J-PAL
ABDUL LATIF JAMEEL POVERTY ACTION LAB

J-PAL GUIDE TO PUBLISHING RESEARCH DATA

Sarah Kopper, Anja Sautmann, and James Turitto
J-PAL Global
January 2020

Abstract: J-PAL promotes the publication of de-identified data from randomized evaluations. This resource provides guidance on doing so in the form of a checklist for preparing data for submission. It also includes sample informed consent language and other considerations during project planning and before publication, and provides a description of trusted digital repositories that can host data. This guide is intended to be read alongside the accompanying J-PAL Guide to De-Identifying Data.

Please contact research-resources@povertyactionlab.org with comments, questions, or feedback.

Acknowledgements: We thank Shawn Cole, Mary-Alice Doyle, Laura Feeney, William Parienté, and Karl Rubio for helpful comments and Antonn Park for copy-editing assistance. Any errors are our own.

TABLE OF CONTENTS

DATA PUBLICATION OVERVIEW

Over the last decade, the number of funders, journals, and research organizations that have adopted data-sharing policies has increased considerably. When the American Economic Association adopted its first policy in 2005, it was among the first academic journals in the social sciences to require the publication of data alongside the research paper. Today, data publication is required by most top academic journals in economics and the social sciences. Similarly, foundations and government institutions, such as the Bill and Melinda Gates Foundation,[1] the National Science Foundation,[2] and the National Institutes of Health, have data publication policies. J-PAL, as both a funder and an organization that conducts research, adopted a data publication policy in 2015 that applies to all research projects that we fund or implement.[3]

This guide and the accompanying guide on de-identification are intended to help research teams think about the steps involved in publishing research data. They draw on J-PAL's experience of publishing research data on randomized evaluations in the social sciences for more than a decade.

WHY PUBLISH RESEARCH DATA

Increasing the availability of research data benefits researchers, policy partners who supported the studies, students who learn from using the data, and, importantly, the people from whom the data was collected. Data sharing can provide many benefits and opportunities to the research community, including:

- *Allowing for re-use of the* data by researchers, policymakers, students, and teachers around the world.
- *Creating insights based on multiple studies* through meta-analyses, and answering questions on external validity and generalizability of results.[4]
- *Enabling the replication and confirmation of published results* as well as sensitivity or complementary analyses.

We are still in the early stages of data availability in social science research. Gertler, Galiani, and Romero (2018) found that only a small minority of empirical papers published in the top nine journals in economics in May 2016 contained all the necessary materials (raw data, estimation data, cleaning code, and analysis/estimation code) to successfully reproduce the results of the original study.[5] J-PAL's goal is to make research data from randomized evaluations widely available and accessible.

[1] https://www.gatesfoundation.org/How-We-Work/General-Information/Open-Access-Policy
[2] https://www.nsf.gov/bfa/dias/policy/dmp.jsp
[3] https://www.povertyactionlab.org/sites/default/files/documents/JPAL%20Data%20Publication%20Guidelines.pdf
[4] See, for example, Bandiera et al. (2016); Meager (2019).
[5] https://www.nature.com/articles/d41586-018-02108-9

CONSIDERATIONS BEFORE PUBLISHING

Before publishing your data, you should ensure that you have the legal, regulatory, and ethical authority to do so. Some questions to ask at the outset of the data publication process include:

- *Who is the owner of the data?* Was it collected by the research team through surveys, or is the data part of an administrative dataset that was provided by a third party? See below for additional considerations when publishing administrative data.
- *What did you report to your institutional review board (IRB) at the outset of the study?* How did you describe what would be done with the data? Did you mention data-sharing in your original protocol? Does your IRB have any specific guidelines on how research data should be published?
- *What information was provided to the study participants about how their data would be used?* See below for sample consent form language that allows for data publication.
- *How sensitive is the data that was collected?* Does the data include private health information and biomarkers, information about criminal activity, personal financial information, or agricultural yields?
- *What kind of data do you plan to publish (demographic information, geographic information)?* How identifiable are the study participants based on the data you plan to publish? Will all identifiers be removed, redacted, or masked during the publishing process? If not, have you considered publishing in a restricted data archive? See the accompanying Guide to De-Identifying Data for more information.
- *What jurisdiction(s) or legal authorities govern the data you collected?* Does the jurisdiction have certain legal frameworks you need to consider?
- *Does the donor have specific requirements for publishing research data?* What are their requirements? Is there a particular timeline by when you must publish the data? Is there certain data that must be published? Is there a particular location where the data should be published?

Addressing these questions will help determine where and what data to publish.

PERMISSION TO PUBLISH

Informed Consent for Survey Data

All studies that collect survey data from individual subjects should go through an informed consent procedure. The informed consent procedure should include language that allows for the publication of de-identified data. As with all parts of the consent procedure, this language should be concise and clear while avoiding jargon or technical terms that study participants may not understand. Before collecting data, researchers should review their consent process regarding data sharing.

- Avoid using complicated or confusing terms such as "confidentiality" or "de-identification."
- Avoid making promises that might limit what data can be shared later, such as "only members of the research team will have access to the information or responses that you provide to us."
- Avoid using terms that might not be achievable, such as "anonymous" or "anonymity," as truly anonymous data is a very high bar.

Sample consent form language that could be used (subject to approval by the IRB of record):

EXAMPLE CONSENT FORM LANGUAGE THAT PERMITS SUBSEQUENT DATA PUBLICATION

No one outside the survey team can directly connect your personal details, like your name, address, and cell phone number, with anything you say in this survey. Your survey responses and personal details will be stored in secure international computer storage. Your personal details will be encrypted and password-protected to prevent unauthorized access. Before we share the study with anyone other than the research team, your personal details will be separated from your survey responses. We do this to prevent anyone outside the research team from linking your survey responses back to you.

The Inter-university Consortium for Political and Social Research (ICPSR) has developed a set of recommendations for researchers to think about when drafting informed consent clauses, which you can access here.[6] Remember that informed consent and data-sharing disclosure requirements may vary by host institution and legal jurisdiction in which the data is collected; if in doubt, consult with the IRB of record.

Data Use Agreements for Administrative Data

In addition to the considerations listed above, publishing and sharing administrative data requires permission from the data provider, who will have final determination over what data can be published. Data that is provided by a third party often falls under additional regulatory authorities.

Much of what a research team can do with administrative data is controlled by the data use agreement (DUA) signed with the data provider at the beginning of the study. It is thus important to have a conversation about data-sharing with the data provider at the beginning of the study so that plans for publishing data can be added to the DUA at the outset. If the DUA does not clearly regulate data publication, then it is critical to consult with the data provider, and any research partners or implementing organizations, to determine what data can be published.

Data providers might have concerns about sharing their data because it is often governed by robust regulatory regimes (such as private health information, personal financial information, or criminal activity). They may be concerned about the privacy of participants or other potential effects if certain information is released. For example, businesses may be concerned about competitors using the data to gain commercial advantage, while government agencies may be concerned about political sensitivities, such as the release of data on spending practices or total case numbers.

It is important to discuss with data providers that data can be made available in a variety of formats. Restricted data archives, discussed in further detail below, provide locations where more sensitive and regulated data could be made available.

If a data provider has agreed to make data from the study publicly available, additional considerations include:

[6] https://www.icpsr.umich.edu/icpsrweb/content/datamanagement/confidentiality/conf-language.html

- Is there a review period for data providers? Many DUAs stipulate that data providers must have a chance (often 30 days) to review any papers, presentations, or use of their name prior to publication or submission for publication.
- Do the DUA and IRB allow researchers to publish data alongside the paper? If so, are there any conditions?
- Does the DUA or IRB specify that data must be deleted or returned to the provider?

WHERE TO PUBLISH

Trusted Digital Repositories

Published research data should be stored in a *trusted digital repository* to ensure long-term access to the files and documentation. A trusted digital repository is defined as a data repository "whose mission is to provide reliable, long-term access to managed digital resources to its designated community, now and in the future."[7]

A trusted digital repository is committed to maintaining the repository in perpetuity, ensures minimal to zero data loss or data rot, and allows for version control. For example, the Harvard Dataverse allows users to view exactly what has changed, starting from the originally published version to any subsequent published versions. Users are also allowed to access these versions and see the changes made for that particular dataset. A trusted repository also assigns each dataset and its associated program code with a unique identifier (e.g., a "digital object identifier" (DOI)) to enable citation, cataloging, and search. This identifier is designed to persist even if URLs—or the website itself—change. Most digital repositories capture metadata about the published research materials. This allows future researchers to quickly explore and understand the data without having to download the data and run the code.

Some common trusted digital repositories used by researchers in the social sciences include:

- Harvard Institute for Quantitative Social Sciences (IQSS) Dataverse[8]
- ICPSR[9]
- Mendeley[10] (by Elsevier)
- The UK Data Archive[11]
- Yale Institution for Social and Policy Studies (ISPS) Data Archive[12]

While posting data on personal websites (even if hosted by the university) technically makes it public, the lifespan of personal websites is much shorter than that of a trusted digital repository. Furthermore, it makes the data more

[7] "Trusted Digital Repositories: Attributes and Responsibilities. An RLG-OCLC Report" https://www.oclc.org/content/dam/research/activities/trustedrep/repositories.pdf
[8] https://dataverse.harvard.edu/
[9] https://www.icpsr.umich.edu/icpsrweb/
[10] https://www.mendeley.com/?interaction_required=true
[11] https://www.data-archive.ac.uk/
[12] https://isps.yale.edu/research/data

difficult to search, cite, and explore. Instead, researchers can cite their published data on their personal websites, by using the permanent identifier (DOI) to link to the where the data is stored.

Restricted Data Archives

Some repositories have set up archives for particularly sensitive data. In addition to openICPSR, a free public access repository, the ICPSR has also developed more secure repository options that range from secure downloads to physical on-site storage. More information on ICPSR's restricted repositories can be found on their website.[13]

A CHECKLIST FOR DATA PUBLICATION

This section provides a checklist of steps and best practices for researchers who wish to publish their research data. The checklist draws from the World Bank's Microdata Catalog submission checklist, and further information can also be found in ICPSR's Guide to Social Science Data Preparation and Archiving.

Preparing Data for Submission

After ensuring you have the authority to publish your data, you can prepare the data for publication. The process has two purposes: first, to ensure the dataset is clean and comprehensible to new users, and second, to make sure that the privacy of research subjects is protected.

The best way to publish a set of files related to a research project is to save them in a clear file and folder structure and then compress the entire set of folders into e.g., a zip archive. The folder structure might look something like the following:

Main folder
- Data
- Code
- Output
- Additional documentation
- Readme

Steps to prepare each of these files are described next. Recognizing that preparing data for publication can be a time-consuming and involved process, we differentiate between steps that are absolutely ***essential***, ***important*** (steps that we strongly recommend, as they facilitate re-use of the data), and ***suggested*** (additional steps that facilitate data re-use further but are less essential).

[13] https://www.icpsr.umich.edu/icpsrweb/content/ICPSR/access/restricted/

DATA

1. ***Essential***: Provide the data in a file format that can be used independently of statistical package choice, such as csv files.
2. ***Essential***: Ensure that your data is de-identified.
    1. Check for personally identifiable information (PII).
    2. Follow J-PAL's Guide to De-Identifying Data:
        1. Remove direct identifiers
        2. Make decisions about indirect identifiers
3. ***Essential***: Include all variables, treatment conditions, and observations collected from implemented survey instruments (excluding PII)—if feasible and allowed by the data provider (for administrative data).
    1. ***Important***: Keep enumerator IDs (so that enumerator effects can be controlled for in analysis). These IDs should be randomly generated so that enumerators' identities cannot be linked.
    2. ***Suggested***: Keep interview date (the day, week, or month of interview may be important for certain analyses, such as if asking respondents about consumption over the past seven days).

Less important to include are variables used for quality control during data collection. If using SurveyCTO, this could include text audits and time stamps (start time and end time). Automatically collected information such as the device ID, the subscriber ID, the SIM ID, and the device phone number do not need to be included—not least, they could contain identifying information.

4. ***Suggested***: Keep "raw" variables along with constructed, corrected, or imputed variables (except in cases where the raw variables could be used to identify individuals, as could be the case for variables with outliers).
5. If you have multiple datasets (e.g., administrative data on all households but survey data for only the households you interviewed, or a field-level dataset and a household-level dataset), then do the following:
    1. ***Essential***: Ensure there are ID variables in each dataset that link across datasets (e.g., ensure the household ID is in both the household-level dataset and the field-level dataset).

Important: Ensure there is no overlap between datasets (e.g., if you have panel data on daily electricity consumption and a survey that was conducted only once, do not publish a merged dataset (that includes the daily electricity consumption merged with the survey data) AND the two datasets separately—either publish just the merged file OR publish the two separate datasets. If one data set is much larger than the other, then they are ideally separate. This reduces the risk of inconsistencies and saves computer storage and memory when working with the data.

Additional checks for data:

1. ***Essential***: Does the ID variable uniquely identify observations?
2. ***Essential***: Are all missing values correctly and consistently coded and labeled?
3. ***Essential***: Can variables be matched with the accompanying questionnaire?
4. ***Important***: Is the ID the first variable listed?
5. ***Important***: Are all variables labeled? This can be done automatically by many digital data collection platforms, such as SurveyCTO.
6. ***Important***: Do all categorical variables have value labels? This can also be done automatically in SurveyCTO and similar software.
7. ***Important***: Other than the ID variable, are the variables ordered in a logical way? For example, if the order of the survey questions is potentially relevant for understanding the data, then the variables should be ordered in the same way. Distinct survey modules are often best grouped together. Variables in panel data (e.g., baseline and endline) should be grouped together and included in the same order for each wave (ideally, the variable names are the same for each wave, with suffixes indicating the survey wave).

CODE

1. ***Essential***: Include programs and scripts needed for a push-button replication of all published results:
    1. Make sure code files have headers (including the name of the person who last wrote/edited the code, the date, and the software and version used).
    2. Make sure code has comments or is self-documenting.
    3. Remove unnecessary code that creates tables and figures that are not included in the main results or appendix of the paper.
    4. Remove comments that are not necessary, such as comments and messages between the research team.

2. ***Important***: Data cleaning and variable construction documentation
    1. Codes for missing values in survey instrument and dataset
        1. We suggest using Stata's extended missing values (for example, don't know = .a, refuse to answer = .b, etc.). While researchers sometimes use values that will stand out when looking at the distribution of a variable, such as missing = –999, refused to answer = –888, Stata's extended missing values are preferred because they will be dropped automatically, without the need for a careful eye to look at the distribution of every single variable.
    2. Record of corrections made
        1. Were any variables grouped, top- or bottom-coded, smoothed out, or imputed? What method was used?
    3. If variables are masked for de-identification, report method used (link below).
    4. If applicable, describe how to link observations across data files (e.g., in an agricultural survey, describing how demographics of a plot manager who is a family member could be obtained by linking the plot roster to the household roster using household IDs, plot manager IDs, and family member IDs).
    5. If variables (e.g., consumption, total income) were constructed, how?

OUTPUT

If including output files with your published data, we suggest self-explanatory naming of files (e.g., table 1, table 2, etc.; or main tables, robustness checks, appendix tables, etc.; matching the corresponding publication).

ADDITIONAL DOCUMENTATION

1. Questionnaires
    1. ***Essential***: At a minimum, include the PDF version of the questionnaire in English translation, though we recommend also including the questionnaire in the original language if not in English.
        1. ***Suggested***: Including a pre-programmed SurveyCTO form can assist researchers wishing to conduct a statistical replication of the study in other settings.
    2. ***Essential***: Ensure PII (village names, etc.) is removed from all published questionnaires and documentation.
2. ***Important***: Provide research materials and description of procedures necessary to conduct an independent statistical replication of the research, including the following:
    1. Dates of field work
    2. Number of households visited
    3. Refusal rates
    4. Number of households and individuals in final sample
    5. Problems that occurred during the administration of the survey
    6. Enumerator manuals, preferably in English translation

3. ***Important***: Provide a copy of, or link to, the manuscript (published or working paper) that was written based on the specific data and code, if applicable.

README FILE

A readme file is ***essential.*** The readme file should be in an open, platform-independent format such as ASCII text, markdown, or a PDF and at a minimum include the following:

1. Description of data (e.g., replication data for xyz paper)
2. Date of data collection
3. System requirements
4. How to run the code files
5. List of files with short description
6. Permissions for publishing the data—this ideally includes the IRB approval number and institution, with a reference to or quote of the section of the section that allows publication of the de-identified data. If applicable, it should also include information on the data use agreement (DUA) and quote of the clause in which the data provider approves data publication.
7. Additional notes: for example, you may describe re-use rights, provide a recommended citation format or repository where the latest version of the data is stored, list contact information, etc.

For an example of a published readme document, see Dupas, Huillery, and Seban (2017).

Submitting Data to a Repository

1. ***Suggested***: Execute final versions of the code. Once all the code has been edited to accommodate the changes in the data, run it to ensure it runs without error and that the new results are consistent with those reported in the accompanying published or working paper.
2. ***Suggested:*** Maintain file structure for data users. First check the repository's policy and process guidelines for uploading. Each repository has a different process. We recommend zipping all of the files you have prepared in one folder so that the all files will be downloaded together in the appropriate structure. (NOTE: The Dataverse unzips archive files when data is uploaded so that double zipping will ensure the files remain zipped once on the server. This preserves the file structure, which will make it easier for other researchers to navigate and understand the data and other files, but will also result in some loss of functionality. For instance, the Dataverse Data Exploration function allows users to view and analyze data on the Dataverse site but cannot be used with zipped files.)
3. ***Important***: When submitting a dataset, the inclusion of detailed metadata can facilitate re-use by providing other researchers with key information about the data without needing to download it. Whenever possible, metadata should be added to the fields in the repository, making it machine readable. However, if no fields are provided by the repository and no additional fields can be added to the repository, then you should include metadata in the readme documentation. We strongly recommend including the following:
    1. Study identification: Project title; names of all principal investigators (if available including research ORCID numbers); keywords; and conflict of interest or funder information
    2. Study overview: Target population, intervention type, methodology, whether this is a panel study, anticipated survey rounds (if panel study, then list number of surveys in this round), project description, project period, related publication or working paper
    3. Geospatial metadata: Country, state, district, and other relevant toponyms; geographical coverage (e.g., national, state, school)

4. Implementation: Unit of randomization, if applicable (if clustered, specify level of clustering); unit of stratification, if applicable; unit of analysis (individuals, households, firms, schools, etc.); variables describing units of randomization, stratification, and analysis; sampling method for study inclusion; sampling method for each survey round; and project partners and nature of partnership (e.g., schools provided venue to meet with parents, hospitals gave records of doctor visits)
5. Software/programs: Software of code and program version number
6. Terms of use: Access levels, access conditions, and data contact information

4. *Note*: Different data repositories have different limits regarding file size. The repositories listed above (ICPSR, Dataverse, and the ISPS Data Archive) generally do not limit the number of files that can be uploaded but may limit file size (e.g., the Harvard Dataverse limits file size to 2.5 GB). Support for big data publication and use is under development at some repositories (see the Dataverse's most recent work to this end here).[14] If your file size exceeds the limit for your chosen repository, we recommend compressing the file to below the size limit (if possible). You may be required to split the large data file into two or more files by only including a subset of the variables in each file. Consider creating randomized extracts so that interested users can open one of the files and see a representative subsample. You may contact the repository support team for advice on how to handle large datasets.

Updating Published Data

Trusted data repositories allow for version control, which is especially useful for long-term research data management where metadata and files are updated over time. Versioning is used to track any metadata or file changes (e.g., by uploading a new file, changing file metadata, adding or editing metadata) once you have published your dataset. In most repositories, a new DOI will be issued for each version. When updating versions of published data, it is important to include a note documenting the changes made from the previous version.

A trusted repository will automatically document changes such as metadata changes and addition/deletion of files, but it will not document the specific changes inside the files. So for these data/file-specific changes, it is useful to note down in the readme file the changes that were made. For example:

Version 1: Nairobi eye clinic survey round 1
Version 1.1: Coding error fixed
Version 1.2: Added related publication
Version 2.0: Round 2 data added

Deaccessioning Data

The act of removing a dataset from a digital repository where it has been published is known as deaccessioning. Most repositories will have a process for deaccessioning a dataset. Deaccessioning is an important component of version control that allows the metadata and citation of the data to remain available even if access to the data is removed. You can deaccession one version of a dataset or the entire dataset. Deaccessioning of a dataset could happen for a variety of reasons. For instance, if a research team inadvertently publishes the wrong dataset, incomplete data, or code with an error, then they might want to deaccession the dataset and upload the correct dataset that accompanies their study. When a dataset is deaccessioned, that version will no longer be available, but the citation and metadata will remain

[14] "Big Data Support," http://guides.dataverse.org/en/latest/developers/big-data-support.html

available. This allows other researchers to know that a version of the dataset existed previously in case they come across it in another study.

RESOURCES

- Bandiera, Oriana, Fischer, Greg, Prat, Andrea, and Erina Ytsma. 2016. "Do Women Respond Less to Performance Pay? Building Evidence from Multiple Experiments." CEPR Discussion Paper No. 11724. https://ssrn.com/abstract=2893078
- Berkeley Initiative for Transparency in the Social Sciences. n.d. "Replication." Last accessed January 4, 2019. https://www.bitss.org/research-transparency-mooc/replication/
- Center for Open Science. n.d. "Transparency and Openness Promotion Guidelines." Last accessed January 4, 2019. https://cos.io/top/
- Dataverse. n.d. "Big Data Support." Last accessed December 8, 2019. http://guides.dataverse.org/en/latest/developers/big-data-support.html
- Dillon, Moira R., Kannan, Harini, Dean, Joshua T., Spelke, Elizabeth S., and Esther Duflo. 2017. "Cognitive Science in the Field: A Preschool Intervention Durably Improves Non-Symbolic, but not Symbolic, Mathematics." Harvard Dataverse, V2. doi:10.7910/DVN/LCLKDT
- Dupas, Pascaline, Huillery, Elise, and Juliette Seban. 2017. "Risk Information, Risk Salience, and Adolescent Sexual Behavior: Experimental Evidence from Cameroon." Harvard Dataverse, V1. https://doi.org/10.7910/DVN/MLVGY9
- FAIR Data Principles.
- Gertler, Paul, Galiani, Sebastian, and Mauricio Romero. 2018. "How to Make Replication the Norm." Nature 554, 417–419.
- ICPSR. n.d. "Guide to Social Science Data Preparation and Archiving." Last accessed January 4, 2019. https://www.icpsr.umich.edu/icpsrweb/content/deposit/guide/index.html
- J-PAL. n.d. "Guidelines for Data Publication." Updated June 2015. https://www.povertyactionlab.org/sites/default/files/documents/JPAL%20Data%20Publication%20Guidelines.pdf
- Meager, Rachael. 2019. "Understanding the Average Impact of Microcredit Expansions: A Bayesian Hierarchical Analysis of Seven Randomized Experiments." AEJ: Applied 11, 57–91.
- Research Libraries Group. n.d. "Trusted Digital Repositories: Attributes and Responsibilities. An RLG-OCLC Report." Last accessed December 10, 2019. https://www.oclc.org/content/dam/research/activities/trustedrep/repositories.pdf
- World Bank. n.d. "Checklist: Microdata Catalog Submission." Last accessed January 4, 2019. https://dimewiki.worldbank.org/wiki/Checklist:_Microdata_Catalog_submission