
Data Management Basics

Anca Vlad
UK Data Service
Research Data Services

03 December 2020



Overview

- UK Data Service
- Managing your data – background, why and how
 - GDPR
 - Consent
 - Anonymisation
 - Access controls
 - Documentation
 - Security
 - Encryption
 - Backups

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

Website, data catalogue:
<https://www.ukdataservice.ac.uk/>

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

The screenshot shows the UK Data Service website homepage. At the top, there is a navigation menu with links for 'About us', 'Get data', 'Use data', 'Manage data', 'Deposit data', and 'News and events'. Below the menu is a search bar with the text 'Search data' and a magnifying glass icon. The main content area features a large banner with the text 'Explore the UK's largest collection of social, economic and population data resources'. Below the banner, there are two columns of content. The left column is titled 'About the UK Data Service' and features a video player showing a red double-decker bus. The right column is titled 'Guides and resources' and lists 'Dataset guides', 'Topic guides', 'Methods and software guides', and 'Guides to exploring online'. A 'See more >' link is provided below the list. To the right of the list is a 'Video tutorials' section with a purple button that says 'See our range of training videos'. At the bottom of the page, there is a dark blue banner with the text 'See data from all over the world' and a purple button that says 'Browse our data map'.

UK Data Service

About us Get data Use data Manage data Deposit data News and events

Explore the UK's largest collection of social, economic and population data resources

Search data

About the UK Data Service

Guides and resources

Dataset guides

Topic guides

Methods and software guides

Guides to exploring online

See more >

Video tutorials
See our range of training videos

See data from all over the world

Browse our data map

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 publishers now specify requirements for data handling, including the formulation of a data management plan

Why is data management important?

- 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 plan
- Make sure data are shareable and can be understood:
 - Obtain consent to share
 - Do not disclose identities without consent
 - Use open source and standard formats
 - Provide context and documentation
 - Protect your data at all stages (secure storage, encryption)

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, anonymization, re-use strategies

Copyright/Intellectual Property Ownership

Responsibilities

Management and curation

[Data management plan guidance](#)

Data protection regulation 2018

- Applies to personal data, pseudonymised data and living persons only
- Personal data are 'any information relating to an identified or identifiable natural person'
- Note that not all research data personal data
- Also note there may still be ethical reasons for wanting to protect this information though!

GDPR – processing grounds

There are **six** grounds for the processing of personal data, and one of these must be present in order to process a data subject's personal data:

1. Consent
2. Contract
3. Legal obligation
4. Vital interests
5. Public interest (public task)
6. Legitimate interest

Multiple tools for protecting participants

1. Establish the processing ground that will apply to the data you are collecting. If this is **informed consent**, then ensure this covers data sharing and long-term preservation and curation
2. Protect identities e.g. **anonymization**, and (or) not collecting personal data (only collect data that is necessary)
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

Wording in consent forms and information sheets could be broken down in **three** key areas:

1. Taking part in the study
2. Use of the information in the study
3. Future use and reuse of the information by other

Model consent form (GDPR compliant):

<https://www.ukdataservice.ac.uk/media/622375/ukdamodelconsent.docx>

In practice: Wording

3. Future use and reuse of the information by others

I give permission for the [specify the data] that I provide to be deposited in [name of data repository]
so it can be used for future research and learning.

Specify in which form the data will be deposited, e.g. anonymised transcripts, audio recording, survey database, etc.; and if needed repeat the statement for each form of data you plan to deposit.

Specify whether deposited data will be anonymised, and how. Make sure to describe this in detail in the information sheet.

Specify whether use or access restrictions will apply to the data in future, e.g. exclude commercial use, apply safeguarded access, etc.; and discuss these restrictions with the repository in advance.

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 be used as well. 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 [name of repository]. Then the interview data can be disseminated for reuse by other researchers, for research and learning purposes.

Anonymization

Quantitative

- Remove direct identifiers
e.g. names, address, institution, photo
 - Reduce the precision/detail of a variable through aggregation *e.g. birth year instead of date of birth, occupational categories rather than jobs; and, area rather than village*
 - Generalise meaning of detailed text variable
e.g. occupational expertise
 - Restrict upper and lower ranges of a variable to hide outliers *e.g. income, age*
- ✓ *Keep an anonymization log (and keep it separate from anonymised data files)*

Qualitative

- Remove direct identifiers (e.g names, address, institution, photos) or replace with pseudonyms
 - Avoid blanking out; use pseudonyms or replacements
 - Identify replacements with [brackets]
 - Plan or apply editing at time of transcription
 - Consistency throughout project
 - Avoid over-anonymising – removing information in text can distort data, make them unusable, unreliable or misleading; so balance anonymization with the need to preserve context
- <https://www.ukdataservice.ac.uk/deposit-data/stories/gush>
- ✓ Keep an anonymization log (and keep it separate from anonymised data files)

Anonymization - Audio-visual data

- Data 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(wish access restrictions in place)
 - Avoid mentioning disclosing information during audio recordings

In practice: example anonymisation

Health and Social Consequences of the Foot and Mouth Disease Epidemic in North Cumbria, 2001-2003 (SN 5407)

M. Mort, Lancaster University. Institute for Health Research

NOTE: all identifying info contained in this transcript is fictional

Date of Interview: 21/02/02

Interview with Lucas Roberts, DEFRA field officer

Date of birth: 2 May 1965

Comment [v1]: Replace: Ken

Comment [v2]: Delete

Gender: Male

Occupation: Frontline worker

Location: Plumpton, North Cumbria

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.

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 and who have received training

In practice: data with access conditions

Mort, M. (2006). *Health and Social Consequences of the Foot and Mouth Disease Epidemic in North Cumbria, 2001-2003*. [data collection]. UK Data Service. SN: 5407, <http://doi.org/10.5255/UKDA-SN-5407-1>

- 40 interview and diary transcripts are archived and available for reuse by registered users (**safeguarded access**)
- 3 interviews and 5 diaries were embargoed until 2015 (**Safeguarded – embargoed**)
- Audio files archived and only available with permission from depositor(s) (**Safeguarded – Special agreement**)

Documenting your data

- Enables you to understand the data if/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 and research instruments
 - Ensure accurate processing and archiving
 - Create a catalogue record for a published data collection

Include as documentation

- Data collection methodology and processes: sampling, sample size, fieldwork protocol, experiment protocol, interviewer instructions
- Codebook, user guide (for quantitative data)
- Information sheet, consent form (blank versions)
- Questionnaires, show cards, topic guides
- Transcripts: header with context information: data and place of interview, interviewer, interviewee details (in line with consent form) etc.
- Data list: overview of key information about each interview, a map of the data collection (for qualitative data)
- Links to reports and publications (preferably DOIs where possible)

Data-level documentation

- All structured, tabular data should have adequate variable names, variable and value labels
- Variable names might include:
 - Question number system matching questions in the questionnaire used e.g. Q1a, Q1b, Q2, Q3b
 - 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'
 - For interoperability across platforms, variable names should not be longer than 8 characters and without spaces

Data-level documentation

Similar principles for variable labels:

- Be brief, maximum 80 characters
- Include unit of measurement where appropriate
- Reference the question number of a survey or questionnaire

e.g. variable 'q11hexw' with label 'Q11b: hours spent taking physical exercise in a typical week' – the label gives the unit of measurement and a reference to the questions number (Q11b)

- Coding or classification schemes used, with a bibliographic reference

e.g. Standard Occupational Classification 2000; ISO 3166 alpha-2 country codes

For value labels:

- 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(skipped)', '96= not known', '95=error'

In practice: user guide and documentation

A user guide should contain variety of documents that provide context: interview schedule, methodology, study findings, consent procedures, transcription notes, codebook etc.

[User guide](#) for

Mort, M. (2006). Health and Social Consequences of the Foot and Mouth Disease Epidemic in North Cumbria, 2001-2003. [data collection]. UK Data Service. SN: 5407, <http://doi.org/10.5255/UKDA-SN-5407-1>

In practice: data list

Study Number 6377

Integrated Floodplain Management, 2006-2008

Morris, J.

Floodplain farm survey

Interview ID	Farmer code	Age	Farm scheme	Farm type	Size of farm (hectare)	Number of holdings	Date of interview	Interviewer name	No of pages	Text file name	Audio file name
1	Be1	35-45	Beckingham	Beef	360	1	04.12.2006	Helena	28	6377int001	6377int001
2	Be2	45-55	Beckingham	Arable	364	1	05.12.2006	Helena	21	6377int002	6377int002
3	Be3	45-55	Beckingham	Arable	372	2	06.12.2006	Helena	22	6377int003	6377int003
4	Be4	45-55	Beckingham	Arable	194	3	06.12.2006	Helena	18	6377int004	6377int004
5	Be5	55-65	Beckingham	Arable	108	1	07.12.2007	Helena	21	6377int005	6377int005
6	Be6	45-55	Beckingham	Arable	1254	2	01.02.2008	Helena	19	6377int006	
7	Bu1	55-65	Bushley	Mixed	101	2	13.02.2007	Quentin	29	6377int007	6377int007
8	Bu2	>65	Bushley	Mixed	97	1	15.02.2007	Quentin	15	6377int008	6377int008
9	Bu3	>65	Bushley	Arable	194	4	13.02.2007	Quentin	21	6377int009	6377int009
10	Bu4	55-65	Bushley	Mixed	202	1	15.03.2007	Helena	19	6377int010	6377int010
11	Cu1	35-45	Cuddyarch	Dairy	64	1	08.05.2007	Helena	19	6377int011	6377int011
12	Cu2	55-65	Cuddyarch	Dairy	189	2	08.05.2007	Helena	18	6377int012	6377int012
13	Cu3	55-65	Cuddyarch	Mixed livestock	76	1	08.05.2007	Helena	13	6377int013	6377int013
14	Cu5	45-55	Cuddyarch	Mixed livestock	198	1	09.05.2007	Helena	24	6377int014	6377int014
15	Cu6	55-65	Cuddyarch	Dairy	89	1	09.05.2007	Helena	14	6377int015	6377int015
16	Cu7	>65	Cuddyarch	Mixed livestock	190	4	11.05.2007	Helena	20	6377int016	6377int016
17	Cu8	55-65	Cuddyarch	Mixed livestock	109	2	11.05.2007	Helena	22	6377int017	6377int017
18	Id1	55-65	Idle	Arable	158	3	07.02.2007	Quentin	17	6377int018	6377int018a
18	Id1	55-65	Idle	Arable	158	3	07.02.2007	Quentin	17	6377int018	6377int018b
19	Id1b	55-65	Idle	Arable	158	3		Quentin	22	6377int019	
20	Id2	45-55	Idle	Dairy	150	1	08.02.2007	Quentin	17	6377int020	6377int020

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: data, place, interviewer name, interviewee details, etc.

Other considerations:

- Cover page
- Compatibility with import featured of Computer Assisted Qualitative Data Analysis Software (CAQDAS)

In practice: transcript format

Study Name:
Depositor:

Interview ID:
Date of Interview:

Information about interviewee:

(e.g. Age, Gender, Occupation, Marital Status, Geographic region, etc. as relevant /appropriate)

R= Respondent/Interviewee *(if more than one respondent, use R1, R2, etc.)*

I=Interviewer

R: I came here in late 1968.

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

R: 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?

R: 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?

R: Yes.

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

R: After he worked here for a few years -- in the past, it was quite common for them to go back to Hong Kong to get a wife. Someone introduced us and we both fancied each other. At that time, it was alright to me to get married like that as I wanted to leave Hong Kong. It was like a gamble. It was really like a gamble.

I: You were very brave to think about going abroad as you were so young at that time.

Model Interview Transcript:

<https://www.ukdataservice.ac.uk/media/622380/ukdamodeltranscript.pdf>

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

[UK Data Service optimal file formats](#) for various data types

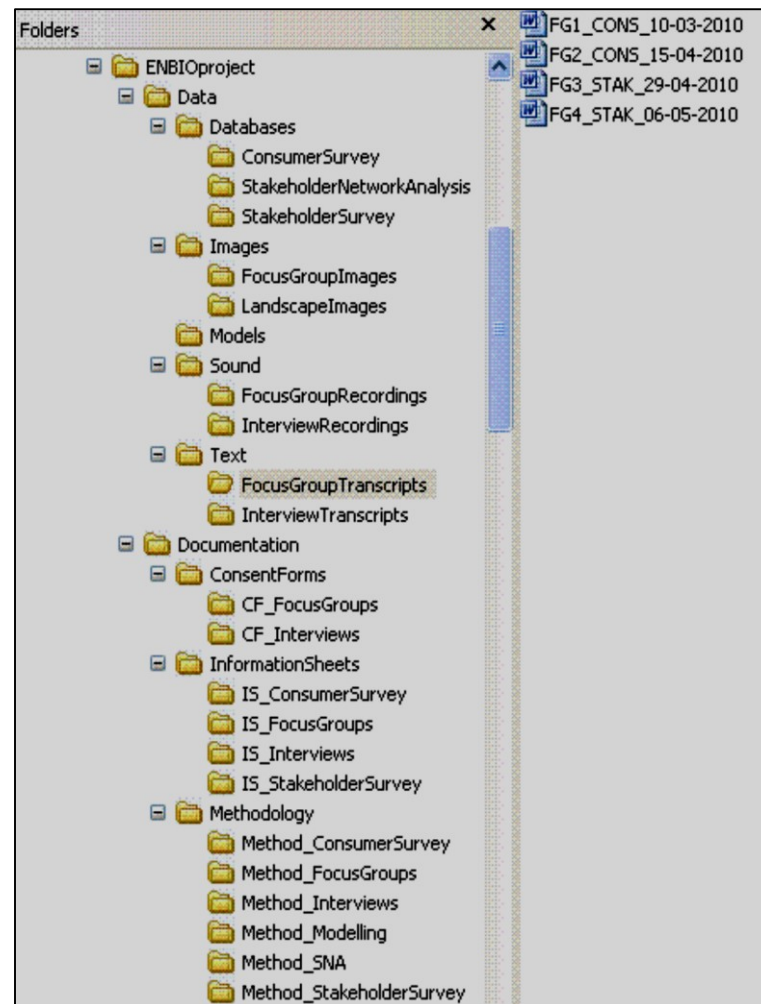
[Digital Preservation Coalition](#) guidance on preservation formats

Organising data

- Plan in advance how to best organise data (project specific)
- 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...

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
 - Restrict access to sensitive materials e.g. consent forms and patient records
 - Personal data need more protection – always keep them separate and secure
 - Utilise encryption
 - on all devices: desktops, laptops, memory sticks and mobile devices
 - at all locations: work, home and travel
- 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](#)



[BitLocker](#)



[Axcrypt](#)



[FileVault2](#)



Data encryption tutorials:

<https://www.youtube.com/playlist?list=PLG87Imnep1SmnFGhAjFVHonQSVmMlpHkV>

Video tutorials

- VeraCrypt: <https://www.youtube.com/watch?v=Ogm9QHQPfQU>
- AxCrypt: <https://www.youtube.com/watch?v=ACcRInsoYZg>
- FileVault 2: <https://www.youtube.com/watch?v=JIZ9EFMS0ic>
- BitLocker: <https://www.youtube.com/watch?v=y4Iosu-Yfsw>
- Time Machine: <https://www.youtube.com/watch?v=hlsQaVj7WtA>
- MD5summer: <https://www.youtube.com/watch?v=VcBfkB6N7-k>

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

- 2Tb 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?)
 - More secure options? [Mega.nz](https://mega.nz) [SpiderOak](https://spideroak.com) [Tresorit](https://tresorit.com)



- 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

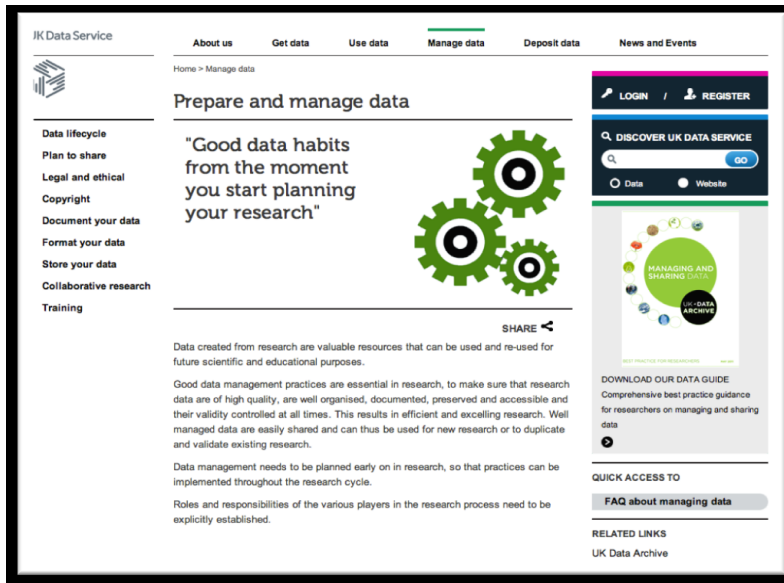
AxCrypt - free open source file and folder shredding

- Integrates into Windows well, useful for single files



UKDS data management guidance

- Best practice guidance: www.ukdataservice.ac.uk/manage-data.aspx
- Managing and Sharing Research Data – a Guide to Good Practice (Sage Publications Ltd)
- Training: www.ukdataservice.ac.uk/news-and-events/events
- Twitter: @UKDSRDM



The screenshot shows the UK Data Service website interface. The main heading is "Prepare and manage data". Below this, a quote reads: "Good data habits from the moment you start planning your research". To the right of the quote are three green gears of different sizes. Below the quote, there is a "SHARE" button and a paragraph of text: "Data created from research are valuable resources that can be used and re-used for future scientific and educational purposes." followed by another paragraph: "Good data management practices are essential in research, to make sure that research data are of high quality, are well organised, documented, preserved and accessible and their validity controlled at all times. This results in efficient and excellent research. Well managed data are easily shared and can thus be used for new research or to duplicate and validate existing research." and a final paragraph: "Data management needs to be planned early on in research, so that practices can be implemented throughout the research cycle. Roles and responsibilities of the various players in the research process need to be explicitly established."



Tools and templates

Model consent form:

<http://www.dataarchive.ac.uk/media/112638/ukdamodelconsent.pdf>

• Survey consent statement:

<http://dataarchive.ac.uk/media/147338/ukdasurveyconsent.doc>

• Transcription template:

<http://dataarchive.ac.uk/media/136055/ukdamodeltranscript.pdf>

• Transcription instructions: <http://dataarchive.ac.uk/media/285633/ukda-example-transcriptioninstructions.pdf>

• Transcription confidentiality agreement:

<http://dataarchive.ac.uk/media/285636/ukda-transcriber-confidentialityagreement.pdf>

• Data list template:

<http://dataarchive.ac.uk/media/2989/UK%20Data%20Archive%20Example%20Data%20List.pdf>

Training

[Recurring workshops and webinars](#)

Webinar: Data management basics

Webinar: Key issues in reusing data

Webinar: Finding and accessing data in the UK Data Service

Webinar: Guided walk through ReShare

Webinar: Key data: UK and cross-national surveys

Keep connected

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

Contact

Enquiries/ Help Desk:

<http://ukdataservice.ac.uk/help/get-in-touch.aspx>

help@ukdataservice.ac.uk

Follow us on:

<https://twitter.com/UKDataService>

<https://www.facebook.com/UKDataService>

<https://www.jiscmail.ac.uk/cgi-bin/webadmin?A0=UKDATASERVICE>

