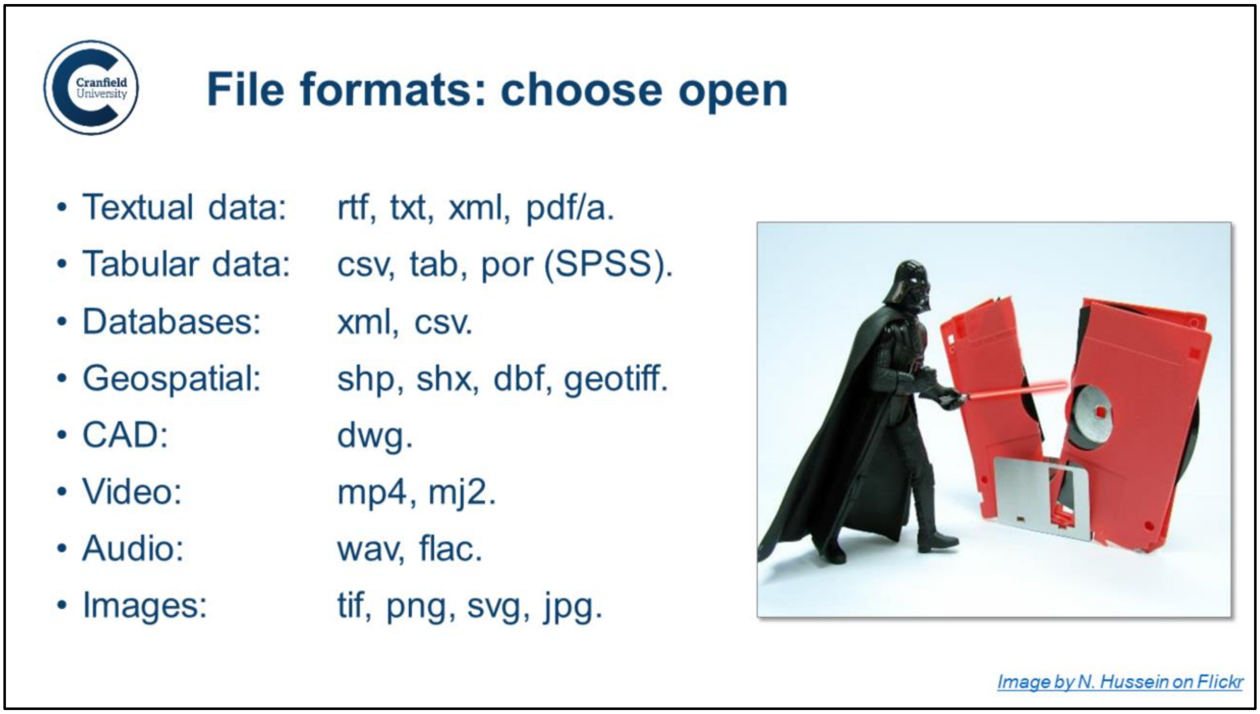
Predicting bioavailability change of complex chemical mixtures in c... Frederic Coulon 04/12/2018

Collection: The Impact Game Georgina Parsons 04/12/2018

[GP]
After your project, though, data shouldn't remain on a network drive, because it isn't easily findable or accessible. Data should always be preserved on a data repository. There are three main types:

1. If your project is funded by a body that has its own repository, you should use that.
2. Otherwise if there's a subject repository in your field, use that because other researchers in this area will know about it. The main one we recommend is the UK Data Archive for social, political, and economic data.
3. Otherwise you can use CORD, our own data repository. It's a great system, easy to use, and any data uploaded will get a DOI so you can link to it permanently, and we will ensure it's maintained in line with funder requirements (for at least 10 years).

Although your data might also appear in your thesis or papers, these days it does need to be published in its own right. It can then be cited independently, and will be preserved in its original editable format, whether that's a spreadsheet or the original high-resolution images or video, etc. You'll also get metrics on its usage which is really nice.



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.

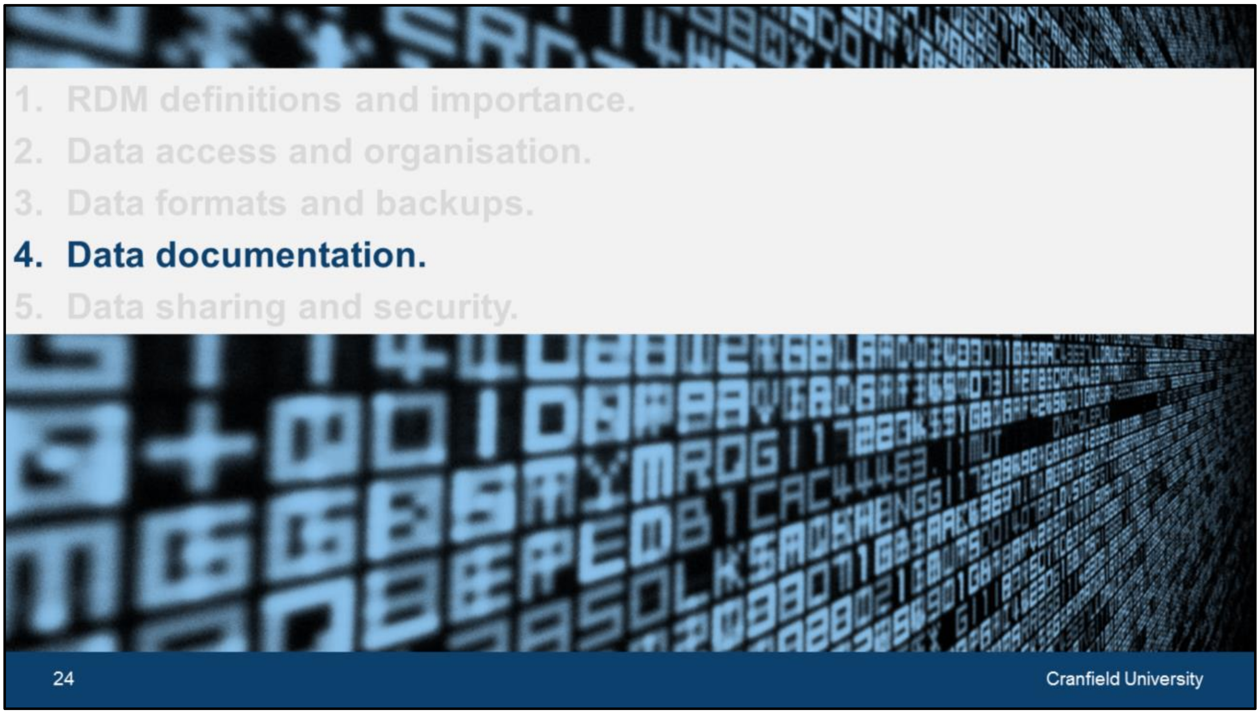
Image by N. Hussein on Flickr

[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 openable in different software/versions. We always ask that files are kept in open formats. An ‘open format’ is one where the specification is published, so new software can be designed to open it. For example, I could write a new program tomorrow that could perfectly reliably open a pdf or a csv file.

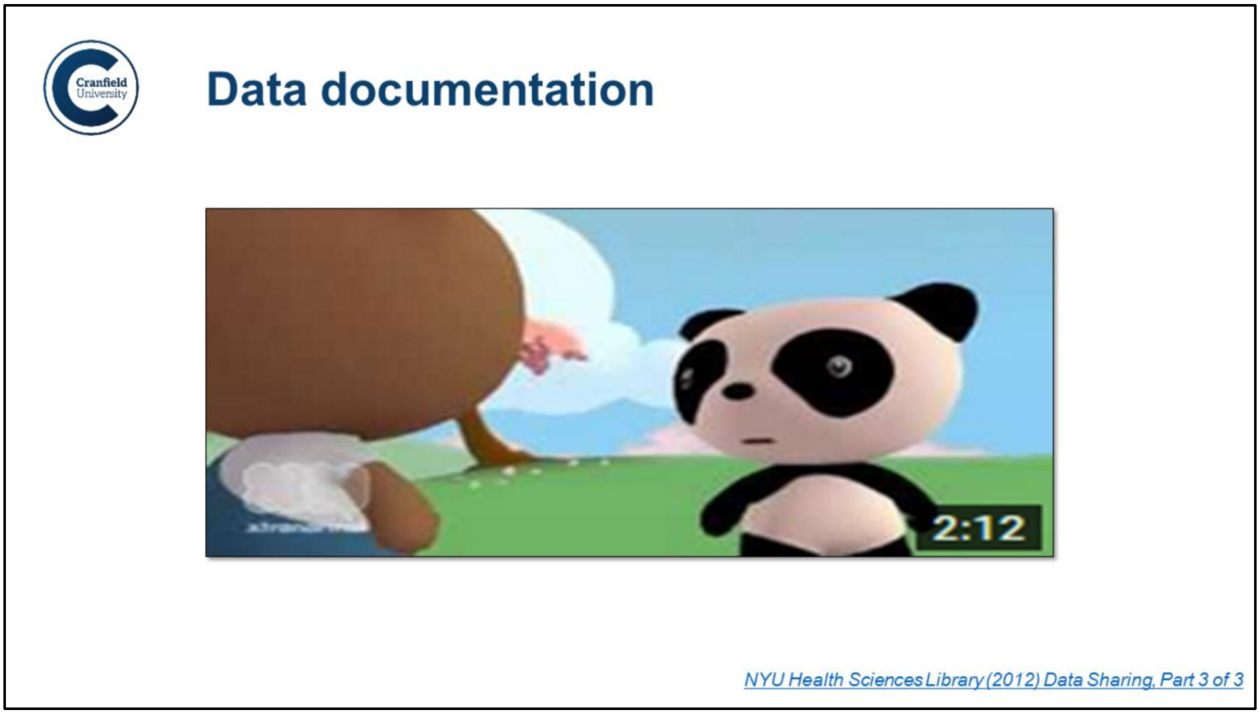
A closed (proprietary) format is one like ‘xls’ or ‘doc’ which relies on particular software to open it - companies such as Microsoft use proprietary formats so that people will have to buy their software. Although you often can open them in other software, it’s never reliable and there sometimes problems between versions of the software. I don’t use Windows at home, I can’t use Microsoft Office, and it’s a real frustration when people share files in formats that make it difficult to reuse.

We need to remove barriers to accessing data, which means using open formats wherever possible. It’s good to be aware of this now so you don’t have to convert files later on.



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

[IS]



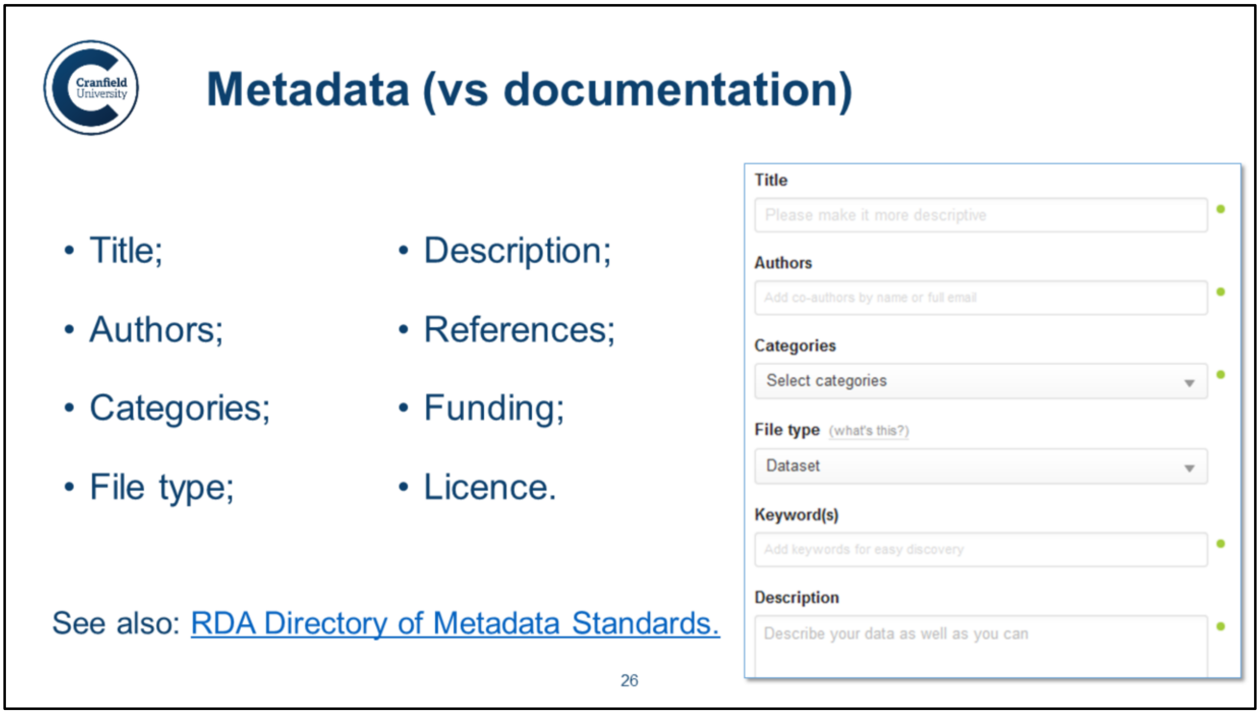
Data documentation

NYU Health Sciences Library (2012) Data Sharing, Part 3 of 3

[IS]
Here’s the final instalment in the ‘how not to do RDM’ videos. Again, 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.

Title
Please make it more descriptive

Authors
Add co-authors by name or full email

Categories
Select categories

File type (what's this?)
Dataset

Keyword(s)
Add keywords for easy discovery

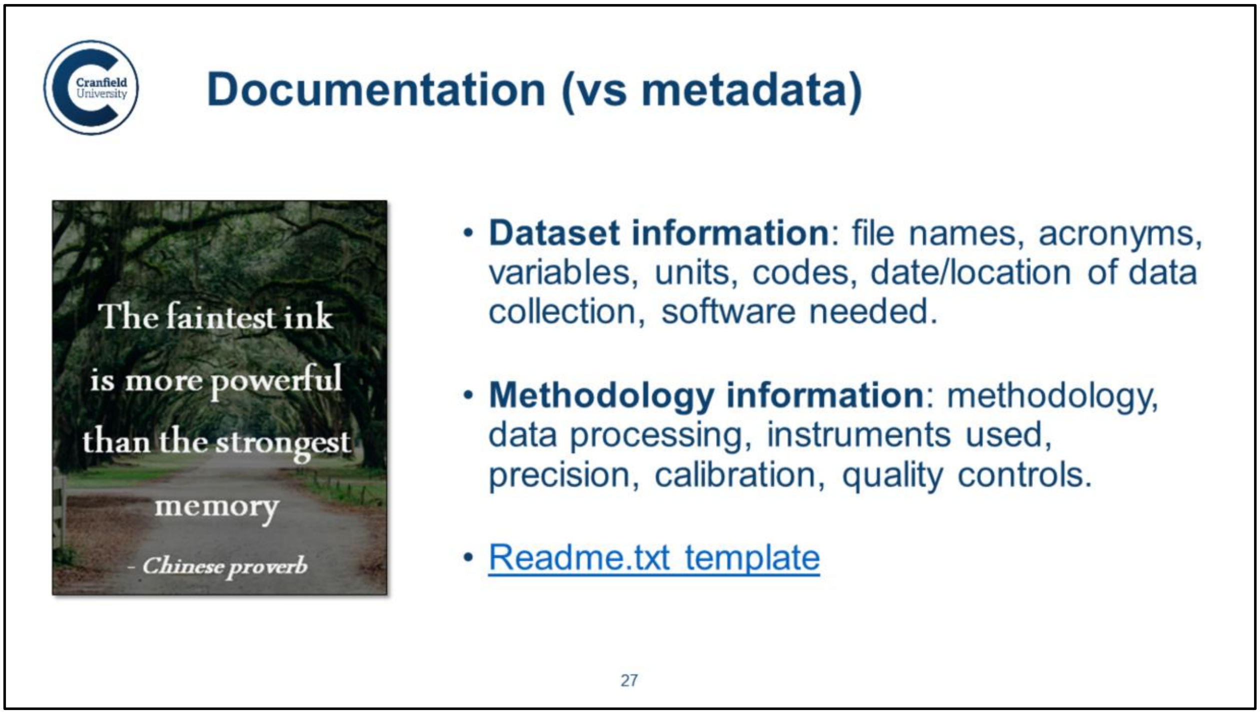
Description
Describe your data as well as you can

[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, especially where that person is “in China” with no forwarding contact details. This extra information is generally considered in two parts: metadata (structured fields for computer use) and documentation (free text for human use).

[GP]
Metadata is simpler than it sounds! If you use CORD, the metadata you have to enter is shown on screen: it’s a really simple form and uses a “schema” (standardised set of choices) for integration and use in search/filter/recommendations/etc. Different repositories have different requirements, relating to the domain, e.g. requiring coordinates for geographic data, start/end dates for longitudinal studies, sampling methods or universe for social science studies, etc. Just know what will be expected in your domain and be prepared to add it when sharing your data.

The other aspect to metadata is using a metadata schema for data collection. For example, you’re doing a survey and want to record someone’s occupation. If you get free text answers, that will be time-consuming and maybe difficult to analyse, but if you want to give them a list of choices, that’s going to be time-consuming and difficult to come up with. But don’t worry – there’s a metadata standard for that! 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 spreadsheet of job titles mapping into these categories). By using this existing standard, you’ll save time in your data collection, and also make your data more valuable because it will be interoperable with other studies that should also be using this standard.

[Ref: another example: if researching species, you would use the taxon ID to name the species, because this is the domain-specific standard (https://www.uniprot.org/taxonomy/9606).]



Documentation (vs metadata)

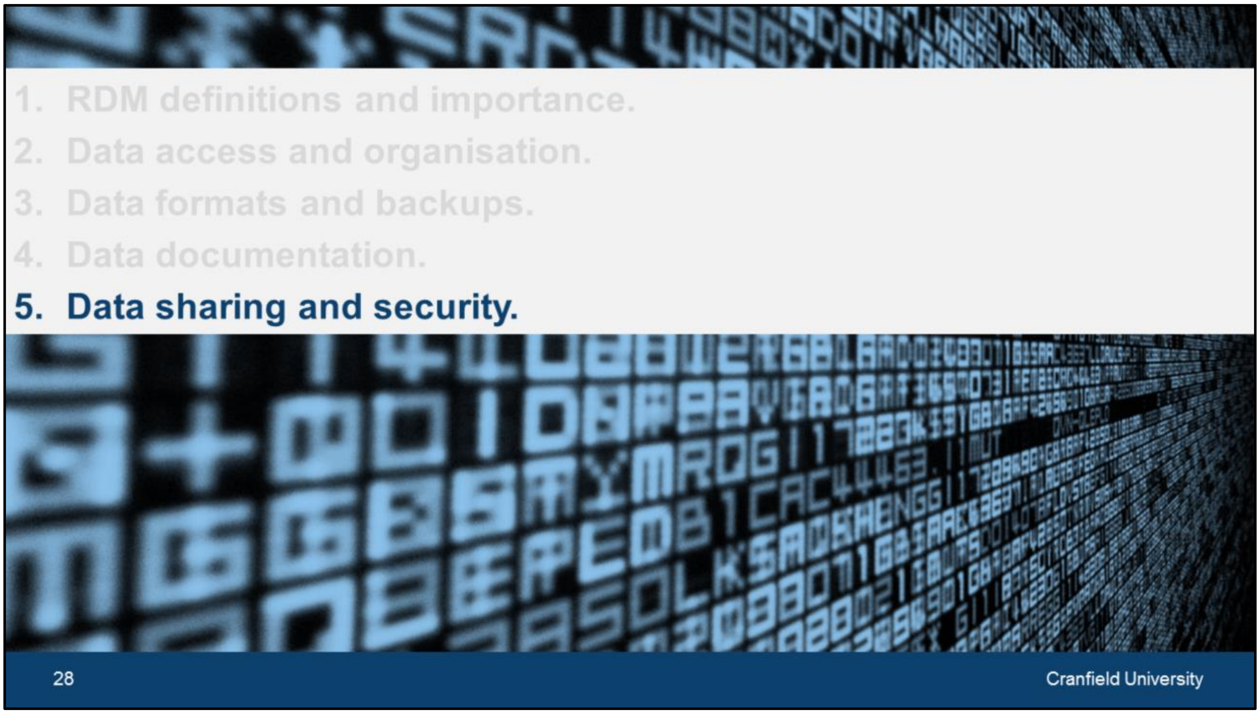
The faintest ink
is more powerful
than the strongest
memory
- Chinese proverb

- **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.
- Readme.txt template

[IS]
The other part of providing contextual information for your data is documentation. Usually, this just means providing one text file (traditionally the file is named "readme") alongside your data on a repository. It might contain any of the information shown on the slide (you can start from our readme template that we link to). You might want to note all this key information as you go. You can also link out from repositories, so if your methodology is clearly explained in your thesis, and your thesis is securely preserved long-term, we wouldn't ask you to redocument this, but your data can link to your thesis.

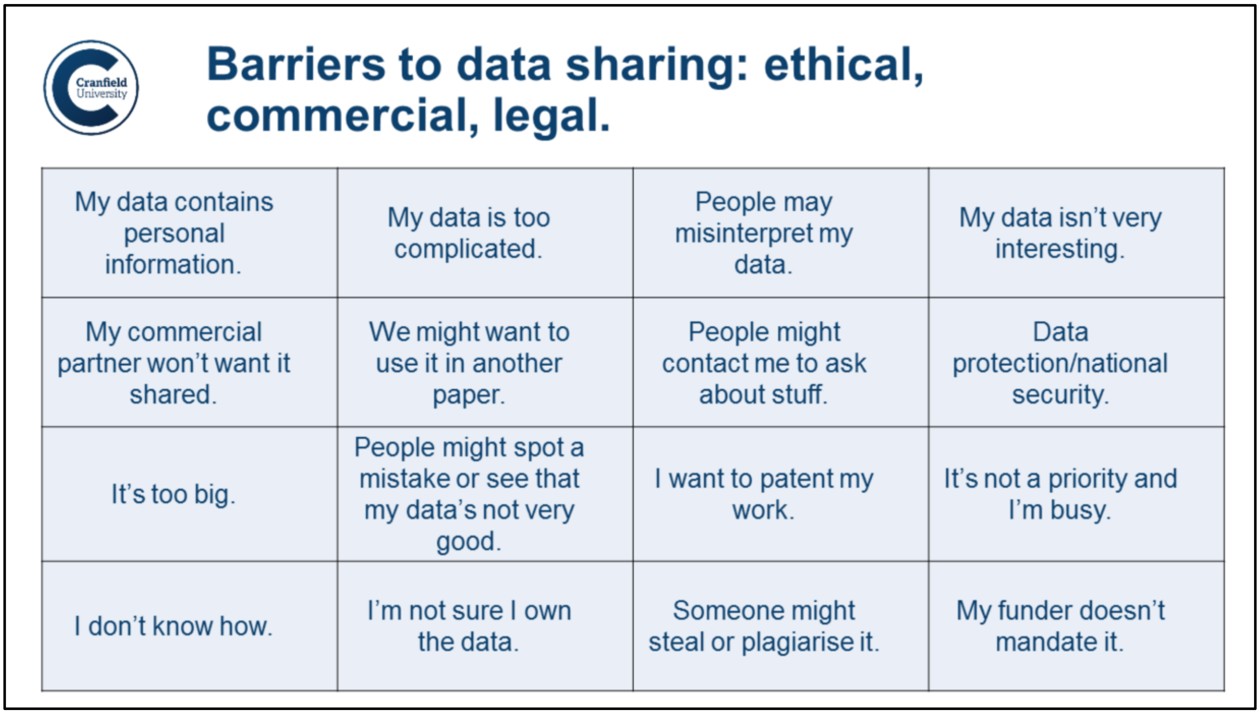
A final note: if you've used SPSS, it has an option to "export the data dictionary" which will give you a text file of variable labels, missing value codes, etc, that you can save with your data.

An example on CORD is https://doi.org/10.17862/cranfield.rd.6270233 – it is one dataset in a spreadsheet, with an accompanying readme.txt file explaining the data: how it was created, etc. There is also an image to better demonstrate the scenarios.



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

[GP]



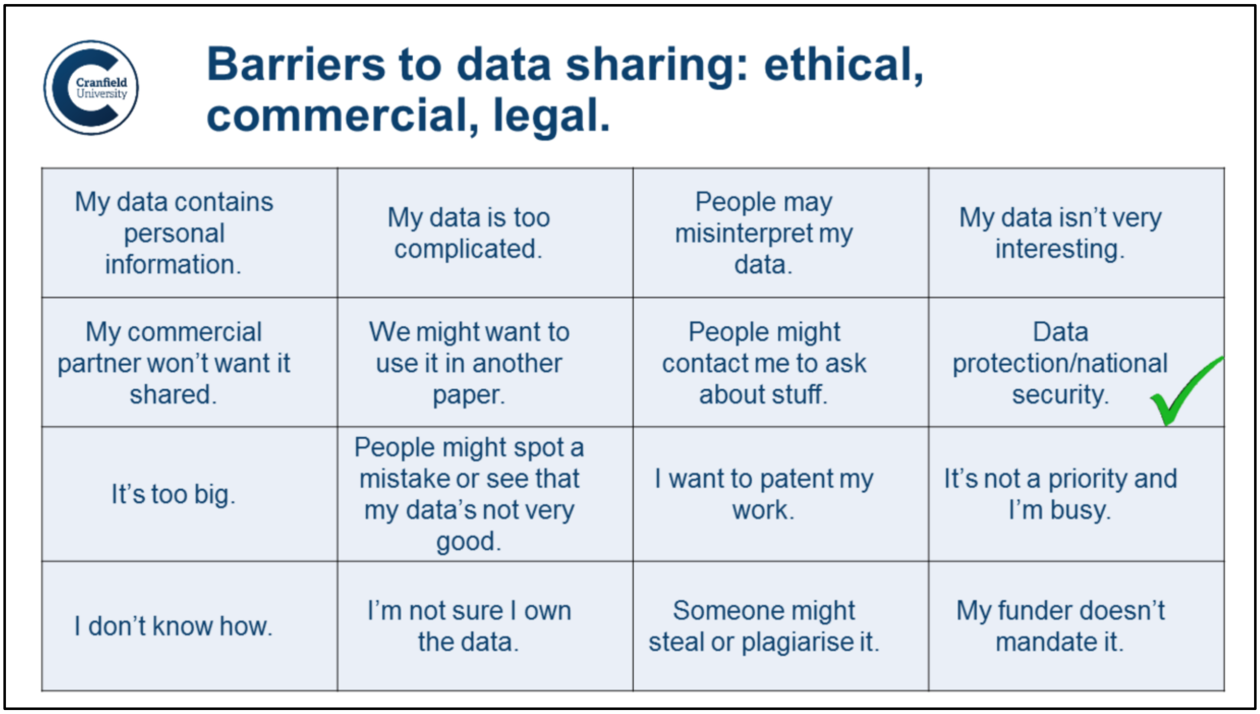
Barriers to data sharing: ethical, commercial, legal.

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]
[NB animations don't work in WebEx so separate slides.]

Wherever possible, you should be planning to make your data openly accessible, for all the 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?

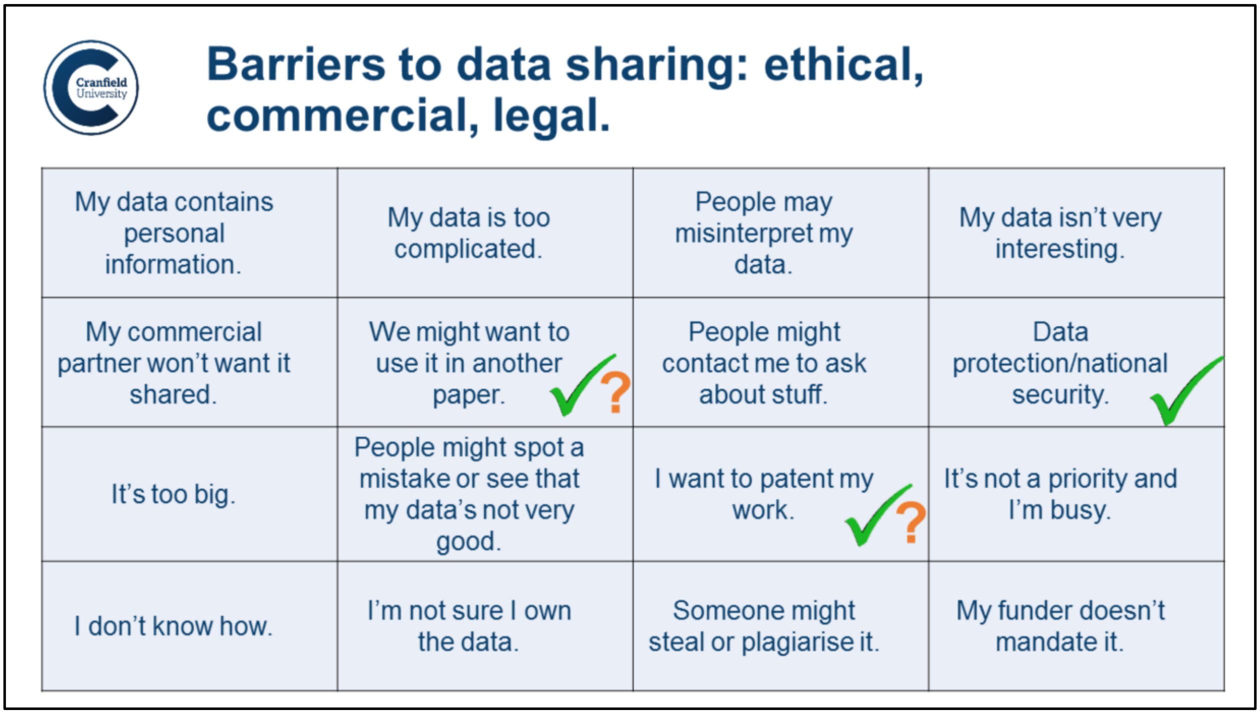


Barriers to data sharing: ethical, commercial, legal.

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]

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

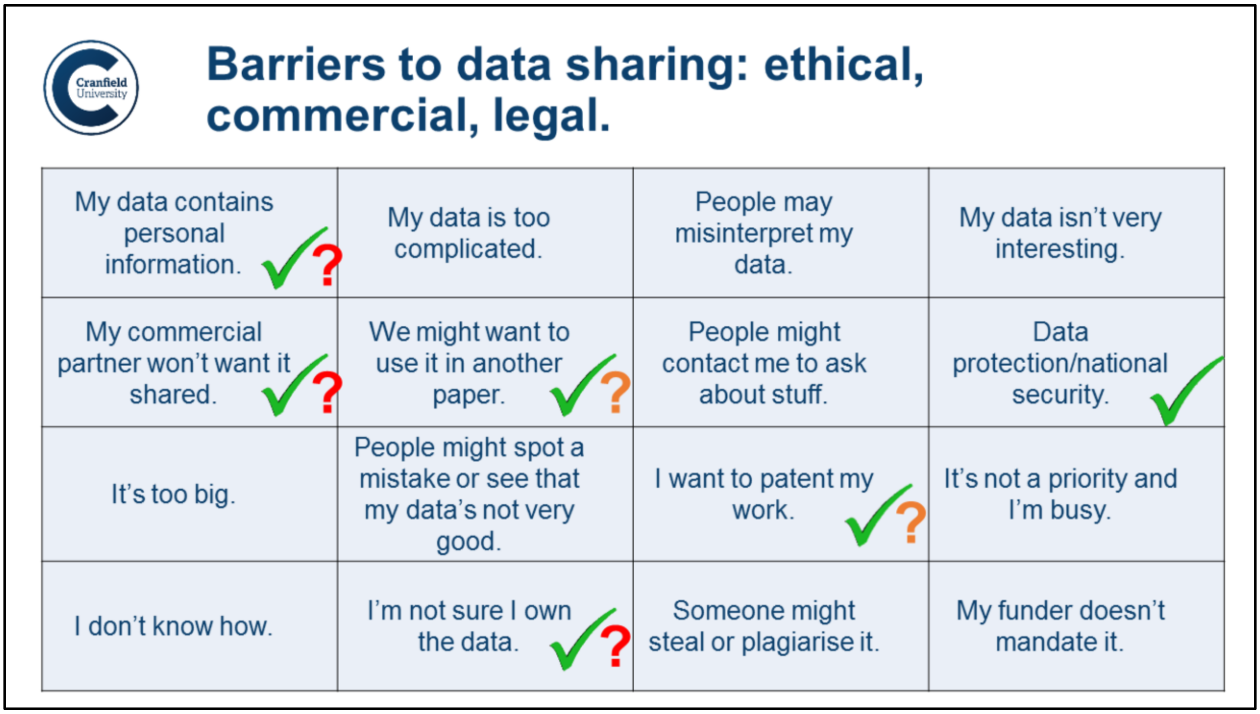


Barriers to data sharing: ethical, commercial, legal.

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]

Tick/orange Q – this is a valid reason to delay access to your data, but not to withhold your data. You have the right to first use, so can embargo your data until you've published or patented, but this isn't a reason not to share your data at all.

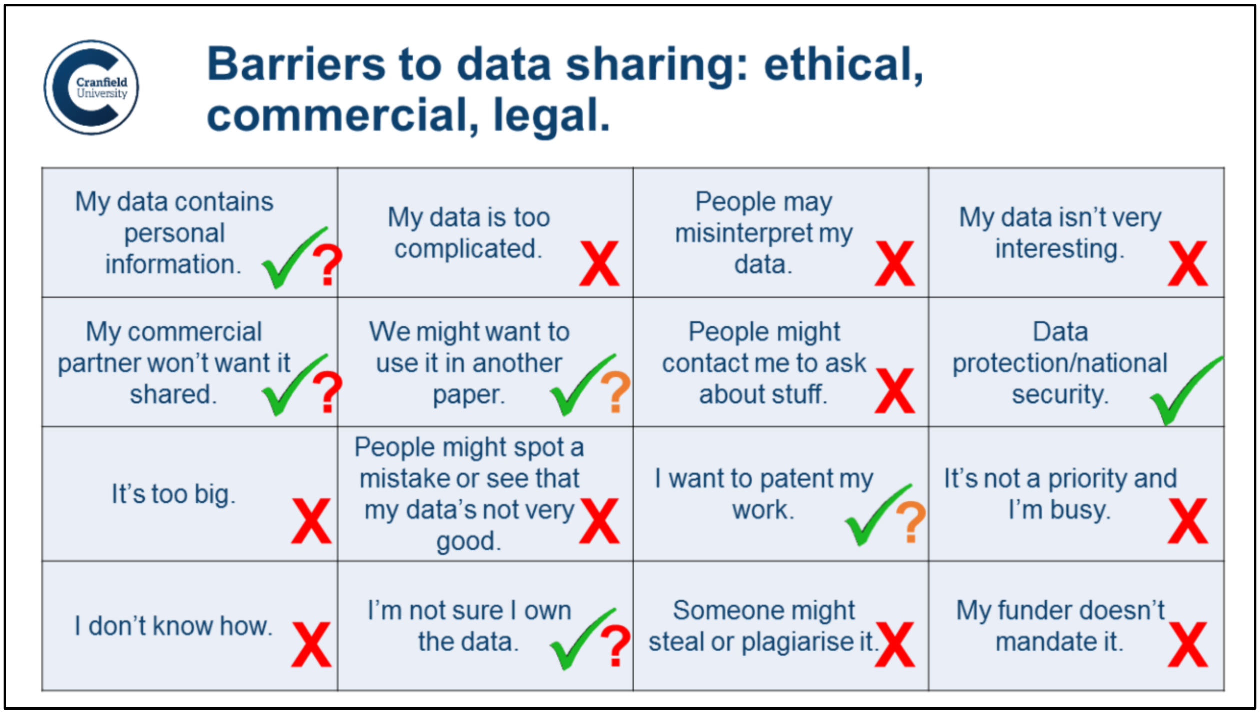


Barriers to data sharing: ethical, commercial, legal.

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]

Tick/red Q – this might be a valid reason, but you might be able to overcome it – we are expected to take steps to try to overcome any potential barriers to open data. If you're collecting personal information, for example, you should plan to anonymise it, and publish the anonymised version. 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? (If not, maybe I can help!)

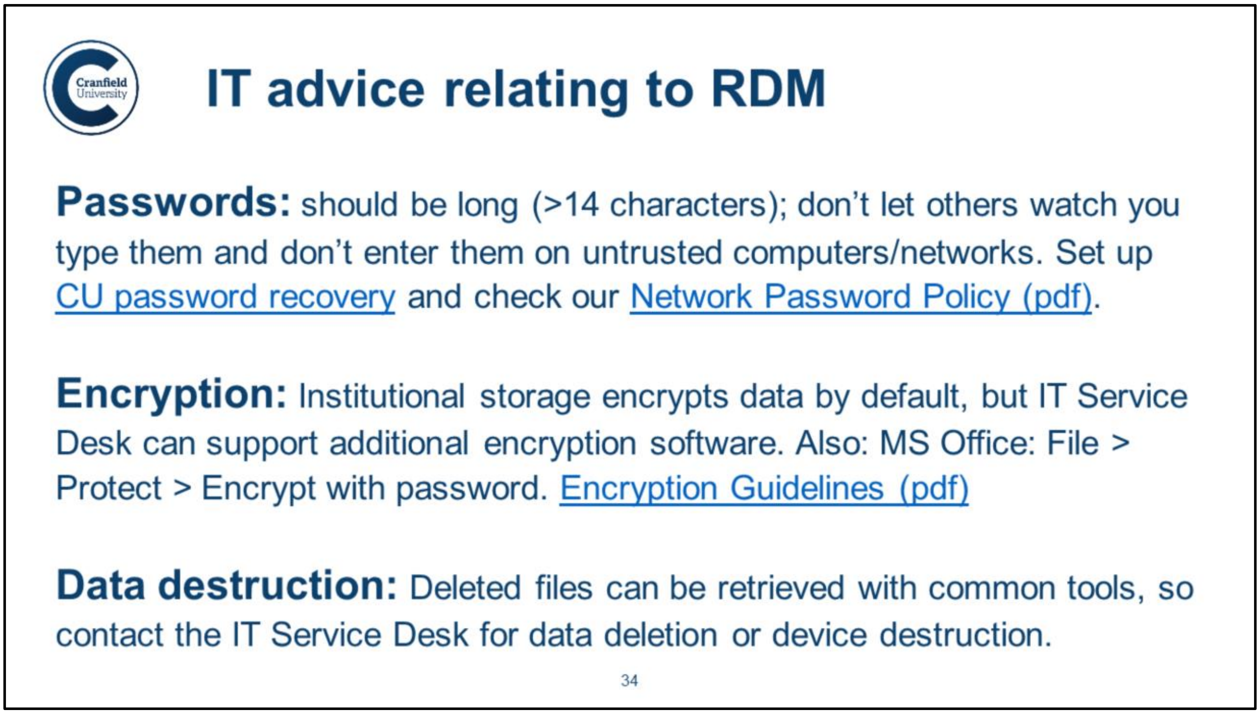


Barriers to data sharing: ethical, commercial, legal.

My data contains personal information. ✓?	My data is too complicated. X	People may misinterpret my data. X	My data isn't very interesting. X
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. X	Data protection/national security. ✓
It's too big. X	People might spot a mistake or see that my data's not very good. X	I want to patent my work. ✓?	It's not a priority and I'm busy. X
I don't know how. X	I'm not sure I own the data. ✓?	Someone might steal or plagiarise it. X	My funder doesn't mandate it. X

[GP]

Cross – these are excuses.



IT advice relating to RDM

Passwords: should be long (>14 characters); don't let others watch you type them and don't enter them on untrusted computers/networks. Set up CU password recovery and check our Network Password Policy (pdf).

Encryption: Institutional storage encrypts data by default, but IT Service Desk can support additional encryption software. Also: MS Office: File > Protect > Encrypt with password. Encryption Guidelines (pdf)

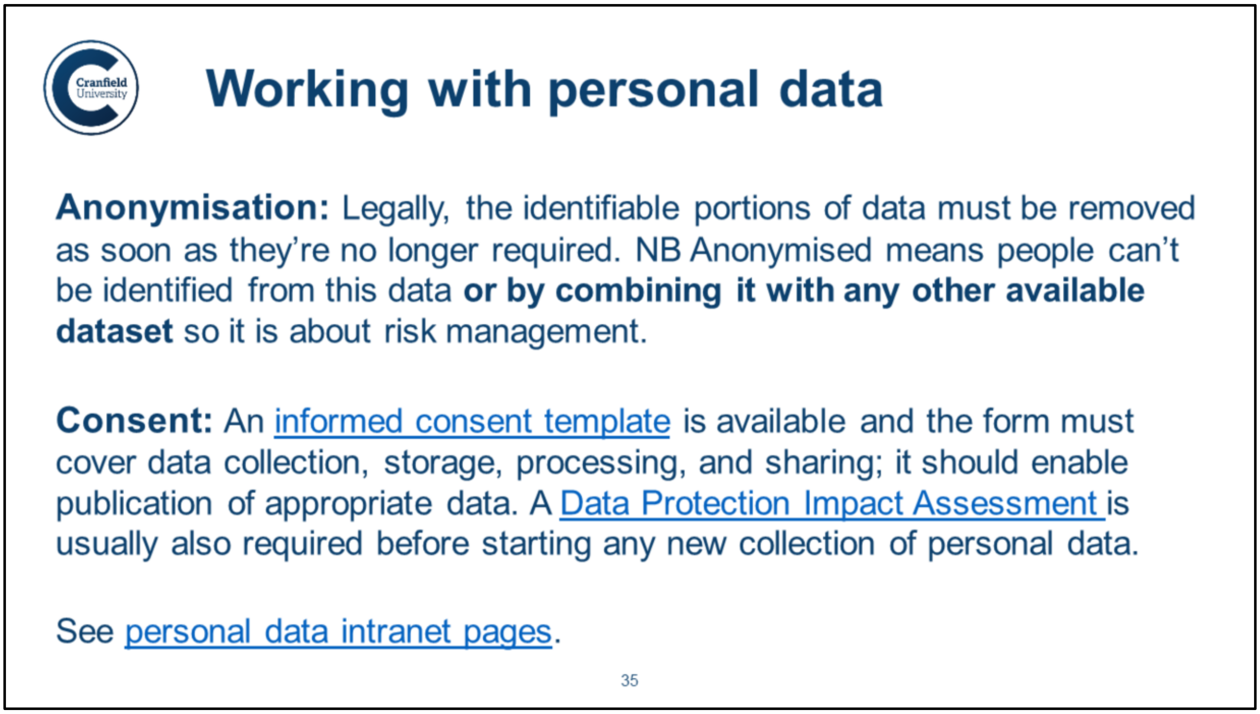
Data destruction: Deleted files can be retrieved with common tools, so contact the IT Service Desk for data deletion or device destruction.

[GP]
Some quick IT aspects now.

Firstly, your network password provides the security for data access so make it a good one! Make sure it's over 14 characters so it can be hacked by brute force attempts, and make sure you do not reuse it for any other service, otherwise that compromises university network security if they are hacked. Do set up password recovery but make sure you use difficult security questions people can't guess or work out.

Encryption shouldn't be needed as our storage meets all core security requirements. Office password protection can be used for extra industry-level encryption and is useful if you need to send someone else a file – password-protect it first and send the password by a different method. Just don't forget the password!

If you collect personal data, anonymise it, and then need to delete the original files, you can't just delete them. They will still be openable (my car would always open deleted files on my USB drive of podcasts or music – if a Hyundai is clever enough to access deleted files, it really can't be that hard!). Send IT the path to your files, or take your device in, and they will securely destroy the files. (See also ICO data deletion guidelines.)



Working with personal data

Anonymisation: Legally, the identifiable portions of data must be removed as soon as they're no longer required. NB Anonymised means people can't be identified from this data **or by combining it with any other available dataset** so it is about risk management.

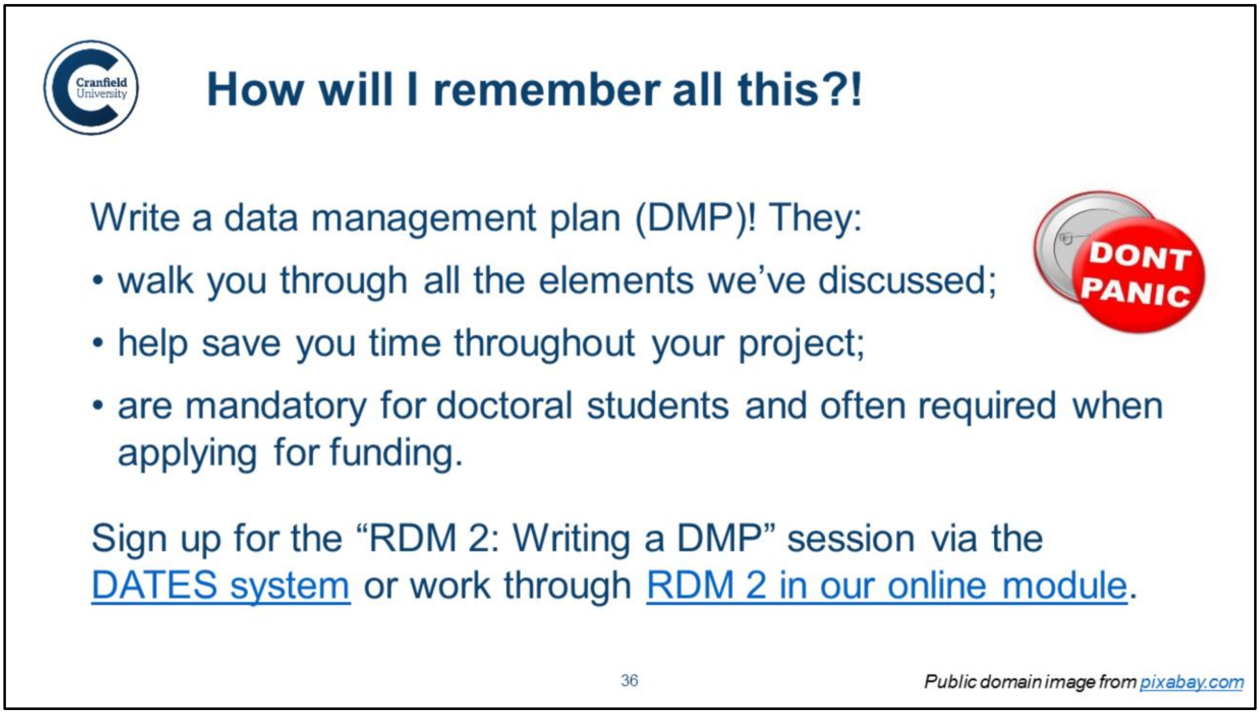
Consent: An informed consent template is available and the form must cover data collection, storage, processing, and sharing; it should enable publication of appropriate data. A Data Protection Impact Assessment is usually also required before starting any new collection of personal data.

See personal data intranet pages.

[GP]
We have some guidance online about working with personal data so I'm just pulling out two key aspects here.

One is anonymization, which means people can't be identified from your dataset on its own or in combination with any other dataset now or in the future. So it's hard to guarantee and is really all about managing risk. There are various techniques to use (removing names, generalising variables such as changing postcodes to counties, and removing outliers because a 101-year-old is going to be easy to identify). If you're not confident that your data is sufficiently anonymised, there are also options to make it restricted access.

We also have a new template consent form, written to meet RDM, ethical, and GDPR requirements. It's good to work from this form and I can help advise on it if needed. If you're collecting personal data, you should also email gdpr@cranfield.ac.uk to ask about filling in a DPIA form, and they can advise further on the legalities of working with personal data and GDPR requirements.


Cranfield University
DONT PANIC

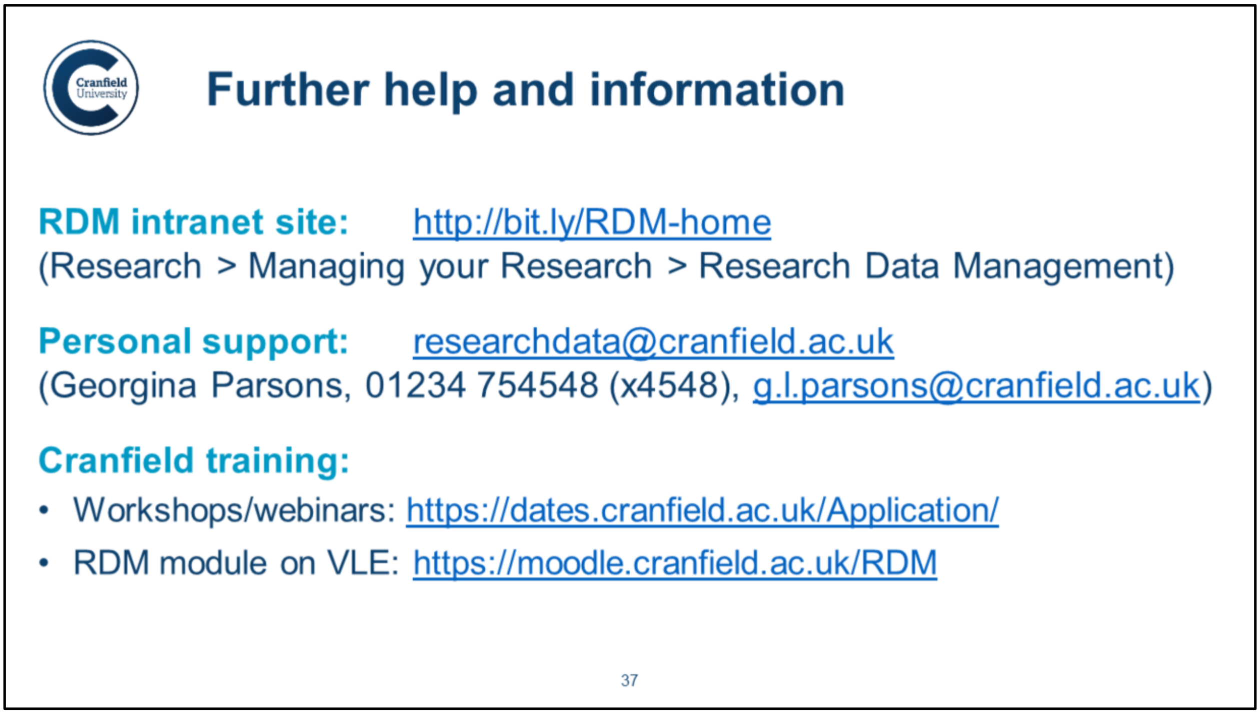
How will I remember all this?!

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 or work through RDM 2 in our online module.

Public domain image from pixabay.com

[GP]
The next RDM2 dates are:
- Workshop: __________
- Webinar: ___________



Further help and information

RDM intranet site: http://bit.ly/RDM-home
(Research > Managing your Research > 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://dates.cranfield.ac.uk/Application/
- RDM module on VLE: https://moodle.cranfield.ac.uk/RDM

[GP]