Trainer's notes for module 5:
Deposit and long-term preservation

Good practice in research data management

# Session Details

## Aims and Objectives / Learning Outcomes

By the end of this module participants will have:

* An understanding of the importance and benefits of depositing data and long-term preservation
* Considered funder mandated data centres and repositories
* A practical understanding of metadata
* Considered Newcastle University's Research Data Catalogue

## Session Topics

* Considering the long-term
* Data Centres and repositories
* Metadata and the Newcastle University Research Data Catalogue

## Structure

This module is planned to be delivered via one session lasting around 1 hour with a group of 12-20 attendees. The expectation is that these are postgraduate students and/or early career academics.

### Indicative timings

|  |  |
| --- | --- |
| Considering the long-term | 10 minutes |
| Data centres and repositories | 10 minutes |
| Metadata (pre activity)Activity: How much metadata is "Sufficient metadata"?The Newcastle University Research Data Catalogue (RDC) (post activity) | 5 minutes20 minutes10 minutes |
| Session review  | 5 minutes |
|  | 60 minutes total |

## Set-up

Slides are provided as detailed in the "notes to accompany slidedeck". Note the use of hidden slides by default for and slides offering further detail on a topic. (Hiding/unhiding slides is best done in 'slide sorter' view; select slides, right-click and toggle 'hide slide')

# Notes to accompany Slidedeck for module 5

## Considering the long-term (slides 3-9)

### Slide 4: Distinctions and definitions

As we think towards the long-term, it's useful to confirm a few definitions. Digital archiving, preservation and curation are commonly used, but do have some distinctions.

Digital archiving addresses concerns to do with backup and maintenance, and digital preservation addresses longer term strategies to ensure on-going access.

Digital curation is more holistic, concerned with the holistic range of processes applied to data over its lifecycle - and a definition from the DCC is the next slide.

### Slide 5: What is digital curation?

This short definition is taken from the DCC; the "Digital Curation Centre". Add emphasis as you see fit; however consider stressing and offering an interpretation of the inclusion of "adding value" and the mention of "trust". Also note the mention of both "current" and "future" use.

This more holistic is useful. Regular backups of data are not enough to ensure long term accessibility and reusability. For example, file formats may become obsolete, storage media may become damaged or your data may not be understandable if it lacks appropriate description and documentation to enable reuse.

A fuller definition is provided on the following slide – unhide it if you intend to use it!

The DCC offers comprehensive materials and briefing papers around topics relating to Digital Curation. See their Digital Curation 101 materials: <http://www.dcc.ac.uk/training/train-the-trainer/dc-101-training-materials>

### Slide 6: What is digital curation? #2

This longer definition is take from the summary definition offered by the DCC in their "What is Digital Curation?" guide: <http://www.dcc.ac.uk/webfm_send/437>

Again, add emphasis as you deem fit.

### Slides 7: DCC Curation lifecycle model

Let's look at the DCC Curation lifecycle model again (examined in more detail in module 4: Data Management Planning). Note the "occasional action: Dispose". Depending on time, you could lead into a few minutes discussion here, asking attendees:

* Do you need to store your data for the long term?
* Will it have future use?
* Do you need to be selective about what you choose to store?

### Slides 8: Key questions

Following on from the previous slide, elicit responses on these key questions:

* Where will you deposit your data for the long-term?
* Will you embargo the use of your data? If so, why? And how long for?

### Slides 9: Why is long-term preservation important?

Answering the question is similar considering reasons for sharing data (See module 4) - Sharing information facilitates new developments (for any subject area).

By depositing research data in an archive or repository you will make them accessible - even for your own future use.

Sharing research data provides benefit to the general subject community, but the benefit is less obvious for the investigator who makes his or her data available. However, there is some evidence publications with associated datasets are more highly cited. See:
[http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0000308](http://www.plosone.org/article/info%3Adoi/10.1371/journal.pone.0000308)

## Data Centres and repositories (slides 10-14)

### Slide 11: Where do I deposit my data?

Researchers should deposit data in recognised data repositories where these exist for particular data types**.**

In the UK the 15 national data centres are funded by one or more of the research councils and/or Jisc, and hosted by universities, national research centres, or laboratories. Examples are listed on the next slide.

### Slide 12: Example funder mandated data centres / repositories

Data centres and repositories bring together data sets from a variety of sources in one place, they curate these collections, and they provide access to data that is ready to reuse.

Some are designated by the relevant research council as the repository for their discipline, e.g. the seven data centres mandated by NERC, the Archaeology Data Service (York) and the UK Data Archive for the social sciences (Essex).

Note the table of National Data Centres on the Newcastle RDM site:
<http://research.ncl.ac.uk/rdm/policyandgovernance/nationaldatacentres/>

You may also wish to reference this report of "Data centres: their use, value and impact":
<http://www.rin.ac.uk/our-work/data-management-and-curation/benefits-research-data-centres>

### Slide 13: Benefits of depositing data in a data centre or repository

The benefits listed on the slide are self-explanatory.

It's also worth noting that many data centres provide extra services, e.g. guidance and training for researchers to help them create data that are ready for curation and reuse, including metadata, collection methods, and file formats.

### Slide 14: Example subject repositories

There are many subject repositories / data centres worldwide.

For example DRYAD (<http://datadryad.org/>) is an open repository of both journal articles and associated datasets for evolutionary biology and a growing range of subjects, hosted at the University of North Carolina. Figshare (<http://figshare.com/>) is a publisher-funded open repository based in London aimed at the biological sciences but open to all disciplines.

To find out if your discipline has a hosted repository, there are a few registries of data repositories.

* DataCite is a British initiative jointly edited by the British Library, BioMed Central and the Digital Curation Centre. Databib is an American initiative based at Purdue University Libraries with volunteer librarian editors.
* Re3data is a German website in English language whose partners are actively involved in the German Initiative for Network Information (DINI), and launched a full registry of data repositories in autumn, 2012 - The re3data.org registry allows the easy identification of appropriate research data repositories, both for data producers and users.

[MIMAS](http://mimas.ac.uk/) connects UK researchers to [Landmap](http://landmap.mimas.ac.uk/index.php/Datasets/) spatial datasets, including incredibly dense and rich satellite, thermal and radar images of the Earth's surface.

* MIMAS: <http://mimas.ac.uk/>
* Landmap: <http://landmap.mimas.ac.uk/index.php/Datasets/>

Some do not hold any original data, but focus on providing access, such as the Chemical Database Service.

But others do; European Bioinformatics Institute and their US equivalent, the NCBI, are developing major national and multi-national data repositories - it is often a grant requirement that academics store their raw data here.

## Metadata and the Newcastle University Research Data Catalogue (slides 15-27)

### Slide 16: What is metadata?

Metadata is increasingly important, allowing for a lot of the "magic" in the information age! Metadata describes data and offers context, and the more description and context there is, the more it can be combined, "mashed up" and reused, often in ways unintended, or unimaginable, by whoever created the data.

It explains the origin, purpose, time reference, geographic location, creator, access conditions and terms of use of a data collection. It conforms to set standards and is machine readable.

Metadata are typically used for resource discovery, providing searchable information that helps users to easily find existing data, and as a bibliographic record for citation.

You may want to show this page on the RDM@Ncl website:
<http://research.ncl.ac.uk/rdm/lifecycle/createandcollect/metadata/>

### Slide 17: Working with metadata

When depositing in an archive or with a data centre, researchers are generally given a template, structured to international standards or schemes, for completion of metadata when depositing in an archive or with a data centre.

Providing detailed and meaningful dataset titles, descriptions, keywords and other information enables data centres to create rich resource-discovery metadata for archived data collections.

### Slide 18: Example funder stipulation: EPSRC Principle 6

Note this principle from the EPSRC relating to metadata, emphasising the variable interpretation of what's "sufficient"? (This is akin to the need for "reasonable adjustments" wrt the Disability Discrimination Act.)

Activity: How much metadata is "Sufficient metadata"?

So, if we understand that metadata is additional data to describe a dataset, let's practice identifying it, and consider how much might be "sufficient".

This activity is intended to run for 20minutes (but could easily run on to 30minutes – if time is short, keep an eye on the clock!)

You will need enough printed copies of the handout (03 Handout metadata activity.docx), one for each attendee.

#### Running the activity

* Distribute the handouts and run through the fictitious metadata scenario (also available on slide 20)
* Allow a couple of minutes for individuals to read through the scenario; ask them to begin to list relevant metadata that could describe the dataset.
* After 5 minutes, encourage discussion in pairs. After 5 minutes discussion re. defining metadata, ask them to turn over the handout and consider different metadata requirements for different stakeholders. Allow up to 10 minutes for group discussion.
* Run a short feedback session, according to how much time you have for the module, reviewing the key questions on slide 21.

An extension of this activity would be to ask attendees to begin to identify metadata for their research data.

### Slide 22: 3 levels of metadata

This is an optional slide, hidden by default.

The three-tier metadata approach has been developed by the DaMaRO project at the University of Oxford:
<http://damaro.oucs.ox.ac.uk/index.xml>

### Slide 23: HEIs and metadata

This is a linking slide, connecting the understanding of metadata to why it's important at Newcastle University.

Universities are now required to make available on request the datasets that underlie the publications of their staff. They achieve this by making the metadata available online.

### Slide 24: Introducing the Newcastle University Research Data Catalogue (RDC)

### In the context of RDM, the Newcastle University Research Data Catalogue (RDC) is a significant step!

It's not an official repository for data at the University of Newcastle, rather a proof of concept metadata catalogue. It allows us to record research data location and important key information about it; it know that many and various repositories exist. Non-digital data sources (filing cabinets, audio tapes) and disconnected data sources (external hard drives and legacy disks) therefore remain in scope for the purposes of discovery at least. The system should allow researchers to indicate the location of their data sources in a manner which may allow the data catalogue to automatically harvest metadata.

### Slide 25: RDC enabling RDM

The RDC has been designed with the intention of being the University’s bespoke system that provides a link between the MyProjects and MyImpact systems, so linking research publications information with the projects that created them.

* My Impact: <http://www.ncl.ac.uk/res/resources/myimpact/>
* My Projects: <http://www.ncl.ac.uk/res/resources/myprojects/>

At this point you may want to login to and demonstrate the RDC.

### Slide 26: RDC visualised

The diagram shows how the RDC takes information from MyProjects and MyImpact, producing metadata connecting the people, projects, publications and data sources, which can be outputted in machine readable formats or for human readable web content.

### Slide 27: RDC: More information

More information is available about the RDC. The main location to reference is the RDM website:
<http://research.ncl.ac.uk/rdm/tools/researchdatacatalogue/>

## Session review (slides 28-30)

### Slide 29: In summary

This is a round up slide, with one summary point for each of the topics covered – you may wish to edit to align with your particular emphasis.

### Slide 30: Acknowledgements

Cited here are acknowledgements for resources used to create this module.

This module itself is licensed under a Create Commons Attribution-NonCommercial-ShareAlike license, meaning:

* You must attribute the work [[ADD HOW!!]]
* You may not use this work for commercial purposes.
* If you alter, transform, or build upon this work, you may distribute the resulting work only under the same or similar license to this one.