CIVICA RESEARCH

DATA MANAGEMENT PLANS FOR OPEN SOCIAL SCIENCE RESEARCH

LSE

Hannah Boroudjou, Research Data Librarian, LSE Library

Helen Porter, Research Support Services Manager, LSE Library

TALK STRUCTURE

- The benefits of writing a data management plan for researchers and for open social science
- Different types of data management plan and when you should write one
- The importance of personal data
- Data storage and security solutions over the lifetime of a project (and a research career)
- Identifying and assessing data
- Data documentation and description
- FAIR data principles
- Choosing a data repository and deciding on an access level
- Copyright and ownership
- Costing open social science and data management
- Supporting documentation

BENEFITS OF WRITING A DATA MANAGEMENT PLAN TO THE RESEARCHER

- Gets you thinking about the ins and outs of your data collection and organisation throughout your project
- Good organisation and data security will save you time during your project
- Helps you keep promises of anonymity and confidentiality with your participants
- Prevents data loss through theft, loss or hardware corruption
- Helps you plan ahead how to manage data at risk points i.e. when in the field and when travelling
- Puts you in a position to publish or archive your data, receive a DOI, share with collaborators and get cited
- Now an integral part of the grant application process for most funders. They are peer reviewed so a good one will strengthen your applications

BENEFITS OF A DATA MANAGEMENT PLAN FOR OPEN SOCIAL SCIENCE

- A DMP helps you plan how you will make your data open from the very start of the project. This is beneficial because:
    - You can plan your time more effectively and prep data for sharing as you go along.
    - Lots of projects leave data sharing till the end of the project but there is some data you may be able to share earlier
    - For larger projects you may have scope to factor costing for data sharing into the funding application
    - Good data management practice makes data easier to handle and organise, this will help you when it comes to prepping data for sharing.
    - Asks you to consider copyright and ownership issues that may prevent data sharing ahead of time
    - You can use the DMP to demonstrate your commitment to open social science to funders

WHEN SHOULD YOU WRITE ONE?

- For your funder. Most funders require a DMP, even for applications where only secondary data is being analysed.
- When you want to or need to plan for publishing or archiving your data to accompany your project or publications
- When you're collecting data from human participants i.e. most standard data collection in the social sciences
- When you're dealing with data that relates directly to individuals i.e. secure datasets that contain variables that could identify individuals
- When you want to ensure your data is kept safe i.e. backed up regularly and in secure storage
- You will likely be required to submit one to a Research Ethics Committee during your career

DIFFERENT TYPES OF DATA MANAGEMENT PLAN

- Funder templates (some have a focus on open science and others on data security)
- Institutional templates
- Iterative data management plans, when you update your plan throughout your project
- Collaborative data management plan (with accompanying data management guidance)
- DMP Online is an online tool where you can access different templates and access guidance as you write
- Most major funder templates are freely available online to view, also often have accompanying guidance or policies that can help you identify funder priorities for data management and open social science

KEY ELEMENTS OF A DATA MANAGEMENT PLAN

- Really thinking about all the data you will use, collect, generate on your project and listing it comprehensively
- Where will your data be stored at different parts of your project and stages of data analysis and manipulation?
- What security measures do you need to put in place during your project?
- How are you describing, documenting and organising your data as you go along?
- Are you following the FAIR data principles?
- Have you cleared any copyright or ownership issues in your data?
- How specifically will you publish your data – which repository and on what level of access?
- Can you use any of your funding to cover data management tasks?
- If you are working collaboratively who has responsibility for what?

PERSONAL DATA AND SPECIAL CATEGORY DATA

- Why is personal data different, why does it matter?
- Personal data is a protected category of data under GDPR. GDPR is data protection legislation that governs data handling in the UK/ EEA/ EU. It continues to apply in the UK after Brexit.
- As such, there are consequences for mishandling personal data and causing a data breach:
    - Harm to participants, possible repercussions of a breach could include; identity theft, risk of reputational damage and financial loss, imprisonment
    - Fines for the institution (possibly very significant fines).
    - Reputational damage and embarrassment to the institution and yourself
    - Loss of access to research participants, possibly for the whole research community.

PERSONAL DATA AND SPECIAL CATEGORY DATA

- Personal data is information that relates to an identified or identifiable individual (ICO definition)
- This can include direct identifiers like names and email addresses, but can also include indirect identifiers i.e. voice recordings, job titles, location etc.
- Special category data is personal data that requires more protection because it is sensitive (ICO definition)
- Examples: ethnicity, religion, trade union membership, sexuality, political affiliation etc.
- Any personal data you're working with needs to be mentioned in the plan and you need to outline a strategy for handling it.
- A Data Protection Impact Assessment (DPIA) will reassure funders you're taking your ethical commitments seriously
- Tip: The ERC are especially likely to follow up if they feel you have not defined when you're collecting personal data and how you plan to handle it.

DATA STORAGE & SECURITY

- Every university will have its own solution, please follow guidance from IT/ cyber security teams at your own institution in the first instance.
- At a minimum data storage should be GDPR compliant (i.e. located in UK/ EEA/ EU) and regularly backed up
- You will need to think about how to store anonymised data in a separate location to identifiable data and anonymisation logs
- Your data will need to be stored securely, get best practice guidance from your IT teams i.e. encryption software, screen locks, keeping anti-virus up to date. Generic guidance from your IT team can be linked in your support plan.
- Data requiring an additional level of security i.e. special/ secure license agreement data, data related to sensitive topics like terrorism etc. Will need to be discussed on an individual basis with your local cyber security/ data services team.
- Tip: Many popular solutions i.e. DropBox* and Google Drive have been investigated by our cyber security team and found not to be GDPR compliant.

DATA STORAGE AND YOUR RESEARCH CAREER

- As a researcher you will likely be moving between institutions over the course of your career, or be collaborating with researchers from other institutions
- Most institutions will revoke your access to their IT systems as soon as your contract ends. This can mean losing access to all of your research data.
- Building open science and good data archiving into your data management plan can ensure that you (as well as your collaborators and other researchers) will be able to access your research data long after specific contracts and funding runs out
- For larger projects, a data management plan can ensure that long term institutional support is in place, or that institutions collaborate with each other on how data is stored and accessed

IDENTIFYING AND ASSESSING ALL OF YOUR DATA

- Include the secondary data that you will use and that you have considered permissions associated with re-use
- Show the full range of data e.g. secondary data; scripts or codes; derived data; workshop recordings or minutes; interview guides or questions; recordings and transcripts of interviews
- In what format will the data be? Will that change as you transform the data? For example, for an interview mp3. to word to tabular data
- Consider recommended formats when you plan your data collection
- Can you estimate the volume of data e.g. number of interviews, file sizes?
- Which part of your data will include personal data?

DATA DOCUMENTATION AND DESCRIPTION

- Effective data documentation is an important part of open social science because it will ensure your data can be read and understood by any user.
- It is much easier to document your data as you go along than at the end of the project.
- Can include data indexes or catalogues, interview guides and transcription templates, anonymisation logs, codebooks, Readme files, details of file formats and standard structures used etc.
- It may be possible to share some aspects of your data documentation when you archive your data.
- Should definitely include a data index or catalogue that describes each part of the data when you deposit a dataset
- Previous CIVICA session covered this in more detail

WHY FAIR DATA IS IMPORTANT FOR OPEN SOCIAL SCIENCE

- In 2016, the '**FAIR Guiding Principles for scientific data management and stewardship'** were published in *Scientific Data*.
- Refers to data that is Findable, Accessible, Interoperable and reusable.
- Established a set of criteria for measuring commitment to open science.
- Most data management plan templates will ask how you plan to make your data FAIR. Some will refer directly (ERC) others will not address directly but will ask about data sharing, metadata and description etc (UKRI)
- Detailed breakdown of each principle at go-fair.org

FAIR DATA: FINDABLE

- First principle of FAIR is that data should be FINDABLE i.e. it should be discoverable to both human and computers.
- In practice:
  - data linked from a publication via a persistent identifier (DOI)
  - Metadata (i.e. data description) needs to be high quality and searchable without the DOI (Many data repositories include this as standard)
  - Metadata and dataset might be different files so the DOI needs to be mentioned in the metadata
  - Need to be indexed in a searchable source i.e. a data repository that can be discovered via search engines.

FAIR DATA: ACCESSIBLE

- Second principle of FAIR is that data should be ACCESSIBLE i.e. once the user has found it, they need to know how to use it.
- In practice:
  - Users should be able to retrieve data by clicking on a link i.e. manual intervention should not be needed*
  - Ideally should be viewable and downloadable by anyone with an internet connection i.e. open access
  - However, there are exceptions., i.e. when data needs to be made available on a safeguarded or restricted basis. You can still make data FAIR by choosing a repository that can make intervention as minimal as possible.
  - Metadata should exist even if data degrades or disappears over time i.e. it is still useful to know the people and institutions involved in the research

FAIR DATA: INTEROPERABLE

- Third principle of FAIR is that data should be INTEROPERABLE. This means that it should work with existing systems and be easy to integrate with other data.
- In practice:
  - Humans should be able to read data but so should machines i.e. use commonly controlled vocabularies, thesauri, etc.
  - You should mention if dataset builds on another dataset, if additional datasets are needed to complete the data or if complimentary or relevant information is stored in a different dataset. Ideally, you should be able to link to any other datasets from your own record.
  - It is important to use or convert files to interoperable formats e.g. not attached to a particular software subscription. This may not always be possible. The UKDS has a good reference list

FAIR DATA: REUSABLE

The Fourth principle of FAIR is that data should be REUSABLE. This means that data should be so well described and documented that it can be reused and replicated in other settings.

In practice:

- Should include metadata about the context the data was created within i.e. include the data scope (purpose/ question), any limitations in the dataset, the names and versions of any software used to generate the data, lab conditions (where applicable), versioning.
- The user needs to know where the data came from (origin/ history) who to cite and how you wish to be acknowledged
- Use well established file formats that others will be able to reuse. Most data repositories will have a table of accepted formats.
- Applying a licence that tells researchers under what terms your research data can be re-used

COPYRIGHT AND OWNERSHIP

- You may be asked a question along the lines 'clarify the copyright and intellectual property ownership of the data.'
- The first place to refer is your institutions intellectual property policy.
- Other stakeholders are your funder, any owners of any secondary data you may have used, other researchers on your team.
- Most researchers give a vague answer to this, nobody really likes dealing with copyright!
- This can be ok for individual researchers but can lead to problems on large collaborative projects.
- It's worth considering in advance who will own the data on the project team and negotiating this ahead of time I.e. will the person who collected the data have special ownership of it? Don't assume that everyone is on the same page!
- Consider authorship: how will you credit those involved and how will you be credited?
- You can also plan how to license the open dataset I.e. what creative commons license you use.

WHICH REPOSITORY AND WHAT ACCESS LEVEL?

- We always recommend you choose a dedicated data repository, they will supply you with a DOI and a citation for your data, can offer guidance for access levels – most of the FAIR criteria become easier to meet when you use a data repository.
- Many of you will have an institutional repository you can use (not at LSE).
- We usually recommend the UK Data Service, Zenodo, Harvard Dataverse, Github (for code).
- There are different access levels within repositories:
  - Open – as it sounds, completely open to everybody, users don't even need an account to view and download data
  - Safeguarded – users need to create an account to access data.
  - Restricted – data access via application for special license only. Probably requires a secure access route and an organisational signature.

COSTING

The following costs can be included in research grants and funding bids (although check with your specific funder)

- Assistance with data management activities e.g. preparation for data deposit
- A data management role
- Additional storage costs
- Any software or equipment required for data management
- Data Management training or consultancy
- Transcription or translation services

Your Library, Research Division or IT division should be able to help you identify realistic costs to include

ROLES AND RESPONSIBILITIES

Some plans will ask you to include roles and responsibilities. This is because it is a key factor in data integrity, data security and data sharing:

- Have you developed a workflow to show who is involved and how e.g. in data collection, writing code, developing databases, monitoring data standards
- Is a single person responsible for data storage and security? (everyone should see the DMP regardless!)
- Who is responsible for maintaining standards and quality control I.e. file naming, file formats – can get out of control in a shared space
- Processes for sharing data with paid consultants and collaborators.
- Who can monitor who can access what data? Making sure those who have left can't access data, restricting data access on a 'need to know' basis.
- Who will be responsible for authorising access requests from a repository where dataset access has been restricted.

SUPPORTING DOCUMENTATION

- Some plans will give you space at the end to provide any supporting documentation.
- Very often this is left blank. However, you might want to consider:
  - University IP policy
  - Password policy
  - Any device security policies
  - Data protection policy
  - Data management policy
  - Policy for working with any external collaborators/ consultants outside the UK/ EU/ EEA
  - Any other local policies that may be relevant

COULD YOU SHARE YOUR DATA MANAGEMENT PLAN?

- Research ideas and outcomes journal shares data management plans
- Can also be deposited with the dataset – people want to see good examples of data management plans!

REFERENCES & USEFUL LINKS

From the session:

- UKDS recommend formats
- Previous CIVICA sessions
- FAIR data principles from go-fair.org

Useful links:

- LSE Webpage Research Data Management & Open Data
- UK Data Service YouTube channel (lots of data management skills workshops).
- UK Data Service data management guidance
- UKRI guide to publishing data (including data management policies for each council)
- LSE Research Data Toolkit (LSE staff and students only – CC version is being prepped to go on Zenodo)

CIVICA RESEARCH

Thanks for listening

DATA MANAGEMENT PLANS FOR OPEN SOCIAL SCIENCE RESEARCH

LSE

Hannah Boroudjou, Research Data Librarian, LSE Library

Helen Porter, Research Support Services Manager, LSE Library