HHS Public Access

Author manuscript

Am Psychol. Author manuscript; available in PMC 2018 May 24.

Published in final edited form as:

Am Psychol. 2018 ; 73(2): 146–156. doi:10.1037/amp0000258.

Responsible Practices for Data Sharing

George Alter and

Inter-university Consortium for Political and Social Research, University of Michigan

Richard Gonzalez

Department of Psychology and the Research Center for Group Dynamics, University of Michigan

Abstract

Research transparency, reproducibility, and data sharing uphold core principles of science at a time when the integrity of scientific research is being questioned. This paper discusses how research data in psychology can be made accessible for reproducibility and reanalysis. We describe ways to overcome barriers to data sharing, such as practical measures for protecting the confidentiality of research participants and improving documentation of the research process. We also advocate policies that recognize research data and program code as important scientific contributions.

Keywords

Data sharing

The current debate over access to research data is occurring in an environment in which the integrity of scientific research is being questioned. Well-publicized examples of fraud undermine the credibility and authority of research findings. Reports showing that many published results cannot be replicated or reproduced (e.g., Herndon, Ash & Pollin, 2014; Open Science Collaboration, 2015) cause scientists, media, and the public to question standards and incentives inherent in the way that science is conducted. In this paper we use the term “reproducible” to refer to the ability to verify published findings using the same data set and the term “replicable” to refer to the ability to find similar results in a new study (see Stodden, Leisch & Peng, 2014).

The time when scientists can count on the uncontested respect of the public is past. The legitimacy of scientific research now rests upon demonstrating that core principles of science are being practiced. One of these core principles is that all research results are open to challenge through reexamination, reanalysis, reproducibility and replication.[1] This implies that the methods used in research should be well-described and transparent. It also implies that data, codebooks, relevant analytic and processing scripts and files used to create tables and figures presented in the published article should be available for re-analysis (Asendorpf, 2013; Miguel et al., 2014; Nosek et al., 2015). Open access to such materials has other benefits in addition to re-analysis, including facilitating replication efforts and meta-

Correspondence concerning this article should be addressed to George Alter, ICPSR, University of Michigan, PO Box 1248, Ann Arbor, MI 48106-1248. altergc@umich.edu.

[1]The “Climategate” affair shows that lack of transparency can be used as a weapon against science. See United Kingdom (2011).

analyses. For example, having access to data can facilitate computation of metrics not easily derived from typical summary statistics that appear in published papers.

Our goal in this paper is to examine how research data can be made accessible for reproducibility and reanalysis, i.e., the A in the FAIR principles (Findability Accessibility Interoperability and Reusability; Wilkinson, et al. 2016; see also Martone, Garcia-Castro, and VandenBos in this volume). We do not address other issues that contribute to the integrity of scientific research such as the replication of findings, preregistration of a study, particular method used in a study, etc. We review several key points surrounding the sharing of data such as who owns research data, how to protect the confidentiality of the research participant, how to give appropriate credit to the data creator, how to deal with metadata and codebooks, how to address provenance, and other specifics such as versioning and file formats. We have found that while many scientists appreciate the rationale for data sharing, we often hear various arguments that are presented as barriers for putting that rationale into practice (e.g., “I could be scooped”). By addressing these barriers directly we hope that scientific psychology can move in the direction of more open data sharing.

Most of our comments are shaped by our own experiences working with quantitative data from questionnaires, experimental lab studies, government statistics, and administrative records, but our understanding of “data” is much broader. Psychologists and other social and behavioral scientists now work with many other sources, such as interview transcripts, images, genomics, and social media. In our view any kind of information used in support of “substantive” (American Psychological Association, 2010) or “evidence-based” (American Political Science Association, 2012) claims in an academic publication qualify as data.

Data Sharing: Who, What, and How

Who Owns Data?

While academic scientists often act as if research data are their own intellectual property, this is usually incorrect. In general, universities assert ownership of any data created by a project with external funding, and some universities extend their ownership to any data created by faculty or students. At our institution data from sponsored research is owned by the University under the University of Michigan Technology Transfer Policy, which defines “research tools and data” as “Intellectual Property” (Office of the Vice President for Research, 2007). These policies are justified by the university’s obligations to comply with sponsored project agreements, protect research subjects, and comply with regulations surrounding research misconduct. Sponsored projects are a particular concern, because research grants are awarded to the institution not to the principal investigator (PI). For simplicity in exposition we focus on universities as the institution but the grantee institution could be a company or other non-university entity.

In most cases, federal agencies like the National Institute of Health (NIH) and the National Science Foundation (NSF) cede the ownership of research data to the grantee (e.g., the university), but the grantee is required to comply with federal and agency-specific regulations, which may include data sharing. The PI is considered custodian, or steward, of the data, who is, among other things, responsible for its maintenance and retention. Further,

nonfederal research funders, such as industry funders, may have different agreements with the university around intellectual property and ownership of data and these agreements may not resemble those from federal research funders. In principle, researchers could be required to obtain university permission whenever they publish results from a sponsored project or share their data publicly or privately, but this is rarely enforced and many universities do not appear to have clear and transparent policies to guide faculty and students on this matter.

Research data may also fall under a variety of federal or state laws (National Academies, 2005; Waldo et al., 2007). Data generated by federal statistical agencies are governed by the Confidential Information Protection and Statistical Efficiency Act (CIPSEA). Educational records are protected by the Family Educational Rights and Privacy Act (FERPA). Data collected during health care fall under the Health Insurance Portability and Accountability Act (HIPAA). The Data Access Act of 1999, also known as the “Shelby Amendment,” makes university research data subject to Freedom of Information Act (FOIA) procedures if it has been used to make federal policy.

Federal and university policies are increasingly driven by concern about privacy and the protection of confidential information provided by research subjects. As we are writing this article, the Department of Health and Human Services has issued revisions of the “Common Rule,” which governs the protection of human subjects in research funded by most federal agencies (U.S. Department of Health and Human Services, 2009; Federal Policy for the Protection of Human Subjects, 2017). The revisions address concerns about informed consent for secondary use of research data and protection of data including identifiable personal information.

Similar debates are occurring internationally. In 2016 the European Union revised its General Data Protection Regulation to include additional protections for individual privacy (Regulation, E. U. 2016). The new regulation requires that personal data are only to be processed for purposes covered by consent of the data subject. There was considerable concern that early drafts of the regulation would have limited the re-use of data in scientific research. The final version of the regulation extends previous exemptions for scientific research data, but it re-emphasizes the responsibility of data managers for the accuracy and security of personal data (Chassang, 2017).

Where does this leave the research psychologist? Given that in most cases the university—not the lab, not the PI, not the student—owns the research data, researchers should seek clarification from their university about its policies surrounding data stewardship. Some universities such as Columbia University (http://ccnmtl.columbia.edu/projects/rcr/rcr_data/foundation/#2) have clear policies on their websites, but this does not appear to be the norm. It is important to be clear on data stewardship responsibilities prior to data collection so appropriate protections can be set in place early in the research process (e.g., inclusion of relevant statements in the consent form, a clear and transparent data management plan). Labs should have clear policies for collaborators, research assistants, students and staff about keeping data files on shared drives and copies of data files on their own computers, including access to such files when they complete work on the project and policies that protect privacy.

Reproducibility, Production Transparency, and Analytical Transparency

While public attention is often focused on the issue of reproducibility, we believe that discussion should begin with the broader issue of research transparency. In many types of research the procedures used to create and analyze data are at least as important to a research outcome as the data themselves. For example, Wicherts et al. (2016) list 34 “researcher degrees of freedom” in psychological research. Experience in economics and political science shows that sharing original data is usually not enough to assure that results are reproducible (e.g., Hendron et al, 2014). Replication studies obtain the published results in a small minority of cases (Dewald, Thursby and Anderson, 1986; Duvendack, Palmer-Jones and Reed, 2015; McCullough, McGeary and Harrison, 2006; Open Science Collaboration, 2015). As a consequence, leading journals in both economics and political science require authors to supply not only data but also the program code or other procedures used to produce the published results.

Elman and Lupia (2014) emphasize the need for both “production transparency” and “analytical transparency.” Many types of data are refined, corrected, and manipulated in a number of ways before they are analyzed, and production transparency refers to the steps that were taken to create and process the data. In psychology examples of production transparency would be to document and make available not only original data files but also all preprocessing scripts in an fMRI study, provide a detailed codebook for each variable in the data set, or provide all code used in data cleaning and subject exclusion. Analytical transparency refers to the algorithms and statistical procedures that produce the results (statistical tests, tables, figures, etc.) reported in a publication. In psychology examples would be to provide relevant SPSS, SAS or R code for all statistical tests and documentation (including code, if available) for figures and tables included in the publication. Documentation related to analytical transparency should be organized and keyed to the final publication such as page and paragraph number rather than a disorganized set of statistical commands. Ideally, authors should provide documentation of all the steps that lead to their findings and maintain these documents as part of the regular research pipeline rather than attempt post-publication to piece together the documentation for the analyses. Clear and transparent documentation as part of the standard research workflow will reduce the time and cost burden to the researcher and could lead to better quality control for research teams.

Research transparency obligates authors to describe the process that produced their empirical results. While critics see this as an onerous burden, we believe that conscientious documentation is a hallmark of good science. Indeed, anyone who has tried to update old computer code knows that good documentation saves time in the long run. Many of the practices that contribute to good data management and better documentation are simple, but they also facilitate proper data sharing. For example, filenames should include a version number or date to make it easy to return to earlier versions. Ideally, documentation should cover the entire “data lifecycle” beginning with the links between theoretical concepts and empirical data as well as coding contained in the data file (e.g., male = 1, female = 2).

Advice about best practices in documenting research data is available from:

- The *ICPSR Guide to Social Science Data Preparation and Archiving* (2012) is now in its fifth edition, which is available both online and in pdf.
- J. Scott Long's *The workflow of data analysis using Stata* (2009) offers a wealth of good advice from an experienced social scientist.
- *Project TIER*: *Teaching Integrity in Empirical Research* grew out of a data analysis course for undergraduates developed by Richard Ball and Norm Medeiros at Haverford College (Ball & Medeiros, 2016). They offer a step-by-step guide to data management for reproducible research.
- The *Open Science Framework* (http://help.osf.io/) offers several guides and best practices on topics such as version control, file naming, and making a data dictionary.
- The *Databrary Project* (https://nyu.databrary.org/) has developed a model for sharing video data from psychological research.
- Advice for managing and sharing qualitative data is available from the *Qualitative Data Repository* (https://qdr.syr.edu/guidance).
- The *International Association for Social Science Information Services & Technology* (IASSIST) maintains a directory of resources about data management across a wide range of disciplines (http://www.iassistdata.org/resources).
- The OpenfMRI Project maintains a repository of raw functional magnetic resonance imaging data sets (https://openfmri.org/) and provides guidelines and best practices (e.g., BIDS format) for data management.
- Animal researchers will find useful information on data management at the National Institute of Health directory of data sharing archives specific to the domain of research (e.g., cancer imaging) or species (e.g., worms such as C. elegans).

Data Repositories

Data repositories are organizations that specialize in maintaining and distributing data over time. There are three main kinds of data repositories: domain-specific repositories, general repositories, and institutional repositories. All data repositories make preservation commitments to maintain and distribute data over time. Data repositories make data citable by exposing a citation and by assigning persistent identifiers, such as Digital Object Identifiers (doi).

Most domain-specific data repositories have emerged to provide services for a specific scientific field (Ember and Hanisch, 2013). Since they are closely tied to a scientific community, domain repositories focus on a limited type of data, and they invest heavily in curating data for re-use. Data curation involves documenting and managing data so that they are discoverable, meaningful, usable, and persistent over time. Domain repositories have been leaders in developing standards and formats for data and metadata, like the Data Documentation Initiative, which is an international standard for documenting data that result

from observational methods in the social, behavioral, economic, and health sciences (Vardigan et al., 2008). Domain repositories often invest considerable effort in preparing metadata, which is necessary for finding data and for preserving it over time. Examples of domain-specific data repositories are the Inter-university Consortium for Political and Social Research (ICPSR), which maintains a large archive of quantitative data for social and behavioral research; Databrary, which developed to share videos used by developmental psychologists; the Qualitative Data Repository, which specializes in qualitative and multi-method social inquiry.

General data repositories, like Figshare, Mendeley Data, and the Dataverse Network, serve a broader range of disciplines and provide fewer data curation services than domain repositories. These repositories are designed to be self-service, and they depend upon the depositor to provide documentation and metadata. Metadata is “data about data”, and is usually stored in a different file than the original data. Examples of metadata could include details in a codebook (e.g., wording of survey questions), placement of electrodes, manufacturer and model number of equipment used, identifying codes for technician or research assistant involved in the project, versions of software used to collect data, etc.

Many universities now have institutional repositories, most of which are operated by libraries. These repositories often have a broad mission to document and preserve all of the research produced by faculty, staff, and students. Most of these repositories started by focusing on faculty publications and working papers, but they are now upgrading their technical capacities to handle data as well as text files. The data services offered by institutional repositories vary with the resources of the university. Universities with large research missions have been hiring staff with skills in data management, documentation, and preservation, but it is difficult for any institution to have experience with all types of research data.

There is currently no consensus in the U.S. on long term funding for data repositories (Ember and Hanisch, 2013).[2] ICPSR, the oldest data repository for social and behavioral sciences in the U.S., relies on both membership fees and research grants. The ICPSR consortium began with 21 universities in 1962 and has now grown to more than 750 worldwide. The Qualitative Data Repository (https://qdr.syr.edu/) and Databrary (https://nyu.databrary.org/) are currently funded by grants from NSF and NIH, but they may transition to other funding models in the future. The Dataverse Network (https://dataverse.org/) is supported by Harvard University and research grants. Figshare (https://figshare.com/) is part of Digital Science, a firm owned by Macmillan Publishing, and Mendeley Data (https://data.mendeley.com/) was recently acquired by Elsevier.

In general, there is no cost to researchers for depositing data in a repository, but there may be a cost for professional data curation. For example, anyone at an ICPSR member institution may deposit data in the OpenICPSR (https://www.openicpsr.org/openicpsr/) open-access repository, but additional ICPSR data curation services (e.g. full metadata generation,

[2]In contrast to the lack of long-term planning in the U.S., European funding agencies have identified data archiving as critical research infrastructure (European Strategy Forum on Research Infrastructures, 2016). The Consortium of European Social Science Data Archives is in the process of forming a European Research Infrastructure Consortium (ERIC).

bibliography search, conversion into multiple statistical packages) are available for a fee. The cost of professional data curation to improve the findability and usability of data can be included in research grants.

We recommend that researchers deposit their data in a domain-specific repository whenever possible. Since domain-specific repositories are based on research communities, they are most likely to provide data curation services that will enhance the value of the data for future re-use. For example, domain-specific repositories will migrate data to new formats and standards as software changes. We also advocate "trusted digital repositories" that adhere to standards for data discovery, documentation, and preservation. The Data Preservation Alliance for the Social Sciences (Data-PASS; http://www.data-pass.org/) is a partnership of eight U.S. repositories who are committed to archival standards and mutual support to assure the long term preservation of research data.

Supplementary Materials

The advent of the Internet and electronic publication has created new possibilities for journals to provide access to materials that supplement a research publication. In some fields publications have become so brief that critical information is only available in a supplement. After initial enthusiasm for supplementary materials some journals have turned away from the practice. The *Journal of Neuroscience* referred to concerns about the reviewing process (e.g., guaranteeing that supplemental material be held to the same peer review standard as the primary article) in explaining its decision to stop accepting supplementary materials in 2010 (Maunsell, 2010).

We discourage the practice of attaching research data as supplementary materials associated with a publication. Publishers are generally not in a position to manage research data in the ways that a data repository would. Data files may be converted to text or pdf files, which lack the functionality in their original formats. Supplementary materials may also lack the metadata for discovery by researchers who do not know about the publication to which they are attached. In addition, there is a risk that a publisher will decide that the costs of maintaining supplementary materials outweigh the benefits and discontinue access.

Publication-related data along with relevant metadata and documentation should be deposited in a data repository and associated with a publication through a citation. In particular, we advocate depositing program code and scripts for statistical packages in a repository, so that they can be cited in future publications. Placing materials and files relevant to production and analytic transparency (as described earlier) in a repository gives these objects an identity separate from the publication and allows them to be cited as contributions to science in their own right.

Protecting Research Subjects

One of the most frequently expressed objections to sharing research data is concern that confidential information about research subjects will be compromised. In return for their cooperation we promise participants anonymity, and we assure them that any confidential information will be protected. Respecting this promise is essential, and government agencies

and data repositories have many years of experience safely sharing data including confidential information. The key to protecting research subjects is planning for responsible data sharing at the beginning of a project. Psychologists should anticipate potential risks to subjects in their research designs, inform subjects of these risks, and use appropriate protection when data are shared. Under current rules all of these plans should be approved by Institutional Review Boards (IRBs) before research begins.

Evaluating Disclosure Risks

It is helpful to think of the risks to subjects from the disclosure of information in a data set as the product of two dimensions: potential harm and probability of re-identification (National Research Council, 2014). Data that are low on both harm and re-identification can be shared with minimal provisions for security. For example, national opinion polls usually include little personally identifiable information and ask innocuous questions. Data that include more sensitive or identifiable information require more protection.

The degree of harm that would be suffered by the subject if the information became public varies widely among data sets. Research subjects are often asked about topics with a high potential for harm, such as questions about mental health, drug use, criminal activity, sexual behavior, etc. On the other hand, many research projects collect information that would have little or no impact if it were made public, such as perception tasks in an experimental setting.

The probability that a participant will be re-identified depends upon how they are selected and the kinds of personal information they provide. Direct identifiers (name, phone number, social security number) are usually removed from data sets, but some types of research pose particular problems. Research designs increasingly use geospatial locations (GPS), longitudinal designs with repeated interviews, and contextual information (e.g. grade, school, school district) that make individuals more identifiable. Some data cannot be completely anonymized without destroying its research value.

Assessing the disclosure risk for a data set can be a time-consuming process, especially for lengthy questionnaires, and a formal statistical analysis of the risk of re-identification can be very expensive. HIPAA regulations provide guidance on anonymization of health care data, but these provisions are usually not sufficient for protecting other kinds of social and behavioral data (National Research Council, 2014). The Internet has greatly increased disclosure risks, because there are so many places where individual attributes (age, sex, residence, occupation, etc.) can be found.

We suggest that researchers think about two questions. First, would subjects suffer any harm if information about them were released? Could disclosure of the data result in embarrassment, damage a subject's reputation, or endanger a subject's financial situation? Second, can a subject be identified from the information in the data? Are people sharing observable characteristics (sex, age, race, occupation, etc.) reported in the dataset common or rare? Neither of these risks can be completely eliminated, but they can be compared to the risks experienced in everyday life. According to the Common Rule, "Minimal risk means that the probability and magnitude of harm or discomfort anticipated in the research are not greater in and of themselves than those ordinarily encountered in daily life or during the

performance of routine physical or psychological examinations or tests" (U.S. Department of Health and Human Services, 2009). When the risks of harm and re-identification are both minimal, the data can be shared without additional security precautions. If either type of risk may be consequential, the data protection methods described below should be applied.

Informed Consent

Since the Belmont Report, one of the central tenets of ethical research has been respect for persons, which has been institutionalized in the practice of obtaining informed consent from research subjects (United States, 1978). Under the federal code governing protection of human subjects known as the Common Rule researchers must provide subjects with an explanation of the purposes, risks, and benefits of the research in which they are asked to participate. Participants should be informed that the data they provide will be shared with other researchers and used in other ways.

Not long ago, it was common for Institutional Review Boards to recommend informed consent language that would preclude data sharing and reuse. Informed consents often included wording like "Your data will only be available to members of our research team…" or "At the end of our project all of your data will be destroyed…" NIH has issued guidance rejecting statements of this kind in favor of language that allows data sharing and reuse (National Institutes of Health, 2004). The ICPSR *Guide to Social Science Data Preparation and Archiving* offers advice on informed consent language that is compatible with data sharing (Inter-university Consortium for Political and Social Research, 2012). The new version of the Common Rule includes guidelines for "broad consent" covering storage and future secondary research with data that includes identifiable private information (Federal Policy for the Protection of Human Subjects, 2017). Subjects should be given a general description of the types of research that may be conducted with information collected from them. They should be informed about the types of identifiable private information that will be kept and the types of researchers who may have access to that information. The informed consent should also include a way for subjects to obtain more information about their rights and a person to contact if any harm results from their participation in the study. We expect that IRBs and repositories will develop examples of broad consent documents before the new rules come into effect in January 2018.

The text below was part of the informed consent obtained from teachers for re-use of classroom videos collected in a large study of teaching practices. Informed consent was also obtained from the parents of students. Since individuals are inherently identifiable in videos, ICPSR makes these data available under a data-use agreement through a streaming service that prevents the videos from being downloaded. Most people today understand that no one can promise absolute security for digital data, but they respond positively to assurances that every effort will be taken to protect their information, as they did in this study.

> As a participating teacher, I agree to allow videos of my classroom and associated commentary and materials to be accessed and used for research purposes only. These materials will be housed in a secure database maintained by the University of Michigan. I understand that while I will not be identified by name in this database, it might be possible for someone to identify me from the video content, or by

> combining that content with other materials in the database. The chances of being identified are very small because access to the database will be rigorously controlled. Before researchers can access this database, the University of Michigan will require them to sign pledges of confidentiality promising to strictly adhere to research policies and honor the confidentiality of all students, teachers, and schools. The University of Michigan will pursue sanctions against anyone responsible for releasing information that might be used to identify any person in the study.

A recent development in bio-medical research has been the emergence of "dynamic consent" (Stein and Terry, 2013), which allows patients to decide who can use their health data for research. Systems like Sage Bionetworks (Bot, et al., 2016) send messages to patients through their mobile phones to request permission for their data to be included in new research studies.

Sharing Confidential Data

Understanding disclosure risk as a continuum is useful for deciding how data will be shared. As discussed below, a variety of measures can be applied to share data safely. However, there is a tradeoff between data security and convenience. Most data security measures impose burdens on researchers who reuse data or limit the scientific value of data. Science is best served by balancing the intrusiveness of data protection measures to the disclosure risks inherent in the data.

Felix Ritchie (Ritchie, 2005; Desai et al., 2016) offers a useful framework for describing measures used to share confidential data: safe data, safe projects, safe places, safe people, safe outputs. Data stewards use a combination of these methods that is adjusted to the disclosure risks in a particular data set.

Safe data—"Safe data" refers to measures that remove identifying information from a data set or blur the information in ways that make individuals more difficult to identify. Converting continuous variables like age and income into categories is a simple procedure that reduces the identifiability of individuals. Income is usually "top coded" by putting the wealthiest individuals into an open ended category like "$250,000+" to conceal the identities of extremely wealthy people. More intrusive methods of masking data include adding random noise and swapping responses between otherwise similar respondents. Methods have also been developed to create simulated data based on correlations among variables in the original data and new statistical models based on distributed likelihood optimization that allow data to remain local to the participant (Boker et al., 2015).

Safe projects—Researchers who request restricted-use data are usually required to submit a research plan describing their intended use of the data. These plans are reviewed by the data provider to assure that the proposed analysis is consistent with the informed consent under which the data were collected. In keeping with the Belmont Report (United States, 1978) the UK Data Service asks researchers to "explain how their research will benefit society." (See http://blog.ukdataservice.ac.uk/access-to-sensitive-data-for-research-the-5-safes/) Some data providers can offer advice about the proposed methodology, but a full

evaluation of the scientific merit of the research requires a peer review process, which is expensive and usually unnecessary.

Safe places—Highly sensitive and identifiable data are shared in "safe places." Data enclaves, such as the US Census Bureau's Research Data Centers, allow researchers to analyze data in secure facilities where they can be monitored. The data are never connected to the Internet, and outputs are reviewed before they are removed from the enclave.

Remote execution systems allow researchers to submit programs that will be executed within a secure data center. Unlike a data enclave, the researcher does not need to travel to the place where the data are held. But software choices and functionality offered through remote submission are usually limited, and these systems may involve cumbersome delays. If there is a mistake in the program or an anomaly in the data, the researcher must correct the program and submit it again. Some remote execution systems offer publicly available artificial data that can be used for debugging programs.

"Virtual data enclaves" provide controlled remote access to sensitive data. Researchers remotely control a "virtual computer" with a variety of software that is running in a secure data center. The user establishes a connection to the data center through a "thin client" or by installing client software that disables functions on their local computer like saving to local disks and printing. Virtual data enclaves provide a high level of security, because the data never leave the secure data center, but they do not require users to travel to the location where the data are stored. The principle weakness of virtual data enclaves is that a user can view individual cases and copy the information from the screen. For this reason, virtual data enclaves usually require a signed data use agreement.

Safe people—"Safe people" have been trained to minimize disclosure risks and committed to safe research practices by data use agreements. The training provided by some data repositories helps researchers to understand how they can reduce the risks to subjects in their analyses and publications. For example, authors should avoid releasing crosstabulation tables containing cells with a small number (e.g., less than five) of subjects, who could be re-identified in other databases. Some statistical techniques (e.g., analysis of variance) pose lower risks than others, but special care is necessary when the study population includes individuals with unique characteristics.

Data use agreements (DUAs) are legal agreements that commit the recipients of research data to protection of confidential information. DUAs describe the data covered by the agreement, limitations on the use of the data, and the expectations regarding protection of the sensitive information. Some DUAs include examples of potentially unsafe publication practices, and others specify security measures, such as encryption and system requirements. Private companies often provide data to researchers under "nondisclosure agreements," which are intended to protect their intellectual property. Kanous and Brock (2015) discuss the shortcomings of these agreements for sharing data ("e.g., a failure to clearly articulate the allowable uses of the data, treatment of the provided data as intellectual property, or an attempt to define research results as derivatives of the provided data and thus controlled by the provider", p. 1).

In academic settings DUAs are usually agreements between the institution of the data provider and the institution of the data recipient. Universities typically insist on signing DUAs on behalf of their faculty and students, especially when there are legal consequences to violations of the agreement. In doing so, the university assumes responsibility for supervising the conduct of the researchers who will analyze the data.

Safe outputs—Data providers often require an expert review of outputs derived from highly sensitive data. DUAs may obligate researchers to submit all publications and presentations for review before they can be published. Data enclaves, virtual data enclaves, and remote execution systems allow outputs to be inspected before they are removed from the secure facility. Evaluating disclosure risks in statistical analyses is largely a manual process, which may involve examining program code as well as results, and researchers are often discouraged from submitting hundreds of pages of output for review.

In short, we have a number of tools for sharing data in a safe and responsible way. Since there are trade-offs between data protection and ease of access, security measures should be balanced against risks. For example, datasets with sensitive questions (e.g., depression or anxiety scales) can be shared with moderate data protections if they come from national surveys, like the National Health Interview Surveys. On the other hand, interviews with patients in substance abuse treatments centers might be restricted to a data enclave, because subjects could be re-identified from contextual information in the data.

Promoting Recognition and Collaboration

In this section of the paper we focus on three areas where new practices around data sharing will have additional benefits for the research community: recognizing the contribution of data creators, sharing and citing program code linked to publications, and producing documentation in standardized machine-actionable formats. We believe that changing the practice of psychology in these directions will acknowledge important contributions to research and increase collaboration across the discipline.

Recognizing the Contributions of Data Creators

Research transparency cannot be implemented without demonstrating respect for the contributions of data creators. If data are essential products of scholarship, those who create data must be appropriately acknowledged and rewarded.

Authors frequently express concern that sharing their data will compromise future publications and that they might be “scooped” with their own data. Journals can accommodate authors with a limited period of exclusive use of their data (an “embargo” period). For example, the Ethics Guide of the American Political Science Association declares:

> Researchers who collect or generate data have the right to use those data first. Hence, scholars may postpone data access and production transparency for one year after publication of evidence-based knowledge claims relying on those data, or such period as may be specified by (1) the journal or press publishing the claims, or (2)

> the funding agency supporting the research through which the data were generated or collected. (American Political Science Association Committee on Professional Ethics Rights and Freedoms, 2012, p. 10)

Similarly, the policy recommended by the Guidelines for Transparency and Openness Promotion (TOP) in Journal Policies and Practices is: “Dissemination of these materials may be delayed until publication. Under exceptional circumstances, editors may grant an embargo of the public release of data for at most one year after publication” (TOP Guidelines Committee, 2016). Moreover, journal replication policies only require authors to provide the data used to generate their empirical results, which may be a small part of a larger data collection.

The most fundamental form of scholarly recognition is citation, which has been the accepted practice of acknowledging scholarly contributions for centuries (Uhlir, 2012). Creators of new data should receive the same kind of acknowledgment as authors of books or articles. The “Joint Declaration of Data Citation Principles” (Data Citation Synthesis Group, 2014) is one of several international and interdisciplinary efforts to establish citation of data. Data citations should be included in “References” or footnotes in the same way that other publications are cited.

A basic data citation consists of only five elements: author, title, date, publisher/distributor, and location. Guidelines are being developed for citing more complex types of data, such as data that are updated dynamically (Task Group on Data Citation Standards and Practices, 2013). The preferred form of citing the location of a dataset is a ‘persistent identifier.’ Unlike a Uniform Resource Locator (URL), which points to the network location of a specific machine, a persistent identifier points to a service that resolves it into URL. Since machines are often replaced or moved, URLs become obsolete, and links are broken. When the network location of a dataset changes, the URL associated with the persistent identifier can be updated so that the persistent identifier continues to point to the data. Persistent identifiers, such as ‘digital object identifiers’ (DOIs), are widely used by publishers to point to journal articles, and DataCite was created to provide DOIs for data. Most established data repositories provide citations for data that include persistent identifiers.

In the inaugural issue of *Archives of Scientific Psychology* Cooper and VandenBos (2013 p. 4) argued that acknowledgment of data creators should go further than “simple citation of the data’s origin.” Authors who reuse data provided to *Archives of Scientific Psychology* agree to offer authorship to the originators of the data in any subsequent publication. This is an unusual policy, which is not common in other disciplines, and we are wary about imposing this requirement in psychology. The APA Ethical Principles of Psychologists and Code of Conduct (section 8.12) requires authorship and publication credits to reflect contributions, but it does not mandate how contributions should be credited (American Psychological Association, 2010). If data creators are entitled to be authors of publications that use their data, what will prevent them from suppressing analyses and interpretations that disagree with their own previous publications? We believe that citation is the appropriate way to credit data creators, and that science is better served by open debate than by forcing authors to negotiate with data creators.

Sharing and Citing Code Tied to Publications

As with data, analytic scripts should also be documented. Ideally, the scripts associated with published papers should be organized following the order of the results and properly commented (e.g., “regression in 2nd paragraph of page 541”). This makes it easy for people to follow the analyses as reported in the published paper. The same goes for figures and tables that appear in published work; the code that produced them should be included and well-documented. This should also include code written for mathematical or computational models and any simulations (including power analyses) as part of the research. Modern statistical software such as R include features for commenting code and enabling reproducible data analysis (e.g., the Rmarkdown package in R). We believe it is not sufficient merely to share data. For complete transparency of the results reported in a published academic article, one should also include the scripts that operated on those data to produce the results.

When authors share program code and scripts used with statistical packages, they should receive recognition in the form of citations. Code and scripts are original intellectual products that are as essential to scholarly process as publications, and they should not be treated as incidental by-products of research. Novel methods of processing raw data and computing important measures should evaluated, recognized, and re-used. There has been extensive discussion about the importance of citation for data, but we strongly believe that citation of program code is just as important.

Sharing program code and scripts poses additional problems, because they are embedded in complex computing environments. Changes in the version of software, like a statistical package or operating system, may affect results. Runmycode.org (http://www.runmycode.org/) and Docker (https://www.docker.com/) are recent attempts to make software sharable and preservable over time. At minimum, the documentation of the scripts should state the software packages used, their version numbers, operating system, etc.

Standards-based Documentation

Earlier in the paper we described ways of reducing time and cost burden to researchers by documenting the research process as it progresses rather than after the project has been completed, and we also discussed costs associated with data repositories. The benefit/cost ratio for data documentation could be greatly increased by tools that automate the capture and production of machine-actionable metadata, the “data about data” that attaches meaning to otherwise inscrutable 0’s and 1’s. While most disciplines are still struggling to get scientists to produce minimal data documentation, some fields, notably genomics and astronomy, have already been transformed by data sharing. In these fields discoveries are occurring through the analysis of the collected data of entire research communities. For the behavioral and social sciences to achieve the full benefits of data sharing, data and metadata must available in standard formats that will enable discovery, interoperability, and automated analysis.

Almost all research is eventually translated into digital data, but tools that help researchers create metadata are nearly non-existent. In the social and behavioral sciences a wide gulf has

developed between data repositories, which rely heavily on metadata standards, and researchers, who are unaware that such standards exist. The standard for describing data about individuals was developed by the Data Documentation Initiative (DDI), which is currently in use by all of the major social science data repositories around the world (Vardigan et al., 2008). The benefits of using DDI are beginning to be apparent. For example, the ICPSR website allows researchers to search and compare survey questions in a Social Science Variables Database containing 4.7 million variables.

Unfortunately, researchers currently have little incentive to produce metadata in DDI format or any other standard and few tools to help them. For the most part, DDI metadata has been aimed at the internal processes of data archives. It is much easier for a researcher to create the codebook for a dataset in a text document or spreadsheet than to produce DDI metadata in XML. This should not be the case. Research data is often "born digital," and standardized metadata should be created automatically as it is created and analyzed.[3]

This situation contrasts with the broad adoption of collaboration tools in the software development community. Facilities like Github for collaboration, versioning, and tracking software have become essential to software developers, and they are being used more widely for managing documents, projects, and even data. Similar tools are beginning to emerge for research projects. The Open Science Framework (http://osf.io/) and SEAD (http://sead-data.net/) offer researchers collaborative workspaces in which they can manage, share, and document their data and work processes.

Conclusions

All sciences are under pressure to increase transparency and openness, and the psychology research community needs to balance the obligations and the rewards for those who create original data and methods. As journals and funders adopt more stringent requirements for sharing data, methods and program code, authors should be assured of the professional recognition that they deserve.

Sharing data, methods and program code will require new procedures and workflows. Some simple changes in data management, such as versioning data and program files, are nearly costless and have obvious benefits. New tools, like Github and the Open Science Framework, can offer both efficiencies and expanded capabilities. As a community, we need to encourage discussion of best practices in research and publication and training for our students.

References

American Political Science Association Committee on Professional Ethics Rights and Freedoms. A Guide to Professional Ethics in Political Science. 2012. Retrieved from http://www.apsanet.org/media/PDFs/ethicsguideweb.pdf

[3]George Alter leads a team that has been awarded NSF funding to develop software for "Continuous Capture of Metadata for Statistical Data" (NSF ACI-1640575). This project is creating tools that will update DDI metadata files by extracting data transformation information from scripts used with standard statistical packages (SPSS, SAS, Stata, R).

American Psychological Association. Ethical principles of psychologists and code of conduct. 2010. Retrieved from http://www.apa.org/ethics/code/

Asendorpf JB, Conner M, De Fruyt F, De Houwer J, Denissen JJA, Fiedler K, Wicherts JM. Recommendations for Increasing Replicability in Psychology. European Journal of Personality. 2013; 27(2):108–119. DOI: 10.1002/per.1919

Ball, R., Medeiros, N. Project TIER: Teaching Integrity in Empirical Research. 2016. Retrieved October 19, 2016, from https://www.haverford.edu/tier

Boker SM, Brick TR, Pritikin JN, Wang Y, Oertzen TV, Brown D, Lach J, Estabrock R, Hunter MD, Maes HH, Neale MC. Maintained individual data distributed likelihood estimation (MIDDLE). Multivariate behavioral research. 2015; 50(6):706–720. [PubMed: 26717128]

Bot BM, Suver C, Neto EC, Kellen M, Klein A, Bare C, Friend SH. The mPower study, Parkinson disease mobile data collected using ResearchKit. Scientific Data. 2016; 3:160011. [PubMed: 26938265]

Chassang G. The impact of the EU general data protection regulation on scientific research. Ecancermedicalscience. 2017; 11:709. http://doi.org/10.3332/ecancer.2017.709. [PubMed: 28144283]

Cooper H, VandenBos GR. Archives of scientific psychology: A new journal for a new era. Archives of Scientific Psychology. 2013; 1(2):1–6. DOI: 10.1037/arc0000001

Data Citation Synthesis Group. Joint Declaration of Data Citation Principles. 2014. Retrieved from http://www.force11.org/datacitation

Desai T, Ritchie F, Welpton R. Five Safes: Designing data access for research. Working Paper. 2016

Dewald WG, Thursby JG, Anderson RG. Replication in Empirical Economics - The Journal-Of-Money-Credit-And-Banking Project. American Economic Review. 1986; 76(4):587–603.

Duvendack M, Palmer-Jones RW, Reed WR. Replications in Economics: A Progress Report. Econ Journal Watch. 2015; 12(2):164–191.

Ember, CR., Hanisch, RJ. Sustaining Domain Repositories for Digital Data. 2013. From http://dx.doi.org/10.3886/SustainingDomainRepositoriesDigitalData

European Strategy Forum on Research Infrastructures. ESFRI Roadmap 2016– Strategy Report on Research Infrastructures. 2016. From http://ec.europa.eu/research/infrastructures

Federal Policy for the Protection of Human Subjects, 82 Federal Register 7149-7274 (Jan. 19, 2017) (to be codified at HHS 45 CFR Part 46).

Herndon T, Ash M, Pollin R. Does high public debt consistently stifle economic growth? A critique of Reinhart and Rogoff. Cambridge Journal of Economics. 2014; 38:257–279. DOI: 10.1093/cje/bet075

Inter-university Consortium for Political and Social Research. Guide to Social Science Data Preparation and Archiving: Best Practice Throughout the Data Life Cycle. Ann Arbor MI: 2012.

Kanous, A., Brock, E. Contractual Limitations on Data Sharing. Report prepared for ICPSR as part of Building Community Engagement for Open Access to Data. Ann Arbor, MI: 2015. Alfred P. Sloan Foundation Grant Number 2012-6-11

Long, JS. The workflow of data analysis using Stata. College Station Tex.: Stata Press; 2009.

Lupia A, Elman C. Openness in Political Science: Data Access and Research Transparency. PS: Political Science & Politics. 2014; 47(01):19–42.

Maunsell J. Announcement regarding supplementary material. The Journal of Neuroscience. 2010; 30(32):10599–10600.

McCullough BD, McGeary KA, Harrison TD. Lessons from the JMCB Archive. Journal of Money Credit and Banking. 2006; 38(4):1093–1107.

Miguel E, Camerer C, Casey K, Cohen J, Esterling KM, Gerber A, Glennerster R, Green DP, Humphreys M, Imbens G, Laitin D, Madon T, Nelson T, Nosek B, Petersen M, Sedlmayr R, Simmons JP, Simonsohn U, Van der Laan M. Promoting Transparency in Social Science Research. Science. 2014; 343(6166):30–31. [PubMed: 24385620]

National Academies, Panel on Data Access for Research Purposes. Expanding access to research data: reconciling risks and opportunities. Washington, DC: National Academies Press; 2005.

National Institutes of Health. Frequently Asked Questions: Data Sharing #25. 2004. Feb 16. 2004Retrieved October 19, 2016, from http://grants.nih.gov/grants/policy/data_sharing/data_sharing_faqs.htm#922

National Research Council. Proposed Revisions to the Common Rule for the Protection of Human Subjects in the Behavioral and Social Sciences. Washington DC: The National Academies Press; 2014.

Nosek BA, Alter G, Banks GC, Borsboom D, Bowman SD, Breckler SJ, Buck S, Chambers CD, Chin G, Christensen G, Contestabile M, Dafoe A, Eich E, Freese J, Glennerster R, Goroff D, Green DP, Hesse B, Humphreys M, Ishayama J, Karlan D, Kraut A, Lupia A, Mabry P, Madon T, Malhotra N, Mayo-Wilson E, McNutt M, Miguel E, Levy Paluck E, Simonsohn U, Soderberg C, Spellman BA, Turitto J, VandenBos G, Vazire S, Wagenmakers EJ, Wilson R, Yarkoni T. Promoting an open research culture: author guidelines for journals could help to promote transparency, openness, and reproducibility. Science. 2015; 348(6242):1422–25. [PubMed: 26113702]

Office of the Vice President for Research. University of Michigan Technology Transfer Policy. University of Michigan; 2007. University of Michigan. 303.04

Open Science Collaboration. Estimating the reproducibility of psychological science. Science. 2015; 349:943.

Regulation, E. U. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). Official Journal of the European Union (OJ). 2016; 59:1–88.

Ritchie, F. Access to business microdata in the UK: Dealing with the irreducible risks. Geneva, Switzerland: 2005. UNECE/Eurostat Work session on statistical data confidentiality 2005

Stein DT, Terry SF. Reforming biobank consent policy: a necessary move away from broad consent toward dynamic consent. Genetic Testing and Molecular Biomarkers. 2013; 17(12):855–856. [PubMed: 24283583]

Stodden, V., Leisch, F., Peng, D. Implementing Reproducible Research. CRC Press; Boca Raton, FL: 2014.

Task Group on Data Citation Standards and Practices. Out of Cite, Out of Mind: The Current State of Practice, Policy, and Technology for the Citation of Data. Data Science Journal. 2013; 12:CIDCR1–CIDCR7.

TOP Guidelines Committee. Guidelines for Transparency and Openness Promotion (TOP) in Journal Policies and Practices. 2016. Retrieved 2016-10-17, from https://osf.io/9f6gx/

Uhlir, PE. For Attribution: Developing Data Attribution and Citation Practices and Standards: Summary of an International Workshop. National Academies Press; 2012.

United Kingdom. Parliament. House of Commons. Science and Technology Committee. Eighth Report of Session 201012 Peer review in scientiic publications. Ordered by the House of Commons to be printed 18 July 2011. 2011. Available: http://www.publications.parliament.uk/pa/cm201012/cmselect/cmsctech/856/856.pdf

United States. The Belmont report: Ethical principles and guidelines for the protection of human subjects of research: appendix. Washington, D.C.: Dept. of Health, Education, and Welfare, National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research; 1978.

U.S. Department of Health and Human Services. Code of Federal Regulations - Title 45 Public Welfare CFR 46. 2009. From http://www.hhs.gov/ohrp/regulations-and-policy/regulations/45-cfr-46/

Vardigan M, Heus P, Thomas W. Data documentation initiative: Toward a standard for the social sciences. International Journal of Digital Curation. 2008; 3(1):107–113.

Waldo, J., Lin, H., Millett, L. Engaging privacy and information technology in a digital age. Washington, D.C: National Academies Press; 2007.

Wicherts, Jelte M., Veldkamp, Coosje LS., Augusteijn, Hilde EM., Bakker, Marjan, van Aert, Robbie CM., van Assen, Marcel ALM. Degrees of Freedom in Planning, Running, Analyzing, and Reporting Psychological Studies: A Checklist to Avoid p-Hacking. Frontiers in Psychology. 2016; 7(1832)doi: 10.3389/fpsyg.2016.01832

Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Musen MA. The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data. 2016; 3:160018. [PubMed: 26978244]