11-2016

Beyond the Matrix: Repository Services for Qualitative Data

Sebastian Karcher
Syracuse University, skarcher@syr.edu

Dessi Kirilova
Syracuse University, dpkirilo@syr.edu

Nic Weber
University of Washington, nmweber@uw.edu

Recommended Citation
Karcher, Sebastian; Kirilova, Dessi; and Weber, Nic, "Beyond the Matrix: Repository Services for Qualitative Data" (2016). *Moynihan Institute of Global Affairs*. 1.
https://surface.syr.edu/miga/1

Sebastian Karcher, Dessi Kirilova, Nic Weber

Beyond the Matrix: Repository Services for Qualitative Data

Paper accepted for publication in IFLA Journal (Fall 2016)

Abstract

The Qualitative Data Repository (QDR) provides infrastructure and guidance for the sharing and reuse of digital data used in qualitative and multi-method social inquiry. In this paper we describe some of the repository's early experiences providing services developed specifically for the curation of qualitative research data. We focus on QDR's efforts to address two key challenges for qualitative data sharing. The first challenge concerns constraints on data sharing in order to protect human participants and their identities and to comply with copyright laws. The second set of challenges addresses the unique characteristics of qualitative data and their relationship to the published text. We describe a novel method of annotating scholarly publications, resulting in a "transparency appendix" that allows the sharing of such "granular data" (Moravcsik et al., 2013). We conclude by describing the future directions of QDR's services for qualitative data archiving, sharing, and reuse.

Introduction

Social science has a rich history of archiving, sharing, and reusing data for secondary analysis. Institutions such as the Inter-university Consortium for Political and Social Research, the Roper Center for Public Opinion Research, UK Data, and the data banks of various international organizations have for decades provided quantitative, matrix based[1], datasets for a variety of empirical studies across the social sciences. However, a large number of primary data collections created by social scientists still remain invisible to the broader research community (Tenopir et al, 2011: see esp. Tables 10 and 11). And, until recently, the absence of dedicated repositories for qualitative materials made the likelihood that qualitative social scientists would either share their own data in a way that increases research transparency or make use of others' for secondary analysis almost nonexistent (see e.g. Medjedović and Witzel, 2010, focusing on Germany; Yoon 2014).

The Qualitative Data Repository (QDR) provides infrastructure for the sharing and reuse of digital data used in qualitative and multi-method social inquiry (Elman et al., 2010).

[1] We refer to matrix-based data as data arranged in columns (variable) and rows (observations) as commonly used in statistical analysis. For this purpose, even data that are not rectangular by design, such as event-history data, are arranged in matrix form for analysis.

QDR is housed at the Center for Qualitative and Multi-Method Inquiry[2] (a unit of Syracuse University's Maxwell School of Citizenship and Public Affairs[3]), and funded by the National Science Foundation. The repository is guided by four beliefs:

- All data that can be shared and reused should be;
- Evidence-based claims should be made transparently;
- Teaching is enriched by the use of well-documented data; and
- Rigorous social science requires common understandings of its research methods (Qualitative Data Repository, 2016).

QDR operates most explicitly to develop and provide a user-friendly data submission and preservation platform and to publish data projects in the social sciences domain that scholars can use for analytic and pedagogic purposes. Underlying these operations, however, is the deeper mission to supply practical guidance for its user community of qualitative researchers, especially on the challenging issues of legal and ethical sharing; to educate researches in the basics of data management so that they are well positioned to prepare their own projects with the goal of sharing in mind from the early planning stages; and even to promote a different way of thinking about the materials these

[2] http://www1.maxwell.syr.edu/moynihan_cqrm.aspx
[3] http://www.maxwell.syr.edu/

scholars create and collect as “qualitative data.” To achieve these goals, QDR does not just provide technical infrastructure but has dedicated staff working individually with data depositors to curate qualitative data for preservation and reuse. Over the past two years, QDR has made substantial progress in promoting data sharing and research transparency in social science by tackling challenges unique to qualitative data. We have developed strategies for addressing the concrete copyright and human participant constraints that occur when sharing such data and have developed tools that allow for transparent inference accommodating their unique structure.

Copyright and Human Subjects

An early set of challenges encountered by QDR are constraints placed on data sharing in order to protect human participants and their identities and to comply with copyright laws. Qualitative and mixed methods research often draws upon data—such as archival documents, images, video, interviews, etc.—that have a mix of intellectual property rights and personally identifying information (including rich contextual information). Concerns about privacy and ethics in data sharing figure just as prominently for quantitative research (King, 2011; Lagoze et al., 2013; Lupia and Alter, 2014), but qualitative data pose a unique set of challenges, which tools and strategies borrowed from quantitative traditions cannot solve. Copyright concerns, while shared with other digital archives and repositories, are all but nonexistent for quantitative data (most of which cannot be copyrighted *qua* data), so QDR has been treading ground unfamiliar for most data repositories.[4]

Annotation for Transparency

A second set of challenges in sharing qualitative data concerns the nature of the data. In quantitative, matrix-based data, datasets are analyzed using dedicated software (e.g. R, Stata, SPSS) to produce a set of tables and/or figures. In contrast, qualitative researchers will make a multitude of empirical claims throughout their text. Each claim is backed by a piece of data (e.g., an excerpt from an interview or an archival source), which the author has analyzed or interpreted individually. We refer to this type of data and their analysis as *granular*.[5] Building on the work of Andrew Moravcsik on active citations, a form of author-contributed annotations supplementing a formal publication (Moravcsik, 2010, 2014a, 2014b), QDR has developed a set of pilot studies, working with qualitative researchers wanting to share granular qualitative data in the form of a transparency

[4] QDR is the first archive focusing on qualitative social science data in North America, but is able to draw upon a longer, though still young, tradition in Europe, including UK Data/Qualibank, the Irish Qualitative Data Archive, and Qualiservice in Germany.

[5] We realize that the term “granular” may be used by information and data science to mean an indivisible unit of raw data. In this context we understand it simply in contrast to matrix data.

appendix (TRAX).[6] In the process, we have developed guidelines to help researchers share such data (Moravcsik et al., 2013) and also identified key shortcomings of current technologies and avenues for future improvement, both in methodology and technology.

The rest of the paper will proceed as follows: we begin by reviewing previous work on qualitative data archiving, sharing, and reuse, drawing attention to empirical work that sheds light on the the scholarly practices of contemporary social scientists. We then describe curation services developed by QDR to assist researchers in archiving, sharing, and reusing qualitative data. The first set of services describes challenges and solutions in sharing sensitive or copyrighted materials. The second set of services describes an emerging approach to sharing granular qualitative data. We conclude by outlining some of QDR's future activities aimed at improving both the collection and publication of granular data.

[6] Finished Active Citation pilot studies as of mid-2016 include Crawford, 2015; Handlin, 2015; Herrera, 2016; Saunders, 2015; Snyder, 2015. Given both our experience during those pilots and trends in scholarly annotations, QDR is moving away from the somewhat restrictive "active citations" towards "annotations for transparent inference" (ATI). Throughout this paper, we refer to our pilot projects as Active Citation compilations. The final section then introduces the concept of annotations.

Background: Qualitative Data Archiving, Citation, Sharing, and Reuse

Social scientists have, for many years, used normative arguments for why the reproducibility of their research findings are critical for the viability of the field (King, 1995). As a result, social science data repositories are some of the longest standing and most mature infrastructures for preserving, providing meaningful access to, and representing the myriad research objects produced by social scientists; including qualitative and quantitative data, statistical software, images, videos, codebooks, and data dictionaries. Studies of scholarly practices—such as citation behaviors, archiving, and reusing data—have similarly had an important impact on the development of policies and services developed by data repositories to serve social scientists. Below, we review some important studies influencing QDR's development of qualitative data curation services.

Data Citation

Some of the earliest research into data citation behavior has occurred in the social sciences, including Joan Sieber's work on the norms and practices of data sharing and reuse (e.g. Sieber and Stanley,1988; Sieber, 1991). In 1995, Sieber and Trumbo showed that although both qualitative and quantitative social scientists often engage in secondary analysis, only 19% (n=168) of authors explicitly acknowledge the source of their data in a formal publication. In more recent scholarship, Mooney (2011) showed that of 49 studies reusing data archived at ICPSR, 60% (n=30) failed to cite these data in a reference list. Also using ICPSR citation data, Fear (2013) showed that although secondary analysis is often performed with social science data, citation practices are diffuse—many authors use citations to credit both a repository (or source of data) as well as authors who had originally interpreted those data.

Data Reuse

Fear’s study also shows that tracking the impact of data reuse in the social sciences is complicated by hostility towards formal archiving policies (2013). Other practical objections to the viability of meaningful reuse of social science data are whether or not the context of data collection can be meaningfully reinterpreted in secondary analysis (Boddy, 2001; Fujii, 2016), the complications of intellectual property rights held by participants and researchers (Mauthner et al., 1998), and the ethical ramifications of preserving human-subject materials (Bishop, 2009; Broom and Cheshire 2009; Carusi and Jirotka, 2009). Depending on epistemological commitments, for some scholars who use ethnography or participatory research, data can be seen as co-produced by both researcher and participants—both with equal claims to ownership and intellectual control (Mauthner et al., 1998). Parry and Mauthner (2004) compare the tradition of oral history scholarship, where ownership of materials (transcripts and recordings) is assumed to be held by respondents, with qualitative social scientists, where ownership is retained by an institution or individual conducting the research. They argue that while some social science archives offer guidance for researchers navigating these issues (e.g. Corti and Thompson, 1997), the various modes of producing qualitative data will require a broader, more malleable intellectual property framework (Parry and Mauthner, 2005). In studying social scientists who had completed secondary analysis of social science data, Faniel et al. (2015) demonstrate that five qualities were significant factors of satisfaction in reuse—data completeness, data accessibility, data ease of operation, data credibility, and data documentation.

Data Sharing

Few studies have focused specifically on qualitative data sharing. Yoon's (2014) study of 13 researchers who had published a research article based on secondary analysis of qualitative data found that participants had only engaged in this type of analysis 1-2 times previously, that sharing happened solely through advisors or co-workers, and that successful reuse relied upon access to and in-depth conversation with the original data collector. As Yoon notes, ethical objections to sharing qualitative data are often the reasons why researchers withhold data from federal archiving mandates. This is the case in the UK where “ethics-related” exemptions are the most frequent waiver used by social scientists (Van den Eynden, 2008). Ethical objections to sharing qualitative data are typically made for the sake of protecting research participants’ confidentiality, consent, or privacy. However, Bishop argues that studies impacted by these issues represent only a narrow subset of qualitative inquiry and the subject requires more careful consideration by archivists and infrastructure developers (2009). Indeed, while there are many tools to facilitate privacy-protection in archiving research data (Fung et al., 2010), there have been few applications that cater to the unique ethical aspects of qualitative data as described above.

What makes the challenge even more complicated is that the ethical constraints that researchers legitimately worry about do not overlap completely with the legal protections put in place with the intent of preventing exploitation and harm of research participants (cf. Bosk and De Vries, 2004). A researcher wanting to safely observe both sets of considerations, whose only guidance on the issue might come from a local, risk-averse, and tradition-bound institutional review board, will almost always conclude that sharing of the granular data they have collected in interactions with human participants is not a good idea and will thus perpetuate the status quo of putting all these rich materials “under lock” or, even worse, promising to destroy them at the end of the project (Bishop, 2009: 261).

Intellectual Property

Concerns about infringing on intellectual property, which come into play when archival and other records fixed in a tangible medium are used as primary qualitative data, leads to a different dynamic when these materials are thought of as the data which a scholarly project analyzes. Mindful of the need to use representative information in order to arrive at accurate inferences, scholars are often in pursuit of comprehensiveness and breadth when they collect such materials. But the greater the number of items they use, say from an archive or a magazine articles database, and the greater the proportion of each copy they obtain for their purposes, the more problematic further sharing might become according to the stipulations of current U.S. copyright law (Copyright Act of 1976 and subsequent amendments, known collectively as Title 17 of the United States Code). The usual exceptions, primarily the so-called “fair use” allowance for quoting portions of restricted length for the purpose of scholarly analysis, is of little help where the goal is to share complete data, as encouraged by research transparency norms.

Where scholars plan to use a comprehensive selection of materials with third-party ownership of copyright, the best approach may be to request explicit permissions for archiving and sharing with one's wider scholarly community (UK Data Service, 2016). The permissions obtained will then become part of the administrative documentation of a data project. This can be accomplished best when the researcher is aware of the issue in advance and plans for its resolution from the initial stages of the project.

Qualitative Data Sharing in Practice: Lessons from QDR's Pilot Studies

How did the issues explored in the previous session surface during QDR's pilot phase? What were the principal challenges and what some of the solutions encountered by QDR staff? In this section we explore this question focussing on two sets of cases. We first address copyright and privacy concerns that affect nearly all qualitative data. We then explore the experience with sharing granular qualitative data in the form of active citation compilations.

Copyright and Privacy Concerns

In order to test how some of these principles of resolving the two key categories of concerns might be implemented in the form of practical solutions for qualitative data sharing, QDR began by commissioning a number of pilot projects, largely selected to illustrate in practice some of the thornier aspects of both copyright and human participant safety. Processing these early deposits, we learned that a lot of qualitative data can be shared both legally and ethically but that social scientists need to be aware of the strategies and the repository's technical tools that enable that. Most of all, we became convinced that the repository's role must be pedagogical as much as technological. It needs to impart knowledge of basic data management concepts, as derived from the library and information sciences' long-standing efforts in this field, to its intended user community. Once empowered with an understanding of general principles and best approaches in data management, depositors were in a position to collaborate with QDR's staff to identify applicable ways of sharing their data.

One interesting discovery was that while all researchers who used data collected by interviews or participant observation were acutely aware of the potential for problems regarding the privacy and confidentiality of their interlocutors in the field, few of those who had used archival or secondary literature resources considered the legal implications when they were asked to share such data. In fact, the discussions with QDR's staff and legal counsel on the topic was the first time some researchers encountered the concepts of copyright protection and fair use.

Another general point that united all piloteers was the need for guidance in the creation of useful documentation that contextualized both the data collection/creation process and the resulting qualitative materials. As mentioned above, the extremely heterogeneous nature of granular data makes existing categories such as samples, variables, codebooks, descriptive statistics, and survey instruments, routinely used to document matrix-organized data, unhelpful to many qualitative researchers. Understanding both the logic of documenting the process through which data were obtained and the nature of the process of field research itself enabled QDR's staff to "translate" to researchers in language more relevant to them the goals of contextualizing a deposit in a way that makes the data intellectually accessible and useful for secondary users as well as more discoverable and secure. In fact, we believe that most qualitative researchers would welcome the opportunity to provide in-depth descriptions of the path that started with their initial research designs, passed through the excitement of entering their field sites for the first (or nth) time, proceeded through the vicissitudes of exploring the site, led to the compilation of much more diverse data than could be usefully processed in a single project, and resulted in the subset of better organized materials that actually underwent analysis for publication and ended up as the coherent data deposit.

In a particularly gratifying instance (by a team who engaged in multi-method work that used information gleaned from interviews and extensive following of the local press's coverage on the topic of interest to develop a rich case study, supplemented by quantified aggregations of some of the same facts), the researchers had prepared what they called "data narratives," which essentially captured all the expected details about the research process, enabling traceability of their whole research process, independent methodological learning, and any secondary reader's understanding of their conclusions.

In dealing with human participants, the majority of pilot project creators—whether active citations or more conventional stand-alone data projects—did resort to the default position of qualitative researchers of not sharing the data. This was where the challenge of working through data management issues only retroactively (i.e., belatedly from the perspective of a data manager) proved most detrimental to transparency. In some cases, the provision of at least some excerpts from interviews proved to be an acceptable technique. Researchers felt comfortable using short extracts (in some cases anonymous, in others not, especially when public figures were the source of information) at discrete points in their publications where they deemed that sufficient to illustrate the empirical contention they were making.

Compared with the alternative of no access to any elements of the primary data, this partial solution seems preferable. But in at least one case the researchers decided to undertake a more ambitious step and carefully anonymize a full set of 100 interview transcripts submitted as a stand-alone data collection. The process required close iterative discussions between their team and QDR's curation staff, combing through the basics of direct identifiers first and subsequently the ever more fine-grained details that taken together could serve as indirect identifiers—an aspect that will inevitably continue to bedevil the richest examples of qualitative data projects created in interaction with participants. The fact that the language of the original data collection was not English (also a characteristic we expect will continue to be common to the type of international fieldwork data QDR attracts) introduced an additional wrinkle in the process. As a general solution, QDR currently requests all documentation in English, regardless of the source language of the data. Despite all these difficulties, the anonymization protocol arrived at by these researchers and the extensive set of materials processed according to it are both valuable scholarly products that others can learn from both methodologically and substantively.

While the exact anonymization choices had to be agreed through considerable discussion and applied laboriously, other strategies QDR could offer to enhance participant protections were more easily borrowed from existing data management best practices long used for quantitative surveys. In addition to de-identifying the data, the researchers made sure to select only a sample of their full set of interviews[7] and availed themselves of one of the more stringent access control options QDR offers. They also agreed to revisit the last choice within a few years, as QDR's developing practices suggest more precise ways to assess both the level of risk that the anonymized transcripts might somehow be de-identified by others familiar with the context (currently estimated as very low to low for this project), and the severity of any potential repercussions for the participants, should their identity become known (currently estimated as low to medium, with the imagined repercussions being of a reputational and professional nature).

The lessons learned about copyright-related issues can similarly be grouped into "technological" and "sociological." Only one of the early depositors had considered the copyright status of the archival materials employed in their project before being asked about it by QDR. Their initial assessment was that they would not be able to share the

[7] In order to not lose information about social connections among actors, they intentionally chose not to remove data completely at random, but to sample from within one of the three localities where interviews had taken place—a choice documented in the data narrative.

pages digitally photographed during data collection. The terms of use of the relevant archive were the principal cause of concern. However, it seemed to the QDR staff that a more differentiated assessment was needed for the different types of materials the depositor had used, since a large portion of them were created by officials of the United States as part of their official duties and so most likely were in the public domain (another one of the exceptions that the copyright law allows for). Through discussions with the copyright counsel QDR uses, we established both that this was true for a subset of the data files and that within the context of an active citation type project which this piloteer was compiling, the fair use exception would apply to all the source materials in any case.

The fair use rule does apply extremely well to the suggested model of annotating one's work as done in active citation projects. This technique checks all the boxes for the four factors considered in establishing fair use: 1) the use of the original materials is transformative and adds original value; 2) it is done for non-commercial, explicitly scholarly purposes; 3) the portion of the material used is, relatively speaking, not substantial; 4) the use does not negatively affect the existing market for the original work. While active citations were developed primarily with the goals of analytic and production transparency in mind (Moravcsik 2014a), employing this novel scholarly technique could be one major way of providing at least partial secondary access to textual or audio and visual data, whose direct sharing might otherwise be prohibited by copyright ownership.

As an interesting aside that highlights the fluid nature of the regulatory and technical context within which QDR operates, in the course of curation work for this project, the archive from which the data had originally been collected underwent its own digitization initiative very much driven by motivations similar to some of QDR's broadest goals, i.e. to protect qualitative data in the form of historical assets and facilitate wider discoverability and access for further research. This changed the way the various materials could be linked to the points in the publication where they undergirded particular empirical claims. For a future project that might use the same archive's materials, a scholar will only need to provide the hyperlinks to the items now seamlessly available online.

A second use case is more typical of how QDR's intended depositors carry out their work and the copyright concerns that may commonly arise. A researcher had collected copies of over a thousand video recordings (some digital, others digitized from videotapes, or, in some cases, videotaped from television broadcasts and then digitized), created over almost two decades and in several different countries. He had inventoried all of them diligently (and so had created a lot of the substantive metadata we would need for repository purposes), but had not once wondered about the copyright ownership of any of them. In fact, for the pre-sharing academic uses he put these expansive data to (research, analysis and citation), he did not need to. Once the question of storing them with QDR and making them more widely available for other researchers came into play, however, that consideration became paramount. Once again, the diversity of sources (and, relatedly, potential copyright claimants) meant that a blanket assessment was neither useful nor feasible. But the amount of work it would take to investigate the situation for each recording was prohibitive, not even taking into account the different unfamiliar national jurisdictions that might need to be considered and the multiple foreign languages in which legal communication would have had to take place.

The solution in this case was two-fold: 1) in the short-term, QDR presents only a small number of items (fewer than fifty) from the rich collection, which the researcher had obtained directly from the production companies, giving at least implicit permission for further use by the copyright-holders; 2) during a second phase, QDR is planning to make five-to-ten-second previews of all videos available via a dedicated viewer. Anyone wanting to use the full set of materials will need to visit on-site and access the files on a computer disconnected from the internet (data enclave). In this instance, the intentional "diminishment" of the data could happen purely in quantitative terms, with fewer items or a small sample of each item being made directly available to comply with copyright constraints. Unlike in the anonymization example above, where the diminishment via considered aggregation of qualitative characteristics was substantive and could only be applied by the original researchers familiar with their subject matter, in this case the solution was technological and, once developed, can be applied to future cases that exhibited the same copyright constraint.

Another lesson presents a recurring theme in QDR's work: had the researchers planned for data archiving and sharing from the beginning of their work and been aware of these concerns[8] and the various possible solutions to them, they would have been in a much better position to tackle them during the course of their data collection and found themselves with much less to correct after the fact. This common sense conclusion emerged from QDR's work on every aspect of each pilot project, but it is hardly unique to qualitative data. Archivists and data management professionals across the disciplinary

[8] That is, human participant protection and copyright, for which external legal requirements need to be met as well and data documentation and file naming, where early preparation helps to prepare data for later sharing logistically.

spectrum have made the case for early data management consultations for a significant time.

The challenge for QDR is to create useful guidance texts to prepare the members of our user community, who are not accustomed to think in “data management” terms when embarking on research projects. We believe that the recent requirement by many of the leading funding agencies in the social sciences to submit a data management plan (DMP) along with all grant proposals will publicize the broader need for researchers to seek out such guidance and think through the data management issues most relevant to their type of data collection. And indeed, QDR has received some early inquiries based on grant-required DMPs. Additionally, the creation of a handful of new DMP databases[9] that cover a wide variety of research contexts will advance the opportunities for qualitative researchers in particular to see examples relevant for their approaches to data.

As an encouraging sign, QDR has already started to see a number of inquiries by scholars heading out to the field, asking for consultation exactly on the points of preparing informed consent forms for human participant protection and template

[9] Such as http://dmptool.org and http://rio.pensoft.net/browse_user_collection_documents.php?collection_id=3&journal_id=17.

language they can use to request copyright permissions for data sharing via a scholarly digital repository. We believe that thanks to appropriate planning and preparation for such issues, future project deposited with QDR will involve smoother curation and faster completion to publication.

Working with Granular Data

Annotating publications with excerpts, additional notes, and access to primary materials to create a transparency appendix (TRAX) is one of the most interesting newer techniques for making qualitative research more transparent. It is endorsed by the flagship journal of the American Political Science Association (2016: Note 5). For qualitative researchers, having a data sharing methodology that conforms to their actual practice rather than a poorly adapted quantitative template makes active citations particularly attractive. Browsing the Active Citation compilations currently available on QDR, enriching publications with materials ranging from old Soviet publications (Snyder 2015), diary entries from John F. Kennedy (Saunders 2015), to interviews with African judges (Ellett forthcoming) provides a sense of the promise of the methodology.

In working with a group of eight piloteers (beyond the five cited in footnote 1, three more are scheduled for publication during 2016), we have collected insights about best practices and possible impediments to adoption of active citations. The two principal obstacles to wider adoption are the mode of publication and the workload active citations impose on depositors.

All active citation compilations published by QDR contain the full text of the article or book chapter in question, with annotations unfolding in the text on click. This allows readers to move seamlessly between data and publication. However, given that so called "gold" open access, allowing for free republication of works, is virtually nonexistent in political science (Atchison and Bull, 2015: 130), this poses a dilemma. For each of the works published on QDR, authors needed to obtain special permission from their publishers, who hold the copyright. Publishers were willing to grant such permission, but undoubtedly they would be less inclined to do so if active citations or similar models become commonly used. Moreover, QDR, as a data repository, is not in a position to serve as a publisher of scientific literature, even where copyright allows. For active citations to become more widely used, they need to work with content published on publisher's websites.

Concerns about the additional workload data sharing places on researchers looms large in the debate about data policies and is not limited to qualitative data or the social sciences. In an early study on data sharing in genetics, Campbell and Bendavide (2003: 242) find that 80 percent of researchers who had withheld data reported the work required to produce the requested material as a reason. Among skeptics of data sharing policies for qualitative social science, the fact that such standards “impose a genuinely burdensome amount of new work” (Lynch, 2016: 37) is the most commonly cited objection. Such objections are not limited to “outsiders.” Two of QDR’s piloteers, reflecting on their experience note that “there is no getting around the basic fact that there remain significant costs to transparency that will be borne by individual scholars,” (Saunders 2014: 694) and that “At times it seemed that a yawning active citation sinkhole was about to open up and swallow all of my free research time and my assistant’s.” (Snyder 2014: 714).[10]

A closer look at feedback from piloteers suggests that the effort is due to a combination of factors. Several authors found instructions lengthy and hard to follow. They also encountered occasional technological issues which added to their frustration. Finally, even where technology and instructions worked, compiling the required information

[10] In spite of this honest assessment of the challenges, both authors do see significant benefits to active citations—all the more reason to take their concerns seriously.

into required forms and formats proved time consuming. Even where, for example, an author has a digital image of a document as well as a transcribed excerpt, attaching it to a given passage in an already authored text takes time. QDR has reacted to concerns about unsustainable amounts of work required for active citations by providing a streamlined workflow and continuously testing both technology and instructions so that researchers encounter a minimum of frustrations.[11]

These short-term efforts, however, provide only a partial solution. They provide no solution to the dilemma of publication. Moreover, they still leave a lot of "manual" labor for researchers to annotate their documents, even where they have all data present. In the conclusion of this article, we will discuss our vision forward that will, following the advice of Elizabeth Saunders, leverage "scholars' existing practices for capturing, storing, organizing, and maintaining data" (Saunders, 2014: 697) to achieve "transparency without tears."

[11] During the time of this writing, we are in a second series of user testing of the simplified instructions, which includes both feedback protocols and debriefing focus groups.

Future Directions

A lot of what QDR has managed to do on the basis of these pilot projects was to identify ways in which the original vision for tools and strategies could be made more complete and better integrated with solutions being worked on outside the immediate social science milieu that the repository serves. Through institutional and staff membership in organizations such as the Data Preservation Alliance for Social Science, the Research Data Alliance, the International Association for Social Science Services and Technology and the Annotating All Knowledge coalition, QDR is pursuing new ideas and new partners on several fronts.

Automating Annotations via Reference Managers

In promoting annotations for transparent inference (ATI), the “next generation” of active citations, the single most important issue will be to improve usability and decrease the additional burden annotations place on qualitative researchers. From countless conversations with researchers as well as several published accounts (such as Saunders 2014: 696-697), we know that researchers already store all information required for annotations in various software products. Any workflow that forces researchers to abandon their preferred tools will face stiff resistance and make adoption unlikely. What are the tools used by qualitative researchers to store data? From simple text documents and spreadsheets, to image organizing software like Picasa, to dedicated database products such as Microsoft Access, the variety of products used by researchers is enormous. Nevertheless, by far the most common tool used by researchers for storing their notes are reference managers such as Mendeley or Zotero.

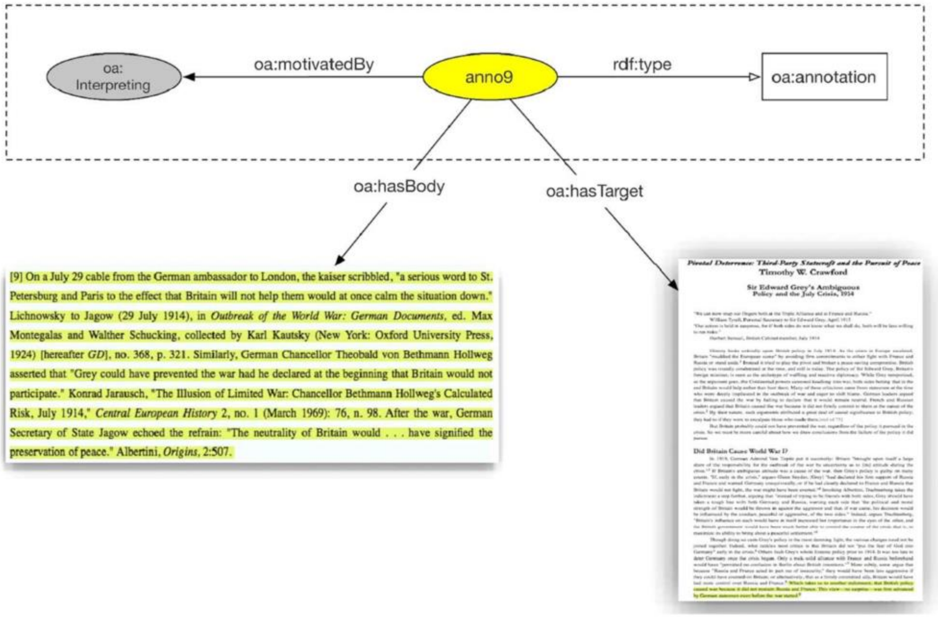
This is fortuitous, as these tools also interact with text in a way very similar to annotations (Moravcsik 2014 and Tonnesson 2012 both make this point independently). In other words, all that is missing is a tool that connects the information users already store in their reference managers, such as notes and source documents, with the references they are inserting into their manuscript using these tools. QDR is currently working on building such integration. While initial prototypes focus on Zotero,[12] any tools and workflows developed by QDR will be sufficiently flexible to extend easily to other tools.

Scholarly Annotation Tools

In future work, we also hope to leverage a number of emerging standards and tools for creating and representing the content of ‘annotations for transparent inference’ in

[12] The choice for Zotero is mostly pragmatic: it is not only one of the most widely used reference managers, but it is also open source and has a well-documented API, which makes integration into other tools particularly easy.

machine readable form. For example, the Open Annotation data model is a standard endorsed by the W3C consortium (Haslhofer et al., 2011). This model allows a number of different domains to - in a standardized way - express the relationship between a user-generated annotation and a web-based text. In the example below, we demonstrate how QDR could use the Open Annotation model to express the relationship between an annotation applied by an author to their writing:


oa:
Interpreting
oa:motivatedBy
anno9
rdf:type
oa:annotation
oa:hasBody
oa:hasTarget
Pivotal Deterrence: Third-Party Statecraft and the Pursuit of Peace
Timothy W. Crawford
Sir Edward Grey's Ambiguous
Policy and the July Crisis, 1914
Did Britain Cause World War I?

[9] On a July 29 cable from the German ambassador to London, the kaiser scribbled, "a serious word to St. Petersburg and Paris to the effect that Britain will not help them would at once calm the situation down." Lichnowsky to Jagow (29 July 1914), in *Outbreak of the World War: German Documents*, ed. Max Montegalas and Walther Schucking, collected by Karl Kautsky (New York: Oxford University Press, 1924) [hereafter *GD*], no. 368, p. 321. Similarly, German Chancellor Theobald von Bethmann Hollweg asserted that "Grey could have prevented the war had he declared at the beginning that Britain would not participate." Konrad Jarausch, "The Illusion of Limited War: Chancellor Bethmann Hollweg's Calculated Risk, July 1914," *Central European History* 2, no. 1 (March 1969): 76, n. 98. After the war, German Secretary of State Jagow echoed the refrain: "The neutrality of Britain would . . . have signified the preservation of peace." Albertini, *Origins*, 2:507.

Open Annotations for Transparent Inference
Figure 1

In this case, annotation 9 (anon-9) has a target in the text, and an explanation of the reason why the author has provided this annotation (to further interpret the main text). QDR could then model the motivation of this annotation using Open Annotation model's definition for "interpreting" as the reason for the annotation. In the future, we envision creating machine readable knowledge graphs (such as the figure above) that can be included alongside new qualitative social science publications or as a supplement to previously published manuscripts.

Similar to the Zotero extension described above, we also hope to provide tools that allow authors to easily apply annotations to existing publications. This will be done by building upon existing browser-based tools like hypothes.is[13], which allow users to apply annotations to any published text on the internet (e.g. blogs, scientific articles, newspapers, etc.). QDR strongly believes that in combining these tools with standards like the Open Annotation model we can provide a practical and less-burdensome path towards making annotation for transparent inference (ATI) a common practice among qualitative researchers.

[13] Hypothes.is: https://hypothes.is/about/

Our short time spent learning from and working with the existing data and annotation communities has provided us with confidence that the technological solutions are in an exciting emergent state where QDR's vision of ATI can fruitfully contribute to the development of new digital tools that will facilitate researchers' personal annotating for the purposes of greater transparency. The sociological challenge remains a greater hurdle however, which will take longer to solve. The current training approaches and professional incentives for social scientists do not immediately allow, even for those interested in adopting the available new practices and tools for data sharing, to dedicate the necessary time and attention to prepare conventionally collected data for sharing as a public good. The mandates that journals and funders are starting to create on that front however, might change the current calculations. While necessary, these changes may not be sufficient in the absence of a broader reconsideration of data sharing activities in hiring, tenure, and promotion decisions.

Outreach and Guidance

While this broader academic landscape is beyond QDR's control, the repository is committed, also on the basis of these early lessons learned, to sharpen and apply its own strengths in close interaction with each future depositor. QDR continues to design strategies and services for qualitative data sharing, to collaborate with journal publishers, to write up practical guidance materials and present data management classes at appropriate venues[14], thus educating its user community with the goal of lowering the sociological barriers for an individual researcher. In this ongoing process, QDR's staff are themselves learning from the established data management and information science communities, refracting all new knowledge through the prism of social scientists' needs. Whereas all researchers need to add data management skills to their work and apply relevant techniques for data sharing more broadly and research transparency more specifically, having a dedicated venue for qualitative and multimethod data which provides the relevant expertise for curation tailored to them and advocates on behalf of the scholars who create and use them should make this task easier for this user community.

[14] As an example of a new training initiative, we are currently preparing a teaching module which we will share with methodology instructors in Political Science graduate programs, so that they can introduce to their students the practical foundations of data management in a class session.

Works Cited

American Political Science Association (2016) APSR submission guidelines 2016 in brief. Available from: http://www.apsanet.org/PUBLICATIONS/Journals/APSR-Submission-Guidelines-2016-in-Brief#5 (accessed 16 May 2016).

Atchison A and Bull J (2015) Will open access get me cited? An analysis of the efficacy of open access publishing in political science. *PS: Political Science & Politics* 48(1): 129–137.

Bishop L (2009) Ethical sharing and reuse of qualitative data. *Australian Journal of Social Issues* 44(3): 255–272.

Boddy M (2001) *Data policy and data archiving: Report on consultation for the ESRC Research Resources Board*. Bristol: University of Bristol.

Bosk CL and De Vries RG (2004) Bureaucracies of mass deception: institutional review boards and the ethics of ethnographic research. *The ANNALS of the American Academy of Political and Social Science* 595(1): 249–263.

Broom A and Cheshire L (2009) Qualitative researchers' understandings of their practice and the implications for data archiving and sharing. *Sociology* 43(6): 1163–1180.

Campbell EG and Bendavid E (2003) Data-sharing and data-withholding in genetics and the life sciences: results of a national survey of technology transfer officers. *Journal of Health Care Law & Policy* 6(2): 241–255.

Carusi A and Jirotka M (2009) From data archive to ethical labyrinth. *Qualitative Research* 9(3): 285–298.

Corti L and Thompson P (1997) Latest News from the ESRC Qualitative Data Archival Resource Centre (QUALIDATA). *Social History* 22(1): 83-86.

Crawford T (2015) Data for: Pivotal deterrence and the Chain Gang: Sir Edward Grey's ambiguous policy and the July crisis, 1914. Active Citation Compilation, Syracuse, NY: Qualitative Data Repository [distributor]. Available from: http://dx.doi.org/10.5064/F6G44N6S (accessed 16 May 2016).

Elman C, Kapiszewski D and Vinuela L (2010) Qualitative data archiving: Rewards and challenges. *PS: Political Science & Politics* 43(1): 23–27.

Faniel IM, Kriesberg A and Yakel E (2015) Social scientists' satisfaction with data reuse. *Journal of the Association for Information Science and Technology* 67(6): 1404–1416.

Fear KM (2013) Measuring and anticipating the impact of data reuse. PhD Thesis, Ann Arbor, MI: University of Michigan. Available from: http://deepblue.lib.umich.edu/handle/2027.42/102481 (accessed 16 May 2016).

Fuji LA (2016) The dark side of DA-RT. *Comparative Politics Newsletter* 26(1): 25–27.

Fung B, Wang K, Chen R, et al. (2010) Privacy-preserving data publishing: A survey of recent developments. *ACM Computing Surveys (CSUR)* 42(4): 14.

Handlin S (2015) Data for: The politics of polarization: governance quality, left factionalism, and party systems in Latin America. Active Citation Compilation, Syracuse, NY: Qualitative Data Repository [distributor]. Available from: http://dx.doi.org/10.5064/F66Q1V52 (accessed 16 May 2016).

Haslhofer B, Simon R, Sanderson R, et al. (2011) The open annotation collaboration (OAC) model. In: *Multimedia on the Web (MMWeb), 2011 Workshop on*, IEEE, pp. 5–9.

Herrera V (2016) Data for: Commercialization and decentralization of local services provision. Active Citation Compilation, Syracuse, NY: Qualitative Data Repository [distributor]. Available from: http://dx.doi.org/10.5064/F6F769GQ (accessed 16 May 2016).

King G (1995) Replication, replication. *PS: Political Science & Politics* 28(3): 444–452.

King G (2011) Ensuring the data-rich future of the social sciences. Available from: https://dash.harvard.edu/handle/1/12724029 (accessed 16 May 2016).

Lagoze C, Willliams J and Vilhuber L (2013) Encoding provenance metadata for social science datasets. In: *Metadata and Semantics Research*, Springer, pp. 123–134.

Lupia A and Alter G (2014) Data access and research transparency in the quantitative tradition. *PS: Political Science & Politics* 47(1): 54–59.

Lynch M (2016) Area studies and the cost of prematurely implementing DA-RT. *Comparative Politics Newsletter* 26(1): 36–39.

Mauthner NS, Parry O and Backett- K (1998) The data are out there, or are they? Implications for archiving and revisiting qualitative data. *Sociology* 32(4): 733–745.

Medjedović I and Witzel A (2010) Sharing and archiving qualitative and qualitative longitudinal research data in Germany. *IASSIST Quarterly*: 42–46.

Mooney H (2011) Citing data sources in the social sciences: Do authors do it? *Learned Publishing* 24(2): 99–108.

Moravcsik A (2010) Active citation: A precondition for replicable qualitative research. *PS: Political Science & Politics* 43(1): 29–35.

Moravcsik A (2014a) Transparency: The revolution in qualitative research. *PS: Political Science & Politics* 47(1): 48–53.

Moravcsik A (2014b) Trust, but verify: The transparency revolution and qualitative international relations. *Security Studies* 23(4): 663–688.

Moravcsik A, Elman C and Kapiszewski D (2013) *A guide to active citation*. Syracuse, NY: Qualitative Data Repository. Available from: https://qdr.syr.edu/guidance/acguide (accessed 19 August 2016).

Parry O and Mauthner N (2005) Back to basics: Who re-uses qualitative data and why ? *Sociology* 39(2): 337–342.

Parry O and Mauthner NS (2004) Whose data are they anyway? Practical, legal and ethical issues in archiving qualitative research data. *sociology* 38(1): 139–152.

Qualitative Data Repository (n.d.) Our Mission. Available from: https://qdr.syr.edu/ (accessed 17 May 2016).

Rachel Ellett (Forthcoming) Democratic and judicial stagnation. Active Citation Compilation, Syracuse, NY: Qualitative Data Repository [distributor]. Available from: http://dx.doi.org/10.5064/F6PN93H4 (accessed 16 May 2016).

Saunders EN (2014) Transparency without tears: A pragmatic approach to transparent security studies research. *Security Studies* 23(4): 689–698.

Saunders EN (2015) Data for: John F. Kennedy. Active Citation Compilation, Syracuse, NY: Qualitative Data Repository [distributor]. Available from: http://dx.doi.org/10.5064/F68G8HMM (accessed 16 May 2016).

Sieber JE (1991) Openness in the social sciences: Sharing data. *Ethics & Behavior* 1(2): 69–86.

Sieber JE and Stanley B (1988) Ethical and professional dimensions of socially sensitive research. *American Psychologist* 43(1): 49.

Sieber JE and Trumbo BE (1995) (Not) giving credit where credit is due: Citation of data sets. *Science and Engineering Ethics* 1(1): 11–20.

Snyder J (2014) Active citation: In search of smoking guns or meaningful context? *Security Studies* 23(4): 708–714.

Snyder J (2015) Data for: Russia: The politics and psychology of overcommitment. Active Citation Compilation, Syracuse, NY: Qualitative Data Repository [distributor]. Available from: http://dx.doi.org/10.5064/F6KW5CXS (accessed 16 May 2016).

Tenopir C, Allard S, Douglass K, et al. (2011) Data sharing by scientists: practices and perceptions. *PloS one* 6(6): e21101.

Tonnesson S (2012) Active citation through hyperlinks: The retarded replication revolution. *International Area Studies Review* 15(1): 83–90.

UK Data Service (n.d.) Copyright scenarios for data sharing. Available from: https://www.ukdataservice.ac.uk/manage-data/copyright/scenarios (accessed 17 May 2016).

Van Den Eynden V (2008) Sharing research data and confidentiality: Restrictions caused by deficient consent forms. *Research Ethics Review* 4(1): 37–38.

Yoon A (2014) ‘Making a square fit into a circle’: Researchers’ experiences reusing qualitative data. *Proceedings of the Annual Meeting for the Association for Information Science and Technology* 12(1): 1–4.