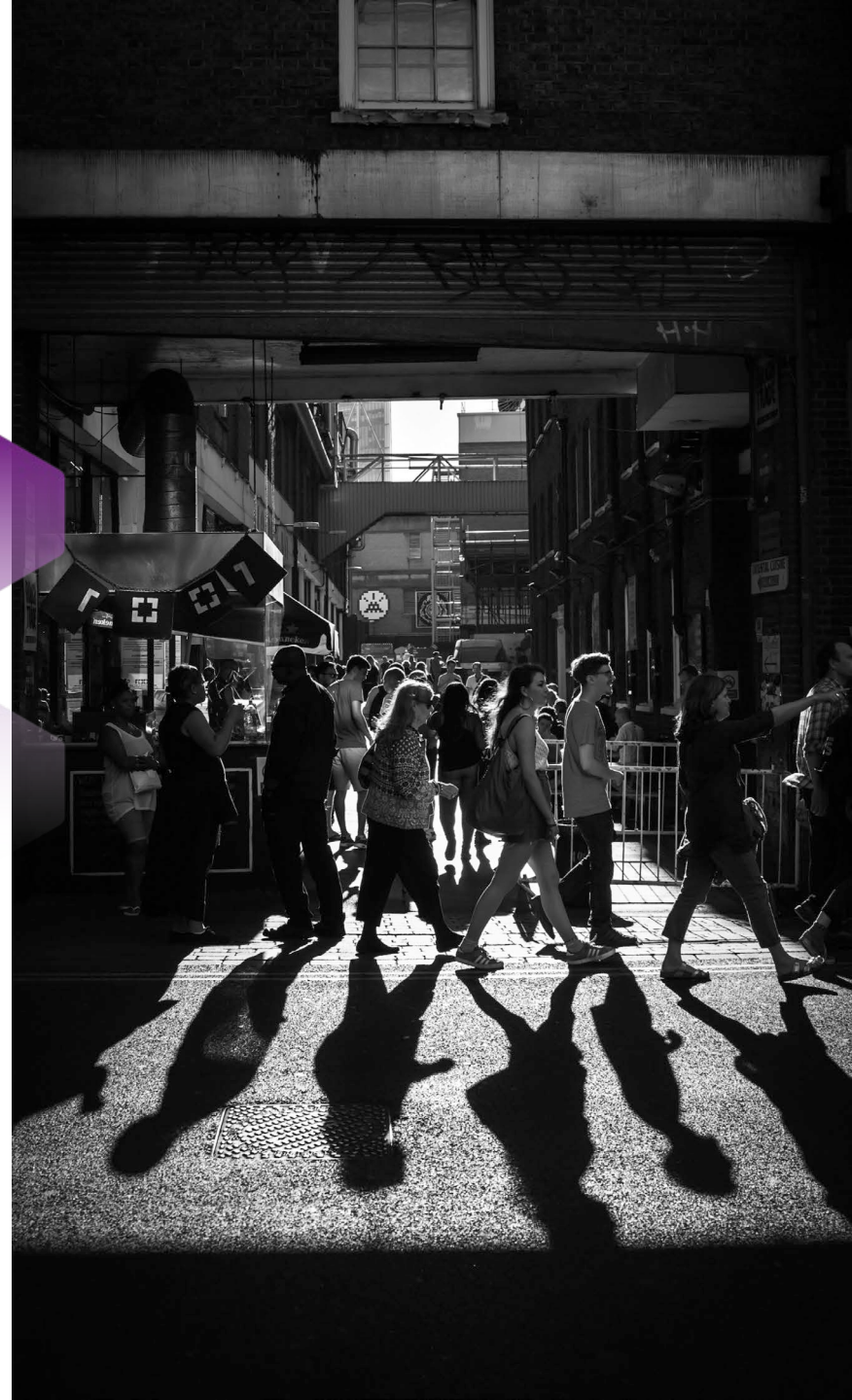
UK Data Service

Introduction to effective and practical research data management

Dr Hina Zahid

Data Producer Support Lead (ethics)



Economic and Social Research Council

Workshop format

- Presentation (slides to be provided after the event).
- Activities: Mentimeter
- Padlet Q&A

Overview

- Research data lifecycle.
- FAIR principles.
- Data management planning.
- Overview of ethical and legal considerations.
- Data security, storage and backup.
- Data curation best practices.
- Data sharing strategies.
- Q&A session.

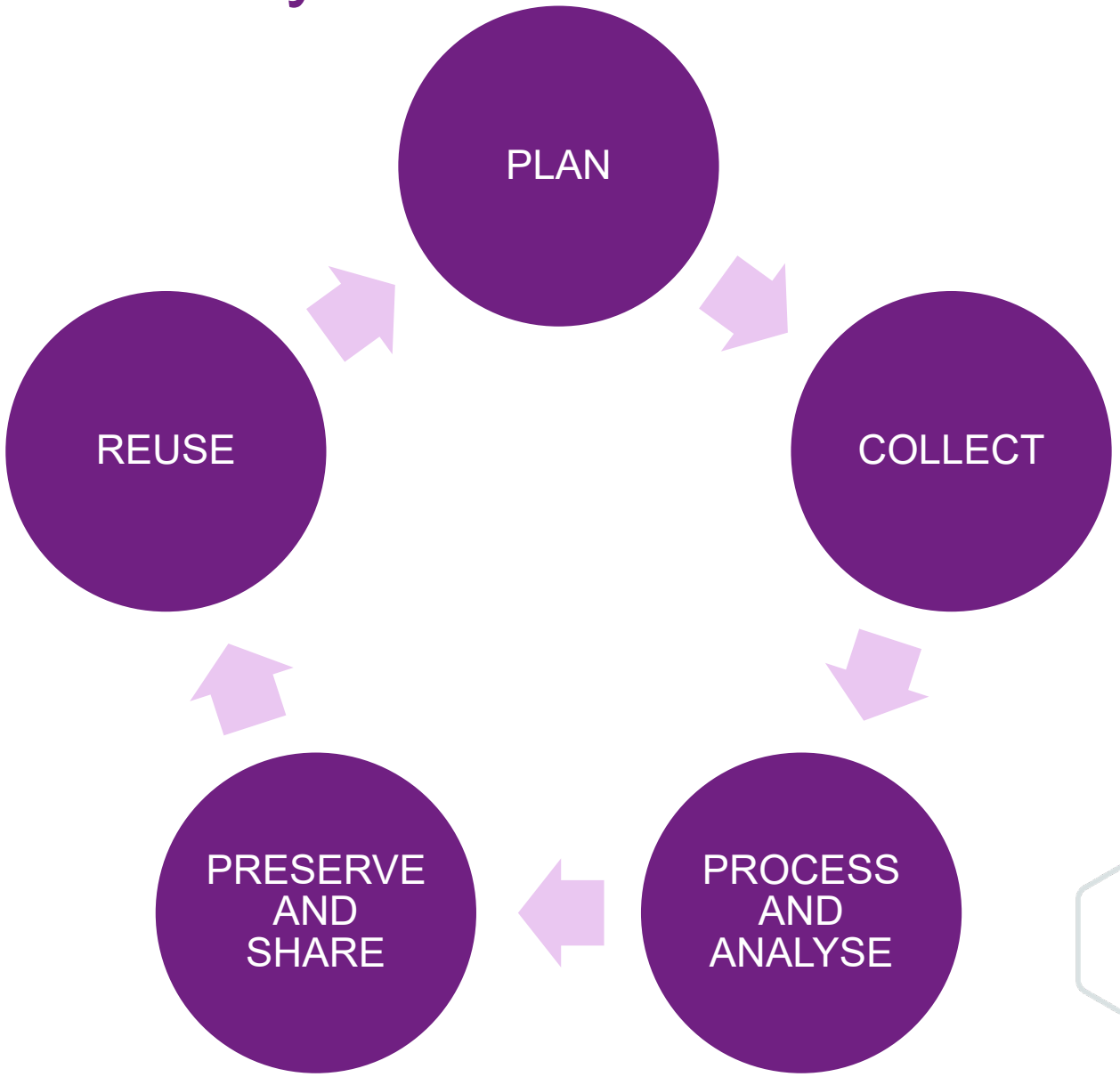
Effective research data management

Ensures data are:

- Compliant with ethical standards and applicable legislation.
- Well-organised, quality controlled and well-documented.
- Safely stored, backed up, processed and analysed.
- Responsibly archived and preserved.
- Appropriately shared for future reuse.

RDM practices safeguard the integrity of the research.

Research data lifecycle



FAIR data principles

Findable

Accessible

Interoperable

Reusable

- Published in 2016 in Scientific Data.
- Guidance to help define good data management.

Find out more about the FAIR initiative.

How FAIR aware are you?

Explore the DANS FAIR-Aware assessment tool to evaluate and improve your understanding of the FAIR principles for managing research data effectively.

Answer a series of questions to reflect on your current practices.

Receive feedback to help align your data management with FAIR principles.

FAIR-Aware Assessment Tool

How to be FAIR

Findable

- community-endorsed discovery metadata standards
- machine readable discover metadata
- unique persistent identifiers e.g. DOIs.

Accessible

- data licensing and availability statements (including restrictions)
- methods/tools to access the data
- metadata preserved indefinitely.

Interoperable

- standard vocabularies/ontologies
- standard metadata schemas.

Reuseable

- community-endorsed data licensing
- provenance information in metadata
- established data quality assurance processes
- open format files for long-term preservation.

UK Data Service

Data Management Plans (DMPs) overview

DMP topics overview

From a generalised perspective DMPs should cover:

1. Data description: new and existing data.
2. Ethical and legal considerations and compliance.
3. Curation of data: organising, formatting, and documenting.
4. Data security, storage and backup.
5. Data sharing strategies.
6. Responsibilities and resources.

We provide in-depth guidance on our website for the ESRC DMP.

Always remember to check funder requirements and that a DMP is a living document, as research evolves the DMP should be reviewed and updated as necessary.

Why is data management planning essential?

- Anticipate and prepare.
- Keep on track.
- Secure necessary resources.
- Think ahead about storage and safeguard data.
- Share FAIR data and ensure reproducibility.
- Meet funders' expectations.

DMPonline

- DMPonline, a web-based platform by the Digital Curation Centre.
- To help researchers create, review, and share DMPs.

UK Data Service

Ethical and legal considerations overview

Ethical considerations

- Maximise benefits and minimize risks.
- Voluntary participation and informed consent.
- Respect individual rights and dignity.
- Integrity and transparency.
- Clear responsibilities and independence.

Legal considerations

- Lawful data handling.
- Data minimisation and anonymisation.
- Secure storage and access controls.
- Data retention and disposal.
- Data sharing protocols.
- Fairness, transparency and accountability.
- Intellectual property rights.

How do ethical and legal considerations in data management affect the integrity and impact of research?

- Protect participants.
- Maintain and build trust.
- Enable sharing and collaboration.
- Ensure long-term impact.
- Enable appropriate risk management.

UK Data Service

Data security, storage and backup

Data security and storage

Data must be protected from unauthorised:

- Access
- Use
- Change
- Disclosure
- Destruction.

Digital back-up strategy

Control access to computers:

- Use pasphrases 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.
- Always keep personal data separate, secure and encrypted.
- 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.

Digital back-up strategy

- Backup of files.
- Regular backups help protect against **accidental** or **malicious** data loss due to:
    - human error
    - hardware failure
    - software or media faults
    - virus infection or malicious hacking
    - power failure.

Three+ copies of the data, with at least one being stored offsite.

Further information on storing data.

Data disposal

- Simply deleting files and reformatting a hard drive will not securely erase information.

- Available software to help erase files from hard disks (BCWipe, WipeFile, DeleteOnClick and Eraser for Windows platforms; and Permanent Eraser for MacOS platforms).

- Certified shredders certified should be used for destroying paper and optical media.

Further information on disposing data.

UK Data Service

Data curation: formatting organising and anonymising

File formats strategy

- What format is best suited for data creation?
- What format is best suited for data analyses and other planned uses?
- What format is best suited for long-term sustainability and sharing of data?
- Should you choose an open versus a proprietary format?
- Should the format be lossy or not?
- Is the format suitable for conversion?

Further information on formatting data.

Recommended formats

Type of data	Recommended formats
Quantitative tabular data with extensive metadata. A dataset with variable labels, code labels, and defined missing values, in addition to the matrix of data.	Proprietary formats of statistical packages e.g. SPSS (.sav), Stata (.dta), .sas7bdat. Delimited text and command ('setup') file (SPSS, Stata, SAS, etc.) containing metadata information. Some structured text or mark-up file containing metadata information, e.g. DDI XML file.
Qualitative data. Textual.	eXtensible Mark-up Language (XML) text according to an appropriate Document Type Definition (DTD) or schema (.xml). Rich Text Format (.rtf). Plain text data, ASCII (.txt).

UK Data Service recommended formats

Best practices for version control

Data management processes inevitably create a number of edits to the data and documentation.

- Identify milestone versions to keep
- Uniquely identify different versions
- Record changes
- Record relationships between items
- Track the location of files
- Regularly synchronise files
- Identify a single location for the storage of milestone and master versions.

Best practices for file naming

- Develop a file naming strategy (minimal elements version number, description of content, publication date, project number).
- Create meaningful but brief names.
- Use file names to classify types of files.
- Use dates in the format YYYY-MM-DD.
- Avoid capitalisation where possible, as some computer platforms may be case-sensitive (e.g. Unix)
- Avoid using spaces, dots and special characters (& or ? or !).
- Use hyphens (-) or underscores (_) to separate elements in a file name.
- Reserve the 3-letter file extension for application-specific codes of file format (e.g. .docx, .xlsx, .mov, .tif).
- Include versioning within file names where appropriate.
- Review file names for archived versions to ensure they do not contain any confusing, irrelevant information (e.g. versioning, misleading description).

File naming examples

File name	Meaning
pn6614_ukhls_wave2_e10_2024-04-15.sav	Project number 6614, Wave 2 of the UKHLS SPSS data 10th edition, last edited on 15 April 2024.
pn018_int127_js_v1_2024-03-02.rtf	Project number 18, transcript of interview with participant 127, conducted by JS on 2 March 2024, first version.
shes_21_dataset_documentation_v2.pdf	Scottish Health Survey 2021 dataset documentation second version.

Variable formatting and measurement levels

- Incorrect variable formatting can lead to incorrect data use.
- Important to **determine whether the data are to be treated as string or numeric.**
- Check numeric variables to ensure the measurement level is correctly defined.

Variable measurements examples

Variable	Stata	SPSS
Ethnicity	Categorical	Nominal
Annual income (banded)	Categorical	Ordinal
Marital status	Categorical	Nominal
Age (banded)	Categorical	Ordinal
Monthly income (£)	Continuous	Scale

STATA: Categorical, Continuous
SPSS: Nominal, Ordinal, Scale

Quality assurance

Check and document any changes made to your data. This will provide a history, version control and provenance trail to help quality assure your data.

Quality assurance checks may include:

- Double-checking coding of observations or responses and out-of-range values.
- Checking data completeness.
- Adding variable and value labels where appropriate.
- Statistical analyses, such as frequencies, means, ranges or clustering to detect errors and anomalous values.
- For qualitative interview data, correcting errors made during transcription.

QAMyData

An open-source tool developed by the UK Data Service to perform 'health checks' on numeric data. It automates the detection of common issues such as missing values, duplicates, outliers, and direct identifiers, ensuring your datasets are accurate and reliable.

- Identifies missing data, duplicates, outliers, and potential direct identifiers.
- Customise assessments to align with your project's specific data quality standards.
- Generates detailed summaries highlighting areas that may require attention.

Tools for qualitative and linguistic data

CLARIN-NL: A hub for linguistic data and tools integrated into the European CLARIN infrastructure; linguistic datasets, analysis tools, demonstrators, and applications.

Explore CLARIN-NL.

CLARIN Switchboard: An intuitive platform linking datasets to suitable CLARIN tools; suggests tools for tasks like tokenisation, parsing, and annotation.

Explore CLARIN Switchboard.

ELAN: A flexible tool for annotating and analysing audio and video data; multi-tiered, time-aligned annotations, ideal for working with interviews, focus groups, or observational data

Explore ELAN.

Three-prong approach to protecting participants

- **Consent:** participants must be informed about risks and benefits of *any* data sharing.
- **Anonymisation:** treat the data reducing the risk of re-identification recognising data utility will be reduced, therefore a balanced approach must be taken
- **Access:**
  - Who? How? For how long?
  - Access levels and user agreements
  - Leverage legislation and existing frameworks such as the Five Safes Framework

Useful semi-automated anonymisation tools

- sdcMicro – R package (free) – has a user-friendly interface so minimal coding skills needed.
- QAMyData - UK Data Service developed a free (GitHub) easy-to-use open source tool, that provides a health check for numeric data. The tool uses automated methods to detect and report on some of the most common problems in survey or numeric data, such as missingness, duplication, outliers and direct identifiers.
- ARX - a comprehensive open-source software for anonymizing sensitive personal data. It supports a wide variety of (1) privacy and risk models, (2) methods for transforming data and (3) methods for analysing the usefulness of output data.
- μ-Argus – developed by Statistics Netherlands; User Manual
- QuailAnon – open-source tool developed by QualiService
- Text anonymisation helper tool – Word Macros tool developed by the UK Data Service
- Textwash - open-source tool uses Python to identify and replace direct identifiers
- FAMTAFOS – in development; open-source desktop app that utilises AI technology to anonymise text at scale; operates on principle of Named Entity Recognition (NER), and can be set to search for names, locations, occupations, etc. They will then tag them for subsequent human editing.
- De-ID - HIPPA-compliant tool to flag potentially identifiable data; only available to organisations

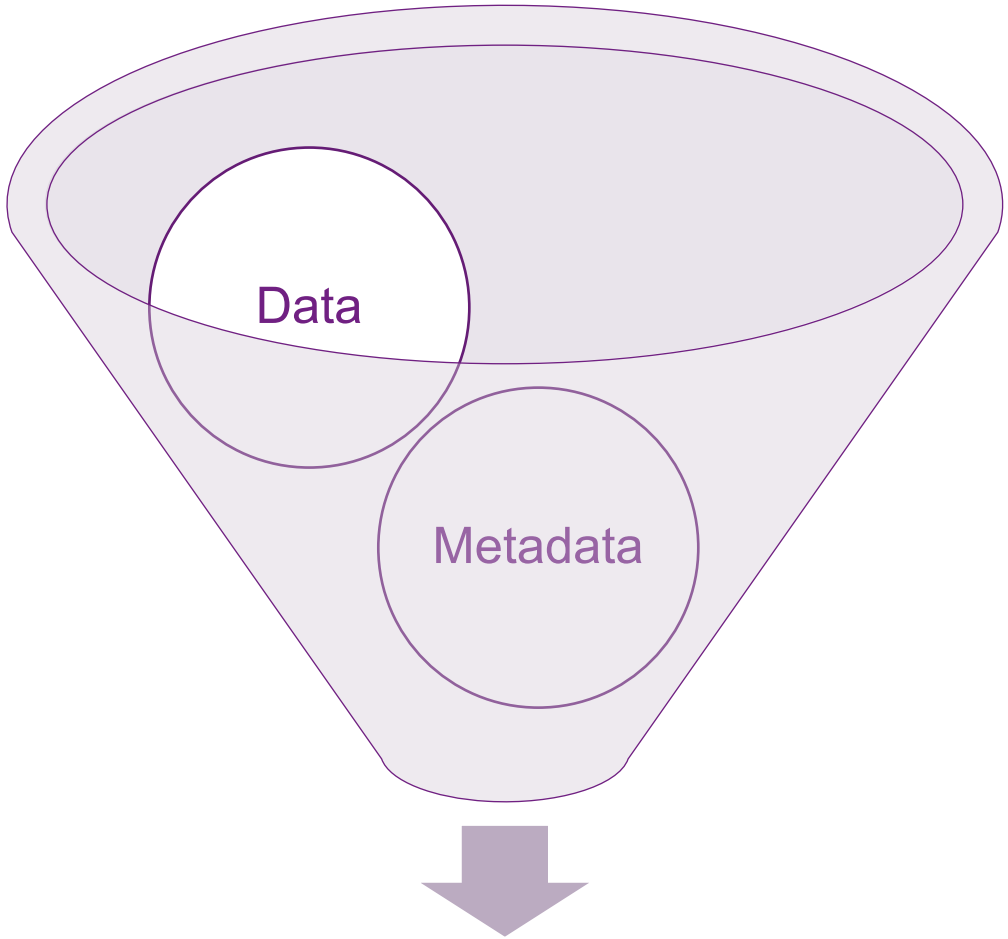
UK Data Service

Data curation: metadata and documentation

Metadata

For data to become information, you need to understand the context in which the data are situated.

Metadata is what provides this essential context.



Why is it essential to document data?

- Efficiency and accessibility.
- Ethical and accurate data reuse.
- Reproducibility and validation.
- Compliance and ethical standards.
- Long-term preservation.

Types of data documentation

Data-level documentation

- provides information on the individual data objects, such as a variable in a data file or an interview transcript. It can be embedded in the data file, such as variable or value labels in a data file, or participant information added in the header or an interview transcript.

Study-level documentation

- provides high-level information on the research context and design, the data collection methods used, any data preparations and manipulations, plus summaries of findings based on the data.

Further information on documenting data.

UK Data Service

Data sharing strategies

Data sharing strategies overview

There are various way of sharing research data including

- Domain specific repositories (data service providers, archives, centres)
- Institutional repositories
- Self-preservation and dissemination
- Commercial data sharing platforms
- Direct submission with journal publications

All data sharing methods have advantages and disadvantages.

Key consideration for data sharing strategies

- Purposes of data sharing
- Data findability, accessibility, interoperability and reusability
- Data sensitivity and confidentiality
- Ethical and legal implications
- Long-term preservation
- Security measures
- Costs and resources
- Technology and infrastructure
- Stakeholder engagement and support

Responsible repositories

- Host facilities that adhere to established standards and best practices in data management.
- Ensure the integrity, preservation, and accessibility of the data they hold.
- Provide reliable, secure, and accessible environments for storing and disseminating research data.

Source: Research Data Alliance TRUST principles

Deposit licence agreements

Protect the rights of the data owner and the repository.

Ensure that data users are aware of their rights and responsibilities.

Facilitate ethical and legal sharing and use of data, enhancing its value to the research community.

What if I use secondary data?

Always check the licence under which the data are made available.

While you might not be able to share derived data you can always share your code.

- code/syntax file are clean, well formatted and do not contain any data
- avoid including unnecessary personal information
- code/syntax file are well commented
- always include the full citation (including the persistent identifier) for the data used
- provide in-depth metadata describing the files and methods used
- provide a ReadMe/Methods document for ease of use for secondary users

Registry of data repositories

- Re3data.org is a comprehensive registry of research data repositories.
- Provides a curated index of more than 3,000 repositories worldwide.
- Covers all academic disciplines.

Source: re3data

Other tools, templates and tutorials

- Research data management learning hub
- Data management checklist
- Data management costing tool and checklist
- Model consent form and survey consent statement
- Transcription template
- Transcription instructions
- Data list template
- Data skills modules
- UKDS events

Get connected

UK Data Service

Jisc mail group

UK Data Service YouTube channel

Powerpoint slides will be available on our website in due course and you can catch up on the recording on our YouTube channel.

UK Data Service

Thank you ever so much!

datasharing@ukdataservice.ac.uk

https://beta.ukdataservice.ac.uk/help

Case Study: Balancing teens' privacy with the desire to share data

- A study involving data about teenagers had to address ethical and legal challenges to ensure participants' privacy was protected while allowing data sharing.
- Sensitive information was anonymised to protect individual identities.
- Researchers balanced privacy with the need for transparency and data reuse.
- Adhering to ethical guidelines and employing anonymisation techniques enabled appropriate data sharing without compromising participant trust.

Read more about this.

Case Study: Repurposing inspection data for research – the HMIP survey journey

- His Majesty's Inspectorate of Prisons (HMIP) collects data during inspections to monitor prison conditions and treatment.
- By repurposing this data for research, it has been possible to provide broader insights into prison reform and policy.
- Data was effectively documented, archived, and shared for secondary use.
- Researchers used the data to explore themes like institutional environments and prisoner wellbeing.
- Comprehensive metadata and proper anonymisation ensured the data's usability while respecting confidentiality.

Read more about this.

Case Study: The data life cycle of an archived qualitative study used for teaching

- A qualitative study, originally conducted to explore social behaviours, was preserved in a data archive and later repurposed as a teaching resource.
- Detailed documentation and careful curation ensured the study's relevance long after its original use.
- The archived data has been used to teach students about research methodologies, data analysis, and ethical considerations.
- Effectively managed data can serve multiple purposes, maximising its educational and research value.

Read more about this.

Even more success stories

Using Understanding Society, researchers looked at the way engaging with arts, culture and sports can lead to greater satisfaction with life.

Studies like the English Longitudinal Study of Ageing can give insight into how enjoying later life can be linked to living longer and the impact of social isolation and loneliness on mortality in older people.

The Millennium Cohort Study was used to investigate the conditions associated with parental involvement with children for policy recommendations.

Family Resources Survey to analyse migrants' experiences of poverty and to compare them with the experiences of UK-born people.