
Data Management Basics

Scott Summers
UK Data Service
Research Data Management Team

Webinar
2nd November 2017

UK Data Service



Overview of this session

- UK Data Service
- Managing your data – background, why and how
 - Consent, anonymisation, documentation, etc.
 - Security, backups, encryption, etc.
- More resources available (this webinar is *highlights* only)
- Your questions



Data Management at the UK Data Service

- support and training for data creators with accessing, managing, and using data
- one-stop-shop for social science data

<https://discover.ukdataservice.ac.uk/>

- more webinars available

<https://www.ukdataservice.ac.uk/news-and-events/webinars>

The screenshot shows the UK Data Service website homepage. At the top, there is a navigation bar with links: About us, Get data, Use data, Manage data, Deposit data, and News and Events. Below this is a large header section with the text 'Welcome to the UK Data Service' and 'Your resource for quality social research data'. To the right of the header is a large graphic with colorful bars. Below the header, there are three main sections: 'LATEST TWEETS' on the left, 'LATEST NEWS' in the middle, and 'OUR DATA COMMUNITY' on the right. The 'LATEST TWEETS' section shows three tweets from UKDataService, @Barnard17, and @Censusacuk. The 'LATEST NEWS' section has a sub-header 'Call for papers: Opinions and Lifestyle Survey user meeting' and a list of news items. The 'OUR DATA COMMUNITY' section has a sub-header 'The UK Data Service is at the core of a network of trust that includes data owners, producers, funders and users.' and a list of 'Who can most benefit from the data we hold?'. On the far right, there is a sidebar with a search bar, login/register links, and a 'QUICK ACCESS TO' section with links to Key data, Census Support, Information for new users, and Frequently asked questions.

UK Data Service



Background

- Data sharing is fast becoming a new paradigm in research across all disciplines, providing benefits to individual researchers, institutions, funders and more
- Good research data management habits are essential to creating data that are suitable for sharing and reuse
- Many funders and academic publishers now specify requirements for data handling, including the formulation of a data management plan



Why is it important to manage research data well?

- Data creation in research is often expensive
- Data is the cornerstone of research
- Good quality data leads to good quality research
- Data underpins published findings
- Enables compliance with ethical codes, data protection laws, journal requirements and funder policies
- To protect data from loss, destruction and potential exposure



Practical steps researchers can take

- Write a data management or a sharing plan
- Make sure data are shareable and can be understood:
 - Obtain consent to share
 - Do not disclose identities without consent
 - Use open and standard formats
 - Provide context and documentation
 - Protect your data at all stages



ESRC data management plan

Assessment of existing data

Information on new data

Quality assurance of data

Backup and security of data

Difficulties in data sharing and measures to overcome these

Consent, anonymisation, re-use strategies

Copyright / Intellectual Property Ownership

Responsibilities

Management and curation

[ESRC DMP guidance](#)

UK Data Service



Multiple tools for protecting participants

1. Seek **informed consent**, also for data sharing and long-term preservation and curation
2. **Protect identities** e.g. anonymisation, and (or) not collecting personal data for admin
3. **Regulate access** where needed (all or part of data) e.g. by group, use or time period

Consent for sharing – one more small step

- Engagement in the **research process**
 - What activities are involved in participating in the project?
- **Dissemination** in presentations, publications, the web
 - Consent for use of quotes for articles and video publicity
- Data **sharing** and archiving
 - Consider future uses of data

Consent is *always* dependent on the research context – special cases of covert research and verbal consent



In practice: wording in consent forms / information sheets

We expect to use your contributed information in various outputs, including a report and content for a website. Extracts of interviews and some photographs may both be used. We will get your permission before using a quote from you or a photograph of you. After the project has ended, we intend to archive the interviews at Then the interview data can be disseminated for reuse by other researchers, for research and learning purposes.

The interviews will be archived at and disseminated so other researchers can reuse this information for research and learning purposes:

- ☐ I agree for the audio recording of my interview to be archived and disseminated for reuse
- ☐ I agree for the transcript of my interview to be archived and disseminated for reuse
- ☐ I agree for any photographs of me taken during interview to be archived and disseminated for reuse

In practice: wording in consent forms / information sheets

Any personal information that could identify you will be removed or changed before files are shared with other researchers or results are made public.

We ask you to consider the following points before agreeing to participate.

- Your contribution to the research will take the form of a focus group participant. This will be digitally video recorded and transcribed.
- Your name and any information which may directly or indirectly identify you will be altered to protect your anonymity.
- Any recordings of the discussions will be kept securely, and only authorised to other researchers on the condition they preserve your anonymity.
- The transcriptions (*excluding* names and other identifying details) will be retained by the researcher and analysed as part of the study. They will also be deposited with the UK Data Archive which has strict regulations about accessing data for research and protecting participant confidentiality.



Anonymising quantitative data - tips

- remove direct identifiers
e.g. names, address, institution and photos
- reduce the precision / detail of a variable through aggregation
e.g. birth year instead of date of birth; occupational categories rather than job; and, area rather than village
- generalise meaning of detailed text variable
e.g. occupational expertise
- restrict upper lower ranges of a variable to hide outliers
e.g. income and age
- combining variables
e.g. creating non-disclosive rural / urban variable from place variables



Anonymising qualitative data

- plan or apply editing at time of transcription
except: longitudinal studies - anonymise when data collection complete (linkages)
- avoid blanking out; use pseudonyms or replacements
- avoid over-anonymising – removing / aggregating information in text can distort data or make it misleading
- consistency within research team and throughout project
- Identify replacements, e.g. with [brackets]
- keep an anonymisation log of all replacements, aggregations or removals made and keep it *separate* from anonymised data files



Audio-visual data

Digital manipulation of audio and image files can remove personal identifiers

e.g. voice alteration and image blurring (e.g. of faces)

Labour intensive, expensive, may damage research potential of data

Better alternatives:

- obtain consent to use and share data unaltered for research purposes
- avoid mentioning disclosing information during audio recordings

In practice: example anonymisation

Ex 1. Health and Social Consequences of the Foot and Mouth Disease Epidemic in North Cumbria, 2001-2003 (study 5407 in UK Data Archive collection) by M. Mort, Lancaster University, Institute for Health Research.

Date of Interview: 21/02/02

Interview with Lucas Roberts, DEFRA field officer

Date of birth: 2 May 1965

Gender: Male

Occupation: Frontline worker

Location: Plumpton, North Cumbria

Comment [v1]: Replace: Ken

Comment [v2]: delete

Comment [v3]: delete

Lucas was living at home with his parents, "but I'm hoping to move out soon" so we met at his parents' small neat house. We sat in a very comfortable sitting room with an open fire and Lucas made me coffee and offered shortbread. Although at first Lucas seemed a little nervous, quick to speech and very watchful he seemed to relax as we spoke and to forget about the tape.

Comment [v4]: Replace: Ken

Comment [v5]: Replace: Ken

Comment [v6]: Replace: Ken

I will just start by asking you to tell me a little bit about yourself and your background.

Well it is an agricultural background. I grew up on the farm where my brother is now. After I left school I did work on the farm but went to college and did exams, did land use recreation, sort of countryside/ environmental management course. So I obviously left agriculture, did the course and came back [to the farm] at weekends.

Managing access to data

Open

- available for download / online access under open licence without any registration

Safeguarded

- available for download / online access to logged-in users who have registered and agreed to an End User Licence (*e.g. not identify any potentially identifiable individuals*)
- special agreements (depositor permission; approved researcher)
- embargo for fixed time period

Controlled

- available for remote or safe room access to authorised and authenticated users whose research proposal has been vetted and who have received training

In practice: data with access conditions

Health and Social Consequences of the Foot and Mouth Disease Epidemic in North Cumbria, 2001-2003 (study 5407 in UK Data Archive collection) by M. Mort, Lancaster University, Institute for Health Research.

- Interviews (audio and transcript) and written diaries with 54 people
- 40 interview and diary transcripts are archived and available for re-use by registered users ([Safeguarded](#))
- 3 interviews and 5 diaries were embargoed until 2015 ([Safeguarded – Embargoed](#))
- Audio files archived and only available by permission from researchers ([Safeguarded – Special Agreement](#))

discover.ukdataservice.ac.uk/catalogue/?sn=5407

doc.ukdataservice.ac.uk/doc/5407/mrdoc/pdf/q5407userguide.pdf



Documenting your data

- Enables you to understand data when you return to it!
- Sufficient information for future researchers to understand and use the data
- If using your data for the first time, what would a new user need to know to make sense of it?
- The UK Data Archive uses data documentation to:
 - supplement a data collection with documents such as a user guide(s) and data listing
 - ensure accurate processing and archiving
 - create a catalogue record for a published data collection



Include as documentation

- Data collection methodology and processes: sampling, sampling size, fieldwork protocol and interviewer instructions
- Information sheet / consent form
- Questionnaire, showcards and question lists
- Transcripts: header with context information: date and place interview, interviewee name, etc.
- Data list: overview of key information about each interview, as 'at-a-glance' summary of the data collection
- Links to reports and publications



Data-level documentation: variable names

- All structured, tabular data should have cases / records and variables adequately documented with names, labels and descriptions
- Variable names might include:
 - question number system related to questions in a survey / questionnaire e.g. Q1a, Q1b, Q2, Q3a
 - numerical order system e.g. V1, V2, V3
 - meaningful abbreviations or combinations of abbreviations referring to meaning of the variable
e.g. 'oz%=percentage ozone', 'GOR=Government Office Region', 'moocc=mother occupation', 'faocc=father occupation'
 - for interoperability across platforms - variable names should be max 8 characters and without spaces



Data-level documentation: variable labels

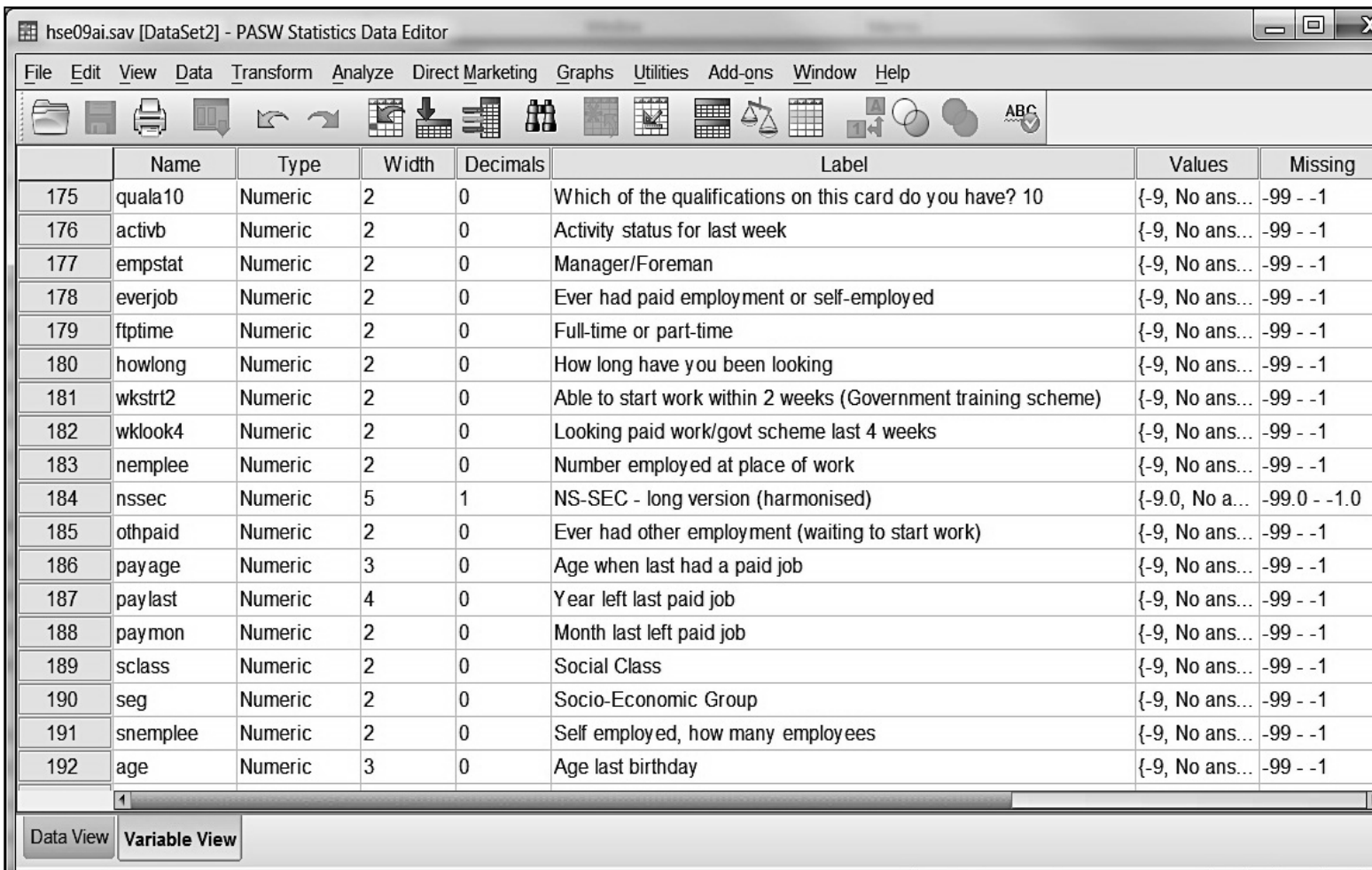
- Similar principles for variable labels:
 - be brief, maximum 80 characters
 - include unit of measurement where applicable
 - reference the question number of a survey or questionnaire

e.g. variable 'q11hexw' with label 'Q11: hours spent taking physical exercise in a typical week' - the label gives the unit of measurement and a reference to the question number (Q11b)
 - Codes of, and reasons for, missing data
 - avoid blanks, system-missing or '0' values

e.g. '99=not recorded', '98=not provided (no answer)', '97=not applicable', '96=not known', '95=error'
 - Coding or classification schemes used, with a bibliographic ref
- e.g. Standard Occupational Classification 2000; ISO 3166 alpha-2 country codes*



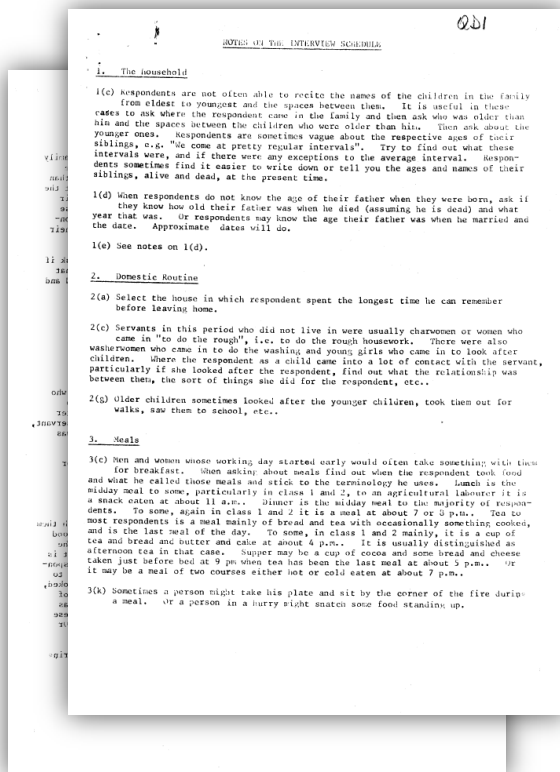
Embedded data-level metadata in SPSS file



	Name	Type	Width	Decimals	Label	Values	Missing
175	quala10	Numeric	2	0	Which of the qualifications on this card do you have? 10	{-9, No ans...	-99 - -1
176	activb	Numeric	2	0	Activity status for last week	{-9, No ans...	-99 - -1
177	empstat	Numeric	2	0	Manager/Foreman	{-9, No ans...	-99 - -1
178	everjob	Numeric	2	0	Ever had paid employment or self-employed	{-9, No ans...	-99 - -1
179	ftptime	Numeric	2	0	Full-time or part-time	{-9, No ans...	-99 - -1
180	howlong	Numeric	2	0	How long have you been looking	{-9, No ans...	-99 - -1
181	wkstrt2	Numeric	2	0	Able to start work within 2 weeks (Government training scheme)	{-9, No ans...	-99 - -1
182	wklook4	Numeric	2	0	Looking paid work/govt scheme last 4 weeks	{-9, No ans...	-99 - -1
183	nemplee	Numeric	2	0	Number employed at place of work	{-9, No ans...	-99 - -1
184	nssec	Numeric	5	1	NS-SEC - long version (harmonised)	{-9.0, No a...	-99.0 - -1.0
185	othpaid	Numeric	2	0	Ever had other employment (waiting to start work)	{-9, No ans...	-99 - -1
186	payage	Numeric	3	0	Age when last had a paid job	{-9, No ans...	-99 - -1
187	paylast	Numeric	4	0	Year left last paid job	{-9, No ans...	-99 - -1
188	paymon	Numeric	2	0	Month last left paid job	{-9, No ans...	-99 - -1
189	sclass	Numeric	2	0	Social Class	{-9, No ans...	-99 - -1
190	seg	Numeric	2	0	Socio-Economic Group	{-9, No ans...	-99 - -1
191	snemplee	Numeric	2	0	Self employed, how many employees	{-9, No ans...	-99 - -1
192	age	Numeric	3	0	Age last birthday	{-9, No ans...	-99 - -1

In practice: user guide and documentation

- A user guide could contain a variety of documents that provide context: interview schedule, transcription notes, even photos



In practice: data list

- Data listing provides an at-a-glance summary of interview sets

Study Number 5407

Health and Social Consequences of the Foot and Mouth Disease Epidemic in North Cumbria, 2001

Mort, M.

The panel respondents for the study were divided into six population groups. The data list for the diary and interviews has been colour-coded accordingly for clarity, using the depositor's original colours:

Group 1: Farmers	Group 2: Rural Business	Group 3: Agricultural related occupations	Group 4: Frontline Workers	Group 5: Community	Group 6: Animal / Human Health Professionals
------------------	-------------------------	---	----------------------------	--------------------	--

1. Interviews

Respondent ID	Population Group	Date of Birth	Gender	Occupation	Interview summary	Place of Interview
PM02	Group 6: Animal / Human Health Professionals	1975	M	Veterinary Surgeon	Family and background, career and work, arrangements during FMD epidemic and perceptions of situation	North Cumbria, respondent's home
PM03	Group 6: Animal / Human Health Professionals	1966	F	Veterinary Surgeon	Family and background, career and work, arrangements during FMD epidemic and perceptions of situation	North Cumbria
PM07	Group 6: Animal / Human Health Professionals	1964	F	Veterinary practice manager	Family and background, career and work, arrangements during FMD epidemic and perceptions of situation	North Cumbria, respondent's home
					Family and background, career and work, arrangements during FMD epidemic and perceptions of situation	

UK Data Service



Transcription template

Should:

- possess a unique identifier
- adopt a uniform layout throughout the research project
- make use of speaker tags - turn-taking
- carry line breaks
- be page numbered
- carry a document header giving brief details of the interview: date, place, interviewer name, interviewee details, etc.

Other considerations:

- cover page
- compatibility with import features of Computer Assisted Qualitative Data Analysis Software (CAQDAS)

In practice: transcript format

Study Name:
Depositor:
Interviewer:

Interview number:
Interview ID: Firstname Lastname
Date of interview:

Information about interviewee

Date of birth:
Gender:
Geographic region:

Marital status:
Occupation:

Y=Interviewee

I=Interviewer

Y: I came here in late 1968.

I: You came here in late 1968? Many years already.

Y: 31 years already. 31 years already.

I: (laugh) It is really a long time. Why did you choose to come to England at that time?

Y: I met my husband and after we got married in Hong Kong, I applied to come to England.

I: You met your husband in Hong Kong?

Y: Yes.

I: He was working here [in England] already?



File formats

Choice of software format for digital data:

- planned data analyses
- software availability / cost
- hardware used – e.g. audio capture
- discipline-specific standards and customs

Digital data is software dependent, so endangered by obsolescence of software / hardware

Best formats for long-term preservation:

- standard, interchangeable and open
- *e.g. tab-delimited, comma-delimited (CSV), ASCII, RTF, PDF/A, OpenDocument format and XML*
- [UK Data Service optimal file formats](#) for various data types
- [Digital Preservation Coalition](#) guidance on preservation formats

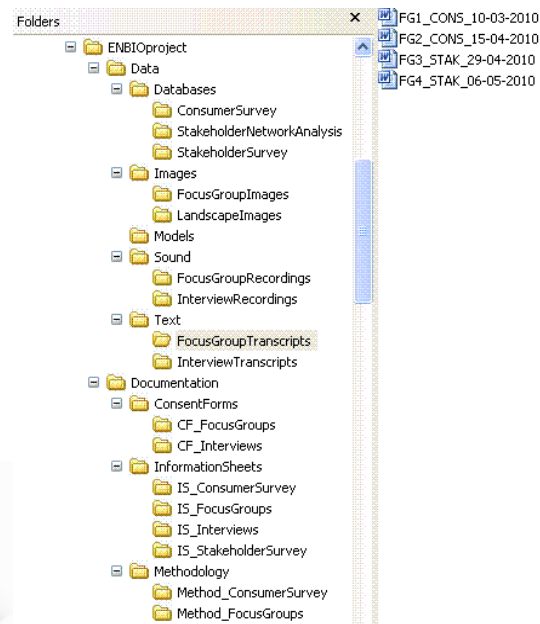
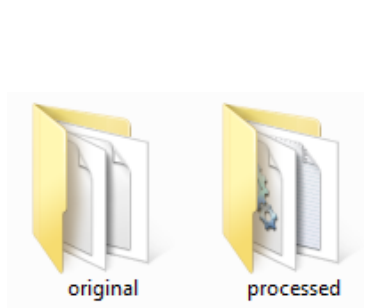


Organising data

- Plan in advance how best to organise data
- Use a logical structure and ensure collaborators understand

Examples

- hierarchical structure of files, grouped in folders, e.g. audio, transcripts and annotated transcripts
- survey data: spreadsheet, SPSS, relational database
- interview transcripts: individual well-named files



Data security and storage

Protect data from unauthorised:

- access
- use
- change
- disclosure
- destruction

Who knows who is watching, listening or attempting to access your data...



UK Data Service



Data security strategy

- Control access to computers:
 - use passwords and lock your machine when away from it
 - run up-to-date anti-virus and firewall protection
 - power surge protection
 - utilise encryption
 - on all devices: desktops, laptops, memory sticks and mobile devices
 - at all locations: work, home and travel
 - restrict access to sensitive materials e.g. consent forms and patient records
 - personal data need more protection – always keep them separate and secure
- Control physical access to buildings, rooms and filing cabinets
- Properly dispose of data and equipment once the project is finished



Encryption software

Encryption software can be easy to use and enables users to:

- encrypt hard drives, partitions, files and folders
- encrypt portable storage devices such as USB flash drives

[VeraCrypt](#)



[Axcrypt](#)



[BitLocker](#)



[FileVault2](#)



Digital back-up strategy

Consider

- **What's backed-up?** - all, some or just the bits you change?
- **Where?** - original copy, external local and remote copies
- **What media?** - DVD, external hard drive, USB, Cloud?
- **How often?** - hourly, daily, weekly? Automate the process?
- **What method / software?** - duplicating, syncing or mirroring?
- **For how long is it kept?** - data retention policies that might apply?
- **Verify and recover** - never assume, regularly test and restore

Backing-up need not be expensive

- 1Tb external drives are around £50, with back-up software

Also consider non-digital storage too!



File sharing and collaborative environments

Sharing data between researchers

- Too often sent as insecure email attachments

Other options:

- Virtual Research Environments
 - MS SharePoint
- Locally managed; ownCloud and ZendTo
- File transfer protocol (FTP)
- Physical media
- Cloud solutions
 - Google Drive, DropBox, Microsoft OneDrive and iCloud (insecure)
 - Securer options? - Mega.nz, [SpiderOak](http://SpiderOak.com) and [Tresorit](http://Tresorit.com)



By David Fletcher
<http://www.cloudtweaks.com/2011/05/the-lighter-side-of-the-cloud-data-transfer/>



tresorit

UK Data Service



- Assess risks of using cloud storage

Data Disposal

Proper disposal of equipment and media

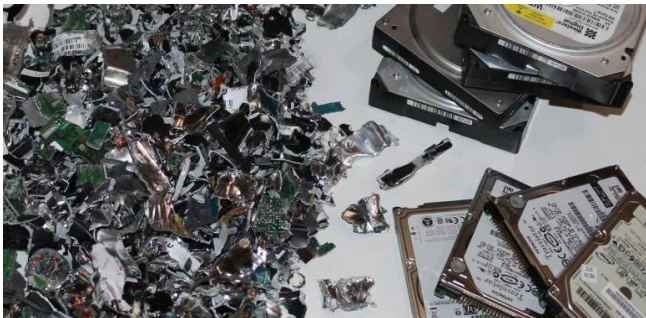
- even reformatting a hard drive is **not** sufficient
- if in doubt, physically destroy the drive



- **BCWipe** - uses 'military-grade procedures to surgically remove all traces of any file'
 - Can be applied to entire disk drives

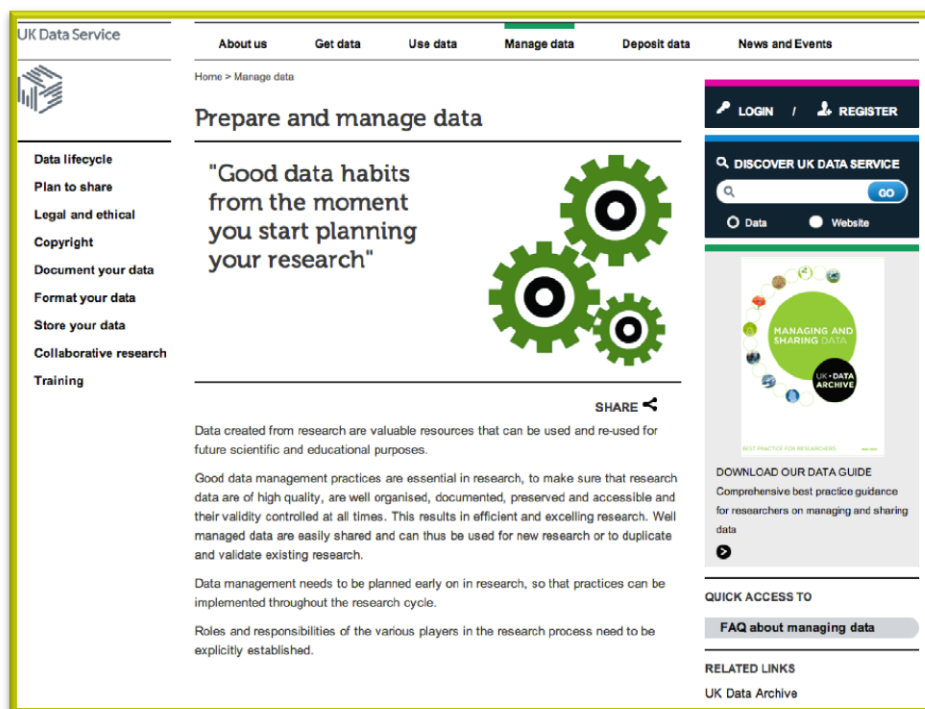


- **AxCrypt** - free open source file and folder shredding
 - Integrates into Windows well, useful for single files



Our data management guidance

- online best practice guidance: ukdataservice.ac.uk/manage-data.aspx
- [Managing and Sharing Research Data – a Guide to Good Practice: \(Sage Publications Ltd\)](#)
- helpdesk for queries: ukdataservice.ac.uk/help/get-in-touch.aspx
- training: www.ukdataservice.ac.uk/news-and-events/events



UK Data Service



Tools & templates

- Model consent form: <http://www.data-archive.ac.uk/media/112638/ukdamodelconsent.pdf>
- Survey consent statement: <http://data-archive.ac.uk/media/147338/ukdasurveyconsent.doc>
- Transcription template: <http://data-archive.ac.uk/media/136055/ukdamodeltranscript.pdf>
- Transcription instructions: <http://data-archive.ac.uk/media/285633/ukda-example-transcription-instructions.pdf>
- Transcription confidentiality agreement: <http://data-archive.ac.uk/media/285636/ukda-transcriber-confidentiality-agreement.pdf>
- Data list template: <http://data-archive.ac.uk/media/2989/UK%20Data%20Archive%20Example%20Data%20List.pdf>
- RDM costing tool: www.data-archive.ac.uk/media/247429/costingtool.pdf
- Encryption tutorials: <https://www.youtube.com/watch?v=y4losu-Yfsw&list=PLG87Imnep1SmnFGhAjFVHonQSVmMlpHkV>



Keep connected

- Subscribe to UK Data Service list:
www.jiscmail.ac.uk/cgi-bin/webadmin?A0=UKDATASERVICE
- Follow UK Data Service on Twitter: @UKDataService
- Facebook
- Youtube: www.youtube.com/user/UKDATASERVICE



Questions?

UK Data Service

University of Essex

ukdataservice.ac.uk/help/get-in-touch.aspx

UK Data Service

