

**I've collected my data, so  
what do I do with it now?**

# **Research data management**

Session 2

Data Curation Lifecycle

Tutor Notes

# DATUM for Health

[www.northumbria.ac.uk/datum](http://www.northumbria.ac.uk/datum)

Project funded by JISC

Copyright holder: Northumbria University, School of Computing, Engineering & Information Sciences, 2011

Materials made freely available under a Creative Commons Attribution-NonCommercial-ShareAlike 2.0 UK: England & Wales license

## Session 2 Data Curation Lifecycle Notes for Tutors

### SESSION DETAILS

#### Aims and Objectives / Learning Outcomes

By the end of this session participants will have:

- an appreciation of how data curation can support and safeguard research
- understanding of the data curation lifecycle
- identification of the processes and activities involved in good practice for research data management
- awareness of the free services and tools available to support data curation

#### Session Content

- Introduction to the data curation lifecycle model
- Coverage of each stage of the lifecycle in detail
- The roles PGR students and others play in data curation
- Checklist of things to consider at each stage
- Links to tools and resources
- Group exercise and discussion on appraising and selecting data and records
- Group exercise and discussion on the students' own data management plans

#### Structure

10.00 -10.03	Introduction to the session
10.03 -10.10	What is data curation
10.10 -10.13	DCC curation lifecycle model
10.13 - 10.35	Conceptualise stage
10.35 - 11.00	Create stage
11.00 - 11.15	Refreshment break
11.15 - 11.30	Appraise stage
11.30 - 12.05	Group exercise - Appraise and select
12.05 - 12.10	Ingest stage Preserve stage
12.10 - 12.15	Store stage Access stage Transform stage
12.15 - 12.30	Group Exercise - Your data management plan
12.30	Directed learning tasks

#### Directed Learning Tasks

- think of the problems you've experienced with managing your research data, to share in Session 3
- think of any good tips or systems you've used for managing your research data, to share in Session 3

## Handouts

- PPT slides, 3 per page (provide the students with a hardcopy for use during the session)
- See the DCC's Digital Curation 101 training materials which are made freely available<sup>i</sup>. The ones made available to the students for this session comprise:
  - What is digital curation
  - Conceptualising data
  - Creating data
  - Appraising data
  - How to guide on selecting and appraising data
- An additional useful DCC resource is the Legal Watch Paper 'Sharing medical data'<sup>ii</sup>
- Appraisal guidance for Appraise Group Exercise (provide the students with a hardcopy for use during the session)
- Evaluation form for session (provide the students with a hardcopy for use at the end of the session)

## NOTES TO ACCOMPANY POWERPOINT SLIDES

### INTRODUCTION TO THE SESSION (Slides 0-3)

#### Slide 1: History of the development of this session

**Tutor:** not to be covered in the session. The slide credits the sources involved in producing this session.

#### Slide 2: Overview

I will introduce the data curation lifecycle model, and then cover each stage of the lifecycle in detail. I will indicate the roles that you and others play in data curation. I will provide checklists of the issues to consider at each stage as well as giving links to tools and resources. This session is providing you with a lot of necessary information, so it is not as interactive as Session 1. So I would encourage you to ask questions or make comments at any time during the session. There will be two exercises for you to carry out later on in the session.

#### Slide 3: Learning Outcomes

By the end of this session you should have:

- an appreciation of how data curation can support and safeguard research
- understanding of the data curation lifecycle
- identification of the processes and activities involved in good practice for research data management
- awareness of the free services and tools available to support data curation

I hope you leave able to explain why data curation is so important.

### WHAT IS DATA CURATION? (Slides 4-6)

#### Slide 4: What is Data Curation?

A definition of data creation is:

“the active management and appraisal of data over the lifecycle of scholarly and scientific interest”<sup>iii</sup>

Data provides the evidence base for the conclusions of research work.

It can't be over emphasised that curation is part of good research practice

#### Slide 5: Why Curate: -Requirements

Just a brief reminder of why you need to manage - curate - your research data. This has been covered in detail in Session 1.

Both international bodies (e.g. the OECD) and national bodies (e.g. RCUK) have as the principles and practice of good research that data should be kept and made publicly available. To do this requires data to be managed.<sup>iv</sup>

And research funding bodies require that in research projects funded by them, data is (i) managed by the use of data management plans, and (ii) made publicly available for use by others.

### **Slide 6: Why Curate: Rewards**

But managing data is not just something you do because others require you to. Good management of research data brings you benefits and rewards:

- it makes it easier and more efficient to do your research
- it can prevent data loss
- it provides validation of your research

and evidence shows that placing your data into the public domain increases your reputation and the impact of your research and opens up new research opportunities.<sup>v</sup>

## **THE DATA CURATION LIFECYCLE MODEL (Slide 7)**

### **Slide 7: DCC Curation Lifecycle Model**

This slide introduces the Data Curation Lifecycle Model produced by the Digital Curation Centre (DCC).

The model comprises eight stages, in sequence:

- Stage 1: Conceptualise - planning what to do
- Stage 2: Create - collecting and analysing data
- Stage 3: Appraise - selecting what to keep
- Stage 4: Ingest - transferring data to a custodian
- Stage 5: Preserve - keeping data over time
- Stage 6: Store - keeping data safe and secure
- Stage 7: Access - finding and using data
- Stage 8: Transform - generating new data
  - from this stage you start the cycle again

In this session I will concentrate on the first three stages - they are the most important for your PhD study - and will cover the rest of the stages briefly.

## **CONCEPTUALISE STAGE (Slides 8- 15)**

### **Slide 8: Conceptualise Stage - Planning What to Do**

This is the first stage of the lifecycle, where you plan what data you are going to collect for your PhD study. This is in some ways the most important stage. Decisions made at the Conceptualise Stage have an impact on every other stage of the lifecycle, so it is well worth getting things right from the start!

Data management is inextricably linked with research methodology. As you define your research question and start to plan out your methodological approach, you naturally start to think about what types of data you want to collect and how you will collect and analyse them.

Though you, as the PGR student, will play the main role in your data management, many other individuals and organisations will play a part too. Your supervisory team is the main source of advice and guidance. The IT department in the university will play an important role. The funders / sponsors of your research (if you have them - not every PGR student does) will have their requirements and also provide guidance. And research governance and ethical procedures are vital.

It is at the Conceptualise Stage that you would start to fill in your data management plan: this was covered in detail in Session 1 and you've had the opportunity to develop your own plan as a directed learning task.

**TASK:** On your own / in pairs, in the context of your PhD:

- think about what activities you need to carry out at the Conceptualise Stage

**Tutor:** Give them 2 mins (exactly) for this activity. Their ideas will be discussed after the Checklist has been covered.

### **Slides 9-10: Checklist - Conceptualise Stage**

#### **Start with Slide 9.**

These checklists are designed to help you think about what you need to do at each stage. They are not designed to tell you **how** to do it. The practical issues are going to be covered in Session 3. I'll go through these, then we can see what you thought should be covered at this stage.

Get into the habit of equating data curation with good research

- This is where you start to put into practice good data management so you can achieve this ideal. Like any other practical activity, you have to learn how to do it, but when you've done it once it becomes much easier to do for future research projects.

Get into the habit of documenting everything

- When you make decisions about your data, or changes to your data, and so on, record this information so in the future it's easy to know what you've done. Memory is a poor substitute when you consider how many years a PhD study can take. It's too late at the writing up stage to realise you've forgotten the full details of how you collected and analysed your data at the beginning of your study.
- Data management is not just about the data itself. It also includes managing associated documents and records, e.g. methodology protocols, consent forms, correspondence with participants, progress reports to the university, and so on.

Start filling in your data management plan

The first time you do this it **is** daunting: but you will find in your later research career that every data management plan you have to complete becomes easier and easier.

Find out your university / sponsor / funder requirements

- These bodies will place requirements on what you have to do - or what you are not allowed to do - with your data, particularly at the end of your PhD. So be prepared!

Find out the ethical approval requirements

- What does the Ethics Panel require you to do with your data? What data do you want to collect, and how? What do you want to do with your data, e.g. to re-use it after the PhD

study, or to share it with others? These questions, and all the above requirements, will determine what consent you will need from your participants.

#### Determine intellectual property rights

- Beside yourself, other people might claim IPR over your data, e.g. the university, a funder or sponsor of your research. So for example at Northumbria University the University owns the IPR, though they might come to an agreement to share this with a funder of the PhD as appropriate: the PGR student holds the copyright of the thesis.<sup>vi</sup>
- Participants hold the copyright to 'their words' so in the consent form you need to ask them to assign this to you.

#### Understand legal requirements

- Two laws in particular affect data - Data Protection Act and the Freedom of Information Act. Data protection involves keeping participants' personal details private, unless they consent otherwise. Freedom of information could affect research data held by public bodies, which include universities. A person could request such data to be made publicly available. Session 3 will go into these topics in more detail. Other laws and regulations might also be applicable to you depending on the nature of the research you are carrying out, and whether you are conducting research in other countries.

#### **Move to Slide 10.**

#### Identify roles and responsibilities

- I will be pointing these out as we go along.

#### Can you use existing data? If so, find out the requirements and constraints

- You don't always have to collect new data to undertake research. One of the drivers for research organisations requiring data to be placed in the public domain is that such data can be used by other researchers and this will increase the value derived from the research grant money. Some research funders ask if you have searched for existing data that you could use. If you do use any existing data then there may be requirements and constraints on how this data can be used.

#### Can you reuse / share your data? If so, identify the ethical, legal and methodological constraints

- Similarly you need to decide whether you will re-use your data in the future, after your PhD is completed, or whether you will share your data with other researchers. As I said earlier, this will affect what consent you will need from your participants. If you are planning to share your data with other researchers will you need to place any restrictions on how or when they can use the data?

#### Identify any anticipated publication requirements, e.g. embargoes, restrictions, open access, etc.

- Will a funder / sponsor place any restrictions on what can be published, e.g. because of commercial confidentiality?
- Do you need to place an embargo on the data, so that it cannot be released for a specified number of years?
- Are you going to release the data under open access conditions such as Creative Commons or Science Commons licenses? Sometimes funders require this.



Identify the software, equipment, storage requirements, skills etc. you will need, and how you will obtain them

- Remember to include data management skills in your training plan.

### ***TASK responses:***

**Tutor:** Talk to the students as a group, and take answers and comments / questions from the floor.

- Were there any other things that you thought should be included at this stage?
  - Students might raise issues that should be part of later stages in the lifecycle. Make clear that the Conceptualise Stage is about 'planning' not 'doing'.
- Did you think of most of these points? Were there any issues in the Checklist that surprised you?
- Any questions?

### **Slide 11: Tools and resources**

The Digital Curation Centre (DCC) offers a range of help and guidance, e.g. you can contact their Helpdesk for advice.

In the DCC's Policy and Legal Pages, they have information about all the research funders' data management requirements. This may not be relevant to you now, but will be in your future research career. I've given links to the data management and sharing policies of three funders of health-related research: The Economic and Social Research Council<sup>vii</sup>, the Medical Research Council and the Wellcome Trust.

## **CREATE STAGE (Slides 12 - 15)**

### **Slide 12: Create Stage: Collecting and Analysing Data**

This is the second stage of the lifecycle when you collect and analyse the data for your research. You also create the accompanying administrative records and documents, such as consent forms, protocols, correspondence.

An important part of data collection is the parallel creation of the metadata and contextual information about the data. Without this metadata and context, the data could become meaningless to you in the future, and would certainly be meaningless to others. As data is often expensive to recapture, or in some cases impossible to recapture, it is essential that you make the necessary effort to ensure that your data has context for long-term comprehensibility and reuse.

During this Create Stage you will be putting into action the plans you developed at the Conceptualise stage, but also developing and amending these plans to accommodate any changes in the direction of your research, and your increasing familiarity with the research process and your data.

The same roles are appearing here, but an additional source of support for you could be information specialists from the university library and from the university's repository.

**TASK:** On your own / in pairs, in the context of your PhD:

- think about what activities you need to carry out at the Create Stage

Give them 2 mins (exactly) for this activity. Their ideas will be discussed after the Checklist has been covered.

## **Slides 13 - 14: Checklist: Create Stage**

### ***Start with Slide 13.***

Be realistic and pragmatic – find a balance between what is ideal and what is necessary

- Though it is important to manage your data and document what you do, be realistic. Manage to the level that helps you do your research, and using techniques that you will be able to continue carrying out through the whole of your PhD study. Perfection is not sustainable!

Continue with your data management plan: it's a living document

- You will be adding to your data management plan and amending it throughout the whole of your PhD study. It is there to help you. It's not a tick box exercise that you complete at the beginning of your PhD and then forget.
- The experiences of collecting your data may require you to change and adapt your plans.

Establish what you want to do with your research data and what you want others to do

- You will have thought about this at the Conceptualise Stage, but as you become more familiar with your data and the reality of your research you may want to amend these plans.

Obtain the necessary consent and permissions

Determine what metadata and contextual information you need and how you will collect and record it

- The DATUM data management plan template provides quite a bit of detail about this, to help you in filling out the plan. I'll just note a few examples to illustrate what you need to consider.
- Metadata is information that makes your data understandable and usable.
  - So for an interview you would need: interviewer name, participant name (or code), date of interview, time / duration of interview, name (or code) of location of interview, 'name' of interview schedule used, name(s) of interview record(s) obtained (e.g. audio record, hand written interview notes), confirmation that consent form obtained, name of transcribed record, name of anonymised transcribed record.
  - You can record this information in the files themselves, e.g. in the transcript, and / or in a database.
  - Supporting documents you would need would include the key that links the participant name / interview location to the codes used, the interview schedule, the signed consent form, and so on.
- Contextual information provides a wider, richer picture of your research.
  - This would include your project proposal, project information sheet, methodology description, characteristics of participants, description of research setting, research instruments, organisational / social / political background.
  - This will be recorded in documents, maybe in interview audio records and transcripts, in your research diary and so on.

### **Move to Slide 14.**

Consider all the different types and formats of data and records you will collect and produce

- Just a few examples to get you thinking about what you will collect. So for an interview you may have audio recordings, transcripts as word documents, research diary / interview notes which might be in paper form and then scanned electronically, pictures. Then administrative records would include emails setting up the interviews, project information sheet, signed consent forms, interview schedule, and so on. Records of all the admin to do with your data may be kept in an Access database. Your analysed data may be in software packages such as NVivo.
- How you manage these different types of data and records will also differ. I will pick that up in the group exercise we will do shortly.

Consider the continuum from raw data to published outputs and the many processes your data will undergo

- So for an interview you may have recording, transcribing, anonymising, agreeing the transcript, analysing, synthesising, writing for publication. All these processes generate data and administrative records.

Set up file naming rules, version control etc.

Establish protocols for data protection, anonymisation etc

- These practical issues will be covered in Session 3

Obtain the necessary training and support

### **TASK responses:**

**Tutor:** Talk to the students as a group, and take answers and comments / questions from the floor.

- Were there any other things that you thought should be included at this stage?
  - Students might raise issues that should be part of later stages in the lifecycle.
- Did you think of most of these points? Were there any issues in the Checklist that surprised you?
- Any questions?

### **Slide 15: Tools and resources**

Your university will have handbooks to help you. A very useful source is the UK Data Archive - the national repository of digital research data in the social sciences and humanities. Their create and manage data web pages have a wealth of practical information. A similar source from another country is the Australian National Data Service - there is a lot of information on this topic on the Web. And the DCC also provides resources. The DATUM for Health project produced a customised search engine - RDM Training - linking to such resources.

### **REFRESHMENT BREAK**

## **APPRAISE STAGE (Slides 16-21)**

### **Slide 16: Appraise Stage: Selecting What to Keep**

Appraise is the third stage of the lifecycle model. While storage space and costs may not be an issue for some of you, the 'keep everything' approach may not be viable in the longer-term. As the volume of data you keep increases it becomes harder to easily find what you want.

So you need to decide what you have to keep (legally, or ethically), what you want to keep (for your own purposes, or because the data will be of use to others) and what you can get rid of. And you need to think how you will keep data and conversely how you will get rid of it.

An additional source of help to those noted previously will be your university repository.

After I've covered this topic we will have a longer group exercise.

### **Slide 17: General appraisal criteria**

All records and information need to be appraised and there are general criteria to help with such decisions. These criteria can help you to appraise your data too.

#### **1. Relevance to Mission**

Does the item content fit the organisation's remit and priorities, including any legal requirement to retain the item beyond its immediate use?

#### **2. Scientific, Social, Cultural, Historical Value**

Is the item scientifically, socially, or culturally significant? You need to anticipate future use, from evidence of current value.

#### **3. Uniqueness**

To what extent is the item the only or most complete source of the information it contains?

#### **4. Potential for Redistribution**

Is the item authentic (i.e. what it says it is), with integrity (unchanged), and usable? Does it meet IPR and ethical requirements?

#### **5. Non-Replicability**

Would it be feasible, or financially viable, to replicate the item?

#### **6. Economic Case**

What is the cost of managing and preserving the item, and does the value of the item justify this cost?

#### **7. Full Documentation**

- Is the metadata and contextual information needed to find, access and reuse the item comprehensive and correct?

## **Slides 18 - 20: Checklist - Appraise Stage**

### **Start with Slide 18**

Think about appraisal and selection as early as possible

- Remember, this isn't only about data but also about the associated contextual and administrative items.

Determine the minimum you need to keep for your findings and publications to be supported over time

- Don't hold on to unnecessary items, particularly personal data. It clutters up your files, makes it harder to find the items you want, and makes you more vulnerable to problems, e.g. loss of personal data, responding to data protection requests.
- Deleting or destroying items is an important part of data management.

Think about what your university / sponsors / funders expect you to keep, for how long and where

- For example, the university's retention policy<sup>viii</sup>; guidance from RCUK<sup>ix</sup>

Think about what laws and regulations affect your data, e.g. what you are allowed, or not allowed, to keep

Think about what you want to do with the data – now and in the future after your PhD is completed

### **Move to Slide 19**

Think about what you want others to do with your data

Consider if your data has wider scientific / cultural value

Know what consent and permissions you obtained

- This illustrates the importance of thinking about appraisal early. If you haven't got the correct consent to keep items you either have to destroy the items, or go back to your participants and re-consent to do what you want to do.

Things to keep – plan where, who, how, and for how long

- Once you've identified the items you are going to keep you need to plan where you are going to keep them, who will be responsible for them, how they will be managed, and for how long.

Find out what resources you have access to for storage

Find out if local or national repositories will take your data

### **Move to Slide 20**

Ensure you have all the necessary contextual information, metadata and data documentation

Things to destroy - plan how

- Items that you want to destroy need to be deleted / destroyed securely and confidentially, e.g. by shredding personal details.

- Keep a record when you make major destructions, e.g. if you've stated that you will keep the data and consent forms for 5 years after the end of your PhD, then when you reach that time, note the date that you destroyed them and how you destroyed them.

Find out what resources you have access to for destruction

Continue updating your data management plan

- This is where you could record for example the date etc. of major destructions, or link to such information from your plan.

## Slide 21: Tools and Resources

These are some guidance from the DCC, plus an example of a university policy.

## GROUP EXERCISE - APPRAISE AND SELECT (Slide 22)

**Tutor:** Organise the students into groups of 3 or 4, deliberately arranged to have a mix of years (i.e. stages of the PhD) in each group. Ask them to think of both data and associated records (such as correspondence, consent forms, protocols) that they are using for their own PhD. Then to group these into three categories: they only need to give one or two examples per category. They are to give the reasons for their decisions, and state how long they would keep the item and, as applicable, how they would destroy the items, or where they would keep them. The categories are:

1. Items kept for a short period
2. Items kept very long term
3. Items that fall in between the two categories above

Circulate to respond to questions and ensure participants are focusing on the task as set. Allow 20 minutes for discussion. This is followed by feedback and open discussion of their decisions. This part of the exercise to take 15 minutes.

Illustrative examples to help in the open discussion:

1. Items kept for a short period
  - Audio file of interview; deleted after transcript agreed with participant. Reason: to maintain anonymity. Deleting and overwriting file; physically destroying / breaking audio tape / CD / USB stick.
    - Note: for some research where the actual voice of participants are important audio files might be kept long term. It is the individual research project that determines the nature of the appraisal decisions.
  - Draft versions of an interview transcript; deleted once the final version has been agreed with participant. Reason: Draft versions are unnecessary. Deleting and overwriting file.
  - Emails that arranged a visit/interview; deleted once the event has taken place. Reason: unnecessary, and contains personal data - don't keep personal details once the purpose for using that personal data has been achieved. Delete from email box, and empty 'Deleted items' folder. Note: You have no control over what the recipients of the email will do with their emails.
  - Meeting agenda; deleted once minutes have been prepared. Reason: unnecessary; the agenda details will be contained within the minutes. Deleting and overwriting file;

deleting email with attachment from email box and emptying 'Deleted items' folder; recycling or shredding (if personal details are present) surplus paper copies. Note: You have no control over what the recipients of the agenda will do with their copies.

2. Items kept very long term

- Government datasets collected on a regular basis, e.g. Health Survey for England, Welsh Health Survey, Scottish Health Survey. Available from the Economic and Social Data Service (ESDS)<sup>x</sup> and government websites.
  - Note: these are questionnaire surveys collecting quantitative data
- 'Mothers and Daughters : Accounts of Health in the Grandmother Generation, 1945-1978', 46 anonymised interview transcripts, available at the ESDS<sup>xi</sup>. "The research looked at beliefs and attitudes to health and medical care, inter-generational relationships, and social history of members of a grandmother generation. ... [in particular] They give opinions on the quality of health care before and after the introduction of the National Health Service."
  - This data has historical interest and the information is unlikely to be easily collected again as the generation covered ages and dies.

3. Items that fall in between the two categories above

- Most data, and associated records, fall into this category, and this is where the more difficult decisions will occur. Decisions are often driven by practical and pragmatic reasons, e.g. if at the end of your PhD you have to keep the data and records yourself you might choose to destroy a lot of material at the end of your PhD.
  - Hopefully most of the open discussion will cover this category.
- Data collection instrument: you might not need the instrument for the PhD study once it has been completed. However, you might want to keep it to help in the design of instruments for future projects.
- Researcher diary. Is this redundant once the PhD project has been completed? Or will it contain insights of use to writing articles from the PhD study, or conducting other research projects? If you keep it more long term, you will need to ensure it contains no personal details of the PhD participants. Is this anonymisation a quick and easy task to do?
- Data continuum - raw, anonymised, analysed, synthesised, published - do you keep data from every stage, or only from the later stages, and if so which stages?
- If you keep data, then do you also need to keep consent forms and the key to participant codes? If you are only keeping anonymised, synthesised data then you can destroy the consent forms etc.

## INGEST STAGE (Slide 23)

### Slide 23: Ingest Stage: Transferring Data to a Custodian

This is the fourth stage of the lifecycle. At this point, the data will be transferred to a curation environment. This could be within a university or national repository, or something as simple as a drive on the university's IT system. The organisation will take over some of the data management activities. Repositories take all the responsibility, whilst university IT departments will carry out basic activities such as keeping the data safe and secure.

### Checklist

Know what the university's IT policies and procedures are

Be able to meet the repository's requirements

- Repositories will require the data to be in a specified format, accompanied by sufficient metadata and contextual information and with all the required consent and permissions.

## **PRESERVE STAGE (Slides 24 - 26)**

### **Slide 24: Preserve Stage: Keeping Data Over Time**

This is the fifth stage of the cycle. If you keep data for any length of time - and the period of a PhD study is a long time in technology terms - then you need to ensure that the data remains accessible and usable. And authentic - which means that the data is what it says it is; a true and complete item. And has integrity - which means that it has not changed since it was produced, or if there have been changes then these are known and were made by authorised people; i.e. the data has not been tampered with.

Roles are those of the same people we met before.

### **Slide 25: Preservation**

Preservation is needed because of the rapid pace of technological change. Hardware and software are updated with newer versions, or completely new technologies. So the older versions become obsolescent and ultimately unusable. Storage media have a short life - they can break or degrade - and storage devices are also subject to obsolescence. In fact, paper is far more long lasting! But don't think of preservation as only being a response to these risks. It is also an opportunity to maintain the value of an important asset - your digital data - and therefore ensuring you have the potential for creating new opportunities.

Continuing access to your data is not guaranteed: some action has to be taken. The key approaches to preservation are:

#### **1. Migration**

- Copying or converting data to a new system, e.g. a new version of software or operating system; a new form of data storage.

#### **2. Emulation**

- Replicating / mimicking the functionality of older software on a newer system.

#### **3. Hardware preservation**

- Maintaining access to the computer hardware and peripherals.

#### **4. Exhumation**

- Maintaining access to the software environment.

The Digital Preservation Coalition provides some resources that will be of interest.

### **Slide 26: Checklist: Preserve Stage**

Plan for the long term early not late

Be aware of technological changes during your PhD study



Appreciate that there is a limit to what you can personally do to preserve your data

- Realistically, you personally can only migrate your data, and there is a limit to how practical this will be. Other activities require organisational support, such as the IT department. Repositories will take on responsibility for preserving the data they hold.

Know what organisations will do to preserve your data, e.g. the IT department, a repository – and be clear what your role is in this process

- Repositories usually ask for data in formats that are considered to be more accessible and longer lasting. You would be responsible for migrating your data to that required format.

Document your preservation actions

- This enables you, and other people, to know what has been done to the data over time. You could keep such information in your data management plan, or linked from your plan.

## **STORE STAGE (Slide 27)**

### **Slide 27: Store Stage: Keeping Data Safe and Secure**

This is the sixth stage of the lifecycle. You need to keep data safe - protecting it against corruption and loss. Mechanisms such as backups and anti-virus software help with this. You need to keep data secure – controlling access to prevent unauthorised people getting hold of the data, and keeping personal details and data confidential. Useful mechanisms are id/password access, encryption.

#### ***Checklist***

Data on university IT facilities – know what their policies and procedures are, e.g. amount of storage space, backup, security etc.

Data on your own IT equipment – work out how to keep it safe and secure

## **ACCESS STAGE (Slide 28)**

### **Slide 28: Access Stage: Finding and Using Data**

This is the seventh stage of the lifecycle. Over the time span of a PhD study (which can be many years, e.g. four years for full-time students, seven years for part-time students, plus time after the award of the degree for writing articles) you need to be able to access and use your data. Your data needs to withstand scrutiny – now and over time. Will you be able to re-interpret your findings if you have to respond to academic misconduct allegations? How easy would it be for you to find and access your data if you received a request from a participant to see all the data you had on them, or a participant asked to withdraw from the study, and for all their data to be withdrawn too? Beyond the end of your PhD, if appropriate, you might want to reuse your data for new research projects, or you may want to share your data with other researchers.

### **Checklist:**

Be aware of the many ways you can make your data available to others

- There are many ways of sharing your data with other researchers: informally on request by email or CD / USB stick; placing data on the Web<sup>xii</sup>; putting data into publications, such as your thesis, or using enhanced publications<sup>xiii</sup>; putting data into university or national repositories.

## **TRANSFORM STAGE (Slide 29)**

### **Slide 29: Transform Stage: Generating New Data**

This is the eighth and last stage of the lifecycle. At this point, new data may be generated from original data, for example by:

- migrating data into a different format
- creating new visualisations of the data
- enhancing the data, e.g. by adding in additional information
- integrating the data with other data to create a combined dataset
- applying new analyses and techniques to the data

And thus the cycle starts again at the Conceptualise Stage.

### **Checklist:**

Consider the epistemological, methodological and practical issues of reusing / sharing qualitative data

- There is much discussion in the literature about the epistemological, methodological and practical issues of reusing / sharing qualitative data.
- Issues include:
  - Because of the specificity of research questions and the way qualitative data is obtained (the interrelationship between the researcher and the participants) and analysed (the interaction of the researcher with the data), can it ever be reused by the researcher themselves, let alone shared with other researchers?
  - Secondary analysis of qualitative data requires a different investigative epistemology and investigative practices.
  - Qualitative research projects are not normally 'repeated', nor the data reanalysed to check the accuracy of the work, as is the case with quantitative data.
  - Ethical issues must be considered.
  - Producing sufficient contextual information to make the data 'meaningful' to other researchers is a challenge.

## **GROUP EXERCISE - YOUR DATA MANAGEMENT PLAN (Slide 30)**

### **Slide 30: Group Exercise: Your Data Management Plan**

**Tutor:** Organise the students into the same groups of 3 or 4 as with the Appraise Group Exercise. Ask them to discuss their own data management plans in the light of the information provided in this session - starting to develop a Data Management Plan for their own PhD research was a directed learning task following Session 1. Get them to consider the questions:

- What do you need to add?
- What do you need to change?
- What further questions do you have?

Circulate to respond to questions and ensure participants are focusing on the task as set. Allow 10 minutes for discussion. This is followed by feedback and open discussion of their plans. This part of the exercise to take 5 minutes.

## **DIRECTED LEARNING TASKS (Slide 31)**

### **Slide 31: Directed Learning Tasks**

Session 3 will be covering problems and practical strategies and solutions. So your directed learning tasks are:

- To think of the problems you've experienced with managing your research data
- To think of any good tips or systems you've used for managing your research data

and to share these with the other students in Session 3.

---

<sup>i</sup> DCC, Digital Curation 101 materials, <http://www.dcc.ac.uk/training/train-the-trainer/dc-101-training-materials>

<sup>ii</sup> McGinley M, Sharing medical data, DCC, <http://www.dcc.ac.uk/resources/briefing-papers/legal-watch-papers/sharing-medical-data>

<sup>iii</sup> DCC definition

<sup>iv</sup> OECD Principles and Guidelines for Access to Research Data from Public Funding, 2007, <http://www.oecd.org/dataoecd/9/61/38500813.pdf>

“make relevant primary data and research evidence accessible to others for reasonable periods after the completion of the research: data should normally be preserved and accessible for ten years, but for projects of clinical or major social, environmental or heritage importance, for 20 years or longer” (RCUK 2009, p.8)  
RCUK Policy and Code of Conduct on the Governance of Good Research Conduct, 2009, <http://www.rcuk.ac.uk/documents/reviews/grc/goodresearchconductcode.pdf>

<sup>v</sup> Piwowar, H.A., Day, R.S., and Fridsma, D.B. (2007) Sharing detailed research data is associated with increased citation rate. PLoS ONE 2(3): e308. doi:10.1371/journal.pone.0000308

#### **“Abstract**

**Background** Sharing research data provides benefit to the general scientific community, but the benefit is less obvious for the investigator who makes his or her data available.

**Principal Findings** We examined the citation history of 85 cancer microarray clinical trial publications with respect to the availability of their data. The 48% of trials with publicly available microarray data received 85% of the aggregate citations. Publicly available data was significantly ( $p = 0.006$ ) associated with a 69% increase in citations, independently of journal impact factor, date of publication, and author country of origin using linear regression.

**Significance** This correlation between publicly available data and increased literature impact may further motivate investigators to share their detailed research data.”

<sup>vi</sup> Northumbria University, Regulations for the Degrees of Master of Philosophy (MPhil) & Doctor of Philosophy (PhD), 2011, [http://www.northumbria.ac.uk/static/5007/graduateschool/regs\\_mphilphd11.pdf](http://www.northumbria.ac.uk/static/5007/graduateschool/regs_mphilphd11.pdf)

“The copyright of the submission rests with the student.” (Section 11.2)

“As a condition of enrolment, the student agrees that the University shall own any Intellectual Property (IP) that may result from his/her research activity and in return the student is eligible for a share of the revenue generated under the same procedures as members of academic staff ... In the case where a student is funded wholly or partly by a third party external to the University, the supervisor(s) and the student should clarify intellectual property ownership issues with the third party at the outset of the research project.” (Section 11.3)

<sup>vii</sup> Economic and Social Data Service. The 'data collection' section of the Joint electronic Submissions (Je-S) form, <http://www.esds.ac.uk/aandp/create/esrcfaq.asp>

ESRC research data policy, 2010, [http://www.esrc.ac.uk/images/Research\\_Data\\_Policy\\_2010\\_tcm8-4595.pdf](http://www.esrc.ac.uk/images/Research_Data_Policy_2010_tcm8-4595.pdf)

“Those ESRC grant applicants who plan to generate data are responsible for preparing and submitting data management and sharing plans for their research projects as an integral part of the application. It is expected that an outline data management and sharing plan will include the following points:

- \* an explanation of the existing data sources that will be used by the research project with references;
  - \* an analysis of the gaps identified between the currently available and required data for the research;
  - \* information on the data that will be produced by the research project, including the following:
    - data volume
    - data type, e.g. qualitative or quantitative data
    - data quality, formats, standards documentation and metadata
    - methodologies for data collection
  - \* planned quality assurance and back-up procedures [security/storage];
  - \* plans for management and archiving of collected data;
  - \* expected difficulties in data sharing, along with and causes and possible measures to overcome these difficulties;
  - \* explicit mention of consent, confidentiality, anonymisation and other ethical considerations;
  - \* copyright and intellectual property ownership of the data; and
  - \* responsibilities for data management and curation within research teams at all participating institutions.”
- (ESRC 2010, p.5)

<sup>viii</sup> Northumbria University, Data Protection and Secure Storage of Research Records, 2009  
[http://www.northumbria.ac.uk/static/5007/respdf/data-prot\\_secure\\_storage.pdf](http://www.northumbria.ac.uk/static/5007/respdf/data-prot_secure_storage.pdf)

<sup>ix</sup> “data should normally be preserved and accessible for ten years, but for projects of clinical or major social, environmental or heritage importance, for 20 years or longer” (RCUK 2009, p.8)

RCUK Policy and Code of Conduct on the Governance of Good Research Conduct, 2009,  
<http://www.rcuk.ac.uk/documents/reviews/grc/goodresearchconductcode.pdf>

<sup>x</sup> <http://www.esds.ac.uk/findingData/majorstudies.asp#gov>

<sup>xi</sup> <http://www.esds.ac.uk/findingData/snDescription.asp?sn=4943>

<sup>xii</sup> Creative Commons <http://creativecommons.org/>

<sup>xiii</sup> SURFfoundation, Enhanced publications,  
<http://www.surfoundation.nl/en/themas/openonderzoek/verrijktepublicaties/Pages/Default.aspx>