A Guide to Sharing Qualitative Data[1]

December 2012

Background

Disciplinary norms are changing to require greater access to data and more transparency in research practices.[2] For instance, in October 2012, the American Political Science Association amended its *Guide to Professional Ethics in Political Science* to include new requirements for how scholars should present their research and the evidence upon which it is based. The value of data to science and society increases as the number of scholars with access to them grows. Sharing data allows them to be analyzed from a range of perspectives or conceptual foci and to have different analytic methods applied to them, facilitating assessment of the generalizability of research findings and encouraging new discoveries and comparative research.

This document provides guidelines for preparing qualitative data for archiving and sharing via the Qualitative Data Repository (QDR) at Syracuse University (https://qdr.syr.edu/).[3] The Repository is currently in development, and will go online in the near future.

Storing and sharing data through the QDR guarantees their safe-keeping in a secure environment; ensures that they will continue to be accessible to users over time; and allows for access control. Doing so also augments the impact and visibility of scholars' research.

Qualitative Data: Types, Formats, and Aggregate Data

Qualitative data take a variety of forms including unpublished primary sources, published primary sources, primary sources cited in secondary sources, secondary sources, and other research materials.

This list, while not exhaustive, includes some of the main types of qualitative data:

- Data from structured, semi-structured, or unstructured interviews; focus groups; oral histories (audio/videorecordings; transcripts; notes/summaries; questionnaires/interview protocols)
- Field notes (including from participant observation or ethnography)
- Maps/satellite imagery/geographic data
- Official/public documents, files, reports (diplomatic, public policy, propaganda, etc.)
- Meeting minutes
- Government statistics
- Correspondence, memoranda, communiqués, queries, complaints
- Parliamentary/legislative proceedings
- Testimony in public hearings
- Speeches, press conferences
- Military records

[1] A sister document, a "Guide to Active Citation," supplements this document for scholars preparing qualitative data for sharing in the context of active citation.

[2] See <https://www.apsanet.org/content_2483.cfm>.

[3] The sections of this document concerning preparing qualitative data for sharing draw on: (1) http://www.esds.ac.uk/qualidata/about/introduction.asp (and underlying pages); (2) http://www.data-archive.ac.uk (and underlying pages); (3) "UK Data Archive Qualidata Process Guide" and (4) "UK Data Archive Managing and Sharing 2011."

- Court records; legal documents (charts, wills, contracts)
- Chronicles, autobiographies, memoirs, travel logs, diaries
- Brochures, posters, flyers
- Press releases, newsletters, annual reports
- Records, papers, directories
- Internal memos, reports, meeting minutes
- Position/advocacy papers, mission statements
- Party platforms
- Personal documents (letters, personal diaries, correspondence, personal papers)
- Maps, diagrams, drawings
- Radio broadcasts (audio or transcripts)
- TV programs (video or transcripts)
- Print media (magazine, newspaper articles)
- Electronic media
- Published collections of documents, gazeteers, yearbooks, etc.
- Books, articles, dissertations, working papers
- Photographs
- Ephemera; popular culture visual or audio materials (printed cloth, art, music /songs, etc.)

Data can be deposited in the QDR in various formats: digital (e.g., word-processed, databases, spreadsheets), audio (digital and non-digital), video, and photographic (digital and non-digital). QDR staff can arrange to have data that scholars only have in paper form (e.g., typed and hand-written notes, newspaper clippings, or documents) scanned.

With regard to aggregate qualitative data, it is useful to distinguish between two styles of data deposit: a "qualitative data collection" and a "qualitative data set":

- *Qualitative data collection:* a group of data cited in a publication. Data collections often include an unrepresentative subset of the data a scholar collected/generated, and the data he consulted, when carrying out the research and analysis for the project on which the publication is based. Data collections are often made available in connection with activating the citations in a particular article or book chapter. While the materials may have a logic that unites them, at the limit they might be a set of ephemera with no intrinsic connections other than their employment in the author's narrative.

- *Qualitative dataset*: a coherent group of data that relate to each other in identifiable, describable ways and represent a stand-alone resource that could be useful for scholars (beyond the scholar who collected/generated them) to analyze. A qualitative dataset has some categorization or logic which makes the data more than just an aggregation of used materials.

While the distinction between these two types of data aggregation is one of degree, qualitative datasets are likely to be most useful for secondary analysis. Qualitative datasets may vary widely in structure. For instance, they may contain formalized information gathered in the context of pre-set categories and come in the form of a preconfigured database (i.e., may actually have rows and columns), or may be a specific group of documents or a specific set of interview transcripts. They may contain many different types of data, and may contain data relevant to different aspects of a research project (some data may

measure a variable, other data may be used for process tracing, representing a sequence of information leading up to an outcome). The key is that the scholar can tell a compelling story that connects them.

Whole research projects can also be stored in a Computer-Assisted Qualitative Data Analysis Software (CAQDAS) format. CAQDAS packages such as NUD*IST, ATLAS-ti and WinMax have export facilities that enable scholars to save a whole 'project' consisting of the raw data, coding tree, coded data and associated memos and notes. The raw data, the final coding tree, and any useful memos should be exported before the project is deposited with the QDR.

The QDR assigns persistent identifiers to its holdings. A persistent identifier is a permanent link to a publication or a dataset on the Internet. The publisher of the resource agrees to maintain the link to keep it active. Over time the link behind the persistent identifier may be updated, but the identifier itself remains stable. There are several kinds of persistent identifiers (DOI, URN, Handle, etc.). Persistent identifiers are "machine-actionable" and facilitate the harvesting of data references for online citation databases, like the Thomson-Reuters Data Citation Index. Scholars can easily track the impact of their data from citations in publications. An increasing number of journals are requiring persistent identifiers for data citations.

The Data Supplement

In order for data shared by one scholar to be truly accessible to others – that is, in order for other scholars to be able to understand and interpret shared data, and make informed and effective use of them – scholars must include documentation that describes those data and the process of collecting/generating them.

Scholars sharing qualitative data in connection with Active Citation should see the "Guide to Active Citation" for guidance on documentation and the creation of a Research Appendix.

Scholars sharing qualitative datasets for secondary analysis should supply a Data Supplement that includes and describes the following:

- *Research project description* – discussion of the purpose of the research project to which the data are connected; the research design (e.g., context [time period and geographic location{s}], analytic methods, case selection/sampling design, etc.) as well as the central findings and conclusions;
- *Data listing* – list of the data being shared; related items (e.g., the audio file and written transcript of a particular interview, or documents and field notes concerning the same case) should carry identifiers that are consistent across data forms so links can be made among the different types of items.
- *Data narrative* – description of:
    - The type and structure of the data, how the data relate to the research design, and how the data relate to each other;
    - Why these data were consulted in connection with the research project and what other data potentially relevant to the project were not consulted and why;
    - Why these data were selected for sharing and how they relate to other data connected to the project that were *not* shared.
- *Data-collection methods* – description of:
    - The procedures used to access the shared data. For instance, for a set of interview transcripts, a description of how interview respondents were selected and how, when, and where interviews were conducted.

    - The research instruments employed, for instance, the interview protocols or topic guides, interviewer instructions, consent forms used in connection with collecting the data. Ideally the actual instruments can be shared as well as the data.
- *Data-capture methods* – description of whether data were photographed, scanned, photocopied, recorded, or whether notes were taken on consulted data
- *Data codebook* including, for each variable, variable description; Instrument, question text, or computation formula; Valid values and their meanings; Cases to which this variable applies; Methods for imputing missing values
- *Data preparation* – description of steps taken to prepare data for sharing (for instance, anonymization)
- *Data confidentiality* – description of:
    - data sensitivity (how personal, specific, etc.)
    - information on access conditions and terms of use where applicable

Although documentation is often supplied in text files or spreadsheets, the standard for documentation in the social sciences is the Data Documentation Initiative (DDI). DDI is an XML markup standard designed for social science data. Since DDI is machine actionable, it can be used to create custom codebooks and to enable online search tools. A list of tools for creating DDI is available at the DDI Tools Registry.

Sharing Data Ethically: Informed Consent and Anonymization

Data stored in the QDR are not in the public domain. Only registered users can view the data, and the use of stored data is restricted to specific purposes. Users sign a legally binding access agreement in which they agree to conditions such as respecting guarantees of anonymity consistent with the original investigator's undertaking, not attempting to identify any individuals from the data, and not sharing data with unregistered users.

Important legal and ethical obligations, including concerns about confidentiality, impinge on the gathering, employing, and sharing of data concerning people. These types of data can often be shared legally and ethically, however, if informed consent from project participants is requested and granted and resulting confidentiality guarantees are maintained; if anonymization strategies are effectively deployed (when needed), and if access to data is carefully controlled (when needed).

Informed consent

To date, the main goal of soliciting informed consent from potential project participants was to have them formally agree to have data collected from them and to have those data used, in an agreed-upon form, in the relevant scholar's work. Scholars who wish to share legacy data concerning people – that is, data they collected without human subjects giving their consent for them to share those data – should determine to what degree the Institutional Review Board agreement and protocols associated with collecting those data would cover such sharing, and discuss the process for gaining permission retroactively with IRB staff.

Practices for requesting informed consent, and the content of informed consent, are changing as a norm of data sharing emerges. For scholars who plan to share their data, acquiring informed consent must include coming to an agreement with potential project participants about how the information collected from them will be used in the scholar's research; how those data will be stored; how confidentiality (to

the level agreed upon) will be maintained; and where, how, with whom, and under what conditions data will be shared.

Put differently, potential study participants must be provided with enough information about the study in which a researcher wishes to involve them and how the information they provide will be used and shared (and anonymized and otherwise protected) that they can make an *informed* decision about whether or not to participate; about how confidential they would like the information they provide to be kept; and about whether and how much of the information they provide will be shared, with whom, and how.

Eliciting *informed* consent thus requires active, open communication between researchers and potential participants. Researchers might create an information sheet with answers to questions such as the following:[4]

- What is an archive?
- Why put information in an archive?
- How do I know my data will be used ethically?
- What does anonymizing mean?
- How might data be used?
- Who owns the data and what is copyright?
- How do archives store my data safely?

Anonymization

Anonymization of data sources (for instance, of documents or interview transcripts) may be needed for ethical reasons (for instance, to prevent identification of individuals or organizations) or for legal reasons. Scholars depositing data in the QDR should make sure those data conform to ethical and legal guidelines with respect to the preservation of anonymity (where requested). Pre-anonymized data can be stored in but not shared via the repository.

Absent the granting of specific consent for personal information to be included (as occurs sometimes with well-known elites or life story material), data obtained through interaction with people must be anonymized before they can be shared. That is, all personal data must be removed so that no information that breaches the confidentiality of any respondent or any other person or entity is present in any data that will be shared. Note that personal identity can be disclosed both directly (through disclosure of a project participant's name, address, or telephone number) or indirectly (such that particular pieces of information about the person, when linked with publicly available information, could reveal his/her identity).

The appropriate level of anonymization for a certain set of data depends upon the agreement the researcher and project participants settled upon during the process of obtaining informed consent, and is closely related to the nature of the research. The goal is to create data that effectively and accurately represent the research process and participants' contributions while protecting the latter's safety.

Anonymization of qualitative material entails:[5]

[4] Based on list at http://www.data-archive.ac.uk/create-manage/consent-ethics/consent?index=3
[5] Draws on UK Data Archive Managing and Sharing Data, p. 26.

- Removing major (direct) identifying details (e.g., real names, locations); replacing them with pseudonyms, replacement terms (e.g., "paternal grandfather"), vaguer descriptors or some coding system (where appropriate) consistently throughout the project; and devising and using a cross-referencing system for pseydonyms that will not be made available to users;
- Removing information in a transcript or notes from a human encounter that may reveal the identity of project participants;
- Aggregating or reducing the precision of information or a variable, e.g. replacing date of birth by age groups;
- Generalizing the meaning of detailed text, e.g., replacing a doctor's detailed area of medical expertise with an area of medical specialty;
- Restricting the upper or lower ranges of a variable to hide outliers;
- Noting the replacement of identifying details in text and the removal or modification of information in a meaningful way (for instance, in transcribed interviews, indicating replaced text with [brackets] or using XML markup tags <anon>…..</anon>);
- Creating an anonymization log (stored separately from the anonymized data files) of all replacements, aggregations, or removals (see Table 1);

Table 1 Example anonymization log

Interview and page number	Original	Changed to
Int1		
p1	Age 27	Age range 20-30
p1	Spain	European country
p3	Manchester	Northern metropolitan city or English provincial city
p2	20th June	June
p2	Amy (real name)	Moira (pseudonym)
Int2		
p1	Francis	my friend
p8	Station Road primary school	a primary school
p10	Head Buyer, Produce, Sainsburys	Senior Executive with leading supermarket chain

In some cases, data will not be able to be completely anonymized, or anonymization would lead to too much loss of content or data distortion (both of which compromise the potential for secondary data analysis). In such instances, user-access restrictions of various levels can be established.

Sharing Data Ethically: Copyright

Copyright law has important implications for generating, storing, sharing, and re-using qualitative data. Copyright is an intellectual property right that is automatically assigned to the original author or creator of many kinds of research data, datasets, databases, data sources, and data outputs. Unauthorized copying and publishing of an original copyrighted work is illegal.

Scholars should review and closely adhere to any user agreement they signed with the entity from which they secured the data they wish to share, and follow the relevant rules guiding the use of those data. Scholars wishing to deposit copyrighted data (in any form) for storing or sharing must acquire, from all

persons or entities holding the copyright to any aspect of the materials to be deposited, explicit written permission (and/or a licensing agreement) outlining the terms of and conditions for the reproduction of the materials. Relevant questions to ask those persons/entities include how much of the data they can share (how many pages, how many quotes, etc.); whether original image can be shared; whether text needs to be transcribed; what fall-back options there might be if the initial answers are "no."

The QDR simply publishes data, and has no rights in any of the data collections it stores and/or processes, and or the sharing of which it facilitates.

Sharing Data Ethically: Access and Access Controls

Controlling access to data can help to protect confidentiality and address copyright limitations. For confidential or sensitive data, or data with copyright restrictions, stricter access controls or user regulations may be imposed such as:[6]

- needing specific authorization from the data depositor to access data
- placing confidential data under embargo for a given period of time until confidentiality is no longer pertinent
- providing access to approved researchers only
- providing secure access to data through enabling remote analysis of confidential data but excluding the ability to download data

Mixed levels of access may be established for some data collections and datasets, combining more restricted access to confidential data with less restricted access (or unrestricted access) to non-confidential data. The QDR works in tandem with those depositing data to identify the level of access appropriate for the kind of data and confidentiality involved.

[6] Quoted from: http://www.data-archive.ac.uk/create-manage/consent-ethics/access-control.