University of Essex

Data Management Basics

Scott Summers
UK Data Service
University of Essex

Newcomers
12th January 2018

UK Data Service

UK • DATA ARCHIVE

Presentation Structure

- What is the UK Data Service?
- Why is it important to manage your research data?
- DMPs
- Protecting participants
  - Consent
  - Anonymisation
  - Access controls
- Documentation
- Organising and storing data
- Hands on exercises
- Your questions

What is the UK Data Service?

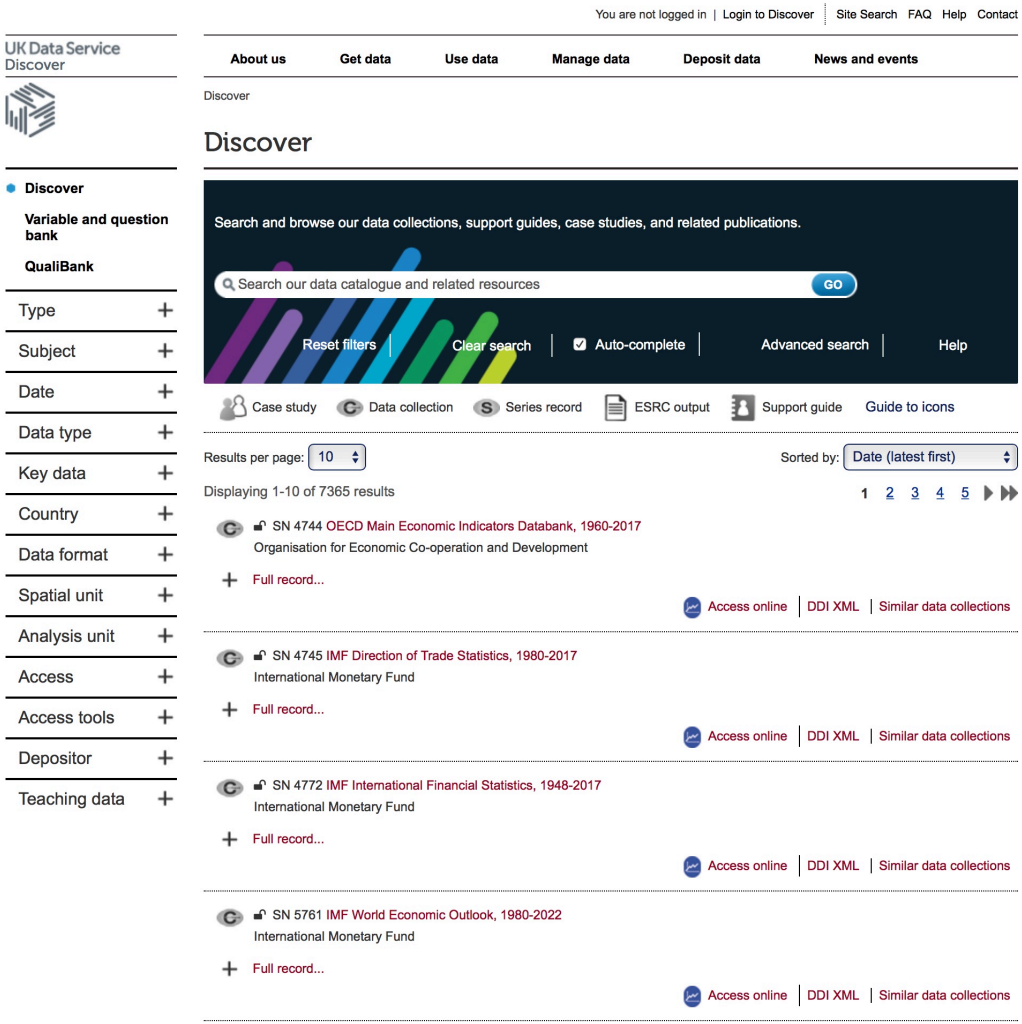
- Funded by the ESRC
- Single point of access to a wide range of secondary social science data
- We provide support and training for data creators with accessing, managing, sharing and using data
- Delivered by staff based at universities across the UK (Essex, Manchester, Leeds, Southampton, Edinburgh & UCL)
- UK Data Archive – manages the UK Data Service and curates the data

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/



UK Data Service

University of Essex

UK · DATA ARCHIVE

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 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

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

Informed consent (broadly)

- Consent needs to be **freely given**, **informed**, **unambiguous**, **specific** and by a **clear affirmative** action that signifies agreement to the processing of personal data.
- The best way to achieve informed consent for data sharing is to **identify** and **explain** the **possible future uses of their data** and offer the participant the option to consent on **a granular level**.
- For example, in a qualitative study, this may involve allowing the participant to consent to data sharing of the anonymised transcripts, the non-anonymised audio recordings and the photographs.
- The GDPR requires that researchers document consent. Therefore, it will be essential to keep documented and accurate records of the consent obtained from participants.

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

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

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 abut the tape.

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.

Comment [v1]: Replace: Ken
Comment [v2]: delete
Comment [v3]: delete
Comment [v4]: Replace: Ken
Comment [v5]: Replace: Ken
Comment [v6]: Replace: Ken

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 an SPSS file

hse09ai.sav [DataSet2] - PASW Statistics Data Editor

File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help

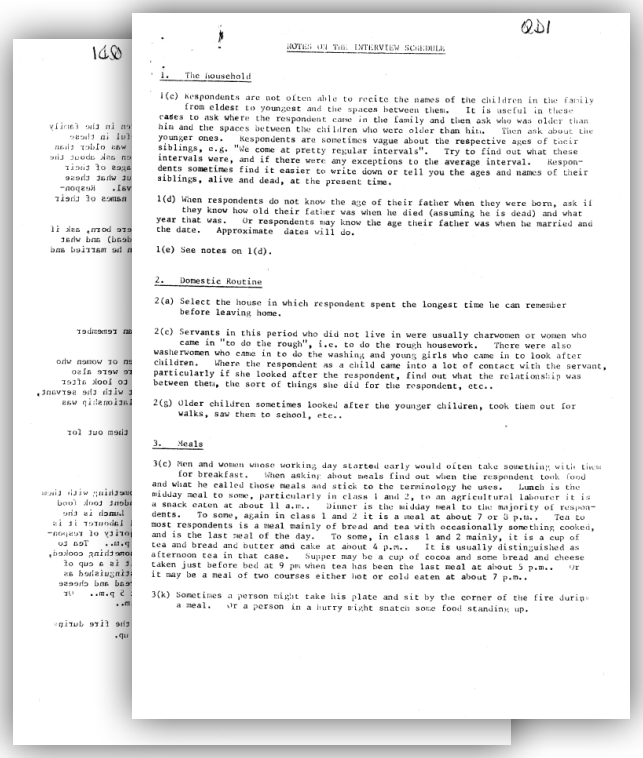
	Name	Type	Width	Decimals	Label	Values	Missing
175	quala10	Numeric	2	0	Which of the qualifications on this card do you have? 10	{-9, No ans...	-99 - -1
176	activb	Numeric	2	0	Activity status for last week	{-9, No ans...	-99 - -1
177	empstat	Numeric	2	0	Manager/Foreman	{-9, No ans...	-99 - -1
178	everjob	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

Data View | Variable View

PASW Statistics Processor is ready

In practice: user guide and documentation

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



QD1

NOTES ON THE INTERVIEW SCHEDULE

1. The household

1(c) Respondents are not often able to recite the names of the children in the family from eldest to youngest and the spaces between them. It is useful in these cases to ask where the respondent came in the family and then ask who was older than him and the spaces between the children who were older than him. Then ask about the younger ones. Respondents are sometimes vague about the respective ages of their siblings, e.g. "We come at pretty regular intervals". Try to find out what these intervals were, and if there were any exceptions to the average interval. Respondents sometimes find it easier to write down or tell you the ages and names of their siblings, alive and dead, at the present time.

1(d) When respondents do not know the age of their father when they were born, ask if they know how old their father was when he died (assuming he is dead) and what year that was. Or respondents may know the age their father was when he married and the date. Approximate dates will do.

1(e) See notes on 1(d).

2. Domestic Routine

2(a) Select the house in which respondent spent the longest time he can remember before leaving home.

2(c) Servants in this period who did not live in were usually charwomen or women who came in "to do the rough", i.e. to do the rough housework. There were also washerwomen who came in to do the washing and young girls who came in to look after children. Where the respondent as a child came into a lot of contact with the servant, particularly if she looked after the respondent, find out what the relationship was between them, the sort of things she did for the respondent, etc..

2(g) Older children sometimes looked after the younger children, took them out for walks, saw them to school, etc..

3. Meals

3(c) Men and women whose working day started early would often take something with them for breakfast. When asking about meals find out when the respondent took food and what he called those meals and stick to the terminology he uses. Lunch is the midday meal to some, particularly in class 1 and 2, to an agricultural labourer it is a snack eaten at about 11 a.m.. Dinner is the midday meal to the majority of respondents. To some, again in class 1 and 2 it is a meal at about 7 or 8 p.m.. Tea to most respondents is a meal mainly of bread and tea with occasionally something cooked, and is the last meal of the day. To some, in class 1 and 2 mainly, it is a cup of tea and bread and butter and cake at about 4 p.m.. It is usually distinguished as afternoon tea in that case. Supper may be a cup of cocoa and some bread and cheese taken just before bed at 9 pm when tea has been the last meal at about 5 p.m.. Or it may be a meal of two courses either hot or cold eaten at about 7 p.m..

3(k) Sometimes a person might take his plate and sit by the corner of the fire during a meal. Or a person in a hurry might snatch some food standing up.

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, respo 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, respo home
					Family and	

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.

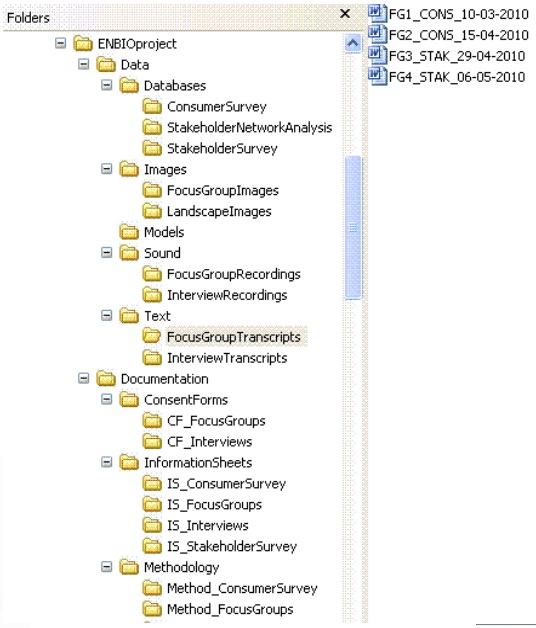
I: He was working here [in England] already?

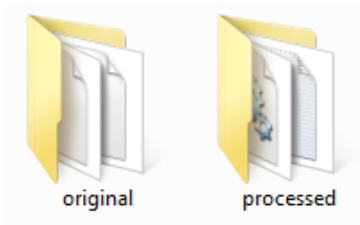
Organising data

- 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





Naming files

Example:

- 20130311_interview2_audio.wav
- 20130311_interview2_trans.rtf
- 20130311_interview2_image.jpg

- Naming of files:
  - Version
  - Dates – YYYY-MM-DD (e.g. 2017-11-28)
  - Creator
  - Description of content
  - Spacing, special characters and dots – (e.g. Interview Transcript 01)
  - Interview20%Transcript20%01
  - Underscores – (e.g. Interview_Transcript_01)
  - Avoid very long names
  - Bulk file renaming

Versioning files

- Version control of files:
  - How many versions to keep? How long for?
  - It can be difficult to identify the correct version of a file if no standard naming practice is implemented
  - Major revisions vs minor revisions
  - 02-00
  - 02-01

Recommended File Formats

Documentation and scripts	Rich Text Format (.rtf) PDF/UA, PDF/A or PDF (.pdf) XHTML or HTML (.xhtml, .htm) OpenDocument Text (.odt)	plain text (.txt) widely-used formats: MS Word (.doc/.docx), MS Excel (.xls/.xlsx) XML marked-up text (.xml) according to an appropriate DTD or schema, e.g. XHMTL 1.0

Image data	TIFF 6.0 uncompressed (.tif)	JPEG (.jpeg, .jpg, .jp2) if original created in this format GIF (.gif) TIFF other versions (.tif, .tiff) RAW image format (.raw) Photoshop files (.psd) BMP (.bmp) PNG (.png) Adobe Portable Document Format (PDF/A, PDF) (.pdf)

Textual data	Rich Text Format (.rtf) plain text, ASCII (.txt) eXtensible Mark-up Language (.xml) text according to an appropriate Document Type Definition (DTD) or schema	Hypertext Mark-up Language (.html) widely-used formats: MS Word (.doc/.docx) some software-specific formats: NUD*IST, NVivo and ATLAS.ti

Data security

Protect data from unauthorised:

- access
- use
- change
- disclosure
- destruction

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

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

BitLocker

Axcrypt

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!

"We back up our data on sticky notes because sticky notes never crash."

Verification and integrity checks

- Ensure that your backup method is working as intended
- Automated services - check
- Be wary when using sync tools in particular
  - Mirror in the wrong direction or using the wrong method, and you could lose new files completely

- You can use checksums to verify the integrity of a backup
- Also useful when transferring files
- Checksum somewhat like a files' fingerprint
- …but changes when the file changes

Checksums

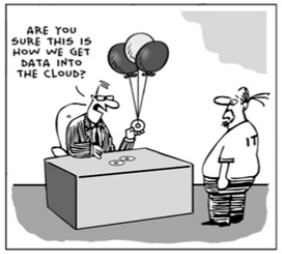
- Each time you run a checksum a number string is created for each file
- Even if one byte of data has been altered or corrupted that string will change
- Therefore, if the checksums before and after backing up a data file match, then you can be sure that the data have not altered during this process
- A free software tool for computing MD5 checksums is MD5summer for windows
- MacOS has the ability built in

Collaborative Storage

Sharing data between researchers

- Too often sent as insecure email attachments
- Physical media?
- Virtual Research Environments
  - MS SharePoint
  - Clinked
  - Huddle
  - Basecamp
  - Google Docs

Cloud storage services



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

- Online or 'cloud' services are becoming increasingly popular
- Google Drive, DropBox, Microsoft OneDrive and iCloud

Google Drive Dropbox OneDrive

- Benefits:
  - Very convenient
  - Accessible anywhere
  - Good protection if working in the field?
  - Background file syncing
  - Mirrors files
  - Mobile apps available

But,

  - These are not necessarily secure
  - Potential DPA issues
  - Not necessarily permanent
  - Intellectual property right concerns?
  - Limited storage?

Cloud storage services

- Perhaps more secure options?

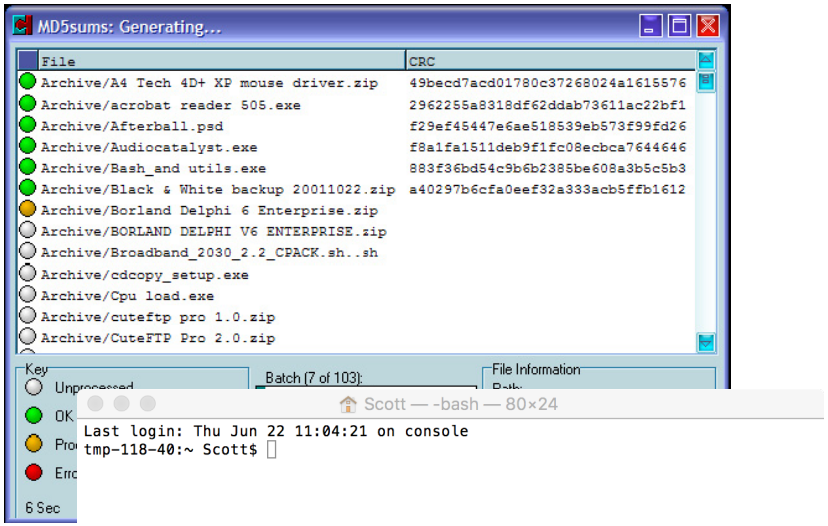
Mega.nz

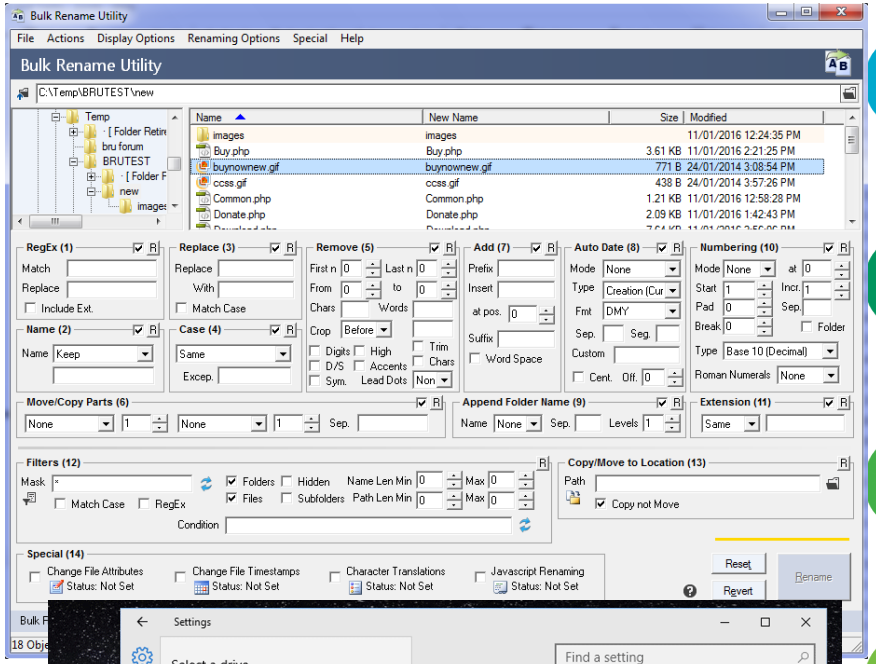
SpiderOak

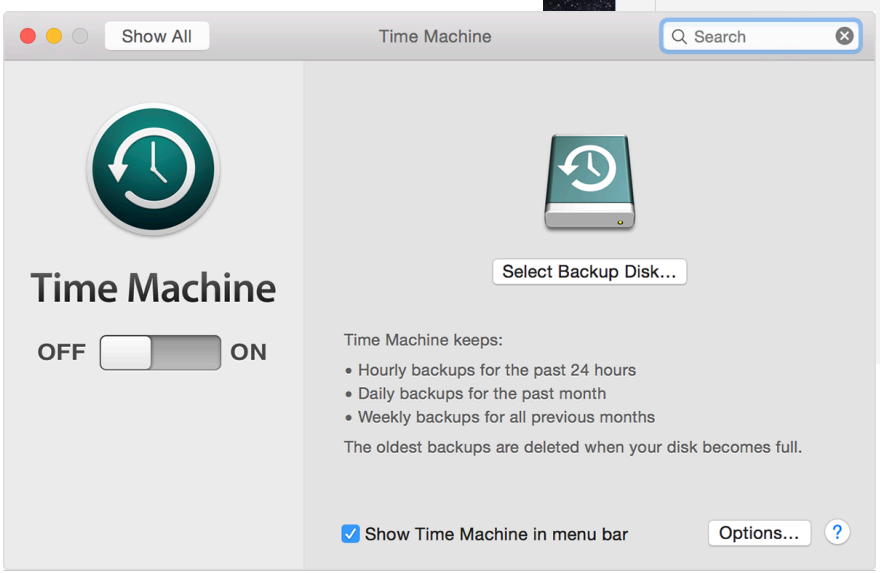
Tresorit

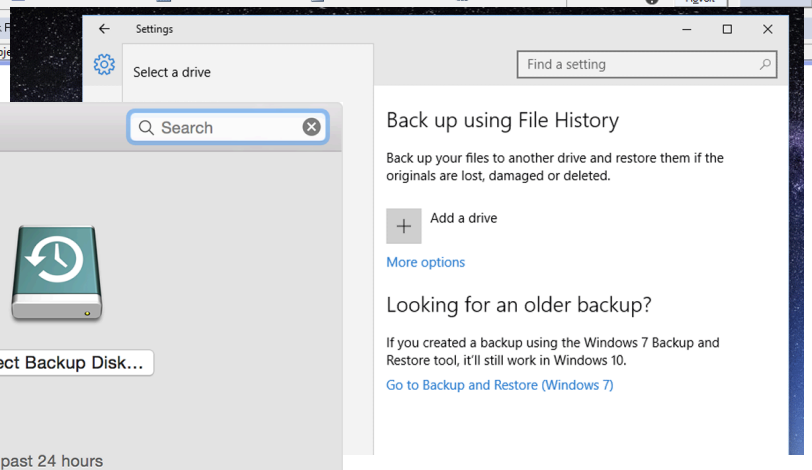
- Cloud data storage should be avoided for high-risk information such as files that contain personal or sensitive information, or information that has a very high intellectual property value.

Hands on Exercises









Contact

Producer Relations team
UK Data Service
University of Essex
ukdataservice.ac.uk/help/get-in-touch