ESPA Working Paper Series

No: 001 / December 2015

ISSN 2058-9875

Sharing social data in multidisciplinary, multi-stakeholder research

Best practice guide for researchers

Authors: Veerle Van den Eynden, Kate Schreckenberg, Louise Corti

Acknowledgement

This guidance is based on current research experiences from the Ecosystems Services for Poverty Alleviation programme[1] (ESPA), a seven year research programme funded by DFID, ESRC and NERC to provide new knowledge on how ecosystem services can reduce poverty and enhance well-being for the world's poor; as well as the expertise of the UK Data Service (UKDS) in curating, disseminating and providing access to a wide variety of social sciences research data for research purposes, in particular ongoing efforts to make available household and health data from the DFID-funded Millennium Villages in Northern Ghana Impact Evaluation[2] project and efforts to archive and share human rights data. The development of this guidance evolved from the ESPA Social Surveys Event[3] held in October 2014, where the challenges of archiving data from social surveys to maximise the opportunities for use by future researchers were discussed, followed up by further discussions between ESPA and UKDS.

Many people contributed towards the writing of this guide, by reviewing and commenting on draft versions, making useful suggestions and providing examples as case studies. We thank in particular Chris Barnett (Itad), Libby Bishop (UK Data Service), Gisella Susana Cruz Garcia (CIAT), Nicole Gross-Camp (UEA), Lynne Henderson (DFID Ghana), Katherine Homewood (UCL), Mahesh Poudyal (Bangor University), Rick Stuart (Environmental Information Data Centre), Carlos Torres Vitolas (University of Southampton), Paul van Gardingen (ESPA) and Simon Willcock (University of Southampton).

This document has been produced by the Directorate of the Ecosystem Services for Poverty Alleviation (ESPA) Programme in partnership with the UK Data Service (UKDS). ESPA is a programme funded by the Department for International Development (DFID), Economic and Social Research Council (ESRC) and Natural Environment Research Council (NERC).

The ESPA Directorate is hosted by Research Into Results Limited, a wholly-owned subsidiary company of the University of Edinburgh, responsible for the delivery of research and project management services in the area of international development. The UK Data Service is a national data service that provides research access to a range of social and economic data, funded by ESRC.

Abstract

There is an increasing drive for openness and sharing of data, with funders and other stakeholders expecting publicly-funded data to be available for further use. Science benefits from data being maximally available as a resource for new and future research and technological advances make it easier for digital information and data to be discoverable and accessible to a very wide audience. Equally, sharing information and data amongst stakeholders is fundamental in collaborative and multi-stakeholder projects. Yet sharing data collected from human participants (e.g. through surveys, questionnaires, interviews, focus groups, participatory methods, video) can present ethical challenges as they often contain personal or confidential information. In multi-disciplinary projects, the collection of social data in conjunction with geospatial information may make it very difficult or impossible to conceal the identity of participants or fieldwork locations. Appropriate procedures are needed to maximise opportunities for future use. This guide seeks to support researchers, consultants and evaluators in sharing their data widely by highlighting key considerations and providing helpful tips, from the planning stages of research and evaluation through to the possible deposit of data with a data repository.

.

Introduction

This guide is for researchers, consultants and evaluators collecting data through research that involves people as participants (human subjects), and where sharing of those data will be desirable or required. Methodologies such as surveys, questionnaires, interviews, focus groups,

or participatory methods in social science and multidisciplinary research capture qualitative or quantitative data through audio and video recordings, photography, in writing or in numerical format. The resulting data will often contain personal or confidential information and the identity of participants or fieldwork locations may be difficult or impossible to conceal.

There is an increasing drive for openness and sharing of data, with funders and other stakeholders expecting publicly-funded data to be available for further use. Journals and publishers may expect data to be at hand as evidence for published findings and as proof of research integrity. Science benefits from data being maximally available as a resource for new and future research and technological advances make it easier for digital information and data to be discoverable and accessible to a very wide audience. Equally, sharing information and data amongst stakeholders is fundamental in collaborative and multi-stakeholder projects.

While some data collected from human participants can pose challenges for sharing in an ethically sound way, appropriate procedures can maximise opportunities for future use.

Challenges may arise from the complex and multidisciplinary nature of projects, such as those involving the collection of social data in conjunction with geospatial information (e.g. on natural resource use), or the collection of sensitive social or health data, which may pose particular challenges for data sharing and confidentiality. The multi-stakeholder nature of many projects which may include natural resource managers, government officials, donors and other non-academic stakeholders as project partners can bring further complexities. This can necessitate intricate data-sharing arrangements even within the project partnership if, for example, collected data describe potentially illegal or security-related activities of individuals that cannot be disclosed to certain partners. Projects working in multiple countries and cultures may also face challenges in meeting different social and legal expectations for data sharing.

The guide outlines key recommendations, from the planning stages of research and evaluation through to the possible deposit of data with a data repository. Maximising re-use value depends on providing sufficient annotation and documentation to make data independently understandable and the need to maintain confidentiality of participants.

Using this guide

In writing this document we recognise that research practices and fieldwork circumstances vary widely across research domains and geographic situations. The guide sets out best practices in a concise way, providing examples from different contexts. Readers can find extensive detailed guidance, templates and examples on the UK Data Service Manage Date[4] website, and may have to adapt practices to fit the context of their particular research scenarios.

WHAT TO DO BEFORE FIELDWORK BEGINS

Plan for sharing

In research with people as participants, plans for sharing and providing access to data should be considered from project design, when commissioning projects or during ethical review. Measures needed can apply to the sharing of research data within the project partnership or the future sharing with other researchers / stakeholders. Discuss and agree measures in advance, as it may strongly influence the protocols for data collection (containing personal information or not), for data storage, what needs to be discussed and agreed with participants as well as with stakeholders in the research.

Consider:

- whether participants are happy for the information they contribute to be attributed to them, or would prefer their identity be kept hidden
- whether research will touch on sensitive personal topics, health-related issues or illegal activities that participants do not want to be widely known
- how informed consent for data use and sharing can be discussed and gained with participants
- how to protect the identity of participants by not collecting personal information in the first place (e.g. using pseudonyms during data collection), by keeping files with respondent identities and codes separate from the data files, or through redaction or anonymisation of data
- how to regulate access to part or all of the data files where necessary, for example by a membership group, by the purpose of the use or for limited time periods
- how to ensure secure storage and transfer of data where needed, through physical protection (locked cabinets, secured offices), security of computer systems (firewalls, password protection, system access controls) and file-level protection (encryption, password protection)
- which expectations or policies from funders, publishers, host organisations or team members apply to access and sharing of data
- contacting a data centre to plan long-term preservation and access to the data
- making a plan for data sharing to help clarify these issues for all stakeholders involved

And take action:

- train enumerators and translators to ensure full awareness of ethics and the need for confidentiality
- use signed confidentiality agreements for team members
- include data access controls in tendering agreements
- put in place technical solutions, such as encrypting hard disks, establishing secure systems for data transfer, or firewalls across a project team
- allocate appropriate resources
- define who has responsibility for each set of activities
- nominate a dedicated person responsible to oversee all confidentiality measures, ensuring the entire team adheres to instructions when data is shared within the project team

If identities are to be kept confidential, the risks of disclosure need to be considered before, during and after the data are collected. Risks may occur during data collection if participants reveal information that may be used against them. For example, in the context of research on people's use of the environment, there is a strong likelihood that data will be collected on uses that may be considered illegal. Protecting the identity of research participants is therefore of paramount importance.

In multi-stakeholder projects, it may be necessary to have project-internal data firewalls to ensure that information cannot be traced back to specific individuals or even groups of participants, where revealing that information to certain stakeholders can pose a risk to participants. This may mean thinking carefully about data storage (e.g. if using shared project folders) during the project's lifetime. Consideration may also be needed of how biophysical data are likely to be used in project publications and whether spatially explicit information could be linked to social data from the same projects.

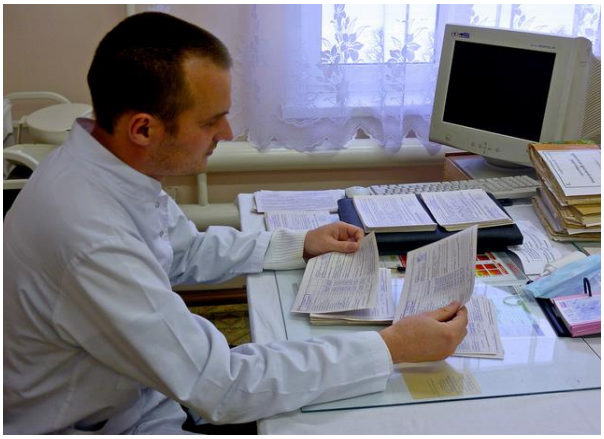
Case study: Madagascar p4ges project[5]

The p4ges project in Madagascar investigates whether paying for global ecosystem services can reduce poverty in a newly created protected area (CAZ). The organisation managing this protected area is a key research partner in the p4ges project, as this is an effective way for research to have impact.

This can also be potentially problematic, especially where research delves into the activities of poor local households, such as clearing forest for swidden agriculture, that are against the law or against protected area rules. The project was set up in such a way that members of the CAZ management organisation are not directly involved in any data collection concerning potentially illegal or banned activities. But to achieve impact, information must be easily shared among the project partners (and with other stakeholders). P4ges have implemented a framework for data management with strict internal firewalls to limit who has access to certain information generated through the social surveys. For example, all information that can be used to identify specific individuals, such as personal identifiers and GPS locations of the households, are removed and the data anonymised before sharing within the project. To allow for repeat surveys to be carried out, the identifier information from each survey is separated from other data. The keys linking the anonymised data and the corresponding personal information are stored securely and separately with access only to the members of the team involved in gathering and managing the data.

More complex is archiving and sharing the data in the long-term, providing transparency about the research and making the data accessible for future use, without the risk of repercussions to individuals. P4ges will produce data for sharing with the main socio-economic data separated from the personal identifier data; with the latter placed under restrictive access requirements to allow future repeat studies.

Assessing disclosure risk



Dimitry Kirillov/World Bank

Assessing risk is done by evaluating key characteristics or variables in data files that are the most risky for leading to participant identification in a specific project. These can be direct identifiers (e.g. a person's name, picture or detailed geographic location), indirect identifiers (e.g. extreme household size, specialised profession, unusual health conditions), or a combination of various characteristics that jointly result in disclosure. Disclosure risk analysis can involve running frequency analyses of variables to determine low-frequency responses and extreme outliers. This needs to be complemented by qualitative analysis of risk characteristics based on local knowledge of field work, its design and the population and individuals studied. A single house with glazed windows in a small rural village may be highly disclosive if this is not a common feature.

Risks to participants remain once data has been archived as users may be able to infer the identity of participants from the information provided. It is important to decide whether confidentiality is sufficiently served by concealing individual and household identities, or whether community, district or other location-specific identifiers also need to be concealed. Risk may, for example, apply to a whole community if data reveals that community members engage in illegal activities such as hunting or are in conflict with neighbouring communities or indeed local authorities or private companies (such as mines).

Anonymisation procedures

Anonymisation is a valuable tool that allows data to be shared, whilst preserving privacy. The process of anonymising data requires that identifiers are altered in some way such as being suppressed, substituted, distorted, generalised or aggregated.

A cautionary note is that the process of anonymisation may impact on the usefulness of data. Removing key variables in quantitative data files, applying pseudonyms, generalising and removing contextual information from textual files, and blurring image or video data could result in important details being missed or incorrect inferences being made. In this case, overprotection of data can be undesirable.

Examples from the UK Anonymisation Network Guidance Report[6]

Example 1. Consider the attributes age and marital status. At first glance these attributes are not obvious identifiers but let us imagine a case where one of the respondents in our dataset is a sixteen year old widow. Our implicit demographic knowledge tells us that this is a rare combination which means that if we were to publish this information for this respondent, then she could potentially be identified.

Example 2. Consider the attribute gender, which is not an obvious identifier. Let us imagine a case where we have a dataset in which there is only one female respondent. The gender attribute then would be identifying for this female respondent.

In such cases, the data could be anonymised. In example 1, the respondent age could be grouped into age interval groups (e.g. blocks of 10 years) or data could be base-coded (e.g. all ages below 20 become a single group whilst other ages remain as they are).

However, if anonymisation reduces the usefulness of data (e.g. for researchers interested in gender), then it may be better to control who can access the data (e.g. only researchers) and under which conditions (e.g. sign a data use agreement).

D McCourtie/World Bank

Case study: Disclosure review of household survey data for the Millennium Villages Impact Evaluation project[7], northern Ghana.

In order to make microdata collected as part of the Northern Ghana Millennium Villages project evaluation publicly available via the UK Data Service, to encourage research and ensure openness and transparency, all variables in the household survey were assessed for disclosure risk, with recommendations for action. Table 1 shows examples of some assessed variables, with risk and actions taken. It shows examples of variables commonly assessed for disclosure risk (age, community) as well as variables for which local knowledge is essential to indicate risk (fuel type use, house wall material).

Table 1. Extract of household survey variables assessed for disclosure risk

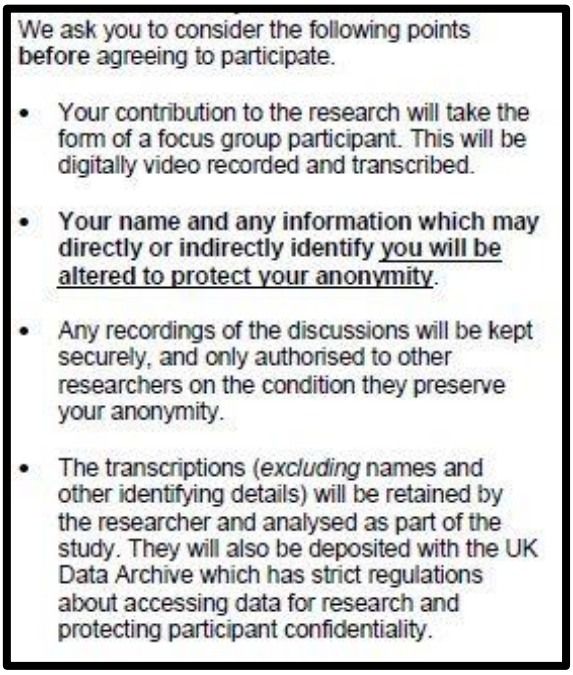
Variables	Disclosure risk	Action
Community	Low frequency counts for all named communities, respondents who gave answers very easily identifiable (especially in combination with other variables)	Exclude variable from dataset
Age	Low counts of older respondents over 75 years old	Top-code age >= 75 as '75 and over'
Main occupation during last 12 months	Low counts of very specific occupations	Occupations aggregated into standard occupation codes
Ethnicity of the Household Head	Low counts of specific ethnicities.	Recode the low-frequency responses (all responses but 'Mamprusi' and 'Builsa') into 'Other'.
Household's primary type or energy/fuel used for cooking	Very low counts for 'Gas/LPG' and 'Electricity-solar panel' responses may lead to household identification (especially if combined with other datasets)	Recode all responses into the following main categories: 1 - 'Firewood'; 2 - 'Electricity-based'; 3 - 'Charcoal'; 4 - 'Other', 5 - 'Don't know'; 6 - 'NA/missing'.
Main material of the wall of the house	A number of low-frequency responses; exterior features (households/buildings easily identifiable)	As the main material of the wall refers to the exterior of a building, it may be advisable to recode the low-frequency and 'Other' variables into 'Other (incl. wood-based and stone-based') and retain the remaining groups
Crops grown on plots	A number of low-frequency specific responses for each variable	Variables are recoded into crop categories

Consent procedures

Discussions with participants should not only consider the terms of their participation and the use of the information gathered for the primary research purposes, but also implications of sharing data more widely. Potential future uses other than by the primary research team and the nature of publishing data in a reputable repository should be noted. Consider this early on during ethical review and when planning consent procedures and developing information and consent forms.

Information on plans for sharing and archiving data for the longer term can be provided:

- on an information sheet
- when seeking consent, explaining how people's confidentiality may be maintained by, for example, anonymising data. Whether consent is gained in writing or verbally will depend on the nature of the research and the cultural context.


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.

Figure 1. Example wording in a consent form

Gaining consent for participation in the research, for the primary use of the data in research outputs and for future sharing of data, can be a one-time event or can be phased using process consent. Since it may be difficult for participants to understand when research starts what kind of information will actually be produced, it may be more effective to gain permission for data to be archived after the data have been collected, e.g. for qualitative research, after interviews have taken place and been transcribed. A transcript destined for sharing can even be shared with participants for their approval.

For surveys or other methods where coded data are captured and personal administrative information will not be made available, it is usually made clear that consent for sharing is implied from participation. The destination repository for the shared data can be mentioned too.

Example consent statement for a survey or questionnaire

The information provided by you in this questionnaire will be used for research purposes. It will not be used in a manner which would allow identification of your individual responses.
Anonymised research data will be archived at the UK Data Archive in order to make them available to other researchers in line with current data sharing practices.

See also: http://ukdataservice.ac.uk/manage-data/legal-ethical/consent-data-sharing/surveys.aspx

Extract from the UK Data Service model consent form[8] for qualitative research, showing various options for data sharing:

Use of the information I provide for this project only	**Yes**	**No**
I understand my personal details such as phone number and address will not be revealed to people outside the project.	☐	☐
I understand that my words may be quoted in publications, reports, web pages, and other research outputs.	☐	☐
*Please choose **one** of the following two options:*		
I would like my real name used in the above	☐	
I would **not** like my real name to be used in the above.	☐	
Use of the information I provide beyond this project		
I agree for the data I provide to be archived at the UK Data Archive.	☐	☐
I understand that other researchers will have access to this data only if they agree to preserve the confidentiality of the information as requested in this form.	☐	☐
I understand that other researchers may use my words in publications, reports, web pages, and other research outputs, only if they agree to preserve the confidentiality of the information as requested in this form.	☐	☐

Case study: Consent for participatory video in research

ESPA fellow Nicole Gross-Camp is utilising participatory video to help capture local perceptions of wellbeing in rural communities of Tanzania with and without community-based forestry. From inception to film production, the community largely controls the process and retains the ownership of the video, and the researcher's use of the video hinges entirely on the permissions of the community. Gross-Camp's project has created three films in three communities, with each granting her permissions to use the video for non-commercial purposes as well as placement on her institution's website. Overarching considerations of consent in the creation of a participatory video are: (1) consent to participate in the video creation process, (2) consent to be filmed, and (3) consent to distribute the created film. Consent discussions are taken in steps and ideally discussed as a group, but with individual participants feeling secure about each part of the process. The distribution of the film can only be discussed when its content has been agreed and a solid draft of the final film created and approved by the larger community in a public screening. In Gross-Camp's experience, discussions of consent occupy a large percentage of the process, particularly that of distribution.

Challenges of linked data

Additional challenges exist where there is the potential for a well anonymised survey dataset to be linked to public information sources that could still pose a risk for disclosure. This could occur in the case of multidisciplinary research when social and environmental data are held in different data repositories and spatial identifiers of the environmental data need to be kept (else making the data useless).

What other information could be available and linked to an anonymised dataset will be defined by how the data are shared. There are a range of ways in which data can be shared for example through secure access protocols (data labs and remote access), restrictions via licensing agreements or openly. The risk of deanonymisation can be limited by managing the data access environment so that sensitive data can only be shared in a secure and controlled environment whilst non-sensitive and less detailed anonymised data can be shared in a less controlled environment. Protective factors against identification (accidental or purposeful) include training and accreditation on issues such as data management, data storage and security, and ethics, consent and confidentiality.

Secure storage and transfer of data files

Data that contain personal information should be treated with higher levels of security than data which do not. Security can be made easier by:

- separating data content according to security needs (e.g. store participant names and addresses separately from survey files)
- encrypting data containing personal information before they are stored or transmitted

Secure storage of data may need:

- physical protection such as locked cabinets or secured offices
- network security by not storing sensitive data on networked PCs

- security of computer systems through firewalls, password protection, encrypted machines, system access controls
- file or folder-level protection through encryption and password protection

Cloud-based storage such as Google Drive, Dropbox, OneDrive, iCloud or YouSendIt are easy to use, but not necessarily permanent or secure. Such storage is often based overseas and therefore not covered by UK law, e.g. in violation of the UK Data Protection Act 1998, which states that personal and sensitive data should not be transferred to other countries without adequate protection.

Cloud data storage should be avoided for high-risk information, such as files that contain personal or sensitive information, information that is covered by law or that has a very high intellectual property value. While file encryption can be used to safeguard data files to a certain degree, it would still not meet the requirements of data protection legislation.

Alternatives are secure FTP servers or content management systems set up and controlled by an institution; or secure workspaces that come into existence (Basecamp, huddle).

At the same time, the practical aspects of sharing information in large multi-partner projects with partners in various countries and jurisdictions, as well as with varying degrees of connectivity needs to be considered. Accessing secure FTP services remotely may be more complex than using email or Dropbox when connectivity is limited. Here, removing sensitive information upfront or encryption can provide extra safeguards.

Finally, researchers working overseas and transferring data to another country for analysis need to act in accordance with the relevant Data Protection legislation in the countries where they collect the data.

DURING AND AFTER FIELDWORK

Preparing data files

When preparing data for sharing and for deposit and publishing in a data repository, the following measures are recommended:

- use consistent and meaningful file naming that reflects the file content, avoiding spaces and special characters, e.g. ProjectX_HHsurveydata_Phase1.csv
- check that the level of detail included in the data is suitable for the agreed access arrangements and licensing
- use open and standard file formats for long-term validity (Table 2)[9]
- if converting data across file formats, check that no data or internal metadata have been lost or changed
- for data collections that consist of multiple files, group your data files in folders / bundles according to their content or file format, e.g. a zip bundle of interview transcripts, a zip bundle of audio files

Table 2. Recommended file formats for common data types[9]

Type of data	Recommended formats	Acceptable formats
Tabular data with extensive metadata variable labels, code labels, and defined missing values	SPSS portable format delimited text and command file (SPSS, Stata, SAS, etc.)	proprietary formats of statistical packages: SPSS, Stata, MS Access
Tabular data with minimal metadata column headings, variable names	comma-separated values tab-delimited file delimited text with SQL data definition statements	delimited text MS Excel, MS Access, dBase, OpenDocument Spreadsheet
Textual data	Rich Text Format plain text, ASCII eXtensible Mark-up Language	Hypertext Mark-up Language MS Word NUD*IST, NVivo and ATLAS.ti
Image data	TIFF 6.0 uncompressed	JPEG TIFF RAW image format Photoshop files Adobe Portable Document Format
Audio data	Free Lossless Audio Codec	MPEG-1 Audio Layer 3 (mp3) Audio Interchange File Format Waveform Audio Format
Video data	MPEG-4 (mp4) motion JPEG 2000	

Preparing quantitative data files (surveys)

- use meaningful and self-explanatory variable names, codes and abbreviations
- ensure variable and value labels are complete and consistent, and included in the data file or a code book
- use standard measurement units where possible and explain context-specific units, e.g. ‘bundles’ of firewood or ‘bags’ of charcoal
- ensure internal consistency checks are done

- remove your own temporary, administrative or dummy variables created for internal purposes
- ensure no repetition of variables, especially redundancy in derived variables
- apply an appropriate level of anonymisation where necessary, for example
    - remove administrative information such as names and addresses
    - reduce the precision of potentially disclosive variables
    - apply top and base coding to hide outliers
- check that any textual variables included are suitable for dissemination, e.g. no disclosive information or internal comments in free-text variables
- ensure consistent treatment and labelling of missing values
- include weights as variables but do not apply them in the deposited data files

Preparing qualitative data files

Much qualitative data takes the form of audio recordings that are transcribed, for example, interviews, focus groups and some observations. Good practice calls for researchers to:

- apply an appropriate level of anonymisation:
    - use pseudonyms instead of names
    - replace detailed location information,
- remove names and disclosive information hidden in 'file properties', e.g. using MS Word Document Inspector
- beware of hidden tracked changes in text files
- apply a uniform transcription format to interview transcripts
- note that transcripts will be archived in their original language, with a data repository possibly unable to check these for disclosure risk
- include basic contextual information in each transcript: date and place of interview, names/pseudonyms of interviewer and interviewee,….

Note that ethnographic, observational and some other qualitative data may lose their value for future researchers without the necessary contextual and individual-level information that may typically be removed during anonymisation. In such cases confidentiality is better assured through carefully worded access conditions (see later).

For audio-visual data such as photos, videos or participatory diagrams, anonymisation is typically very costly, difficult or impossible. Again confidentiality may be better assured through access conditions.

Transcript extracts illustrating anonymisation adapted from Richardson et al (2013)[10]

Returnee Women Interview 32

Interviewer = I
Respondent = R

[Interview preamble removed]

I: Sailiji [pseudonym], are you ready to share some of your experience to me?

R: Yes, I'm ready.

I: What are you doing these days?

R: I'm staying at home and involve in agriculture.

I: Are you looking for any work?

R: I'm not looking for any work but when I came to WOSCC [Women's Skill Creation Centre] and I attended a meeting on income generation in which [staff member] mentioned about the ways for income generation. Now, I'm interested to get involve in income generation activities.

I: What do you like to do?

R: I'm busy at home and agriculture.

I: What do you want to do for income generation?

R: For income generation…we've started mushroom farming…

/////

I: Do you think you can get the work in [town in Makwanpur District] if you look for it?

R: I can…WOSCC.

I: Here in factories and industries in [town in Makwanpur District]?

R: I may get it if I look for…um…don't know exactly…but don't feel looking for.

I: Do women like you who returned from abroad get the work easily on their return?

R: I don't know this and I don't ask for it. Beside WOSCC, I go nowhere since I returned from abroad.

I: Why?

R: Earlier also I was busy in agriculture and same is the situation after I returned.

I: Okay.

R: I went abroad with an aim to earn and buy land here because we did have little land before. My husband had also earned and collected some amount during the time I was in abroad work. We also sold the land we had in Kulekhani area. We, by investing the money we collected from our work or land bought lands which we think is enough for our sustenance. I think this would be enough to sustain my life.

////

Preparing data documentation

Data documentation should give future researchers sufficient information to be able to understand and reuse the data. Consider what kind of research, data, and other contextual documentation, can explain what the data mean, how they were collected and which methods were used to create them.

Julio Pantoja/World Bank

Documentation for a data collection typically includes:

- a ReadMe file for the data collection, with:
  - for each filename a short description of what data the file includes
  - any relationships between the data files
  - for tabular data, definitions of column headings and row labels (variables and records), data codes (including missing data) and measurement units
  - for textual data, a data list of all interviews, focus groups, etc. (Table 3)
- clear variable descriptions / definitions and code labels in each data file
- syntax for derived variables
- a questionnaire form or data dictionary for surveys
- interview schedules and question lists for interviews or focus groups
- methods descriptions, including sampling, fieldwork methods
- consent forms and information sheets
- references to reports, publications and places where linked data (e.g. biophysical data) are stored
- information about any known errors in the data

Example of an existing ReadMe file for a collection in the ReShare[11] repository

ESRC Project (RES-360-25-0062)
Title: Governing 'New Social Risks': The Case of Recent Child Policies in European Welfare States
The study focused on parenting support as policy and practice in England

Files uploaded to the archive comprise a) accompanying documentation and b) 39 interviews (primary data)

Accompanying Documentation as follows:
Study Outline: A 1 page overview of the purpose and nature of the study that was circulated to prospective participants and interested colleagues.
Research Design and Methodology: A 7 page description of research questions, approach and methodology.
Consent Form: A 2 page document containing information provided to participants and the consent form signed (as approved by the University of Oxford's Ethical procedures: CUREC)
Interview Guide: A 15 page document containing the guides used when interviewing the 3 respective groups of respondents.
Classification sheet: An excel spreadsheet listing all interviews with details of interviewees (anonymised)

Primary data consists of 39 interviews (16 with service providers, 14 with experts, 9 with decision-makers)
Filenames reflect the respondent group, for example:
interview1LocalDM1 is an interview with a local decision maker
interview10ExpertE1 is an interview with an expert
interview23ProviderP1 is an interview with a service provider

Table 3. Example data list for a collection of qualitative data items, adapted from Russell (2015)[12]

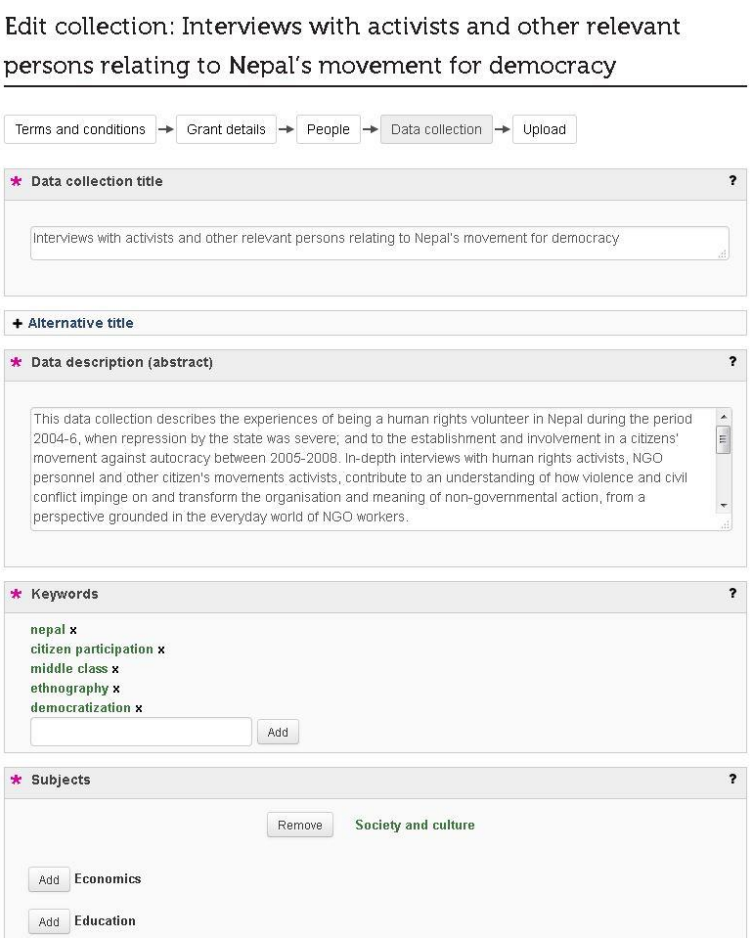
Interview ID	Pseudonym	Age	Gender	Occupation	Country	Place of Interview	Date of Interview	No of Pages	Text File Name
4929	Bridge	42 years	Male	Transporter	Uganda	Home	28/06/2011	37	4929_JK_Bridge
756	Fred	47 years	Male	Casual Labourer	Uganda	Health centre	25/07/2011	31	756_JK_Fred
632	Jacob	32 years	Male	Fisherman	Uganda	Home	01/06/2011	49	632_JK_Jacob
4718	Vincent	74 years	Male	None	Uganda	Home	23/06/2011	46	4718_JK_Vincent
34	Davis	43 years	Male	Farming	Uganda	Home	15/04/2011	38	34_RM_Davis
5208	Matthew	51 years	Male	Security Guard	Uganda	Workplace	25/07/2011	33	5208_RM_Matthew
918	Tom	44 years	Male	Guest House attendant	Uganda	Workplace	05/04/2011	30	918_RM_Tom
4218	Happy	46 years	Female	Bar business owner	Uganda	Workplace	26/07/2011	16	4218_SN_Happy
4321	Naome	26 years	Female	Municipal Council cleaner	Uganda	Home	26/06/2011	46	4321_RN_Naome
1104	Julie	37 years	Female	Farmer	Uganda	Health centre	29/06/2011	53	1104_SN_Julie
58	Gloria	29 years	Female	unemployed	Uganda	Various	01/06/2011	42	58_SN_Gloria

Data collection record: metadata

When depositing the data with a data repository, structured information should be provided about the data collection via a form. This typically includes:

- title, abstract, keywords, temporal and spatial coverage of the data
- who created the data
- the data collection methods
- access conditions
- related project information
- documentation describing the data files themselves

To make such metadata entry easy and structured, controlled vocabulary lists are used, and some information is harvested from existing external research knowledge systems.



Providing metadata

The online ReShare data submission uses metadata from the Data Documentation Initiative[13] (DDI), an international XML-based descriptive metadata standard for research data used by many social science data archives across the world.

DISCUSS DEPOSIT WITH A DATA REPOSITORY

In the UK, social data resulting from Economic and Social Research Council grants (and cross-council-funded research) are expected to be deposited in ReShare[14], the UK Data Service's repository for research data; where such data are made available for future use in research and learning worldwide. The UK Data Service specialises in managed access regulation for data containing potentially personal or sensitive information and will accept data from any projects, providing the data fall within the scope of its Collection Development Policy[15]. In general, this includes all data for the social science research and teaching communities that have potential for secondary use and analysis for research; teaching and learning use; or serve for replication and validation of research.

Alternative repositories specialising in publishing social science data are the USA-based and the Dataverse[16] and the German datorium[17]; or data can be deposited with a generic data repository such as Zenodo[18], an EU-developed system hosted by CERN, or an institutional repository. Using formal repositories to deposit data helps protect data owners' and participants' rights effectively and can lead to increased transparency and visibility of the research. Data repositories typically assign a DataCite Digital Object Identifier (DOI) to each dataset, which can be used to reference and cite a published dataset in a journal paper.

Access and licensing

The level of detail to be included in the data and the type of access permitted can be determined as appropriate for the data concerned. At the UK Data Service data may be licensed under one or more of these access levels[19]:

- open data – under open licence without any registration, either Open Government Licence (OGL) for Crown Copyright data or Creative Commons Attribution- 4.0 International Licence for other data (ShareAlike or not)
- safeguarded data – requiring an End User Licence, users to be authenticated and, where appropriate, special conditions agreed to
- controlled data – requiring user accreditation and registration through training and approval by a data access committee, and users to be authenticated

In addition, publication of a data collection can be delayed up to 12 months to allow publication of research findings.

Other repositories specialising in publishing social science data will operate similar data access levels: open data, restricted data and the ability to grant access to certain user groups.

Whilst a data repository can regulate access to safeguard sensitive or confidential data, it is always transparent about any access restrictions, by publishing which data exist in the collection, and who can access the data for which purpose and under which conditions. In all cases, the original researchers (usually project Principal Investigators or a designated person) remain the copyright holders. Depending on the access requirements, they may need to be consulted to negotiate specific use permissions.

Sharing sensitive data that cannot be anonymised

Data that present a risk of participants (human subjects) being identified where such identification poses a risk of harm, should only be made available under strict access protocols such as via an accredited researcher route. Such a protocol relies on a combination of:

- SAFE PEOPLE: researchers intending to use these data should undergo vetting or training in how to treat data safely
- SAFE PROJECTS: projects need to be approved by a committee overseen by the data owners
- SAFE SETTINGS: computer or physical access to the data may need to be locked down.
- SAFE OUTPUTS: in this case, outputs of analysis can be checked by experts for any possible disclosure

TRANSFER DATA TO A DATA REPOSITORY

Procedures for submitting a data collection to ReShare

- Register with the UK Data Service[20]
- Log into: ReShare[14]
- Create the metadata record for your data collection
- Upload data and documentation files
- Set access and licence conditions
- Submit the collection

When data are collected in multi-country projects by multi-country partnerships, agreement needs to be reached within the project as to where data will be preserved in the long term. It could be that one country takes full responsibility, or each country takes responsibility for their data, or multiple copies of the same data can be kept in separate countries, taking care that strict version control is in place, to avoid the risk of data being updated in one place and not the other.

WHAT THE UK DATA SERVICE DOES WITH YOUR DEPOSITED DATA

In-house data review checks

The UK Data Service reviews[21] each data collection submitted to the ReShare repository, before publishing it, for:

- disclosure risk
- copyright breaches
- validity of file formats
- accuracy of metadata
- level of documentation

Any concerns, changes needing to be made to data files, or requests for additional documentation are communicated to the data depositor for further action. Review comments and actions taken are noted in the data collection metadata record, as a record of provenance for that collection.

Safeguarding your data

Since 2010, the UK Data Service, the organisation that houses and hosts the data based at the University of Essex, has been certified under the international ISO 27001 standard for information security. A number of government departments have also carried out surveillance visits to the Service to clarify its information security regime.

The Service's Information Security Management System supports all aspects of digital curation carried out in-house, such as the selection, acquisition, ingest, archiving, provision of access to data, and the management and planning of digital curation. The ISO certification allows the Service to handle secure data on site and supports secure remote access to these research-rich data.

In house, data are classified according to their level of detail, sensitivity and confidentiality and appropriate data handling and access safeguards are in place.

Disseminating your data

The UK Data Service makes data available for research and teaching purposes according to the access and licence conditions decided by the depositor. Users are required to agree to the terms and conditions pertaining to the use of data.

References

[1] Ecosystems Services for Poverty Alleviation programme: http://www.espa.ac.uk/
[2] Millennium Villages in Northern Ghana Impact Evaluation: http://r4d.dfid.gov.uk/Project/61006/
[3] ESPA Social Surveys Event: http://www.espa.ac.uk/news-events/events/thu-2014-10-23-1000/espa-social-surveys-event
[4] UK Data Service Manage Date: http://ukdataservice.ac.uk/manage-data.aspx
[5] Madagascar p4ges project: http://www.p4ges.org
[6] UK Anonymisation Network Guidance Report: http://www.ukanon.net/wp-content/uploads/2013/10/About-Anonymisation-for-data-about-people-OCT-2013-1.pdf
[7] Millennium Villages Impact Evaluation project: http://www.ids.ac.uk/project/millennium-villages-in-northern-ghana-impact-evaluation
[8] UK Data Service model consent form: http://data-archive.ac.uk/media/210661/ukdamodelconsent.doc
[9] UK Data Service recommended file formats: http://ukdataservice.ac.uk/manage-data/format/recommended-formats.aspx
[10] Richardson, D. et al. (2013). Post-Trafficking Livelihoods in Nepal: Women, Sexuality and Citizenship, 2010-2012 [dataset]. Colchester, UK Data Archive: http://dx.doi.org/10.5255/UKDA-SN-7358-1
[11] Daly, M and Bray, R. (2015). State and civil society actor interviews on parenting support in England [Dataset]. Colchester, UK Data Archive: http://dx.doi.org/10.5255/UKDA-SN-851776
[12] Russell, S. (2015). Life on antiretroviral therapy: People's adaptive coping and adjustment to living with HIV as a chronic condition in Wakiso District, Uganda. [Dataset]. Colchester, UK Data Archive. http://dx.doi.org/10.5255/UKDA-SN-851094
[13] Data Documentation Initiative: http://www.ddialliance.org/
[14] ReShare: http://reshare.ukdataservice.ac.uk
[15] UK Data Service Collection Development Policy: http://ukdataservice.ac.uk/media/398725/cd227-collectionsdevelopmentpolicy.pdf
[16] Dataverse: http://dataverse.org/
[17] Datorium: https://datorium.gesis.org/
[18] Zenodo: http://www.zenodo.org
[19] UK Data Service access conditions: http://ukdataservice.ac.uk/get-data/how-to-access/conditions.aspx
[20] UK Data Service registration: http://ukdataservice.ac.uk/get-data/how-to-access/registration.aspx
[21] ReShare review procedures: http://reshare.ukdataservice.ac.uk/reshare-review-procedures/

espa ecosystem services for poverty alleviation
A research programme co-funded by DFID, NERC & ESRC and accredited by LWEC
UKaid from the British people
NERC SCIENCE OF THE ENVIRONMENT
E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL