A GUIDE TO SHARING QUALITATIVE DATA[1]
Version 1.3 for Pilot Projects
(November 2013)

Colin Elman (Syracuse University)
Diana Kapiszewski (Georgetown University)

I. Overview

Disciplinary norms are changing to require greater access to data and more transparency in research practices. 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.[2] The value of data to science and society increases as the number of scholars with access to them grows. Sharing data allows scholars to analyze them from a range of perspectives or conceptual foci and apply different analytic methods to them; facilitates assessment of the validity and generalizability of research findings; and encourages new discoveries and comparative research.

This document provides very general guidelines for preparing qualitative data for archiving and sharing via the Qualitative Data Repository (QDR) at Syracuse University.[3] The Repository is currently in development, and will launch in November 2013.

Storing and sharing data through the QDR ensures their safe-keeping in a secure environment; safeguards continuing accessibility to users over time; and allows for access control. Doing so also augments the impact and visibility of scholars' research.

II. Types of Qualitative Data and Formats in which Qualitative Data can be Shared

Qualitative data take a variety of forms including unpublished and published primary sources, secondary sources, and many other research materials.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

[1] A sister document, "A Guide to Active Citation," provides guidelines for scholars preparing active citation compilations. We gratefully acknowledge the assistance of Louise Corti (Associate Director, UK Data Service) and Dessislava Kirilova for helpful comments on earlier versions of this document.

[2] See < http://www.apsanet.org/content_86135.cfm >.

[3] The sections of this document concerning preparing qualitative data for sharing draw extensively 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."

- 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
- 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, yearbooks, etc.
- Books, articles, dissertations, working papers
- Photographs
- Ephemera; popular culture visual or audio materials (printed cloth, art, music /songs, etc.)

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

III. Types of Qualitative Data Projects

We distinguish among three types of data projects (i.e., different forms in which aggregated qualitative data can be deposited in the QDR): active citation compilations, data collections, and topic clusters. While in some senses the distinction among the three is one of degree, we can identify central tendencies in the way the shared data will be employed: active citation compilations are most likely to be used to evaluate the claims offered in published scholarship, data collections are most likely to be useful for secondary analysis, and topic clusters are most likely to be helpful to scholars interested in gaining background information as they develop a related research project.

- *Active citation compilation:* a group of data cited in a publication and made available in connection with "activating" the publication's citations (in-text references, footnotes, and/or endnotes). Active citation compilations often include an unrepresentative subset of the data a scholar consulted, collected, or generated when carrying out the research and analysis for the project on which the publication is based. 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.[4]

- *Data collection*: 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 the data to analyze. A data collection has some categorization or logic which makes the data more than just an aggregation of used materials.
    - Data collections 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 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 logic that connects them be elucidated.
    - QDR can also house data collections in Computer-Assisted Qualitative Data Analysis Software (CAQDAS) format. CAQDAS packages such as NUD*IST, ATLAS-ti and WinMax have export functions to allow 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 QDR.

- *Topic cluster*: an unstructured amalgamation of materials on a particular issue or subject. When conducting research, qualitative political scientists invariably gather considerably more information than they ever carefully organize and analyze. The materials in the "everything else" box can be a windfall for other scholars who are addressing the same or a similar topic or searching for relevant background information.

QDR assigns persistent identifiers to its holdings. A persistent identifier is a permanent link to a publication, data project, or unique metadata instance that points to (and records versioning of) a data project 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.); all 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. QDR uses DOIs, assigned by EZID, a service of the California Digital Library of the University of California.

[4] For additional information, see "A Guide to Active Citation," this guide's sister document.

IV. Requirements for Data Deposit

Shared data need to be understandable and interpretable to scholars beyond those who collected or generated them. So that those subsequent consumers of shared data can make informed and effective use of them, the data must be accompanied by documentation that describes the project to which the data are connected, the data themselves, and the processes by which they were collected or generated.[5]

Scholars wishing to deposit data in QDR will need to complete three forms:

(1) Their first step will be to fill out a Project Proposal form.[6] This form aids scholars in developing a basic (question-guided) narrative about the substance of the research project with which the data are associated, and offering some preliminary information about the data project itself. This form is available here: http://www.maxwell.syr.edu/moynihan/cqrm/About_the_Qualitative_Data_Repository/

(2) When depositors are ready to deposit their data, they will fill out a Data Deposit form. This form walks scholars through the process of documenting their data files. This form is available here: http://www.maxwell.syr.edu/moynihan/cqrm/About_the_Qualitative_Data_Repository/

(3) When depositors are ready to deposit their data, they will also fill out a Depositor License Agreement confirming they have adhered to external constraints and satisfied external requirements for sharing their data, and stipulating access controls and other limits on viewing and downloading their shared data.

QDR staff is currently developing the processes scholars should follow to submit the data they wish to deposit. Please contact QDR staff for more information concerning file transfer procedures.

V. Sharing Data Ethically: Informed Consent and Anonymization

Some, but not all, data stored in QDR are available to all registered users. In order to register, QDR users must provide a few pieces of identifying information and agree to QDR's legally binding "General Terms and Conditions of Use". These terms and conditions include such requirements 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

[5] Scholars sharing qualitative data in connection with active citation should see "A Guide to Active Citation" for a discussion of how to create a Transparency Appendix (TRAX). Information QDR depositors provide via the text files described below will be converted into machine-actionable form by QDR. The resulting metadata will conform with both Dublin Core and Data Documentation Initiative (DDI) standards, two of the most commonly used standards for documentation in the social sciences.

[6] Under typical circumstances, scholars who have an interest in depositing data in QDR will fill out this form, and QDR staff will consider whether the data would be suitable for the repository. Because the QDR pilot projects were solicited, the data have already been approved for deposit, and the Project Proposal form simply serves as a description of the agreed-upon data.

unregistered users. Other data are only available to registered QDR users who are affiliated with institutions that are members of QDR. Data that are only available to registered users, or only accessible to registered users who are affiliates of institutional members, are still "publicly available" in the sense that they can be obtained by researchers other than those who generated them (even though additional steps must be taken or requirements fulfilled in order to access them). Finally, some data are restricted, i.e., are accompanied by more stringent access and use requirements, stipulated by depositors in a Depositor License Agreement, and agreed to by registered users in a Downloader License agreement. Even restricted data can be considered available if the conditions are met.[7]

Important legal and ethical obligations, including concerns about confidentiality, impinge on the gathering, employment, and sharing of data concerning people. Sensitive data, however, can often be shared legally and ethically if one or more steps are taken. These steps include requesting and obtaining informed consent from project participants and maintaining the resulting confidentiality guarantees; effectively deploying anonymization strategies; and carefully controlling access to data (discussed in the final section of this guide).

Informed consent

Until recently, a scholar's main objective in 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 just the scholar's own work. Contemplating the broader sharing and reuse of those data makes the process of soliciting informed consent more complex.

In some cases (including most QDR pilot projects), scholars are sharing data from projects already underway or completed, where the data were collected without participants giving their explicit consent for the data to be shared more widely. For such 'legacy' data, scholars 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 the data they generate through future research projects, 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

[7] Of course, depositors sometimes forbid their data to be shared under any circumstances. If users cannot access the data under any conditions, it cannot be considered publicly available.

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:[8]

- 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 that those data conform to ethical and legal guidelines with respect to the preservation of anonymity. Pre-anonymized confidential data can be stored in but not shared via the repository.

Absent the granting of specific consent for personal information to be included (as, for example, occurs sometimes with well-known elites), 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 (when particular pieces of information about the person can be linked with publicly available information to reveal his/her identity).

The appropriate level of anonymization for any set of data depends upon the agreement the researcher and project participants reached 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:[9]

- 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 pseudonyms that will not be made available to users;

[8] Based on list at http://www.data-archive.ac.uk/create-manage/consent-ethics/consent?index=3

[9] Draws on UK Data Archive Managing and Sharing Data, p. 26.

- 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 or city names by province names;
- 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 Sample anonymization log

Interview and page number	Original	Changed to
Int1		
p1	Age 27	Age range 20-30
p1	Argentina	South American country
p3	Syracuse	American provincial city
p2	20th June	June
p2	Amy (real name)	Moira (pseudonym)
Int2		
p1	Francis	my friend
p8	Pebble Hill elementary school	An elementary school
p10	Senior Vice President, Tiffany's	Senior Executive with luxury retailer

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 (discussed in this document's final section) can be established.

VI. <u>Sharing Data Ethically: Copyright</u>

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 images 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 projects it stores and/or processes, or the sharing of which it facilitates.

VII. 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:[10]

- 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
- using special licensing agreements requiring scholars who wish to use data to do so under additional conditions or only when additional requirements have been fulfilled

Mixed levels of access may be established for some data projects, combining more restricted access to confidential data with less restricted access (or unrestricted access) to non-confidential data. QDR will work in tandem with data depositors to identify the level of access appropriate for the kind of data and confidentiality involved.

Qualitative Data Repository (QDR)
Center for Qualitative and Multi Method Inquiry https://qdr.syr.edu/
Syracuse University
A Guide to Sharing Qualitative Data_v1.3 2013-11-05

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