2020

Making Research Data Accessible

Diana Kapiszweski
Georgetown University, dk784@georgetown.edu

Sebastian Karcher
skarcher@syr.edu

Recommended Citation

Kapiszewski, Diana, and Sebastian Karcher. 2020. “Making Research Data Accessible.” In The Production of Knowledge: Enhancing Progress in Social Science, edited by Colin Elman, James Mahoney, and John Gerring, 197–220. Strategies for Social Inquiry. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781108762519.008.

8. Making Research Data Accessible

Diana Kapiszewski
Sebastian Karcher[1]

Published as

Kapiszewski, Diana, and Sebastian Karcher. 2020. "Making Research Data Accessible." In *The Production of Knowledge: Enhancing Progress in Social Science*, edited by Colin Elman, James Mahoney, and John Gerring, 197–220. Strategies for Social Inquiry. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781108762519.008.

One of the key themes in this volume is that social science takes place in a community setting. As social scientists develop and answer their questions, they adhere to the norms and practices of their respective research communities. Over time, understandings of what being a responsible community member entails change. Today, members of social science communities are increasingly expected to provide access to the data they generate and use in their research (within ethical and legal constraints). Of course, discussions about openness in social science research have deep roots. In 1985, for example, Fienberg, Martin and Straf (1985, 25) called for sharing data to become a regular practice. A decade later, political scientist Gary King (1995) highlighted the importance of making available replication data and associated materials underpinning quantitative and qualitative research publications.

The last few years, however, have seen a marked acceleration in discussions about expanding access to research data across the social sciences—spurred on by broader technological and societal changes, as well as policy interventions by the White House, National Science Foundation, National Institutes of Health, and others. There is currently increasing momentum towards making openness the default position in social science research, and towards requiring that exceptions be based on established grounds. A key motivation for these discussions and this momentum is the belief that making data accessible impacts social science's ability to produce more credible and legitimate knowledge, and catalyzes scientific progress. Data access delivers these benefits in at least three ways: by allowing for secondary analysis of the data, by enhancing pedagogy, and by supporting research transparency.

Data that are accessible can be used by other scholars for different analyses. Indeed, for some large scale data collection projects (e.g., the American National Elections Studies [ANES], and the Varieties of Democracy [V-Dem] project), it is only the prospect of the data that are produced being shared that allows for the considerable investment of resources that data generation requires. Further, shared data can be used to enhance training in the types of techniques and methods used to generate and analyze the data. Students practicing methods on actual research data or on stylized

[1] An earlier version of the chapter was co-authored with Colin Elman; his foundational ideas made a significant contribution to the final piece.

datasets produced for pedagogical purposes has long been the norm in teaching quantitative research methods, and the same practice can and should be used to teach qualitative methods.

Finally, authors can use data they make accessible to more fully show the basis for the claims and conclusions in their work, thus making their research more easily understood. Making available the data that underpin scholarship also facilitates evaluation of that work (e.g., through reproduction, replication, verification, confirmation, and other processes -- see Freese and Peterson, Fairfield and Charman this volume). Such assessment increases the legitimacy and credibility of research publications (and thus the social scientific enterprise), and informs decisions about whether findings and conclusions offer a solid foundation on which subsequent inquiry, and potentially policy making, can be built. Put differently, the evaluation that access to research data facilitates can catalyze the accumulation of knowledge and the emergence of conversations and dialectics around scholarly work, linking innovation to confirmation, and discovery to appraisal.

This chapter argues that these benefits will accrue more quickly, and will be more significant and more enduring, if researchers make their data "meaningfully accessible." Data are meaningfully accessible when they can be interpreted and analyzed by scholars far beyond those who generated them. Making data meaningfully accessible requires that scholars take the appropriate steps to prepare their data for sharing, and avail themselves of the increasingly sophisticated infrastructure for publishing and preserving research data. The better other researchers can understand shared data and the more researchers who can access them, the more those data will be re-used for secondary analysis, producing knowledge. Likewise, the richer an understanding an instructor and her students can gain of the shared data being used to teach and learn a particular research method, the more useful those data are for that pedagogical purpose. And the more a scholar who is evaluating the work of another can learn about the evidence that underpins its claims and conclusions, the better their ability to identify problems and biases in data generation and analysis, *and* the better informed and thus stronger an endorsement of the work they can offer.

We advance this argument in several steps. We begin by clarifying what we mean by "social science data" and briefly considering the contrast between qualitative and quantitative data. In the chapter's third section, we discuss – and seek to de-mystify – the preparatory steps scholars should take so their shared data are meaningfully accessible. We emphasize the utility of careful documentation for those who generate data generator and for those who re-use them, and highlight that few of the preparatory steps we discuss require work beyond what many would consider good scholarly practice. Next we demonstrate the importance of utilizing emerging data infrastructure, offering a brief introduction to some recent innovations, and to the institutions and individuals responsible for these remarkable developments. Of course, the significant benefits that can accrue from making data meaningfully accessible will only materialize if large swaths of scientists do so. Thus in the chapter's penultimate section we consider we consider some institutional initiatives that could encourage more scholars to make their data accessible. We offer concluding thoughts in the final section.

What Are "Data"?

Definitions of "data" or "research data" are plentiful. Some definitions point to content, such as the following widely cited definition by the National Research Council (1999): "Data are facts, numbers, letters, and symbols that describe an object, idea, condition, situation, or other factors." Alternatively, data can be defined by their use, e.g. as "information used in scientific, engineering, and medical research as inputs to generate research conclusions" (National Academy of Science et al. 2009). Given that almost any piece of information can be employed as data for some research project, we believe the latter definition is more promising. For the purpose of this chapter we define data as *any representations of the social world relevant to a particular type of inquiry and rendered in a form suited to the analysis to be undertaken.*[2] In short, data are the empirical building blocks of knowledge production. Data are thought of differently across academic disciplines, and depending upon research substance, goals, and methods. Nonetheless, we believe this definition aptly describes the kinds of data considered in this chapter: those used in social science inquiry.

The second part of this definition serves, in the context of research, to distinguish data from information, and from data sources. Information in multiple undifferentiated forms is all around us constantly. For instance, it is trivially easy for laypersons to both contribute and receive information about the social world, whether on YouTube, Twitter, Facebook, or on their personal webpages. This more general flow of information could certainly be *converted* into data for use in a particular research project. What differentiates data from information is that data have characteristics that make them useful for the generation of knowledge. Scholars gather or create data sources (an archival document, a focus group transcript) that contain information; that information becomes *data* when a researcher transforms it into something they can use to measure or analyze, for instance.

Perhaps the most common way to distinguish *types* of data is to differentiate between qualitative and quantitative data. Rather than entailing precise categories, this distinction is based on loose family resemblance and general bundles of characteristics. Quantitative data tend to be numeric, organized into a matrix, and analyzed holistically using algorithmic/computational methods with the results represented in tabular form. Such data can also correspond directly to different levels of measurement (nominal, ordinal, interval, and ratio). Qualitative data tend to be non-numeric, and are often analyzed individually (e.g., a particular interview quote or passage from an archival document) or in small groupings to underpin particular claims or conclusions that form part of arguments; they are thus often deployed across the span of a book or article.

Qualitative and quantitative data also have different evidentiary strengths and present different analytic opportunities. Consider, for instance, a freeform answer to the question "how do you feel about presidential candidate XY" (qualitative) vs. the numerical score recorded in response to, "I'd like you to rate how you feel about presidential candidate XY on a feeling thermometer using a scale of 0 to 100." The qualitative data offer a richer depiction of the respondent's attitude and can be used to bring alive or nuance the description of a political context. The quantitative response includes measurement of an underlying concept into a one-dimensional core that can be more easily compared across time and respondents. Of course, "quantitative" and "qualitative" are not perfect categories.[3] After all, quantitative data are sometimes generated by quantifying qualitative (non-numeric) information. Likewise, "big" data are often "born" textual (qualitative), but converted to machine analyzable form and made susceptible to algorithmic and computational organization and analysis. Nonetheless, given the familiarity and ubiquity in the social sciences of the quantitative / qualitative distinction, we employ it here.

[2] Our definition is similar to that referenced by Borgman: "A reinterpretable representation of information in a formalized manner suitable for communication, interpretation, or processing" (2009, paragraph 25).

[3] One example of a more precise typology commonly used in the context of cyberinfrastructure categorizes data by origin: observational, experimental (resulting from lab or field experiments), computational (generated through executing a computer model or simulation), and records of public or private life (trace data) (National Science Board 2005).

Making Data *Meaningfully* Accessible

Making research data accessible in a meaningful sense involves ensuring that they are understandable and interpretable by scholars other than those who generated them. Consensus is beginning to develop on a set of core attributes that well-provisioned social and natural (e.g., physical, biological) science data should have and on the kinds of information that should supplement them. Ensuring that research data are marked by these characteristics augments the value, quality, and usability of the data *for the scholar who generated them* as they use and reuse them over time. It also helps that scholar to decide what information to provide with their data to make them meaningfully accessible to others, and facilitates the provision of those materials – as well as increasing others' comprehension of, and thus ability to employ, shared data.

Creating detailed documentation is an indispensable part of making data accessible.[4] One key type of documentation is the Data Management Plan (DMP), generally associated with a particular research project. A DMP, which scholars should begin to write as they start to design a project, is a formal document discussing how the data generated through the project will be handled (e.g., cleaned, verified, formatted, and organized) as they are generated and analyzed, and once the project has concluded. Developing and following a DMP helps scholars both to make the data they generate a stronger empirical base for their own work, *and* to make them more intepretable by others. DMPs are most useful when they are "living documents" – updated regularly to reflect changing needs and record important decisions as a research project develops (see e.g. Michener 2015). Ongoing work also aims to make DMPs machine actionable (see Simms 2018). Machine actionable DMPs would automate communication among key stakeholders such as researchers, research support staff, ethics boards, funders, and repositories, facilitating alignment between IRB documents and data sharing plans, and allowing researchers to more easily notify funders of compliance with the latter's data sharing requirements, for instance. By automating information flow, machine-actionable DMPs can also augment efficiency by helping researchers to avoid duplicate metadata entry and reporting requirements.

More broadly, documenting data entails outlining various types of descriptive metadata (geography, time period, etc.) to characterize the data, detailing and justifying the multiple steps taken to generate the data, and assembling all relevant research tools (e.g., interview protocols, survey questionnaires). Discussing data generation entails clearly describing and rationalizing how information encountered in the social world was selected and collected (e.g., how the scholar decided where to focus, what to ignore, and what information to record and how); how collected information was interpreted and transformed; and what each step in the data-generation process implied for subsequent steps, and the project as a whole.[5]

Documentation almost always includes certain types of information (for example, date of collection). Yet how data are documented depends heavily on the nature of the data and how they were generated. For instance, there is consensus that scholars who produce survey data should provide details such as response rates, relevant patterns of non-response, and detailed description of

[4] Creating documentation is also a critical aspect of data management. As discussed in the next section, data repositories can aid scholars in creating documentation; they also extract metadata from that documentation and generate additional types of metadata (e.g., structural and administrative) that augment the accessibility of the data.

[5] To mention just a few suggestive ideas, scholars might discuss why particular secondary sources (and certain parts of those sources) were read; why particular political actors were interviewed and asked certain questions; or why the contents of certain boxes and folders were explored in an archive, certain documents perused, and certain passages identified as relevant – and whether information was collected by recording it in full (e.g., via audio taping or taking digital pictures), via paraphrasing, or some other way.

the sampling framework (AAPOR 2017). Likewise, there is agreement that experimental researchers should offer information about the context of the experiment, all protocols, and the sampling framework (see, e.g., Druckman, Green, Kuklinski, and Lupia 2011).

Beyond these baselines, however, there are few hard and fast rules; researchers should produce documentation that they believe will best bring their data alive for other scholars, and for their "future selves" as they re-use their data over time. We hasten to highlight again that many of the processes entailed in preparing data so they can be meaningfully accessible to others simply represent good data management practices, and basic aspects of conducting systematic social science research, in which many scholars *already are* engaging.

Exactly which data (and accompanying documentation) a scholar will make accessible depends on the goals of the research, the objective of data sharing, and attributes of the data. In large scale, multi-researcher projects like ANES and V-Dem, the goal of making data accessible is inherent: such projects were intended from their inception to produce data for secondary analysis, i.e., to create a public good. These endeavors almost invariably require substantial external funding, which is often contingent on the data produced through the project being accessible to the social science community at large. In such projects, producing the documentation needed for others to effectively interpret and use the data is a central aspect of the project from the outset.

Much social science research data is not generated in the context of large-scale projects, however. Instead, individual scholars or small groups of scholars generate data with the primary goal of answering their own research question. The data produced through such efforts are often referred to as the "long tail" of science (e.g. Wallis, Rolando, and Borgman 2013). In these scenarios, what data and documentation scholars will make accessible depends in part on the purposes for which they are doing so. The two most common purposes are to allow for secondary analysis, and to increase the transparency of research.

First, scholars may share data to allow for their secondary analysis. Here scholars make meaningfully accessible a coherent set or collection of data relevant to a particular project or theme. The data need not be tightly tied to any particular research publication. Sharing data so they can be analyzed by others represents an invaluable contribution to the production of knowledge. Nonetheless, this objective is sometimes underemphasized in the scholarly debate (particularly discussion about sharing qualitative data), which tends to focus on the feasibility and utility of making data accessible for the purpose of research transparency. Meaningfully accessible data serve as a multiplier for knowledge generation, and are of particular benefit to scholars in the U.S. and other countries who lack the resources to generate their own data: for many, meaningfully accessible data generated by others *are a prerequisite* for conducting research. Happily (if anecdotally), sharing data for secondary analysis seems to be occurring more frequently, perhaps due to increasing disciplinary recognition that data represent distinct products of value, to scholars' complying with funders' data-sharing mandates, and to changing norms about best practices.

Second, scholars also make the data they generated accessible for the purposes of increasing the openness of publications based on those data. Here the purpose is to surface the scaffolding supporting the conclusions offered in the publication, making the analysis understandable and evaluable. In this scenario, the author's obligation is to provide the data that underlie the published analysis; this may be the full dataset that they created or a subset thereof sufficient to evaluate the publication (a "replication dataset"). If the two are different and the former is not publicly available, the author must provide complete documentation about the replication dataset and its creation – *in addition to* providing the supplemental analytic materials that are typically required for analytic transparency.

Which data scholars make accessible, and how accessible those data can be made, also depends on certain characteristics of the data, as discussed in more detail in the next section. Not all shared data are freely accessible. "Open data" are entirely unrestricted, i.e., accessible by anyone and with no or minimal restrictions on their re-use. The benefits for knowledge production of open data notwithstanding, legitimate limits on data access exist. On the one hand are proprietary limitations: access to the data may be limited by the commercial arrangement under which they were obtained, or they may only be available to scholars affiliated with institutional members of a particular repository. Here the restrictions are part of a business model, helping to cover the provider's costs. Removing such limitations, without providing alternative revenue streams, could undermine the sustainability of repositories that rely on a paywall model (Ember and Hanisch 2013, Hodson 2016). On the other hand, the scholar who shared the data may have requested that access controls be imposed. This could occur, for instance, when the data were generated through interaction with human participants, are sensitive, and cannot be fully de-identified, meaning open access to the data could pose a risk to participants. While access controls appear to reduce data availability, in fact they allow data sharing when it would otherwise be impossible.

In sum, no matter for what purpose data are shared, or what limits ethics and the law may place on sharing, making data meaningfully accessible requires thorough and informative documentation of the type we have described. Yet a sobering caveat on the relationship between data accessibility and the production of knowledge bears noting. Research data can only be used to produce knowledge (understood as comprising truth claims) if the data accurately capture/reflect empirical reality (garbage in, garbage out). Absent repeating the data generation process (i.e., without comparing the data to the empirical reality a scholar claims they reflect), the only way to gauge data's validity is by assessing the quality of the processes through which they were generated. But given the particularities of research, objectively evaluating data generation is difficult; moreover, people can describe data generation processes in ways that makes them appear more systematic than they were. Indeed, notorious cases of fraud, such as LaCour and Stapel, involved detailed descriptions of how data were (supposedly) collected. Moreover, even if data generation seems to have been systematic and robust, without seeking to re-generate the data, we cannot be sure that they reflect empirical reality (i.e., whether the building blocks of subsequent analysis capture truth). We return to this point later in the chapter.

Evolving Infrastructure for Data Accessibility

Making data meaningfully accessible requires that scholars avail themselves of the increasingly sophisticated technology and infrastructure that have been developing over the last decade to facilitate the uploading, storing, indexing, browsing and searching, and downloading of data and associated scholarly materials. Using that infrastructure also benefits scholars themselves, as storing data in institutionalized venues helps protect the data from damage and loss, increases the visibility of the data and thus their creators, and integrates them in credit-awarding systems through citation and other metrics. How the complex web of institutions and technicians who build the infrastructure for storing and transmitting data do so has a profound impact on how "FAIR" data can be (findable, accessible, interoperable and reusable, Wilkinson et al. 2016), and on how social science is undertaken, represented, evaluated, and validated. The better scholars understand how that infrastructure looks and works, the more effectively they can use it to make their data meaningfully accessible, and the more quickly data sharing will become routinized and standardized.

Venues for Sharing Data

Until recently, by far the most common way for scholars to make their data accessible was to indicate that the data and analysis code associated with a particular publication were available "by request". A popular blog chronicles corresponding requests to authors, many of them unsuccessful (https://politicalsciencereplication.wordpress.com/category/replication-correspondence/).[6] Another common method was and is for scholars to post data and code for their studies on their personal website. Such posting does initially allow easy access to data. However, this mode of sharing exposes data availability to "linkrot" (i.e., hyperlinks pointing to web resources that have become unavailable). For instance, more than half of the reproducibility links in articles from the *American Political Science Review* between 2000 and 2013 could not be accessed in 2016 (Gertler and Bullock 2017, 167). In part, this is due to individual researchers lacking the technical knowledge and resources to guarantee the accessibility and long-term preservation of their data when, for instance, their institutions change technologies or they change institutions.

Another option is to include data as supplementary material to journal articles, to be stored by the journal's publisher. While publishers have significant experience in preserving digital publications, they rarely extend their preservation promises and practices to supplementary material (see e.g. Smit et al. 2011, 43-44 for the broader landscape, and Butler and Currier 2017 for data in economics journals). They also exert greater effort to make the articles that appear in their journals easy to find and to cite than they do to make supplementary materials accessible. Given that data shared as supplementary materials can often be useful to other scholars, publishers handling those data without due attention to their careful preservation and indexing may do a disservice to the scientific enterprise.[7]

Dedicated repositories for publishing and preserving digital data are designed to avoid these pitfalls, and are multiplying. Broadly speaking, we can distinguish among three types of repositories: **s**elf-service repositories, which are typically open to all research data (and in some cases other materials); institutional repositories, which are operated by universities or other research institutions and accept pre-prints, working papers and, increasingly, research data generated by affiliates of that institution; and domain repositories, which focus on a specific discipline or group of disciplines (e.g., "social science" or "earth science") and provide specialized services for data commonly used in those disciplines.

Self-Service Repositories

Self-service repositories are the newest type of venue; most were founded after 2000. Well-known examples include figshare (a for-profit company), Zenodo (run by CERN, the European Organization for Nuclear Research, and funded by the European Union), and Harvard Dataverse (run by the Institute for Quantitative Social Science at Harvard University). While hard to assess

[6] Of course, "on request" arrangements can sometimes make data accessible. While an individual and potentially unrepresentative example, in Karcher and Steinberg (2013), one of us requested replication materials from six then recent studies and received data and code (which successfully replicated the papers' results) in every case. Nevertheless, even when they work as hoped, such ad hoc arrangements cannot provide the benefits of data sharing via institutional venues with robust routines and systems for making data available.

[7] This is not as controversial a statement as it may seem. In 2010, the *Journal of Neuroscience* stopped accepting supplementary materials, declaring data repositories "vastly superior to supplemental material as a mechanism for disseminating data" (Maunsel 2010). Most large journal publishers, including Elsevier, Springer, Wiley, and PLoS appear to agree, and now recommend data publication in repositories across all their journals (see Kratz and Strasser 2014 for a general discussion of this).

precisely,[8] the holdings of such venues probably comprise the largest number of individual datasets worldwide. For example, as of December 2018, figshare has more than 80,000 datasets, Harvard Dataverse holds 30,251, and Zenodo has 33,794.[9] Self-service repositories allow easy upload of data of any kind for all researchers. Both deposit and download are free of charge. While convenient and inexpensive, self-service repositories rely heavily on the expertise and efforts of depositors. Typically, deposits are either not reviewed / curated or are only minimally reviewed / curated by staff, and depositors are responsible for supplying cataloging information. Self-service repositories are commonly dedicated to "bit-level" preservation of data files, i.e., they guarantee that any deposited file can be accessed "as is" in the long-term, using multiple, geographically dispersed back-up copies. Generally, they neither check that files are valid (i.e., open correctly in the specified software) nor protect against file-format obsolescence (the inability to open old files with currently available software).

Institutional Repositories

Institutional repositories, most run by college and university libraries, have traditionally been more concerned with holding and making available the publications of their institutions' researchers than with facilitating access to the data underlying those scholars' research. However, recently many institutional repositories have begun to accept research data. Such repositories' proximity to the researchers depositing the data may mean researchers have greater trust in the institution, and allows for immediate contact. Moreover, as data librarians are likely to be a first point of contact for scholars with questions about DMPs, such venues can and do often provide researcher-repository contact across the lifecycle. Furthermore, libraries have significant expertise in the preservation of digital formats. Even at large research institutions, however, libraries (and thus institutional repositories) often lack the information technology and subject-specific capabilities to provide curation, preservation, and dissemination guidance and services on par with domain repositories (see Johnston et al. 2017).

Domain Repositories

Domain repositories, which focus on a specific discipline or subject area, have the longest history of the venues considered here. With regard to the social sciences, the Inter-university Consortium for Political and Social Research (ICPSR), which began operations in 1962, is probably the most prominent domain repository of its type in the United States. Other examples include the Roper Center's public opinion archive (Cornell University), which dates to 1957, the Odum Institute's Data Archive (University of North Carolina, Chapel Hill), and the more recently established Qualitative Data Repository (QDR, Syracuse University) with which the co-authors of this chapter are affiliated. Some domain repositories are "trusted data repositories," a status conferred through a certification process. The most common certification among social science data repositories is the CoreTrustSeal (https://www.coretrustseal.org/).[10]

[8] Most venues host materials other than data, making it difficult to discern what exactly should be counted and how multiple versions are counted.

[9] Both Zenodo and figshare also hold figures, presentations, pre-prints, and other materials, so the total number of items in these repositories is significantly higher, e.g., around 3 million in figshare.

[10] To obtain CoreTrustSeal (CTS) certification, repositories provide in-depth documentation about their compliance with best practices in 16 areas under three broad headings: organization infrastructure, handling of digital objects, and technology. Their answers are peer reviewed; if they are approved, the repository is awarded certification. Currently, 39 data repositories hold CTS certification, 39 repositories hold Data Seal of Approval (DSA) certification (the pre-cursor to CTS), and 61 hold World Data System (WDS) certification. Among US social science repositories, ICPSR and the Odum Institute Data Archive hold the DSA certification; QDR and the Roper Center hold CTS

Domain repositories offer the broadest set of services and guidance to depositors, including data curation and preservation, and promoting deposited data and monitoring their use. One considerable strength of domain repositories is their curation services. Curation is a multi-stage process potentially entailing appraising and selecting certain data for publication from a large deposit, processing and storing the data, describing them with metadata, providing access, preserving, and assessing re-use (Johnston 2016, vol. 2, xiii). Curation is facilitated by researchers engaging in proper data management, on which domain repositories often advise them. Curators may also assist researchers during the deposit process. Upon receiving files, curators perform a variety of checks and ensure that the files are in, or convert them to, a format that is suitable for long-term storage; files are then stored, together with information about the curation steps taken. In addition, to make data easy to find (encouraging re-use and citation), curators optimize the "metadata" (information about data creation and content) associated with the data. In the social sciences, the most sophisticated metadata format, Data Documentation Initiative (DDI, http://ddialliance.org/), holds information about every variable in a quantitative dataset, allowing domain repository users to search for datasets by the text of individual variables.

Domain repositories are also commonly best-equipped to store, preserve, and provide access to sensitive data. Making research data accessible can be legitimately constrained by the need to respect the rights and protect the safety of the people who participated in the research (and, more broadly, sites of investigation) when participants have offered sensitive information or when they request that the information they conveyed remain confidential. Fortunately, information scientists are developing strategies, tools, and technologies to facilitate making the information garnered through encounters with human participants more accessible in an ethical way, and domain repositories are capitalizing on these developments. Such repositories can aid researchers in assessing the risk involved in sharing data; and can help them to develop a strategy to de-identify interview or focus group transcripts, field notes, or other data sources resulting from human interaction can help to maintain confidentiality.[11]

Domain repositories also have (to differing degrees) the technological capacity to provide secure access to sensitive data.[12] As mentioned previously, repositories generally allow depositors to place "access controls" on particular data files that limit who may view or download them and how they may do so. For instance, dedicated data security training or secondary IRB approval may be required before access is granted, or access may only be permitted from secure terminals or through dedicated secure connections). Domain repositories worldwide, in particular those overseeing large national studies with detailed observations on participants, also provide secure modes of remote access to sensitive materials.[13] Some offer tools that allows scholars to analyze quantitative data

certification. Other, more demanding certifications exist (e.g., the International Standardization Organization's standard 16363) but are rarely used.

[11] De-identification refers to removing from data and documentation direct identifiers (i.e., information that is sufficient, on its own, to disclose identify) and indirect identifiers (information that in combination with other available information may disclose identity). Some hold that it may be difficult if not impossible to fully de-identify some data; there are also clear tensions between de-identifying data and maintaining their analytic utility.

[12] An interesting development in this regard is the DataTags project (Bar-Sinai et al. 2016), which automates the handling of sensitive data, in particular the assessment of sensitivity and required protections of a given dataset. However, access requests and monitoring of compliance with usage restrictions still require human oversight.

[13] For instance, users of ICPSR's "virtual enclave" can analyze data housed in the enclave but not retrieve data files (and ICPSR staff can review users' analysis outputs for disclosure risk).

without fully accessing them, e.g., by allowing code execution on a remote server and screening output for potential privacy issues such as low cell counts in cross-tabs.[14]

Finally, domain repositories also play important roles in promoting the research data they hold, making them searchable (and findable), monitoring how they are used, and facilitating the awarding of credit to scholars who generate data (and whose data are used by others). Domain repositories link data to publications in which they are used and/or cited and showcase data holdings via blogposts, press releases, and infographics.

Hybrids

Three other repositories—Dryad, UK Data's ReShare, and OpenICPSR—assume a hybrid position. Dryad, which started out as a domain repository for bio-medical data, now accepts data across disciplines, and performs curation services including metadata improvements and file reviews. Dryad curators are generalists, rather than domain specialists, so curation focuses more on the formal elements of data publication such as file integrity and de-identification. Re-Share and OpenICPSR are social science self-publishing repositories run by and alongside fully curated domain repositories. OpenICPSR is minimally curated, performing a metadata review after publication. Data published on ReShare benefit from reduced but still significant curation work, including several project-level checks and checks on a subset of submitted files.

Common Features

Despite these distinctions, all repositories perform a series of critical functions that aid scholars in making their data meaningfully accessible, establishing repositories as superior venues for storing and publishing data. For instance, most repositories can assign persistent identifiers such as Digital Object Identifiers (DOIs) to their data assets, enabling long-term access to digital resources and facilitating searchability.[15] The broader DOI system, i.e., the technical and social infrastructure for the registration and use of DOIs, provides metasearch functionality for repositories that issue DOIs: on search.datacite.org, researchers can search across all datasets that have a DOI. Once fully employed, the DOI system will greatly facilitate data generators receiving credit for the citation and re-use of the data they generated and shared, and also significantly enhance scholars' ability to find data relevant to their work. CrossRef and DataCite are already collaborating to collect data on data citation in scholarly literature (www.scholix.org). Each time a dataset is cited in a work registered with CrossRef, the event is cataloged, establishing an automated citation count; a similar count system is used to catalog mentions of datasets on blogs or social media. Such "altmetrics" increasingly complement more traditional measures such as citation counts to assess the influence of data and the scholarly impact of work more generally (Costas, Zahedi, and Wouters 2015).

Most repositories also allow the "harvesting" of their metadata through a dedicated protocol (Open Archives Initiative Protocol for Metadata Harvesting, OAI-PMH), facilitating the development of search interfaces covering many data repositories – so called meta-catalogs (similar to what Open Worldcat represents for books and Web of Science for journal articles). For example,

[14] Scholars should begin considering ways to address these ethical challenges to data accessibility – as well as legal constraints and proprietary obligations that may also limit the amount of research data they can share – when they first begin to design the research projects, and should discuss them in their DMP.

[15] Outlets that publish data and other research products "register" the DOIs they assign, and deposit standardized metadata, with a registration agency. The oldest and largest such agency, CrossRef, mainly registers journal articles, books, and book chapters. DataCite focuses on datasets, and its metadata catalog (search.datacite.org) thus provides a powerful metacatalog of research data.

the metadata of several social science repositories are included in the Dataverse Catalog, where users can find entries for all of ICPSR's, the Odum Institute's, Roper's, and QDR's holdings.[16]

All repositories have staff, technical, and material costs that need to be covered in order for data to be curated and preserved (Ember and Hanisch 2013, Hodson 2016). Institutional repositories are obviously funded by the institution with which they are associated. Domain and hybrid repositories, as more independent entities, can face significant sustainability challenges. In many European countries such entities are considered fundamental to scientific infrastructure, and social science domain repositories (e.g., the UK Data Archive, GESIS in Germany, and DANS in the Netherlands) are financed by permanent government support. Because government funding is more ad hoc in the U.S., repositories must seek out other sources of support. Figshare, for instance, sells technical services to universities (i.e., it functions as a service provider for institutional repositories). The most common model, however, is for repositories to charge users (directly, or indirectly through establishing institutional memberships). As noted above, some repositories restrict *access* to data based on paid membership (e.g., ICPSR) while others charge depositors for the *curation* of data (e.g., Dryad, the Odum Institute Data Archive, and QDR). Neither option is unproblematic. Charging for access violates increasingly strong norms of "open science" and "open data" by making access contingent on the ability of a researcher's institution to afford membership (thus disadvantaging citizen scientists and researchers from less well-resourced institutions and/or countries). Charging for curation may deter deposits and reinforce the same inequities.

Organizations that Support Data Infrastructure

The data infrastructure just described is embedded in a complex set of interlocking and overlapping stakeholder organizations that seek to develop and promulgate policy consensus on the handling of research data. Perhaps most directly relevant to and useful for social scientists (although less likely to set policy), IASSIST (the International Association for Social Science Information Services & Technology) provides a venue for exchange between social science data specialists, mainly from research libraries and repositories. Via an active listserv, IASSIST members discuss ongoing developments related to data infrastructure and help each other to answer data-related questions posed by members of their respective communities, thus putting an international network of data experts at the service of social science researchers.

A host of other organizations set policies and guidelines for the venues through which scholars make their data accessible, and are therefore also important for social science and social scientists, if indirectly so. The most influential research-data organization is the Research Data Alliance (RDA). Founded in 2013, today RDA has over 5,000 members and holds two well-attended meetings annually. In a sprawling net of interest groups and working groups, RDA is shaping policies that will have a significant impact on the future of research data. For example, two of the above-mentioned initiatives—the recent revision of the CoreTrustSeal guidelines and the Scholix project for collecting citation count data for DOIs—were developed under the auspices of RDA. Other working groups are currently finalizing recommendations for citation of complex data objects and licensing of data for cross-border exchange.

Another group that warrants mentioning is FORCE 11. Founded in 2011, FORCE 11 is dedicated to advancing scholarly communication—improving how "knowledge is created and shared," in the words of the organization's mission statement. Many of FORCE 11's activities are data related. It coordinated an influential declaration of data citation principles (https://www.force11.org/group/joint-declaration-data-citation-principles-final), which has since

[16] On perhaps the largest scale, Share (share.osf.io) harvests metadata on any research output from (as of this writing) 159 different sources and presents it in a metacatalog (where searches can be restricted to datasets).

been endorsed by most scientific publishers, and is shaping requirements for data citation in scientific journals. Elsevier, Springer, and Wiley—the world's three largest academic publishers—are pushing their journals to adapt data citation guidelines based on these recommendations.

Finally, Data-PASS (the Data Preservation Alliance for the Social Sciences) is a voluntary partnership of major U.S. social science domain repositories. The principle goal of the alliance is a mutual guarantee of holdings: if a Data-PASS member ceases operations, other members agree to assume stewardship of the data. To facilitate this process, Harvard Dataverse contains a metadata catalog of the data holdings of most Data-PASS members. Data-PASS members also provide expertise and advocacy around data-related issues for social scientists and their organizations.

Data infrastructure helps researchers make and keep their data meaningfully accessible, and helps researchers find data. It helps scholars receive credit when their data are employed by others, and helps prevent data from being used inappropriately. Familiarity with the contours and function of that infrastructure and with the institutions that underpin it allows scholars to more effectively share and search for data. Built under the leadership of, and through cooperation among, scientist of all stripes – applied, behavioral, natural, and social – that infrastructure is constantly adapting as the requirements of contemporary research evolve.

Catalyzing Access to Research Data

Over the last four decades – and during the last ten years in particular – significant progress has been made toward establishing data accessibility as a scholarly norm in the social sciences. Nonetheless, the substantial benefits than can accrue as a result of greater data accessibility will only begin to accrue if more social scientists – and more scholars who generate and analyze qualitative data in particular – make their data meaningfully accessible. In this section we consider some steps that could be taken to support further movement toward that goal.

Multi-Stakeholder Conversation and Coordination

Continuing and accelerating the increasingly inclusive conversations that have emerged since 2010 about making research data accessible are critical.[17] The more scholars engage each other in conversations about the challenges of sharing research data, the more quickly creative solutions to those challenges can be developed. Data access has been a crucial part of the groundbreaking Qualitative Transparency Deliberations (www.qualtd.net) in political science, and these discussions could form the basis of continuing debate, as other conversations could continue in parallel. Engaging all of the research traditions that comprise each social science discipline will make it possible to develop solutions and standards that are appropriate for the very different styles of research that comprise each one, and to create guidance materials and other resources to aid scholars in meeting those standards.

These conversations should include leadership from social science's multiple academic associations, as well as its main gatekeepers—funders and publishers. Associations are uniquely positioned to reach scholars on all sides of each discipline's various divides from a relatively neutral standpoint. They could capitalize on that position to play a leading role in encouraging greater data access and catalyzing thoughtful discussion about how scholars can be motivated to make their data available. Funders and publishers are also well-positioned to encourage data access given their

[17] Earlier literature from multiple scientific disciplines is a rich resource for such conversations, e.g., Mauthner, Parry, and Backett-Milburn 1998; Bishop 2005, 2009, 2014; Parry, and Mauthner. 2004 and 2005; Fielding 2004; Heaton 2008; Mauthner and Parry 2009; Corti et al 2014; Yardley et al 2014; Broom, Cheshire, and Emmison 2009.

influence over the resources that support our research and the ways in which our results are disseminated (and scholarly work is rewarded).

Outreach to and partnership with the Institutional Review Board (IRB) community will also be decisive,[18] in particular given impending changes in the Common Rule.[19] Scholars from the social science community can help those involved in designing and implementing research review – members of university leadership, faculty who sit on such boards, and administrators alike – to visualize the difference between medical research, hard science research, and social science research. Such partnerships can also help IRB personnel to see the value of making research data meaningfully accessible. Discussions around providing template informed consent protocols that offer varied options for handling and sharing the information conveyed through interactions with human participants, and around when and how the risks of sharing data can be mitigated, could be important first steps toward generating a culture of data sharing.

As they take on these challenges, stakeholders will benefit greatly from partnering with data repositories (and, through them, connecting with other stakeholders in the data management and infrastructure community). Repositories specialize in storing, publishing and preserving shared data and can thus offer invaluable assistance as stakeholders consider where and how they should require scholars to make their data accessible. Moreover, as repositories are experienced at interacting with scholars over their research data, association members, grantees, and authors can benefit greatly from consulting with data repositories for advice about making data accessible.

The most fruitful conversations may result from all of these stakeholders working together – and with researchers – listening to each other's concerns, appreciating each other's ideas, and developing shared standards and guidelines. There is undeniable value to stakeholders harmonizing their standards, for instance, by coordinating on a pre-existing solution such as the Center for Open Science's (COS) Transparency and Openness Promotion (TOP) Guidelines (https://cos.io/our-services/top-guidelines/). This sort of harmonization reduces the developmental burden on stakeholders, allows them to learn about implementation from each other, curtails inequities with the potential to skew researchers' incentives in unhealthy ways, and makes it easier for scholars to develop practices to meet those consensual standards.

Shifting Incentives: Awarding Credit for Making Original Data Products Accessible

While these multi-stakeholder conversations will be wide-ranging, one key topic should be how disciplinary incentive structures can be shifted to revalue the processes of generating data and making them accessible. In most social science disciplines, textual publications (articles, chapters, books, etc.) remain the coin of the realm. Acts of data generation, by contrast, are (implicitly) undervalued – perhaps due to a generalized bias toward appraisal over innovation. Certainly datasets (in particular quantitative datasets) that scholars develop can be and are listed on their CVs. However, datasets and research publications are not valued equally in merit review and promotion processes; in fact, datasets are sometimes simply included in the assessment of the publication they underlie rather than being considered a separate scholarly product.

As such, for many scholars, investing time and money in making the data they have generated meaningfully accessible may seem to simply syphon resources away from career-advancing

[18] One avenue to do so is through interacting with the PRIM&R (Public Responsibility in Medicine and Research) organization; founded in 1974, PRIM&R seeks to strengthen the community of research administration and oversight personnel, and offer opportunities for education and professional development (https://www.primr.org/).

[19] The Common Rule (1981) concerns ethics in biomedical and behavioral research involving human participants. It has been undergoing revision, and an amended version is expected to go into effect in 2019.

activities.[20] More broadly, it may seem to introduce inefficiencies into the research process that disproportionately handicap some scholars, and put a drag on the collective production of knowledge (in the form of textual publications). In short, the perceived imbalance between the professional pay-off from, and the practical demands of, making data accessible discourages scholars from sharing their research data.

Social science should take steps to foment greater appreciation for data generation activities (and their fruits). Data are the fuel that powers our research – the lifeblood of our scholarship. Considerable resources (both time and money), expertise, and effort are required to generate data, and doing so represents an immense contribution (Lupia and Alter 2014). Shifting disciplinary reward and credit structures to revalue data generation will encourage individual scholars and teams of researchers to generate more data, and organically incentivize providing greater access to data (i.e., showcasing data generation). The greater availability of more research data could, in turn, trigger an evolution in views of the data lifecycle so that seeking to maximize data's reuse potential becomes standard practice. Additional credit could then be awarded to those scholars who render their original data products meaningfully accessible.[21]

Since the possibility of measurement is a prerequisite for reward, perhaps the easiest transition would involve creating mechanisms for data-related activities and achievements to register on familiar accomplishment scales. For instance, published datasets (quantitative and qualitative) could be counted in review and evaluation processes as stand-alone research products rewarded independently from any published articles, chapters, and books that they accompany. Further, bibliometrics (see the chapter by Gerring, Apfeld, and Karcher) for datasets – systematic counts of how many times they are cited, downloaded, and used – are being developed (e.g., www.scholix.org, makedatacount.org). There are also promising initiatives for recognizing when scholars share data in tandem with a publication; for instance, COS has created "badges" to acknowledge practices such as making the research data underlying a publication accessible in a persistent location.[22]

Incentivizing scholars to make their data meaningfully accessible, however, requires more than simple counts. It entails establishing clear processes and flexible metrics to evaluate the quality of shared data and datasets and their documentation. Establishing clear criteria for distinguishing between high- and low-quality datasets and developing ways to fairly and systematically evaluate data quality – while undoubtedly complicated and challenging – could have multiple positive effects. For instance, it could help scholars to understand how to produce robust data and incentivize them to do so; it could also allow for the establishment of new awards and prizes for the generation and sharing of quantitative and qualitative datasets.[23]

Another way to draw attention to shared data would be to establish more "data journals" (such as Brill's *Research Data Journal for the Humanities and Social Sciences*) featuring short sophisticated essays discussing the generation and analysis of particular kinds of data or particular datasets, or critiques thereof. Journals publishing more replications (or venues specifically designated to do so) could also incentivize the generation and sharing of high-quality data.

[20] This is particularly concerning when scholars fear that by sharing their data they will be "scooped" – that is, that another researcher will use those data as the basis of their own scholarship before the person who collected them has time to develop all of the scholarly products she wished to produce using the data.

[21] Data access could be encouraged via punishment (sticks) rather than incentives (carrots). We doubt the former would be salutary or effective: sticks often yield high levels of poor quality compliance. We thus focus here on incentives.

[22] For those skeptical of the utility of using badges to signal open scholarship, COS references two studies that argue that such a system augments rates of data sharing; see Kidwell et al. 2016 and Rowhani-Farid et al. 2017.

[23] For instance, the APSA comparative politics section has been awarding the "Lijphart/ Przeworski/ Verba Data Set Award" since 1999.

Surfacing Data Reuse through Citation and Co-Authorship

If a key reason to make data meaningfully accessible is to facilitate their reuse by other scholars, another way to encourage researchers to share their data is to make that reuse patent. When authors use shared data to produce conclusions in a research publication, they should provide a full citation to the data they analyzed (and to the scholars who generated those data). This practice is the analogue of authors referencing the papers, articles, and books on whose theories and conclusions their research builds. Citing research data generated by others acknowledges that data are a product of value themselves, linked to but distinct from publications that draw on them. Alternatively, the data re-user might offer the data generator co-authorship on the research publication based on secondary data analysis (a practice more prevalent in the natural sciences). An intriguing compromise proposal that recently emerged in the medical field is to list "data authors" (Bierer, Crosas, and Pierce 2017) on publications – a special category of contributors who receive credit for data generation but who did not collaborate on the publication nor necessarily agree with its conclusions.

Community standards are also evolving with regard to what *additional* steps authors who use in their published work data generated by others need to take. Scholars whose work draws on large-scale data projects typically use just a subset of the data produced by the project. Generally, in addition to the complete citation to the original data,[24] authors need to describe in full what elements they extracted for analysis (the analysis dataset) and the process of extraction (typically as software code). Some journals require that authors provide detailed instructions for reconstructing the analysis dataset from the original data; others call for a copy of the author's analysis dataset (a replication dataset) to be deposited in a place specified by the journal, regardless of whether the original dataset is available elsewhere. We believe the first of these options is preferable as it simplifies giving credit to the creators of the original data, avoids partial and incomplete data duplication in multiple venues, and militates against a replication dataset becoming obsolete as updates are made to (or errors found in) the original dataset.

New Initiatives: Infrastructure and Instruction

Significant progress has been made in building the tools and infrastructure that scholars need to make their data accessible, and more can be done. One major set of efforts seeks ways to integrate data and data management into researchers' regular workflows. In some areas, such integration already exists and is used by a growing number of researchers. Employing tools such as knitr and R-Markdown, for example, researchers can combine statistical code and academic writing in a single, "reproducible" document (see Xie 2014). Similar tools exist for other popular languages such as python and Stata. COS's Open Science Framework (osf.io) is designed to integrate different types of tools—storage like Dropbox and Google Drive, code repositories like GitHub or bitbucket, and data repositories such as Dataverse— into the scientific workflow, mainstreaming good data and document management.

Several other initiatives also hold promise. One of the most anticipated developments is the "Roadmap" project (Simms et al. 2016), spearheaded by the California Digital Library and the Digital Curation Center. Both centers currently offer popular tools for writing DMPs, the DMP Tool (https://dmptool.org/) and DMP Online (https://dmponline.dcc.ac.uk/) respectively. "Roadmap" is a next generation tool that provides significantly more guidance to researchers on data management and automates components of DMPs (see above).

[24] Journals rarely require authors to provide information about how the original dataset was generated.

Some efforts seek to bring similar benefits to qualitative researchers. As qualitative researchers increasingly use Computer Assisted Qualitative Data Analysis (CAQDAS) software such as NVivo or atlas.ti, repositories, developers, and expert users are collaborating to develop ways to facilitate the sharing of projects organized in such software in data repositories (Karcher and Pagé 2017; cf. Corti and Gregory 2011 for UK Data's groundbreaking work on the topic). Also, building on Moravcsik's (2010) "active citation," QDR has developed an approach to making qualitative research more transparent that anticipates the sharing of relevant research data – "Annotation for Transparent Inquiry" or ATI (see Karcher et al. 2016).

Finally, integrating the teaching of data management skills – and an understanding of the technologies that are available to store and preserve data and keep them safe – into graduate training will be critical. Topics might include how to interact with IRBs, how to conduct research in a way that does not preclude sharing research data or make doing so prohibitively difficult, and how to manage data with sharing in mind from the start of a research project. Instructing young scholars on these topics will empower them to engage in and shape ongoing disciplinary debates, and will ensure that the next generation of researchers is well-equipped to share their data as new standards are introduced that call on them to do so.

Conclusion

Access to research data is expanding across the social sciences as more scholars become aware and convinced of the benefits of sharing the data that scholarly inquiries generate. The very real benefits that this trend can produce are augmented and multiplied when scholars make their data *meaningfully* accessible: when they render them interpretable and useable for multiple purposes by other scholars, and take advantage of the ground-breaking institutional and technical developments in data infrastructure. The achievement of these benefits will likewise accelerate as more scholars make their data accessible to others, and we considered various initiatives that scholarly communities might undertake to catalyze their doing so.

Making such changes will be challenging. Academic disciplines have deeply embedded practices that are difficult to amend. Stasis dominates due in part to human nature, yet there is also a path dependent aspect to the stickiness of academic praxis: scholarly norms, expectations, and infrastructure have been erected around existing practices and procedures, serving to lock them in place. Moreover, movement toward sharing data will engender unintended consequences that will be important – and hard – to address. To offer just one example, social science disciplines will need to reconsider how they interpret mistakes and error. The assessment of data, replication of findings, and other forms of evaluation that greater data accessibility facilitates will almost certainly bring more errors to light – not because more mistakes are being made but because greater transparency raises the likelihood of error discovery. Such errors should not and cannot be interpreted as research failures, but precisely how *should* be considered, and what *should* be done on the basis of their discovery?

The challenges associated with making data accessible notwithstanding, doing so is in line with, if not an implicit or explicit pre-requisite for, many of the other proposals made throughout this volume. Reproducibility (Christensen and Miguel) relies on open data, as does Appraisal/Re-appraisal (Gerring). "Same data replication" (Freese and Petersen) clearly requires access to data, and "new-data replication" also benefits from the ability to compare newly collected data with the data used in the study in question to understand differences and their potential causes. Carefully evaluating measurement (Reiter) and Reliability of Inference (Fairfield and Charman) perhaps most crucially rely on what we call *meaningful* access to data: Without careful documentation of data collection, transformation, and analysis, it is all but impossible to evaluate measurement or investigate details of inferential claims. Finally, generating improved metrics for evaluating research (Gerring, Apfeld, and Karcher) will require the use of the data infrastructure we discuss here and should increase the incentives for data sharing.

The centrality of data access to so many of the proposals mentioned in this volume is no accident: data are the building blocks of knowledge. The fact that so many stakeholders invested in improving the production of knowledge rely on the accessibility of these building blocks bodes well for the future of meaningfully accessible data.

References

AAPOR. 2017. "AAPOR Transparency Certification Agreement." Oakbrook Terrace, IL: American Association for Public Opinion Research. https://www.aapor.org/AAPOR_Main/media/MainSiteFiles/AAPORTransparencyCertificationAgreement-Revised-October-2017.pdf.

Bar-Sinai, M., L. Sweeney, and M. Crosas. 2016. "DataTags, Data Handling Policy Spaces and the Tags Language." In 2016 IEEE Security and Privacy Workshops (SPW), 1–8. doi:10.1109/SPW.2016.11.

Bierer, Barbara E., Mercè Crosas, and Heather H. Pierce. 2017. "Data Authorship as an Incentive to Data Sharing." New England Journal of Medicine 0 (0): null. https://doi.org/10.1056/NEJMsb1616595.

Bishop, Libby. 2005. "Protecting Respondents and Enabling Data Sharing: Reply to Parry and Mauthner." Sociology 39 (2): 333–36. doi:10.1177/0038038505050542.

Bishop, Libby. 2009. "Ethical Sharing and Reuse of Qualitative Data." Australian Journal of Social Issues 44 (3): 255–72.

Bishop, Libby. 2014. "Re-Using Qualitative Data: A Little Evidence, on-Going Issues and Modest Reflections." Studia Socjologiczne, no. 3: 167.

Broom, Alex, Lynda Cheshire, and Michael Emmison. 2009. "Qualitative Researchers' Understandings of Their Practice and the Implications for Data Archiving and Sharing." Sociology 43 (6): 1163–80. doi:10.1177/0038038509345704.

Butler, Courtney, and Brett Currier. 2017. "You Can't Replicate What You Can't Find: Data Preservation Policies in Economic Journals." presented at the IASSIST 2017, Lawrence, KA, May 25. http://iassist2017.org/program/s107.html.

Corti, Louise, and Arofan Gregory. 2011. "CAQDAS Comparability. What about CAQDAS Data Exchange?" Forum Qualitative Sozialforschung / Forum: Qualitative Social Research 12 (1). http://www.qualitative-research.net/index.php/fqs/article/view/1634.

Corti, Louise, Veerle van den Eynden, Libby Bishop, and Matthew Woollard. 2014. Managing and Sharing Research Data: A Guide to Good Practice. Los Angeles: SAGE.

Costas, R., Z. Zahedi, P. Wouters. 2014. "How well developed are altmetrics? A cross-disciplinary analysis of the presence of 'alternative metrics' in scientific publications." Scientometrics 101(2): 1491–1513.

Druckman, James N., Donald P. Green, James H. Kuklinski, and Arthur Lupia. 2006. "The Growth and Development of Experimental Research in Political Science." *American Political Science Review* 100 (4): 627–35. https://doi.org/10.1017/S0003055406062514.

Ember, Carol and Robert Hanisch. 2013. "Sustaining Domain Repositories for Digital Data: A White Paper, December 11." DOI: 10.3886/SustainingDomainRepositoriesDigitalData.

Fielding, Nigel. 2004. "Getting the Most from Archived Qualitative Data: Epistemological, Practical and Professional Obstacles." International Journal of Social Research Methodology 7 (1): 97–104. doi:10.1080/13645570310001640699.

Fienberg, Stephen E., Margaret E. Martin, and Miron L. Straf, eds. 1985. Sharing Research Data. National Academy Press. http://psycnet.apa.org/psycinfo/1997-36511-000.

Gertler, Aaron L., and John G. Bullock. 2017. "Reference Rot: An Emerging Threat to Transparency in Political Science." PS: Political Science & Politics 50 (1): 166–71. doi:10.1017/S1049096516002353.

Heaton, Janet. 2008. "Secondary Analysis of Qualitative Data: An Overview." Historical Social Research / Historische Sozialforschung 33 (3 (125)): 33–45. doi:10.2307/20762299.

Hodson, Simon. 2016. "Sustainable Business Models for Data Repositories." RDA Plenary, Tokyo, March 3, https://rd-alliance.org/sites/default/files/attachment/S.%20Hodson%20Business%20Models%20Presentation.pdf

Johnston, Lisa R., Jake Carlson, Cynthia Hudson-Vitale, Heidi Imker, Wendy A. Kozlowski, Robert Olendorf, and Claire Stewart. 2017. "Data Curation Network: A Cross-Institutional Staffing Model for Curating Research Data." http://hdl.handle.net/11299/188654.

Johnston, Lisa, ed. 2016. Curating Research Data. Chicago: Association of College and Research Libraries.

Karcher, Sebastian, and Christiane Pagé. 2017. "Workshop Report: CAQDAS Projects and Digital Repositories' Best Practices." D-Lib Magazine 23 (3/4). doi:10.1045/march2017-karcher.

Karcher, Sebastian, and David A. Steinberg. 2013. "Assessing the Causes of Capital Account Liberalization: How Measurement Matters." International Studies Quarterly 57 (1): 128–37. doi:10.1111/isqu.12001.

Karcher, Sebastian, Dessislava Kirilova, and Nicholas Weber. 2016. "Beyond the Matrix: Repository Services for Qualitative Data." IFLA Journal 42 (4): 292–302. doi:10.1177/0340035216672870.

Kidwell, Mallory C., Ljiljana B. Lazarević, Erica Baranski, Tom E. Hardwicke, Sarah Piechowski, Lina-Sophia Falkenberg, Curtis Kennett, et al. 2016. "Badges to Acknowledge Open Practices: A Simple, Low-Cost, Effective Method for Increasing Transparency." PLOS Biology 14 (5): e1002456. doi:10.1371/journal.pbio.1002456.

King, Gary. 1995. "Replication, Replication." PS: Political Science & Politics 28 (03): 444–52. doi:10.2307/420301.

Kratz, John, and Carly Strasser. 2014. "Data Publication Consensus and Controversies." F1000Research, October. doi:10.12688/f1000research.3979.3.

Lupia, Arthur, and George Alter. 2014. "Data Access and Research Transparency in the Quantitative Tradition." PS: Political Science & Politics 47 (01): 54–59.

Maunsell, John. 2010. "Announcement Regarding Supplemental Material." Journal of Neuroscience 30 (32): 10599–600.

Mauthner, Natasha S, Odette Parry, and Kathryn Backett-. 1998. “The Data Are out There, or Are They? Implications for Archiving and Revisiting Qualitative Data.” Sociology 32 (04): 733–745.

Mauthner, Natasha S., and Odette Parry. 2009. “Qualitative Data Preservation and Sharing in the Social Sciences: On Whose Philosophical Terms?” Australian Journal of Social Issues 44 (3): 291–307. doi:10.1002/j.1839-4655.2009.tb00147.x.

Michener, William K. 2015. “Ten Simple Rules for Creating a Good Data Management Plan.” *PLoS Computational Biology* 11(10): e1004525. doi:10.1371/journal.pcbi.1004525

Moravcsik, Andrew. 2010. “Active Citation: A Precondition for Replicable Qualitative Research.” PS: Political Science & Politics 43 (01): 29–35.

National Academy of Sciences, National Academy of Engineering (US) and Institute of Medicine (US) Committee on Ensuring the Utility and Integrity of Research Data in a Digital Age, and Institute of Medicine. 2009. Research Data in the Digital Age. National Academies Press (US). https://www.ncbi.nlm.nih.gov/books/NBK215259/.

National Research Council. 1999. *A Question of Balance: Private Rights and the Public Interest in Scientific and Technical Databases*. Washington D.C.: The National Academies Press. https://doi.org/10.17226/9692.

National Science Board. 2005. *Long-Lived Data Collections: Enabling Research and Education in the 20st Century*. Arlington, VA: National Science Foundation.

Parry, Odette, and Natasha Mauthner. 2005. “Back to Basics: Who Re-Uses Qualitative Data and Why ?” Sociology 39 (2): 337–342. doi:10.1177/0038038505050543.

Parry, Odette, and Natasha S. Mauthner. 2004. “Whose Data Are They Anyway? Practical, Legal and Ethical Issues in Archiving Qualitative Research Data.” Sociology 38 (1): 139–152.

Rowhani-Farid, Anisa, Michelle Allen, and Adrian G. Barnett. 2017. “What Incentives Increase Data Sharing in Health and Medical Research? A Systematic Review.” Research Integrity and Peer Review 2 (1). doi:10.1186/s41073-017-0028-9.

Simms, Stephanie, Sarah Jones, Kevin Ashley, Marta Ribeiro, John Chodacki, Stephen Abrams, and Marisa Strong. 2016. “Roadmap: A Research Data Management Advisory Platform.” Research Ideas and Outcomes 2 (March): e8649. doi:10.3897/rio.2.e8649.

Simms, Stephanie. 2018. ‘Scoping Machine-Actionable DMPs’. *DMPTool Blog* (blog). 9 July 2018. https://blog.dmptool.org/2018/07/09/scoping-machine-actionable-dmps/.

Smit, Eefke, Jeffrey Van Der Hoeven, and David Giaretta. 2011. “Avoiding a Digital Dark Age for Data: Why Publishers Should Care about Digital Preservation.” Learned Publishing 24 (1): 35–49. doi:10.1087/20110107.

Wallis Jillian C., Elizabeth Rolando, and Christine L. Borgman. 2013 If We Share Data, Will Anyone Use Them? Data Sharing and Reuse in the Long Tail of Science and Technology. *PLoS ONE* 8 (7): e67332. doi:10.1371/journal.pone.0067332

Wilkinson, Mark D., Michel Dumontier, Ijsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. 2016. “The FAIR Guiding Principles for Scientific Data Management and Stewardship.” Scientific Data 3 (March). doi:10.1038/sdata.2016.18.

Xie, Yihui. 2014. “Knitr: A Comprehensive Tool for Reproducible Research in R.” In Implementing Reproducible Research, edited by Victoria Stodden, Friedrich Leisch, and Roger D. Peng, 3–32. Boca Raton, FL: CRC Press.

Yardley, Sarah J., Kate M. Watts, Jennifer Pearson, and Jane C. Richardson. 2014. “Ethical Issues in the Reuse of Qualitative Data Perspectives From Literature, Practice, and Participants.” Qualitative Health Research 24 (1): 102–113.