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

Research data management

Session 1

Introduction to Research Data Management

Tutor Notes

DATUM for Health

www.northumbria.ac.uk/datum

DATUM for Health – Healthy research needs healthy data

Session 1 Introduction to Research Data Management
Notes for Tutors

SESSION DETAILS

Aims and Objectives / Learning Outcomes

By the end of this session participants will have:
- an understanding of research data management needs and issues
- a template to start planning for managing their own research data

Session Content

- Introduction to the overall research data management programme
- Scene setting presentation (with interaction) to motivate students, highlighting the range of data to be managed and demonstrating why research data management is important
- Introduction to a framework for managing research data (the data lifecycle) and a key tool (the research data management plan)
- Practical exercise on developing a data management plan based on a real scenario
- Group discussion about the issues arise in managing research data, particularly for qualitative and health data

Structure

10.00 -10.10	Introduction to the overall programme
10.10 -10.45	Setting the scene - What is research data? Where is your research data? - Why RDM is important? Reasons for managing research data
10.45 -11.00	How do I manage my research data? - the data lifecycle - the research data management plan (DMP)
11.00 -11.15	Refreshment break (alternatively before the previous short session)
11.15 -12.05	Group activity: Developing a sample data management plan based on a real scenario
12.05 -12.20	What issues arise in managing research data, particularly for qualitative and health data?
12.20 -12.30	Introduction to directed learning tasks Any questions?

NB Allow adequate time for DMP and discussion exercise.

Directed Learning Tasks

- Start developing a personal research data management plan, using template provided in preparation for Session 2
- Watch the 3-minute video of Louise Corti, UKDA, Essex University on 'How can researchers ensure that they'll be able to share, archive or re-use sensitive data?'

Handouts

Handouts for the three sessions in this programme come from a range of sources. Materials written by the DATUM Project Team will be provided as PDF / Word electronic files

accessible from the Project's website http://www.northumbria.ac.uk/datum. Handouts from other sources: URLs to access these sources will be given in the Endnotes. These sources are freely available. In the Tutor notes we will indicate if materials need to be provided to the students in hard copy for use during the session.

- RDM Training programme details in full [provided in Word format so the tutor can tailor it for their own programme]
- PPT slides 3 per page handout (provide the students with a hardcopy for use during the session)
- DCC's DMP checklist leaflet – gives key points about each section of DMP only [i]
- Judge Project case example (provide the students with a hardcopy for use during the session)
- Partially completed DMP for Judge project (provide the students with a hard copy to be used to complete the session exercise)
- Fully completed DMP for Judge project (provide the students with a hardcopy for use during the session))
- DATUM DMP template with guidance notes [a Word file provided electronically format so students can download a copy and use it throughout their PhD study]
- Resources list
- UKDA *Managing and Sharing Data: a best practice guide for researchers* booklet [ii]
- Evaluation form for session [provided in Word format so the tutor can tailor it for their own programme / session] (provide the students with a hardcopy for use at the end of the session)

NOTES TO ACCOMPANY POWERPOINT SLIDES

INTRODUCTION TO THE OVERALL PROGRAMME AND SESSION (Slides 0 - 3)

This part of the process includes a general explanation about the session, the entire training programme, its purpose (aims and objectives), the presenter(s) and that the sessions build on each other, though can standalone.

Slide 1: Programme Introduction

Introduce the overall programme i.e. three sessions about managing research data (RDM) focusing on the health studies discipline (covering generic and discipline-specific issues); qualitative, unstructured data; and the whole of what we call the data management lifecycle. Target audience is PGR (doctoral) students / early career researchers; suitable for students at any stage. (Refer to programme handout)

Aims to provide postgraduate research students with the knowledge to manage their research data at every stage in the data lifecycle, from its creation to its final storage or destruction. You will learn how to use your data more effectively and efficiently, how to store and destroy it securely, and how to make it available to a wider audience to increase its use, value and impact.

Session 1 will introduce RDM, Session 2 will look at the data curation lifecycle and Session 3 will look at RDM problems and practical solutions and strategies. Follow-up activities / directed learning tasks after each session will help you practically with your own RDM and help prepare you for the next session so that you can get the most out of them. Various handouts; all of the materials will be made available on the VLE (or equivalent).

Since RDM has many elements and there is limited time, tutor may wish to make a note of any questions that come from discussion in each session and the answers. If they can't immediately answer (e.g. need to check on policy / follow up) then note those too and get back with answers. E.g. build up a list of FAQs that can put be onto the VLE.

Slide 2: Overview

This session introduces the overall programme, then poses a series of questions and helps you to answer them for your own research now and in the future. (Questions are as on the slide). At the end I will give some directed learning tasks to complete before the next session.

Format – aiming for this to be an interactive session; discussion, activities, as well as information sharing presentation elements. Feel free to ask any questions. Will either answer immediately, flag that they will be answered later or capture for follow-up outside the sessions.

Slide 3: Learning Outcomes

By the end of this session you should:
- understand why you need to manage research data
- be aware of the issues surrounding research data management in the health discipline
- be able to complete an initial plan for managing your own research data

WHAT IS RESEARCH DATA? (Slides 4 - 5)

Note: Student handout has 2 icons only (slide 4). Response has a build of many icons (slide 5).

Show Slide 4

What is research data - an obvious question but an important one. It's vital to think about exactly what our research is / is going to be **at the start** to ensure we consider **all** the data we need to manage and we **don't worry** about anything we don't need to worry about. [Note – not thinking here about published information –don't need to protect literature, only the list of references in a database (say).]

TASK: O**n your own / in pairs** in the context of your PhD:
  1. what data are you using (i.e. that you've got from elsewhere)?
  2. what data are you creating?

Give them 2 mins (exactly) to write down two lists then ask for examples and **capture to a whiteboard / flipchart (in 2 lists).**

Examples

1. Data using	2. Data creating
Articles, conference papers etc Books website materials videos, podcasts previous data sets	meeting notes interview / questionnaire data – text; figures researcher diary notes communications (in emails, letters, SMS) blog posts, tweets, test results (machines / systems) measurements scans, x-rays audio recordings, videos, podcasts ..... thoughts in your head analysed data, summaries

NB ensure participants don't focus on the METHODS of collecting data (i.e. interview, survey, focus group) but on the types of data being collected

Move to Slide 5

At this point highlight that data is very varied in terms of types and formats. Two main sources – what you / any co-researchers create and what you use that already exists. It may be that some of the things on list #1 don't need to be managed or at least not in the same way as list #2. Important to be aware of this because it impacts / influences how you decided to manage your data.

WHERE IS YOUR RESEARCH DATA? (Slides 6 - 7)

Note: Student handout has 2 icons only (slide 6). Response has a build of many icons (slide 7).

Show Slide 6

Knowing what data you need to manage you also need to think about where that data is. Thinking about your own research data again – where is it?

TASK: Ask the whole group for responses and capture on flipchart, prompting for any other locations.

Examples:
- Media
    - paper diaries; notebooks, files, articles; traces from a machine
    - tapes, other artefacts (objects)
    - digital - university drive, PC hard drive; laptop hard drive; removable storage device (CD, USB / flash pen, external hard drive); camera, mobile phone, web / the cloud (e.g. Google docs; facebook, blog)
- Different systems:
    - emails, SMS, blog, microblog (tweets), social networking sites (facebook, linkediin, myspace, NIng)
    - spreadsheet, database, Word, NVivo, project management software, the VLE, pdf files
    - an aggregator (eg Google reader, netvibes, pageflakes, iGoogle)
    - if collaborating – wiki (PBworks, Wetpaint, Wikia)
    - social bookmarking (delicious); references / bibliographies (Endnote database; CiteULike)
    - photos (flickr, picasa)
    - videos (YouTube, vimeo)
    - presentations (slideshare)
    - virtual worlds (second life)
    - thoughts in your head....
- Location – in your desk drawer / office, on PCs / laptops / servers – at home / at the university / elsewhere if travelling to collect data / go to a conference etc.

Move to Slide 7

Summary: bring out of these questions / discussions the key characteristics about research data:
- source – yours, others' (published)
- kinds of data – raw, coded, transcribed, analysed, summarised, synthesised, published
- forms of creation - text; sound; image
- storage formats - paper; analogue; digital (different formats); not captured (memory)
- locations – university; home; mobile (travelling to / from home / work; in the field collecting data); in your head!

A lot to manage / protect and not in one format, or in one place. Whose responsibility? Some you have sole responsibility to manage, others you are relying on others (e.g. university; cloud storage provider – Google) to manage).

WHY should you be bothered about managing your research?

WHY IS RESEARCH DATA MANAGEMENT IMPORTANT? (Slides 8 - 18)

This part of the session is about scene setting to motivate students and demonstrate why RDM is important.

Slide 8: Reasons for Managing Research Data

Note: Student handout has blank list 1-3 because asking them to write down 3 reasons to manage research data

TASK: Ask them to write down 3 reasons why RDM is important based on their experience (1 minute).

Ask them for a few reasons in open discussion. Then share the 7 key reasons on the following slides. Bring out risk under all the reasons. [After slide 18 ask if anyone had written any other different reasons for managing research data.]

Slide 9: Reason 1. A Requirement

Funder requires it (e.g. MRC); university requires it (deposit thesis, might be required / able to deposit data). No choice, about compliance.

Examples:
- **MRC** policy on data sharing and preservation
  "Our policy builds on the central principles of the Organisation for Economic Co-operation and Development (OECD) in its report "Promoting Access to Public Research Data for Scientific, Economic and Social Development". These are that publicly-funded research data are a public good, produced in the public interest, and that they should be openly available to the maximum extent possible."
  "Our data sharing and preservation policy applies to all MRC-funded research. It does not prescribe when or how researchers should preserve and share data, but requires them to make clear provision for doing so when planning and executing their research. **"**A data sharing and preservation strategy should be provided at the proposal stage, summarising the type of data to be generated, foreseeable research uses, and plans for preparing and documenting data for preservation for sharing. Guidance is available online with the MRC data policy." [iii]
- **Wellcome Trust**: Data sharing plans should address seven key questions as clearly and concisely as possible, as noted in the Trust's QandA document on the data policy. [iv]

Slide 10: Reason 2. To Work Effectively and Efficiently

Knowing where data is (i.e. can retrieve it); knowing what it means, how collected and analysed (i.e. can understand it); being able to use it when you're writing up / writing articles – now and in 3-5 years time. This is partly about protecting your PhD / research (demonstrating you did the work) and ensuring you can write up your thesis; also in case of any Freedom of Information request

Slide 11:- Reason 3. To Protect It

To keep your data save, to safeguard it. Covers many things: avoiding loss / theft / accidental deletion / corruption of files – virus attack etc / obsolescence (i.e. can't read format).

Some things are **in your control** and others aren't exactly (e.g. obsolescence, virus attack, theft). However, you **can mitigate the risk** and have **mechanisms / actions** for managing the **risks depending on their likelihood and impact**. You can treat, tolerate or terminate them; not easy to transfer RDM risk.

TASK: Ask if they think there is much of a risk of any of these things happening and resulting in data loss? Has anyone suffered data loss?

Give some real examples of risks that have materialised, for example:

- dropped external hard drive with a masters dissertation on it – no exact backup; had to recover by piecing bits together from various places
- PhD thesis chapter 12k words – accidental key strokes brought up a blank document, on exit at the save prompt, saved with same file name and 'lost' the 4 days of fine tuning
- PhD thesis work on laptop in boot of car - stolen
- others in the press.

Slides 12 - 14: Reason 4. For Use and / or Re-use

Show Slide 12

Note the distinction between use (i.e. read / analyse / raising new questions) and re-use (i.e. further analysis to generate new data)

By you and possibly with others (e.g. supervisor, project team members) for writing articles, new bids, new research. Data has value as an enabler of new opportunities, and re-using data makes good economic sense.

Move to Slide 13

Example of possible re-use of data with different visualisation and how data can be 'mashed-up'. Provide an example. For instance (slides 13-14) technology / systems constantly change and now about to combine data in interesting and / or different ways. Dr John Snow was able to trace the sources of the spread of cholera infections in London in the 1840s by plotting the deaths on a map. The source was the pump at the corner of Cambridge and Broad Street. Show part of the original map created by Dr Snow - example of such a map [v]

Move to Slide 14

And this is a mash up of the address data and Google Earth Maps. The red pins indicate water fountains. Show this slide from this presentation [vi]

Stress that if you are going to use / re-use data then you need to know where it came from, what it represents and doesn't represent– i.e. what it is / isn't. This requires information about the context so that its meaning is correctly understood, which also falls in the realms of the next reason for RDM.

Slide 15: Reason 5. To Share it

This is about standing on the shoulders of those who've gone before and YOU sharing something that others can build upon. In the current economic climate where RCUKs / other

funders are wanting the 'best value for investment' they are now explicitly asking the question when assessing grant applications 'why do you need to collect new data? It may already exist.' You need to demonstrate there is a gap and justify your collection of new data in bids.

Provide an example of the benefit of sharing data. For instance, Streptokinase is one of a number of drugs that break up blood clots. Use of such drugs at early stages of a heart attack is now standard treatment. It became standard treatment in the 1990s. Research was published on streptokinase treatment from the late 1950s. If meta-analyses of this data had been conducted using the data as it was published, then the benefits of the drug would have been realised over 20 years earlier and many lives would have been saved. [vii]

But if others are going to use your data / if you're going to use other people's data then they / you need to know where it came from, what it represents / doesn't, what it is / isn't. Again this requires contextual information about the meaning of the data. This is as true for quantitative data as it is for qualitative data: numbers alone aren't always meaningful – feet, inches, miles, degrees C or what?. Also provenance – who created and when.

Show short video clips

Incremental Project video – Stephen Gray, 'Why should researchers share their data?' [viii]
Incremental Project video - Fanar Haddad, 'Tips for using YouTube in research': the perks and perils of using YouTube as a data source to reinforce the points made about use, re-use and sharing. [ix]

Slide 16: Reason 6. For Preservation

If we are going to be able to use / re-use data for these reasons / other reasons then we need to think about preservation as early as possible (especially in the digital environment). Otherwise, as Jeff Rothenberg pointed out 15+ years ago... "digital information lasts forever - or five years, whichever comes first." [x]

Provide examples of lost data. For instance:
- BBC Domesday survey - project to capture information about life in the UK in 1986, 900 years after the Domesday Book, with contributions from thousands of schoolchildren and researchers. Stored on 12" laser discs, with a laserdisc player and a BBC microcomputer to access the information. The original Domesday Book can still be read in The National Archives, but the information on the discs could have been lost. A working copy was found (though fragile) but it took a lot of technical work to complete a 'just in time' rescue. [xi]
- NASA Apollo 9 / 11 space tapes: 200,000 tapes that were going to support further Apollo missions, Apollo-Soyuz and Skylab were reused having been preserved for a time until NASA thought they were no longer needed. Only later did they want some of the data but had to use broadcast images instead. [xii] [Note: demonstrates the difficulty of knowing when data is no longer needed and can be deleted; determining the value now and in the future is difficult.]

Highlight that data can be deposited in a specialist repository which will ensure they are preserved. Point is that unless you think about preservation before you decide to share in a repository there may be no data to share!
Examples of repositories [xiii]:
- UK Data Archive (UKDA) who manage the Economic and Social Data Service (ESDS) which contains health data sets

- Arts and Humanities Data Service (AHDS)
- Office of Population Research (OPR), Princeton University
- Inter-University Consortium for Social and Political Research (ICSPR), University of Michigan

Slides 17 - 18: Reason 7. Because it is good research practice

Show Slide 17

RDM is just part of good research practice. As are the initial literature review, project management; consent; methodological design etc. In fact the Digital Curation Centre's (DCC) mantra is "because good research needs good data" and the DATUM mantra is "healthy research needs healthy data."

RDM is about evidencing your research, about trust / transparency / accountability and the contract / undertaking to the research participants. This is linked to your research reputation and, by association, the reputation of those you work with (supervisor, research group) and your institution / affiliation.

Two examples of what can go wrong:

- Sir Cyril Burt (1883-1971). Heritability of intelligence as measured in IQ tests with twins "the Burt Affair" An example of false data, fraud. [xiv](See for example Wikipedia).

- Andrew Wakefield. Link between the MMR vaccine and bowel disease and autism. Published research in The Lancet in 1998 purporting to show a link which led to many parents refusing to vaccinate their children and a rise in incidence of measles. An Editorial in The BMJ later accused him of fraud. Following the longest ever General Medical Council disciplinary case he was struck off the medical register (breach of fundamental research principles including misleading reporting of work). The Lancet retracted the original article.[xv] (This example is about having the evidence to support conclusions; transparency / accountability).[xvi]

Message is that if all researchers were required to produce a DMP and these were placed in the public domain / audited in some way, then it would be harder for such things to occur.

Move to Slide 18

DATUM for Health project staff views about the importance of RDM.

Close this section by asking if anyone has any other examples that illustrate why RDM is part of good research practice. Any other reasons you have why we should manage our research data?

Summary: good RDM ensures compliance, minimises / manages risks, offers opportunities. Doing RDM first for yourself and potentially for others. Anyone still not convinced they need to manage their data?

HOW DO I MANAGE MY RESEARCH DATA? (Slides 19 - 21)

Slide 19: The Data Lifecycle

How should I manage my research data?

The first thing to think about is the 'big picture' – the data lifecycle and the data management lifecycle. There are lots of variations on the data lifecycle and the data *management* lifecycle. Some are very detailed (e.g. the DCC's) and really meant for *data managers / librarians* rather than for *data creators and users*.

This slide shows a simplified version the aim of which is to highlight the key stages and for it to be easy to remember so that you are always of aware of whether or not you're adequately managing your research data.

Slide 20: How do I manage my research data – the data lifecycle

The UKDA's data lifecycle [xvii].
Refer to their *Managing and sharing data. Best practice for researchers* booklet. Give hard copy or recommend downloading from their website. Recommend looking at the UKDA website. Lots of FAQs and straightforward advice.

Slide 21: How do I manage my research data – the data lifecycle

Back to our simplified version. Managing research data starts with planning [red highlight on slide] and the data management plan (DMP). We're going to look at a template for creating a DMP and you're going to have a go at doing one in groups.

REFRESHMENT BREAK

Either break for refreshments at this point or after introducing the DMP (Slide 24) and before the DMP activity. During break set out groups and tables so they sit down in a different place ready for this activity.

THE DATA MANAGEMENT PLAN (Slides 22 - 24)

Slide 22: The Data Management Plan

What is a DMP?
"a plan for depositing data into a publicly accessible data repository" (Ross Harvey, 2010, p59[xviii]), who later refers to funding bodies requiring "plans for data sharing, curation and preservation" (Harvey p84) citing Wellcome Trust expecting researchers to "plan at the proposal stage how they will manage and share their data".

Curation is data management, covering creation, use, transfer to archive / repository, preservation, future use. It emphasizes the maintenance of data and adding value to the data for current and future use.

A DMP is the place to document what you need to and are going to do to manage your research data throughout its life. So, it needs to contain details about how you will create

and capture, manage (organise, store), share (access, use / re-use), protect and preserve your data:

"a data management plan will help you to properly manage your data for own use, not only to meet a funder requirement or enable data sharing in the future" MIT Libraries [xix]

How do I develop a DMP? Use one of the templates available.

Slide 23: DCC's DMP Online

The Digital Curation Centre (DCC) provides an online data management tool DMP Online [xx]. [Give out the DCC's DMP Checklist Leaflet]

DMP Online: "In addition to the questions included in the DCC's influential Checklist for a Data Management Plan, it also contains useful guidance on how to prepare a data management plan and carry it through to execution. … The tool draws upon the DCC's analysis of funders' data requirements to help project teams create up to three iterations of a data management plan; a **'minimal'** version for use at the grant application stage, a **'core'** version to be developed during the project itself, and towards the end of the project a **'full'** version that addresses issues of longer-term access and preservation. We have mapped the major UK research funders' data-related policies to the DCC Checklist, the idea being that by answering the DCC questions you will de facto meet your chosen funder's requirements."

Main sections are:
1 Introduction and context
2 Legal and ethical issues
3 Access, data sharing and re-use
4 Data standards and capture methods
5 Short-term storage and data management
6 Deposit and long-term preservation

Slide 24: DATUM's DMP Template

The DATUM for Health DMP template - a much simpler DMP template more focused on PhD research. Based on the DCC one (with their permission / agreement) so it has the same 6 sections (sometimes with slightly different names) but a combination of closed (Yes / No / Don't Know) questions and open ones; also has examples of the type of responses / information you need to include.

Stress that the DMP is a living document – not something that is completed at the start of the research and then never revised. Different versions should be captured.

DEVELOPING A DMP (Slides 25 - 26)

This is the main activity of the session. It requires time – to read the case example project outline, consider the elements of the DMP and discuss how to complete the various sections. Best done in groups of 3 or 4, deliberately arranged to have a mix of years (i.e. stages of the PhD) in each group. Circulate to respond to questions and ensure participants are focusing on the task as set. Allow time for discussion. [Replace with a different project, more familiar to the tutor to ensure questions / issues can be discussed in depth during the feedback.]

Show Slide 25

The Judge Project. – a previous project about health information conducted by Northumbria University research staff. Provide the handout containing a brief project outline 'Judge: Web sites for Health' and the partially completed DATUM DMP template for this project.

***TASK:* T**o develop a DMP for this project using the DATUM template:
- *Need to imagine they are the researcher (Sue Childs) at the start of the project and that* the project is starting now (not in 2003). The reason for this is that there have been changes in terms of RDM etc requirements since the project was undertaken (e.g. need for formal contracts between parties / IPR / University's retention policy). Not unusual, common practice - requirements, expectations etc change over time. Need to keep abreast of changes and review the DMP. This is an important issue, because what you do now might not be what you would have done 8 years ago though it was perfectly acceptable / legitimate at the time. However this may mean that you might not be able to do things with the data in the future (this will be illustrated further in Session 3). Access might be opened up or indeed potentially closed down.
- Some of the fields have been populated with the basic project information to save time (Section 1) and a fully completed DMP is available for students to check the 'right' answers during the open discussion
- Try to complete all of the sections – so if you get stuck or aren't sure, either ask or move on and come back to the things that are more difficult to complete.
- Give 30 minutes for the task.

Move to Slide 26

Feedback / discussion about the DMP for the Judge Project in terms of content and process. (*Note: the order can be what seems best depending on observation of the discussion that took place).*

1. Content
- what did they have for.....(a few Qs to select)
- any difficult areas?
2. Process
- how easy / difficult was it to do? Why?
- would be any easier for their own PhD research? Why / why not? (because they know it so well or it's just not easy regardless of the project)
- would you say you had already addressed all of the points in this DMP for your research?
  - If so why? (experience, someone suggested you did, you had to i.e. it was a requirement – who made it a requirement?)
  - If not then which questions / aspect haven't you considered and / or need to consider further?

Provide a copy of the DMP fully completed by the researcher and raise / cover any of the queries. Get them to compare their version and discuss any differences.

Lead into final topic about the issues of RDM.

RESEARCH DATA MANAGEMENT ISSUES (Slides 27 - 28)

Show Slide 27

We've introduced the concept of RDM in this session; thought about the wide range of data you might have, where it is, what you need to do to manage it; and why it's important to manage it. But does this raise any issues in your mind?

Ask them to spend 3-5 min minutes in the same groups or pairs identifying any issues / challenges that the DMP activity (or other thoughts / views) raises about managing research data.

Share and discuss the issues which might include:
- policy / requirements
- research funders / legal requirements (e.g. Data Protection Act) – awareness; sources of advice
- ethics
- epistemological stance; research methodology
- consent to collect, to use / re-use / share
- sharing – what, who?
- what to keep, thinking about the future use (for you / others)
- security
- creation, organisation, storage, retention (how long), protection / backup, preservation
- tools for RDM

Move to Slide 28

How to manage research data starts with the DMP. Next two sessions will share more information, tips, guidance on how you can MRD.

DIRECTED LEARNING (Slide 29)

Purpose of the tasks is (a) to reinforce learning and (b) prepare for the next session of the programme.

1. Start to develop a Data Management Plan for your own PhD research / current research project using template provided in preparation for Session 2. It will be valuable irrespective of the stage of the PhD. Bring to the next session - chance to raise any issues / ask any questions. Use the DATUM for Health template provided (based on the DCC's minimal plan). Provide the e-version, and a hard copy version if required.
2. Watch the 3-minute video on 'How can researchers ensure that they'll be able to share, archive or re-use sensitive data?' Louise Corti, UKDA, Essex University (from the Incremental Project) [xxi]

---

[i] DCC, DMP Checklist Leaflet, http://www.dcc.ac.uk/webfm_send/371

[ii] Van den Eynden V et al (2011) Managing and sharing data: Best practice guide for researchers, 3rd edn, UKDA, http://www.data-archive.ac.uk/media/2894/managingsharing.pdf

iii MRC policy on data sharing and preservation,
http://www.mrc.ac.uk/Ourresearch/Ethicsresearchguidance/Datasharinginitiative/Policy/index.htm

iv Wellcome TRUST DM Policy, http://www.wellcome.ac.uk/About-us/Policy/Spotlight-issues/Data-sharing/Guidance-for-researchers

v Crosier S, John Snow: The London Cholera Epidemic of 1854, Center for Spatially Integrated Social Science, University of California, http://www.csiss.org/classics/content/8

vi Dr John Snow's location of cholera infections in London (1840s) by address – a Google version of the 'mashup'. Slide 2 from: Building blocks to open Gov. Ed Parsons, Geospatial Technologist, Google Europe https://docs.google.com/present/view?id=0AeQTdF5K85B1ZGNjeGRncHhfMTI4aG5qN3Y1YzQ (Presentation to Government 2.0 Taskforce, 2 Oct 2009, http://gov2.net.au/index.html)

vii Mulrow CD (1994) Systematic Reviews: Rationale for systematic reviews, *BMJ*, Vol 309 p 597. http://www.bmj.com/content/309/6954/597.full

viii University of Glasgow, Data management videos, created by the Incremental Project, http://www.gla.ac.uk/services/datamanagement/training/videos/

ix University of Cambridge, Data Management Training Resources, created by the Incremental project and maintained by DSpace@Cambridge http://www.lib.cam.ac.uk/dataman/training.html#Interviews

x Rothenberg J (1995) Ensuring the longevity of digital documents. *Scientific American*, 272 (1), Jan, pp.42-47.
xi Darlington, J. Finney, A. and Pearce, A. (2003). Domesday Redux: the rescue of the BBC Domesday project videodiscs. *Ariadne*, 36, July 2003. www.ariadne.ac.uk/issue36/tna/
xii Lost for good http://www.collectspace.com/news/news-071709a.html
xiii UKDA (UK Data Archive) http://www.data-archive.ac.uk/ who manage the ESDS (Economic and Social Data Service) http://www.esds.ac.uk/
AHDS (Arts and Humanities Data Service) http://www.ahds.ac.uk/
OPR (Office of Population Research), Princeton University http://opr.princeton.edu/archive/
ICSPR (Inter-University Consortium for Social and Political Research), University of Michigan http://www.icpsr.umich.edu/icpsrweb/ICPSR/

xiv Cyril Burt, Wikipedia, http://en.wikipedia.org/wiki/Cyril_Burt

xv RETRACTED: Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children. A J Wakefield, S H Murch, A Anthony, J Linnell, D M Casson, M Malik, M Berelowitz, A P Dhillon, M A Thomson, P Harvey, A Valentine, S E Davies, J A Walker-Smith *The Lancet*, Vol 351 No 9103 28 Feb 1998 p 637-41
xvi See http://www.guardian.co.uk/science/blog/2011/jan/12/andrew-wakefield-fraud-mmr-autism, Brian Deer investigative journalist
xvii UKDA Research Data Lifecycle http://www.data-archive.ac.uk/create-manage/life-cycle
Van den Eynden V et al (2011) Managing and sharing data: Best practice guide for researchers, 3rd edn, UKDA, http://www.data-archive.ac.uk/media/2894/managingsharing.pdf
UKDA 'Create and manage data' webpages http://www.data-archive.ac.uk/create-manage

xviii Harvey R, 2010, *Digital curation: a how-to-do-it manual*. Facet.

xix MIT Libraries, Data management plans, http://libraries.mit.edu/guides/subjects/data-management/plans.html

[xx] DCC, DMP Online http://www.dcc.ac.uk/dmponline

[xxi] University of Cambridge, Data Management Training Resources, created by the Incremental project and maintained by DSpace@Cambridgehttp://www.lib.cam.ac.uk/dataman/training.html#Interviews