New Curation Software:

Step-by-Step Preparation of Social Science Data and Code for Publication and Preservation

by Limor Peer[1] and Stephanie Wykstra[2]

Abstract

As data-sharing becomes more prevalent throughout the natural and social sciences, the research community is working to meet the demands of managing and publishing data in ways that facilitate sharing. Despite the availability of repositories and research data management plans, fundamental concerns remain about how to best manage and curate data for long-term usability. The value of shared data is very much linked to its usability, and a big question remains: What tools support the preparation and review of research materials for replication, reproducibility, repurposing, and reuse? This paper describes key curation tasks and new data curation software designed specifically for reviewing and enhancing research data. It is being developed by two research groups, the Institution for Social and Policy Studies at Yale University and Innovations for Poverty Action, in collaboration with Colectica. The software includes curation steps designed to improve the research materials and thus to enable users to derive greater value from the data: Checking variable-level and study-level metadata, verifying that code can reproduce published results, and ensuring that PII is removed. The tool is based upon the best practices of data archives and fits into repository and research workflows. It is open-source, extensible, and will help ensure that shared data can be used.

Keywords

Data curation, curation software, data sharing, social science, randomized controlled trials

Introduction

Over the past 10 years, many scientific communities have embarked on discussions of data-sharing and reproducibility. From Biology (Vines, 2014) to Epidemiology (Peng, 2006) to Economics (Hammermesh, 2007) to Political Science (King, 1995), researchers are calling for more data sharing. Research funders and journals have been encouraging data sharing and adopting data access policies in greater numbers over the past decade. For example, in the UK, all of the Research Councils have adopted data-sharing policies (see Data Curation Center's useful *summary*[3] of all of these policies). Wellcome Trust in the UK has led a *joint statement*[4] of purpose on data-sharing principles, which includes over 15 funders. In the US, the Office of Science and Technology Policy *memorandum of 2013*[5] stipulated that US funders receiving $100M or more in federal research funds adopt data-sharing polices, and the government is working to facilitate *code sharing*[6]. Major foundations such as the *Bill and Melinda Gates Foundation*[7] and the *Laura and John Arnold Foundation*[8] have also adopted data-sharing polices. A number of journals are instituting policies in which they require researchers to share the data and code underlying the published research results (see *this list of social science journals with a data sharing policy*[9] and this *journal data policy review*[10]).

There is much variety across policies. Funder policies differ in their timeframes, whether data should be made openly available or simply available on request, which materials should be shared, and in many other ways (for an overview, see Wykstra, 2013). Likewise, journals vary in whether data should be available openly. Some journals, for example the *American Economic Review*[11], require researchers to post the data on the journal website, whereas other journals merely ask researchers to note in the article where they shared the data or that they make it available upon request.

> research will be more credible if others can have full access to all aspects of scholarly work

While the language and particulars may vary, a constant theme running through these discussions is the desire for scientists to be able to examine each other's work. Can others dig into the analysis and data; can others understand the study in enough detail to try to repeat it?

In this paper, we focus on an issue which is crucial for examining others' work: that of the usability of shared data. By "data" here, we mean not just the datasets

themselves but the related materials as well: the analysis code, the metadata, documentation, and instruments. We refer to preparing these materials for public use as data curation. After a description of this project and a discussion of the value of data sharing, we discuss the relation of reproducibility, re-use, and data curation, describe key curation tasks, and present new curation software, developed with *Colectica*[12], aimed at helping with review and enhancement of research materials.

Background: Data from randomized controlled trials (RCTs) in the social sciences

The impetus for the collaboration around data curation between the Institution for Social and Policy Studies (ISPS)[13] at Yale University and Innovations for Poverty (*IPA*)[14] is a focus on a particular way of doing social science research: field experiments. Both organizations collect data from social science research that measures the impact of interventions – such as voter mobilization campaigns and microfinance programs – via randomized controlled trials in the real world. ISPS has been involved with close to 100 such studies, mostly in political science, and IPA in about 300 studies, working with researchers in development economics, among other fields. Studies linked with ISPS and IPA have been published in such journals as *The American Political Science Review, Political Analysis, American Political Research, Public Opinion Quarterly, American Economic Review, American Behavioral Scientist,* and *The Quarterly Journal of Economics.*

Data from these studies are mostly quantitative, often gathered from a combination of administrative records, surveys, and observation, and of potentially high value for researchers, educators, policy makers and students. Data are often generated to address a particular research question and linked to a publication that describes the results of a particular experiment. Datasets underlying published articles or books span time periods and continents, and vary in scale in terms of the number of observations and variables.

Since RCTs are relatively new to the social sciences, metadata standards are still emerging. The Data Documentation Initiative (*DDI*[15]), the primary social science metadata standard, now has a *working group* [16] charged with updating the standard to capture the unique characteristics of this research method. High quality descriptive metadata is essential to facilitating the interpretation of social science studies.

ISPS has supported a *Data Archive*[17] since 2010 (Peer and Green, 2012). The Archive includes research output by ISPS-affiliated researchers, with emphasis on experimental design and methods. Research output includes data and code and is typically deposited at the end of the project, coinciding with manuscript publication. Research output is organized as a complex object around a study, with multiple files of various sorts related to each study, including data, code, output, and other files. Study-level metadata are compiled from information provided by depositors (e.g., researchers) via a deposit agreement form, and from associated materials (e.g., published article). For variable-level metadata, ISPS uses Stat/Transfer to produce make available XML files based on DDI version 3.1 for datasets.

IPA has also launched a new repository to share data from RCTs (both from IPA studies as well as RCTs from other groups). The repository is hosted by Harvard's *Dataverse*[18]. IPA shares ISPS' approach to curation but differs in that it also requests that researchers share the full collected datasets, and works to help them prepare the larger datasets, as opposed to only the data underlying the published research results. In addition, IPA is working with research staff on the ground within its country offices, to improve code and data management processes early on in the study workflow and improve later data usability.

The value of data-sharing

There are two primary sources of value from sharing data: reuse and transparency. First, sharing data permits others to use the data for further purposes. It is currently more common to re-use data from large-scale survey-based studies such as *Demographic and Health Surveys*[19], than to re-use data from experimental studies. However, as data sharing becomes more prevalent there is promise that scientists will conduct additional analyses, such as secondary analysis and meta-analysis and formulate new questions. This is the logic expressed in a *2013 OSTP memo*[20] to all government agencies, which states with respect to government-funded studies that, "the results of that research become the grist for new insights and are assets for progress in areas such as health, energy, the environment, agriculture, and national security." There is some evidence that, at least in one field, studies that made data available received more citations than similar studies for which data were not made available, as measured by number of citations (Piwowar, 2013).

The idea driving research transparency is that research will be more credible if others can have full access to all aspects of scholarly work that led to publication. As King (1995) put it, "the only way to understand and evaluate an empirical analysis fully is to know the exact process by which the data were generated and the analysis produced" (p.444). An essential component is the ability to reproduce computations and analyses by using the shared code and data. Re-analysis of this kind is often assigned in methods courses in the social sciences, in which it is also often recommended that "replicators" go beyond simple re-analysis to conduct robustness checks and delve into the analytical decisions made in the published research (King, 1995). Access to code in addition to the data is increasingly recognized as critical in all computational sciences (i.e., those which rely heavily on analysis of quantitative data) and as contributing to the credibility of the research (Stodden et al., 2013).

Data use and data curation

In order to glean full value from shared data, for re-analysis or any future re-use, the data must be usable in the long-term. The usability of data simply means that it can be "independently understandable" by future scientists (Peer, Green and Stephenson, 2014; Peer, 2014a; Peer and Green, 2015).

Preparing files for long-term use starts with good documentation. It is strongly recommended that a standards-based, structured, open and machine-readable metadata scheme, such as DDI for social sciences data, is used (e.g., Starr et al., 2015; U.S. Government, 2012 ; W3C, 2015). Preparing data files includes, but is not limited to, ensuring that variables are clearly named and labeled. Variables created via original data collection should be linked to the source, e.g., survey questions. Numeric data with value codes should be labelled clearly. Code should be commented to indicate which operations the code carries out (e.g., variable-construction and cleaning, producing tables). In the context of a study, additional documentation may be required. It is recommended that researchers provide readme files documenting the files which are shared, with instructions about running the files and any other information about them (see this *useful guide*[21]). Sufficient study-level metadata is critical to understanding the study and its context. Information such as: time period, geographic area,

sampling frame and selection method, sample size, study methodology, and data collection method should be provided. If there is a publication, the data and the published research results should be clearly linked. And, of course, open and persistent access to files is a precondition for long-term usability. Most basically, the files should be in a sustainable location, preferably a data repository which offers long-term preservation. Files should also be available in non-proprietary formats to increase accessibility. If they are in proprietary formats, there may also be a greater chance they will become unusable over time due to software updates.

We refer to the process of reviewing and enhancing research outputs for the purpose of long-term usability as "data curation." According to the *Digital Curation Center*[22] curation involves "maintaining, preserving and adding value to digital research data throughout its lifecycle."

Our goal in undertaking data curation is to ensure that users may have persistent access and be able to correctly interpret and re-use these materials without the need to contact original researchers. We see particular value in curation when the intended re-use of the research materials is to fully evaluate an empirical analysis, that is, to reproduce research results (Peer, 2011)[23]. Specifically, we use the "data quality review" framework to focus on specific curation tasks (Peer, Green and Stephenson, 2014).

Key curation tasks for data from RCTs in the social sciences

To ensure research transparency, long-term usability, and ongoing persistent access to research outputs generated by a specialized community, certain data curation tasks need to take place. Data archives such as the Inter-University Consortium for Political and Social Research (*ICPSR* [24]) and UK Data Archive (*UKDA*[25]) have established practices that are tried and tested to ensure "that data are accurate, complete, well documented, and that they are delivered in a way that maximizes their use and reuse" (Peer, Green and, Stephenson 2014, p.16). The ISPS curation workflow is based on the ICPSR pipeline (Peer, 2014b), and has been adapted for research output from RCTs in the social sciences (Peer and Green, 2012).

The existing *ISPS Data Archive workflow*[26] has gone a long way toward satisfying the curation needs of ISPS, but the partnership with IPA presents an opportunity to build a modular, open-source curation tool that could be adapted by our organizations to changing needs, research methods, dissemination platforms, and preservation solutions.

On the basis of data archives' best practices and the ISPS curation workflow, we have identified eight key curation tasks that are designed to improve the research materials and thus to enable users to derive greater value from the data by, for example, checking variable-level and study-level metadata and

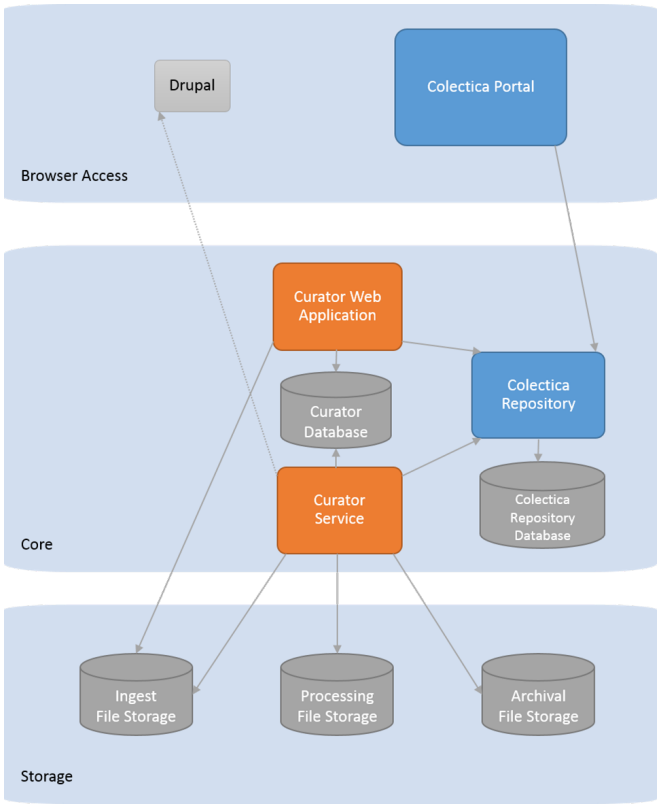


Figure 1

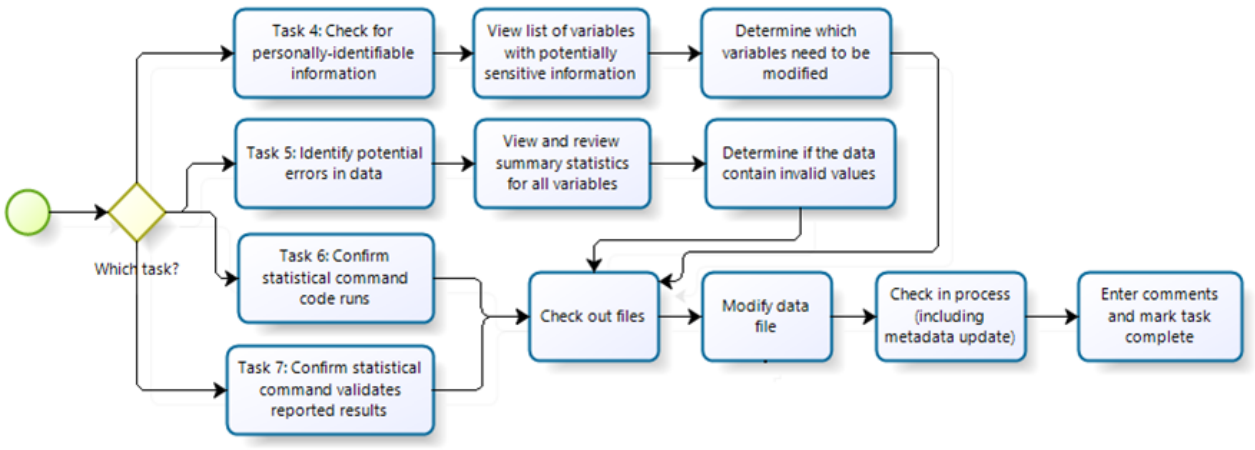


Figure 2

ensuring that personally-identified information is removed. The curation tasks also include the review of code files -- statistical and other programming scripts -- for the purpose of checking whether scientific results can be reproduced with the code and data provided.

The eight tasks are as follows (see Appendix):

1 Check for missing labels.
2 Review observation count.
3 Identify potential data errors.
4 Compare questionnaire, codebook, and data.
5 Ensure there is no personally-identifiable information (PII) in Data File.
6 Confirm code executes.
7 Confirm code replicates reported results.
8 Create preservation and open formats

Digital curators and research teams who strive to meet the demands of these tasks also need to track and confirm the completion of the tasks as well as to capture all the useful metadata generated throughout the processes. The prime objectives for this project are: To automate as many of the curation tasks as possible, to technically integrate these curation tasks, and to do so using a structured but flexible workflow. These eight curation tasks constitute core requirements for the software we describe here[27].

New curation software

Working with Colectica, a software development group specializing in data and metadata tools for social sciences, we have developed new software that structures and tracks the curation workflow, helps automate parts of the data pipeline, captures all metadata throughout the process, and pushes out relevant information to pre-determined destinations (i.e., a user, the archive administrators, a Web based dissemination system, or preservation systems). The tool was developed to fit into repository and research workflows.

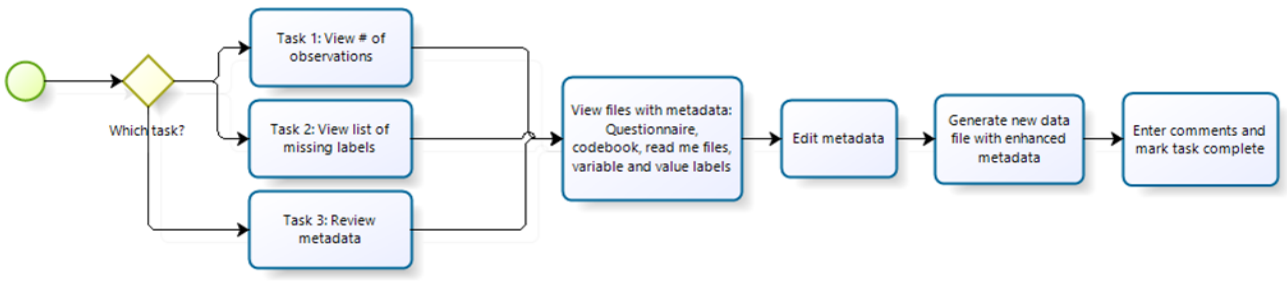
The software is primarily open-source, extensible, and can be easily integrated with other systems. It is written in C# and runs on the ASP.NET MVC framework. It leverages DDI Lifecycle (also known as DDI 3.2) and combines several off-the-shelf components with a new, open source Web application that integrates the existing components to create a flexible data pipeline. Default components include StatTransfer[28] , Colectica Repository, and BagIt file packaging format, but the software is developed so each of these can be swapped for alternatives.

Key curation software characteristics:

1 The software supports specific curation tasks for different file types and provides automatically-generated information helpful in performing each task. The file categories relevant to curation are data files, code file, and other.
2 The software facilitates the production of descriptive metadata at the study, file, and variable levels, and maps and stores all metadata in DDI Lifecycle format. Study-level metadata is used to inform the catalog record. The software identifies file types and, for data files with recognized formats, such as .dta, R, and .sps, produces further metadata for each variable.
3 The software allows editing metadata in the web-based interface and automatically updates file- and variable-level metadata, and the catalog record. The software can produce a new version of a data file with changed metadata (data not changed). All changes to files are tracked using a git platform.
4 The software also allows downloading files for further review or editing for specific curation tasks. The revision management check-out/check-in system recognizes files that have been updated offline and then checked as a new version, and all relevant metadata is updated, with versions of metadata corresponding to new versions of files. All versions are stored. The system records which curation task was completed. For data files, the software also provides the option to create and store revised summary statistics.
5 The software allows viewing other files in a window to compare documentation to metadata in the system.
6 Notes can be entered at every step, at every level of metadata.
7 Each curation step can be approved or rejected. All activity around these curation tasks is tracked, and each task must be performed before a catalog record is published. This allows curation progress to be seen at a glance, and ensures a reliable record of the curation process.

Discussion

We conclude with a few thoughts about this new curation software and its benefits for the scientific research community. The main advantage of this tool is that it helps codify and automate a series of curation tasks that prepare data and code for re-use. Many of these tasks are not new: they have been described in numerous best-practice documents and guides. Yet, as far as we know, there are no tools that facilitate curation of research data prior to their ingest into a repository.[29] The advantages of this tool include its functionality to create a consistent workflow for key curation tasks that can integrate several software environments, automatic metadata production, presentation of missing variable and value labels, versioning of files to control what has been done, and the ability to add notes to document changes and enhancements to files and to track the entire process. We think this unified curation workflow can help bring about many of the recommended



Fiqure 3

curation practices that the research data management community has been advocating for, at scale.

In terms of the application of the software, it is our recommendation that this software be used as close as possible to the research process. In-house trained curation staff at research labs or centers are best positioned to understand the research, the data, and the analyses and can create additional documentation if possible and necessary. In-house staff also typically benefit from access to, and communication with, researchers in case questions come up. Researchers may be asked to provide more information if documentation is incomplete or clarifications are needed. Information and curation specialists, such as data and subject liaison librarians, in conjunction with statistical experts in the researcher's institution or professional society are also in a good position to undertake this type of curation. If no curation was done on research outputs intended for re-use or preservation, we urge repositories, journals, and funders, to facilitate or make use of this software before research outputs are preserved or disseminated.

This curation tool is flexible enough to allow modification. We have described major steps that we have identified within our own research groups, and we have customized the software to aid with these steps. However, the steps may be modified according to the needs of particular research groups, repositories, or researchers using the software. For example, differences between ISPS and IPA in research management and infrastructure led to their using the tool differently. Generally, we foresee circumstances in which some curation tasks may not be relevant (e.g., stand-alone data may not require regenerating the results by running the code to produce tables) and conversely, instances in which additional curation tasks may be added to fulfill specific repository, lab, or discipline requirements and standards.

There are multiple mechanisms for sharing data these days, and most involve some level of curation. This tool can be used by researchers or labs who wish to self-deposit into a general data repository, by established data archives that are looking to automate and integrate disparate curation processes and systems, by journals or funders who wish to review research outputs before they disseminate them along with publications, and by institutional repositories and other archives who plan to preserve these research outputs and ensure they can be persistently accessed and usable. If used by data archives or by general data repositories such as Dryad, Figshare or Dataverse, this tool may be used to support a service model of managing data with a view toward usability and replication.

We have argued that data curation is an essential part of the movement towards open data. Data curation is a key component of usability, without which long-term re-use is not possible. Therefore, we hope that the open source software will be taken up by a number of groups in addition to our own.

Acknowledgments

We thank our partners, Jeremy Iverson and Dan Smith at Colectica, for collaborating with us on this project. We would also like to acknowledge Ann Green and Niall Keleher for their significant contribution to the conception and early development of this entire endeavor.

References

Asendorpf, Jens B., Conner, Mark, De Fruyt, Filip, De Houwer, Jan, Denissen, Jaap J. A., Fiedler, Klaus, Fiedler, Susann, Funder, David C., Kliegl, Reinhold, Nosek, Brian A., Perugini, Marco, Roberts, Brent W., Schmitt, Manfred, van Aken, Marcel A. G., Weber, Hannelore and Wicherts, Jelte M. (2013) Recommendations for Increasing Replicability in Psychology. European Journal of Personality, 27: 108–119. (Available at http://onlinelibrary.wiley.com/doi/10.1002/per.1919/full#per1919 or doi: 10.1002/per.1919

Borgman, Christine L. (2010) Research Data: Who Will Share What, with Whom, When, and Why? RatSWD Working Paper No. 161. (Available at SSRN: http://ssrn.com/abstract=1714427 or http://dx.doi.org/10.2139/ssrn.1714427)

Collberg, Christian, Proebsting, Todd, Moraila, Gina, Shankaran, Akash, Shi, Zuoming, and Warren, Alex M. (2014) "Measuring Reproducibility in Computer Systems Research," Department of Computer Science, University of Arizona, Technical Report. (Available at: http://reproducibility.cs.arizona.edu/tr.pdf)

Hammermesh, Daniel, (2007) Viewpoint: Replication in Economics. Canadian Journal of Economics, 40, no. 3: 715-733.

Holdren, John (2013) Increasing Access to the Results of Federally Funded Research, Office of Science and Technology Policy. (Available at http://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf)

King, Gary (1995) Replication, Replication. PS: Political Science & Politics, 28: 444-452.

Peer, Limor (2011) Building an Open Data Repository: Lessons and Challenges (Available at SSRN: http://ssrn.com/abstract=1931048 or http://dx.doi.org/10.2139/ssrn.1931048)

Peer, Limor (2014a) Why 'Intelligent Openness' is Especially Important When Content is Disaggregated. ISPS Lux et Data blog (Available at http://isps.yale.edu/news/blog/2014/12/why-intelligent-openness-is-especially-important-when-content-is-disaggregated)

Peer, Limor (2014b) Mind the Gap in Data Reuse: Sharing Data is Necessary But Not Sufficient for Future Reuse. LSE Impact blog. (Available at http://blogs.lse.ac.uk/impactofsocialsciences/2014/03/28/mind-the-gap-in-data-reuse/)

Peer, Limor and Green, Ann (2012) Building an Open Data Repository for a Specialized Research Community: Process, Challenges and Lessons. International Journal of Data Curation, 7(1): 151-162. (Available at http://www.ijdc.net/index.php/ijdc/article/view/212)

Peer, Limor and Green, Ann (2015) Research Data Review is Gaining Ground. ISPS Lux et Data blog (Available at http://isps.yale.edu/news/blog/2015/03/research-data-review-is-gaining-ground)

Peer, Limor, Green, Ann and Stephenson, Elizabeth (2014) Committing to Data Quality Review. International Journal of Data Curation, 9(1): 263-291. (Available at http://www.ijdc.net/index.php/ijdc/article/view/9.1.263/358)

Peng, Roger, Francesa Dominici, Scott Segar (2006) Reproducible Epidemiologic Research. American Journal of Epidemiology, 163(9): 783-789.

Piwowar, Heather, RS Day, DB Fridsma (2007) Sharing Detailed Research Data Is Associated with Increased Citation Rate. PLoS ONE 2(3).

Starr J, et al. (2015) Achieving human and machine accessibility of cited data in scholarly publications. PeerJ Computer Science 1:e1 (Available at https://dx.doi.org/10.7717/peerj-cs.1)

Stodden, Victoria et al. (2013) Setting the Default to Reproducible: Reproducibility in Computational and Experimental Mathematics. ICERM workshop paper. (Available at http://stodden.net/icerm_report.pdf).

United States Government (2012) A Primer on Machine Readability for Online Documents and Data data.gov

blog (Available at https://www.data.gov/developers/blog/primer-machine-readability-online-documents-and-data)
Vines, Tim et al. (2014) The Availability of Research Data Declines Rapidly with Article Age. Current Biology, 24(1): 94-97.
W3C (2015) Data on the Web Best Practices. Working Draft. (Available at http://www.w3.org/TR/2015/WD-dwbp-20150625/)
Wykstra, Stephanie (2013), Data Access Policies Landscape, Figshare, (Available at http://dx.doi.org/10.6084/m9.figshare.827268

Appendix

Curation tasks detail

We describe here the eight curation tasks essential to our work, with an explanation of how each is handled by the software. Note that users of the software may choose a combination of one or more of any of the tasks, and may also extend them with APIs to other software. These may include the generation of descriptive statistics, a data dictionary and a set of variable frequency distributions. (See Figures 2, 3.)

1. Check for missing labels.
   a. Rationale: All variables should be properly identified with a name and a description that provides additional information. Nominal (or categorical) variables should have numeric or string labels for each category (see more on *ICPSR guide* [30]).
   b. Description: For data files. The software displays variables and value labels and alerts of any missing labels. Upon ingest, the software analyses data files and extracts variable-level metadata for each column. This metadata is stored in DDI 3.2 format. For any variables without labels, and for categorical data without value labels, the software prompts the curator to enter labels.
   c. Technical: This variable-level metadata extraction is built on the data import functionality found in Colectica Designer and Colectica Repository. The software currently supports reading variable-level information for Stata, RData, and CSV files. Other formats can be supported using Stat/Transfer to convert the file to a supported format, or by extending the data ingest capabilities of the curation software.

2. Review observation count.
   a. Rationale: The goal is to view the number of cases (observations) for every variable and for the dataset as a whole. This is helpful in providing the basis for subsequent curation tasks.
   b. Description: For data files. The software displays # of observations, summary statistics including frequencies. When ingesting data files, the curation software determines the number of observations, calculates summary statistics, and stores this information about the data file with the metadata.
   c. Technical: This task is completed using the curation web application, but may require viewing and reconciling documents in other readers or editors.

3. Identify potential data errors.
   a. Rationale: Unlikely or impossible values for interval variables and undefined or incorrect values for nominal (categorical) variables make it difficult for future users to interpret the data. Also out of range and missing values. This is intended to check the overall integrity of the data. The *UKDA guide*[31] provides examples of such anomalies.
   b. Description: For data files. The curation web application provides a variable-level metadata browser. This shows details of each variable in a data file, including summary statistics and value labels for categorical variables. Curators are responsible for reviewing each variable to ensure there are no obvious errors in the data.
   c. Technical: This task can be completed using the curation web application, but may also require viewing or editing the data in a statistical software package.

4. Compare questionnaire, codebook, and data.
   a. Rationale: This task can be carried out at the same time as other tasks related to data files. Performing tasks 1-3 may be informed by other documentation that was deposited along with the data files. Comparing summary and descriptive statistics along with data dictionary or codebook helps ensure that question text, labels, response categories and value labels are consistent.
   b. Description: For data files. The software allows viewing other files in a window to compare documentation to file- and variable-level metadata in the system, or importing the information from a questionnaire if it can be read by Colectica Designer. For each variable, curators are shown summary statistics and label information, and can review other documents, including links to publications based on the data. Curators are responsible for flagging instances where the observation count reported in publication does not match the observation count in the data file, where variables are missing or transformed, or where label information is missing or incomplete. Curators are responsible for ensuring all information is consistent.
   c. Technical: This task is completed using the curation web application, but may require viewing and reconciling documents in other readers or editors.

5. Ensure there is no personally-identifiable information (PII) in Data File.
   a. Rationale: Human subject data are prevalent in the social sciences. Any entity that shares human subject research data has responsibility to protect respondent confidentiality and to minimize the risk of identifying individuals. Guidance on *confidentiality*[32] is provided by *ICPSR*[33] and the Australian National Data Service (*ANDS*)[34].
   b. Description: For data files. Using the curation web application's variable-level metadata browser, curators are responsible for reviewing each variable to ensure it does not contain personally identifiable information (PII). If a curator finds variables containing names, social security numbers, or other identifiable information, they are responsible for removing the columns and submitting a new version of the data file.
   c. Technical: This task can be completed using the curation web application, but may also require viewing or editing the data in a statistical software package, as well as also viewing other documentation. In addition to visual inspection of the data and documentation, code may be used to identify variable names that indicate PII (social security numbers, phone numbers, addresses, etc). Future development of the software could enable display of a list of predefined types of variables or data and API integration with anonymization software (e.g., *Anonimatron*[35], *QualAnon*[36]). Follow up may

require curator to create new program file that produces a new data file.

6. Confirm code executes.
   a. Rationale: By code we are referring to computational workflows used in the research process, including data collection, cleaning, and analysis. In some disciplines, researchers are just warming up to the idea of sharing their data and are not used to providing code. But even in disciplines such as computer science, where code is commonly shared, problems with code builds exist (Collberg et al., 2014). The goal here is to test the code with the given data to identify any potential errors in the script itself.
   b. Description: For code files. Curators are responsible for ensuring that all source code submitted executes without errors. The curation web application provides links to download the source code and any dependencies, such as data files. A web-based preview with syntax highlighting is also available.
   c. Technical: To complete this step, curators must use the appropriate statistical software (e.g. Stata, R, SPSS, or SAS). The curation software tracks the versions of the code files for changes made by the curator.

7. Confirm code replicates reported results.
   a. Rationale: After confirming that the script is error free, "an assessment is made about the purpose of the code (e.g., recoding variables, manipulating or testing data, testing hypotheses, analysis), and about whether that goal is accomplished." (Peer et al., 2014). The main goal is to check whether the code, in conjunction with the data provided, produces the results reported. The idea is that, "Researcher B… obtains exactly the same results (e.g. statistics and parameter estimates) that were originally reported by Researcher A (e.g. the author of that paper) from A's data when following the same methodology" (Asendorpf et al., 2013). Confirmation that results can be replicated, and any additional annotation created in the process, helps inform future users exactly how results were generated. Note that the focus here is on the regeneration of results and not on the correctness of the methodology, analysis, or interpretation of the results.
   b. Description: For code files. Curators are responsible for ensuring that statistical programs produce the results reported in any related publications. The curation web application provides links to download the source code and any dependencies, such as data files. The curator analyzes the output of these programs, reviews any numbers, tables, and charts included in publications, and ensures they match the output.
   c. Technical: To complete this step, curators must use the appropriate statistical software (e.g. Stata, R, SPSS, or SAS). The curation software tracks the versions of the code files for changes made by the curator.

8. Create preservation and open formats.
   a. Rationale: The goal is to create file formats that easily lend themselves to reuse via technology. Files trapped in licensed formats (e.g., xls, .dta) will not be available for use by non-licensed software mechanisms, and may be less usable over time as software is outdated. The *UKDA guide*[37] explains preservation formats.
   b. Description: For all files. The curation software automatically converts supported data files to CSV format for preservation. This occurs during publication after curation has been performed, reviewed, and approved.
   c. Technical: For file types in proprietary formats that are not specifically supported by the curation software, curators are responsible for creating a file in the appropriate preservation format, and uploading the file to the catalog record. This can be accomplished using conversion software such as Stat/Transfer, or saving documents to text or PDF formats. For code file, curators may write R scripts that replicate any statistical code provided by the researcher in a licensed statistical program.

Notes

1 Limor Peer is Associate Director for Research at Yale University's Institution for Social and Policy Studies, http://isps.yale.edu/. Contact: limor.peer@yale.edu.
2 Stephanie Wykstra is Research Manager of Research Transparency at Innovations for Poverty Action, http://www.poverty-action.org/
3 http://www.dcc.ac.uk/resources/policy-and-legal/overview-funders-data-policies
4 http://www.wellcome.ac.uk/About-us/Policy/Spotlight-issues/Data-sharing/Public-health-and-epidemiology/WTDV030690.htm
5 https://www2.icsu-wds.org/files/ostp-public-access-memo-2013.pdf
6 https://government.github.com/
7 http://www.gatesfoundation.org/How-We-Work/General-Information/Open-Access-Policy
8 http://www.arnoldfoundation.org/sites/default/files/pdf/Guidelines%20for%20Research%20Funded%20by%20LJAF%2011-12-2013%20MA%20-%20July%2016%202015.pdf
9 https://jordproject.wordpress.com/project-data/social-science-journals-that-have-a-research-data-policy/
10 https://docs.google.com/spreadsheets/d/1IWE-HnMJugV9JRG22f9tOg0BFh3fdnUUAD65JVOHUC0/edit?pli=1#gid=160911802
11 https://www.aeaweb.org/aer/data.php
12 http://www.colectica.com/
13 http://isps.yale.edu/
14 http://www.poverty-action.org/
15 http://www.ddialliance.org/
16 http://www.ddialliance.org/alliance/working-groups#Governance
17 http://isps.yale.edu/research/data
18 http://thedata.harvard.edu/dvn/dv/socialsciencercts
19 http://dhsprogram.com/
20 https://www2.icsu-wds.org/files/ostp-public-access-memo-2013.pdf
21 http://data.research.cornell.edu/content/readme
22 http://www.dcc.ac.uk/digital-curation/what-digital-curation
23 To be clear: performing data curation doesn't ensure that studies may be deeply examined and re-analyzed. If only a subset of the data and code is shared, for example, there may easily be limits to how thoroughly others can examine the original researcher's method of arriving at results. Likewise, re-use may be very limited if only a small subset of originally collected data is shared, as is often the case when researchers share their data in accordance with journal requirements. However, we argue that data curation is a very helpful, if not sufficient, condition for reproducibility and re-use.
24 https://www.icpsr.umich.edu/icpsrweb/content/datamanagement/lifecycle/ingest/enhance.html
25 http://www.data-archive.ac.uk/media/54770/ukda081-ds-quantitativedataprocessingprocedures.pdf
26 http://or2013.net/content/repository-data-re-user-hand-curating-replication/index.html

27 Other ISPS and IPA requirements included a workflow management dashboard, version tracking, integrating metadata production with data and code review and cleaning, creating preservation metadata, secure upload, storage and access, persistent identifier assignment, easy transition to public dissemination of content, and preference for open source solutions.
28 Separate licenses may be needed for some software.
29 Software exists that helps with curation of other digital objects, e.g., BitCurator (http://www.bitcurator.net/bitcurator-access/), LadyBird (http://ladybird.library.yale.edu/).
30 http://www.icpsr.umich.edu/icpsrweb/content/deposit/guide/chapter3quant.html#labels
31 http://www.data-archive.ac.uk/media/54770/ukda081-ds-quantitativedataprocessingprocedures.pdf
32 https://www.icpsr.umich.edu/icpsrweb/content/datamanagement/confidentiality/
33 http://www.icpsr.umich.edu/icpsrweb/content/deposit/guide/chapter5.html
34 http://ands.org.au/guides/sensitivedata.html
35 http://sourceforge.net/projects/anonimatron/
36 https://www.icpsr.umich.edu//icpsrweb/DSDR/tools/anonymize.jsp
37 http://www.data-archive.ac.uk/create-manage/format/formats