IDCC14 | *Practice Paper (PRE-PRINT)*

Committing to Data Quality Review

Limor Peer
Yale University

Ann Green
Digital Lifecycle
Research & Consulting

Elizabeth Stephenson
UCLA

Abstract

Amid the pressure and enthusiasm for researchers to share data, a rapidly growing number of tools and services have emerged. What do we know about the quality of these data? Why does quality matter? And who should be responsible for their quality? We believe an essential measure of data quality is the ability to engage in informed reuse which requires that data are independently understandable (CCSDS, 2012). In practice, this means that data must undergo quality review, a process whereby data and associated files are assessed, and required actions are taken, to ensure files are independently understandable for informed reuse. This paper explains what we mean by data quality review, what measures can be applied to it, and how it is practiced in three domain specific archives. We explore a selection of other data repositories in the research data ecosystem, as well as the roles of researchers, academic libraries, and scholarly journals in regard to their application of data quality measures in practice. We end with thoughts about the need to commit to data quality and who might be able to take on those tasks.

PRE-PRINT COPY

Submitted 14 January 2014

Correspondence should be addressed to Limor Peer, The Institution for Social and Policy Studies, Yale University, 77 Prospect Street, P.O.Box 208209, New Haven, CT 06520-8209. Email: limor.peer@yale.edu

The 9th International Digital Curation Conference will take place on 24–27 February 2014 in San Francisco. Please ensure you use the guidance in this template to produce your paper. Please submit your paper in one of the following formats: Microsoft Word (.doc, .docx), Open Document Format (.odt) or Rich Text (.rtf). http://www.dcc.ac.uk/events/idcc14/submissions

Introduction

We are seeing a growing number of tools and services that allow researchers to share their data, their code, their research design, and their analyses, and that's a good thing. Amid this growth and enthusiasm we think it is imperative to ask: What do we know about the quality of these research products? Why does quality matter? And who should be responsible for their quality?

Judgments about the quality of data are often tied to specific goals such as authenticity, verity, openness, transparency, and trust (Altman, 2012; Bruce and Hillman, 2013). Data quality might also consist of a combination of goals. The categories of data quality as defined by Wang and Strong (1996) are often in competition with each other or prioritized differently by stakeholders. As Kevin Ashley (Ashley, 2013) recently observed, some may prize the completeness of the data while others their accessibility. He urges that curation practices "be explicit about quality metrics and curation processes in domain-independent ways." For the purpose of our discussion, we define data quality as a set of measures that determine if data are ***independently understandable for informed reuse***. We argue that this perspective not only complements many of the goals referenced above, it also provides a roadmap for implementing specific quality measures and practices. We then urge the scientific community to subscribe to this vision of data quality by committing to data quality review. The paper explains what we mean by data quality, what measures can be applied to it, and how they are practiced in three domain specific archives.[1] Next, we explore a selection of other data repositories in the research data ecosystem and ask whether there are gaps in the application of quality measures in practice and how they might be addressed. We end with thoughts about the need to commit to data quality and the review of data quality, and who might be able to take on those tasks.

Data Quality: Independently Understandable Data for Informed Reuse

We distinguish between the quality of the research and the quality of research products, including data, metadata, and code. Questions about whether the research questions are important, whether the data are valid (i.e., measure what they are intended to measure) or appropriate, or whether the analysis is correct are critically important questions that are best left to the scholarly community. Our concern here is with the products of the research that are made publicly available for reuse by being placed in archives or repositories. Although our perspective is on social science data, we believe that our recommendations and discussions in this paper could apply across scientific domains.

Data reuse means that the original researchers, or other researchers, may use the data at a future time without predefining what those specific uses might be. Motivations for reuse can be varied and include data verification, new analysis, re-analysis, meta-analysis, and reproducing original analysis and results. In order to enable reuse, data

[1] For the purposes of this discussion, we differentiate between domain specific data archives (e.g., ICPSR, ISPS at Yale, UCLA SSDA, Roper Center for Public Opinion, etc.), institutional repositories (usually based in academic libraries), general data repositories (e.g., Dryad,figshare, and Zenodo), and research collaboration platforms (e.g., GitHub and Open Science Framework).

need to be processed, shared, and preserved in a way that ensures that they are "independently understandable to (and usable by) the Designated Community," and that there is enough information to be understood "without needing the assistance of the experts who produced the information" (CCSDS, 2012). The concept of "informed use" has also made its way into recent efforts to establish common citation principles; among the "first principles" for data citation is the following: "Citations should facilitate access both to the data themselves and to such associated metadata and documentation as are necessary for both humans and machines to make informed use of the referenced data." (CODATA, 2013)

One type of reuse – reproducing original analysis and results – sets an even higher bar for quality. When viewed through the lens of "really reproducible research," data as well as code need to be made available to allow regeneration of the published results (Peer, 2013). According to Peng (2011), such reproducibility "fill[s] the gap in the scientific evidence-generating process between full replication of a study and no replication." All of these materials need to be "independently understandable." In Gary King's discussion of the "replication standard," requirements for replication include the provision of, "sufficient information… with which to understand, evaluate, and build upon a prior work if a third party could replicate the results without any additional information from the author." (King, 1995)

Insisting that data be independently understandable is intended to speak to the credibility crisis in science as described by Jean-Claude Guédon in a keynote address at Open Repositories 2013. Guédon talked about the need for transparency as a self-correcting mechanism that can root out subpar or even fraudulent practices. Victoria Stodden, speaking at the same conference, talked about the central role of algorithms and code in the reproducibility and credibility of science (Stodden, 2013b). The goal is to reduce the risk of having a less-than-perfect scientific record – for example, insufficient information about variables, values, coding, scales and weighting or lack of transparency in descriptions of methodology, sampling, and instrumentation – which makes it difficult to reuse data and to validate results, and hampers the transfer of knowledge and the progress of science. As Jones et al (2006), commenting on the field of evolutionary biology, put it, "It is false economy, and poor scientific practice, not to ensure that the data are present and useful to all users in the future." And, a recent guide lists the rule: "conduct science with a particular level of reuse in mind," as one of ten "steps scientists can take to ensure that their data and associated analyses continue to be of value and to be recognized." (Goodman et al., in press)

Another reason we focus upon independently understandable data is the proliferation of research products of unknown quality. Because there are more ways to share data, and because the scholarly landscape supports and encourages that, there is a proliferation of data files on many different types of systems that do not meet the criterion of quality as we define it in this paper. For example, Ethan White et al (2013) argue that, despite improvements in data sharing, "much of the shared data in ecology and evolutionary biology are not easily reused because they do not follow best practices in terms of data structure, metadata, and licensing."

At issue is not the dearth of guidelines and best practices for preparing data. These are readily available from domain specific data archives in the social sciences, e.g., ICPSR[2] and UKDA,[3] tDAR in archaeology,[4] and DataOne[5] in the ecological and

[2] http://www.icpsr.umich.edu/files/ICPSR/access/dataprep.pdf
[3] http://www.data-archive.ac.uk/create-manage
[4] https://dev.tdar.org/confluence/display/TDAR/Documentation+Home
[5] http://www.dataone.org/best-practices

environmental sciences. The Digital Curation Center offers links to metadata production standards specific to subject disciplines.[6] The article by Ethan White et al (2013) provides useful suggestions directed specifically at research in progress, "making your data understandable, easy to analyze, and readily available to the wider community of scientists." Similarly, Allan Dafoe (2013) offers recommendations for producing "good replication files for researchers engaged in statistical analysis," including preparing all data and analysis in code, following best practices for coding, fully describing variables, and documenting every empirical claim. Goodman et al's list of "10 Simple Rules for the Care and Feeding of Scientific Data" includes adopting format and metadata standards and keeping careful track of versions of data and code (2014). And, a replication oriented set of recommendations from Sandve et al (2013) itemizes "Ten Simple Rules for Reproducible Research," including keeping track of how every result was produced, to record all results in standard formats, and to provide public access to scripts, runs and results.

These guides and best practices are an expression of significant cultural changes in the research community, which is coming to terms with a more open science.[7] They have enormous potential to change how data are prepared for publication and sharing, if they are implemented uniformly. This paper explores ways to validate that best practices have been implemented and that files can truly be considered ready for independently understandable informed reuse.

Data Quality Review

A data quality review is a process whereby data and associated files are assessed, and required actions are taken, to ensure files are independently understandable for informed reuse. This is an active process, involving a review of the files, the documentation, the data, and the code.[8] We strongly believe that data quality cannot be realized without a data quality review. Below we explain what this review entails, who is positioned to carry out such a review, and what it means to commit to such a review.

[6] http://www.dcc.ac.uk/resources/metadata-standards

[7] Others have explored various external incentives for high quality data, including (e.g., Borgman, 2012; Hedstrom et al, 2006).

[8] Future reuse of data typically requires that data are made available to others. Here, we assume that files are published to a system that enables access to the files, however, we do not address access and licensing issues.

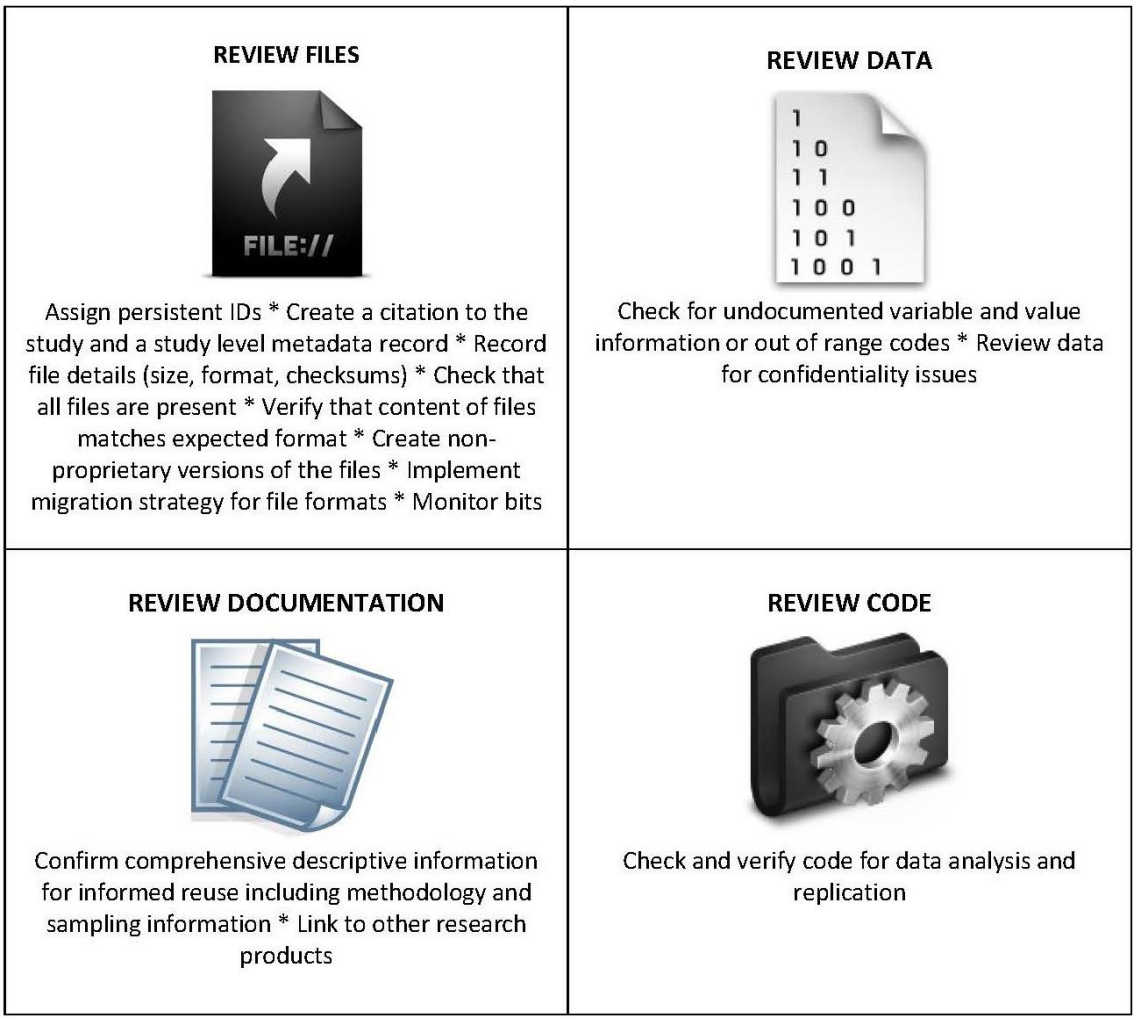

Figure 1. Data Quality Review Actions

Data quality requires that files are clearly identified, and that they are functional and accessible for the long term. A review of these basic aspects of data quality entails generating persistent identification (file level and study level where appropriate), creating a citation, recording file sizes and formats, creating checksums, checking that all necessary files are present, creating a study-level metadata record including file information (where appropriate), and creating non-proprietary file formats for dissemination and preservation. This also includes preservation-oriented steps such as, implementing a migration strategy for file formats, and ongoing bit monitoring.

Data quality also requires that documentation supporting the use of data is comprehensive enough to enable others to explore the resource fully, and detailed enough to allow someone who has not been involved in the data creation process to understand how the data were collected.[9] Files making up a data set (data, code, metadata, contextual materials, etc.) need to be carefully reviewed to establish that there is comprehensive descriptive information about the files and about methods and sampling, and to take corrective actions where this information is missing, including creating documentation compliant with community standards, e.g., the DDI XML

[9] Digital Preservation Coalition Handbook, Introduction – Definitions and Concepts. http://www.dpconline.org/advice/preservationhandbook/introduction/definitions-and-concepts#docu1

specification.[10] All other known related research products (e.g., publications, registries, grants) also need to be explicitly linked to the data.

A data quality review also involves some processing – examining and enhancing – of the actual data. These actions require performing various checks on the data, which can be both automated and manual procedures. The United Kingdom Data Archive (UKDA) provides a comprehensive list: "double-checking coding of observations or responses and out-of-range values, checking data completeness, adding variable and value labels where appropriate, verifying random samples of the digital data against the original data, double entry of data, statistical analyses such as frequencies, means, ranges or clustering to detect errors and anomalous values, correcting errors made during transcription."[11] In addition, data need to be reviewed for risk of disclosure of research subjects' identities, of sensitive data, and of private information (Lyle, Alter & Green, 2014) and potentially altered to address confidentiality or other concerns.

Similar to data files, code files should also be subject to examination and potential enhancement to provide transparency and enable future informed reuse. A data quality review requires that code is executed and checked, that 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. As Victoria Stodden, a long-time advocate for code disclosure, put it, "A research process that uses computational tools and digital data introduces new potential sources of error: Were the methods described in the paper transcribed correctly into computer code? What were the parameters settings and input data files? How were the raw data filtered and prepared for analysis? Are the figures and tables produced by the code the same as reported in the published article? The list goes on." (Stodden, 2013a). Roger Peng, also an advocate of reproducible research, argues that articles that have passed the reproducibility review "convey the idea that a knowledgeable individual has reviewed the code and data and was capable of producing the results claimed by the author. In cases in which questionable results are obtained, reproducibility is critical to tracking down the "bugs" of computational science." (Peng, 2011)

These data review activities are essential for ensuring and enhancing data quality over time. To more clearly illustrate what is involved we briefly report on data quality review practices of three domain specific data archives and three general data repositories (See also the Appendix titled Quality Measures in Practice: A Comparison, at the end of this paper).

Data quality review in domain specific data archives

In this section we describe key data quality review practices in three disciplinary data archives: ICPSR and two small, domain specific data archives, the Social Science Data Archive at UCLA and the ISPS Data Archive at Yale University. These are only three of numerous social science data archives, and we focus on them because we know them best.

Data quality review is embedded in data curation practices. The goal of curation is to maintain, preserve and add value to digital research data throughout its lifecycle, which reduces threat to the long-term research value of the data, minimizes the risk of its obsolescence, and enables sharing and further research.[12] 'Gold standard' curation

[10] http://www.ddialliance.org
[11] http://data-archive.ac.uk/create-manage/format/quality
[12] http://www.dcc.ac.uk/digital-curation/what-digital-curation

processes are carried out by data archives around the globe.[13] Their approach to data processing involves organizing, describing, cleaning, enhancing, and preserving data for public use and includes format conversions, reviewing the data for confidentiality issues, creating documentation and metadata records, and assigning digital object identifiers. In some cases, data, documentation, and code are not only reviewed but they are modified to improve their usability. Some of these actions create new versions of the data and metadata, and some create completely new files. This requires agreements with the data producers about such work, clear policies about what changes can be made, communications with researchers about enhancements to the data, and records kept and made available regarding all changes made to the original files. Most of these curation steps take place prior to sharing of the files.

ICPSR (The Inter-university Consortium for Political and Social Research, at the University of Michigan, Ann Arbor)[14]

ICPSR is a well-known, member-based repository and archive of data used in social science quantitative research. "As a global leader in data stewardship and an international consortium of more than 700 academic institutions and research organizations, ICPSR maintains a data archive of more than 500,000 files of research in the social sciences. ICPSR stores, curates, and provides access to scientific data so others can reuse the data and validate research findings. Through the curation process, data are organized, described, cleaned, enhanced, and preserved for public use… to make the works accessible to the public now and in the future."

Data: ICPSR accepts data from the social and behavioral sciences, including surveys, public opinion polls, census enumerations, and other files produced by government agencies, research organizations, and individual scholars. Data are deposited using an online form[15] that collects metadata such as study description and methodology, including study design, sampling, weighting, geographic details, as well as citations to publications that resulted from analyses of the data. Multiple data formats are generated from deposited files for dissemination and preservation.

Data quality review: At ICPSR, once data are submitted in a submission information package, the data pass through a "pipeline" for processing and enhancement."[16] Steps include reviewing data and documentation for confidentiality issues and completeness, and assessing formats of study documentation and datasets. Depending on the outcome of the initial review, ICPSR staff may, in consultation with the data producer, recode variables to address confidentiality issues, check for undocumented or out of range codes, and standardize missing values. Study documentation is enhanced to ensure that question text, labels, and response categories and value labels are associated with variables. ICPSR documents all changes to files with syntax files and all correspondence with PIs and depositors. Once ICPSR completes processing work, the data collection goes through an internal quality review to insure the data collection is

[13] For example, Social Science data archives include the following examples: The Inter-university Consortium for Political and Social Research (http://icpsr.umich.edu), the Roper Center for Public Opinion Research (http://www.ropercenter.uconn.edu), the Odum Institute (http://www.irss.unc.edu/odum), the UK Data Archive (http://www.data-archive.ac.uk), the Council of Social Science Data Archives in Europe (CESSDA http://www.cessda.net), and the Australian National Data Service (http://www.ands.org.au).

[14] http://www.icpsr.umich.edu/

[15] http://www.icpsr.umich.edu/icpsrweb/deposit/

[16] For an overview of ICPSR's data processing see: http://www.icpsr.umich.edu/icpsrweb/content/datamanagement/lifecycle/ingest/enhance.html

complete and self-explanatory, as well as to insure no unintended changes were made during processing.

Metadata: ICPSR creates a full and complete metadata record for the data collection based on the DDI schema, and produces a DDI-compliant codebook. Any other documentation files are formatted as PDF files.

UCLA SSDA (The Social Science Data Archive, at the University of California, Los Angeles)[17]

The UCLA SSDA is a small domain-specific repository of surveys, polls, enumerations and administrative data used in social science quantitative research. SSDA serves the entire UCLA campus in providing access to publicly accessible data and, providing curation services and long term preservation of data collected by UCLA investigators.

Data: Data collected in survey research by UCLA faculty are initially described by the researcher in a detailed Deposit Agreement. Acceptance of materials is based on the Archive's ability to carry out all phases of the workflow considering the allocation of resources and fitness to collection policy. Some life-cycle curatorial processes (e.g., metadata creation compliant with DDI, distributed replication) are shared among a partnership of archives through the Data Preservation Alliance for Social Sciences (DataPASS).[18] Other processes such as media refreshing and file format migration are carried out by the UCLA SSDA.

Data quality review: Data deposited at the UCLA SSDA undergo many of the same operations as those listed above for ICPSR. We have developed a workflow to address data quality at each step, from initial appraisal, ingest, metadata production, access and preservation. The Archive employs several open source and licensed software tools to carry out these tasks, including statistical software packages, emulation software, and Colectica Designer, Colectica for Excel and Colectica Repository.[19] SSDA works with researchers to resolve inconsistencies in the data and any changes are made with researcher approval.

Metadata: In order to produce complete life-cycle level metadata, Colectica Designer permits us to import statistical package format files, create item level documentation, and export DDI compliant metadata and documentation. Colectica for Excel is useful as an intermediary step to document variables, values, labels and question text. We use Colectica Repository to enable reuse through an item-level search capability, and links to downloadable files

ISPS Data Archive (at the Institution for Social and Policy Studies, Yale University)[20]

The ISPS Data Archive is a small, specialized data repository, dedicated to supporting reproducible research (Peer & Green, 2012). It is meant to capture and preserve the intellectual output of a single unit within the university, to provide free and public access to research materials in line with open access principles, and to be used for reproducing research results through replication, i.e., by using author-provided data, code, codebooks, and other research materials.

Data: Data deposited in the ISPS Data Archive are produced by scholars affiliated with ISPS, with special focus on experimental design and methods. Field or other

[17] http://dataarchives.ss.ucla.edu
[18] http://www.data-pass.org
[19] http://www.colectica.com
[20] http://isps.yale.edu/research/data

experiments (i.e., survey, natural, lab) produce original, often "small," data of high value for researchers, educators, policy makers and students. Datasets frequently combine these data with survey or administrative data.

Data review: Researchers are asked to deposit all research output, including data, metadata, statistical code, codebooks, research materials, and description files, and all files are subject to review before publication. The ISPS Data Archive pipeline closely follows that of ICPSR, including checking data for confidentiality and completeness, and assessing and enhancing study documentation and dataset formats. In addition, the Archive has developed curatorial practices that include verification and replication of the original research results. The ISPS Data Archive pipeline relies on some software, such as Stata, R, and Stat/Transfer, but many steps are manual. ISPS works closely with researchers when changes to the data or code files are made.

Metadata: The specialized nature of experimental data requires high quality documentation and metadata to facilitate replication of, and provide meaning to, each study. The ISPS Data Archive adheres to prevailing metadata standards, including OAI-PMH, Dublin Core, and the Data Documentation Initiative. Study-level metadata are compiled from information provided by researchers (via a deposit agreement form) and from associated materials (e.g., published article). Study-level metadata is made available on the Archive website and depends on content management functionality for search. For variable-level metadata, ISPS uses Stat/Transfer to produce make available XML files based on DDI version 3.1 for datasets.

Data quality review in general data repositories

Next, we describe the practices of three data repositories and data sharing venues. We illustrate varying curation policies and actions, and how measures of the 'quality' of data are reflected in the goals and capabilities of these repositories.[21] Information for the comparison was taken from the websites for these examples; it may be that this information is in flux and there may be features under review for future implementation.

Also note that some general repositories have developed as data publishing services, and in many respects they do not share the curatorial mission of domain specific data archives, who are more closely involved in data preparation and review prior to publicly sharing data. The examples we use (e.g., Dryad, Dataverse, and Figshare) provide secure storage, persistent identifiers, useful guidelines, and they support varying degrees of file inspection. However depositors have to take on the responsibility for preparing data for sharing, with the data, documentation and code properly vetted prior to submission.

Dryad[22]

Dryad services have been set up to provide "long-term access to its contents at no cost to researchers, educators or students, irrespective of nationality or institutional affiliation. Data files associated with any published article in the sciences or medicine, as well as software scripts and other files important to the article" may be deposited in Dryad.

Data: Dryad partners with journal publishers to make available the data behind published articles. "Most data in the repository are associated with peer-reviewed

[21] General data repositories (and research collaboration platforms) play an important role in the research ecosystem, but evaluating the quality of the tool should not be equated with reviewing the quality of the data.

[22] http://datadryad.org

articles, although data associated with non-peer reviewed publications from reputable academic sources, such as dissertations, are also accepted."[23] Data are linked both to and from the corresponding publication and, where appropriate, to and from select specialized data repositories (e.g. GenBank).[24]

Data review: Dryad has a curatorial team that checks files for technical problems and "works to enforce quality control on existing content."[25] Curators check for copyright statements and licenses, and identifiable human subject data. The review improves the odds that the data will be re-usable. In terms of quality review while the Dryad curators may discover problems, they do not verify that the data deposited can be reused to replicate findings in publications. Instead, "[s]ubmitters are *advised* to follow community data standards for the content and format of data files. Submitters should aim to provide sufficient data and descriptive information such that another researcher would be able to evaluate and reproduce the findings described in the publication."[26] Dryad does not limit the types of files that are put into the repository, however a draft preservation policy describes levels of support that will be given to specific file types.[27] The information content of the original file is never intentionally modified or processed, but copies may be made in different file formats to facilitate preservation. "When a data file is submitted in a non-preferred format, a Dryad curator will convert it into the most appropriate preferred format. Both formats will be made available, labeled as original deposited file and transformed file for preservation."[28]

Metadata: Dryad has placed itself on the side of promoting good practice without actually requiring it, and relies on the scholarly community to press for completeness of documentation. Dryad implements an automatic description process and "minimizes the amount of typing/clicking required by submitters. Descriptions are automatically propagated from the article description to descriptions for individual data files."[29] Dryad encourages researchers to "include a ReadMe file that provides additional information to make sense of the files (e.g. instructions for use of software scripts, variable abbreviations, measurement units, and data codes)."[30]

Dataverse[31]

Dataverse describes itself as "a repository for research data that takes care of long term preservation and good archival practices… and [s]upports the sharing of open data and enables reproducible research." Our descriptions here pertain to the Harvard Dataverse Network[32] which is open to all researchers from all disciplines worldwide "to share, cite, reuse and archive research data."

Data: Researchers are advised to "deposit preferred or commonly used file formats in your discipline to ensure that others will be able to more easily replicate your research (and) to remove information from your datasets that must remain confidential."[33]

[23] http://wiki.datadryad.org/Content_Policy
[24] http://datadryad.org/pages/repository
[25] http://wiki.datadryad.org/Curation
[26] http://datadryad.org/pages/faq#
[27] Dryad's preservation policy in development: http://wiki.datadryad.org/Preservation_Policy. Curation tasks are described here: http://wiki.datadryad.org/wg/dryad/images/2/25/Curation_man_2013-12-12.pdf
[28] http://wiki.datadryad.org/Preservation_Policy
[29] http://wiki.datadryad.org/Submission_System
[30] http://datadryad.org/pages/faq#
[31] http://thedata.org
[32] http://thedata.harvard.edu
[33] http://thedata.org/book/replication-guidelines

Data review: It is expected that data review happens prior to submitting data to a Dataverse system. The analytical tools that are part of the Dataverse software can be used to view documentation, to confirm sample size, to run summary statistics for the purposes of checking for missing information, and to review metadata in system files. However, changes to the files need to be made outside the Dataverse and resubmitted as new versions. These need to be done by the data depositor or another designated researcher or curator. There is no disclosure analysis for sensitive data built into Dataverse, but there is a new project to integrate the DataTags.org web application with "Secure Dataverse." This initiative will provide "a standardized framework for sharing when data cannot be 100% open." (Crosas, 2013) Another new feature recently announced is an application for an integrated publishing workflow for open data. The application returns a data citation that can be inserted in publications. When the depositor wants to release the data to the public when the article is published, the metadata and data are released.

Metadata: Upon submitting data to Dataverse, a metadata record is created using a template containing fields selected by the depositor. Additional metadata is generated from statistical data files when they are submitted to Dataverse. The documentation files are compliant with the DDI 2 metadata schema. Review of the metadata record and documentation are not part of the Dataverse services, but an organization may choose to mediate deposits prior to release.

figshare[34]

figshare "allows researchers to publish all of their data in a citable, searchable and sharable manner. All data is persistently stored online under the most liberal Creative Commons licence, waiving copyright where possible. Users of the site maintain full control over the management of their research whilst benefiting from global access, version control and secure backups in the cloud."[35]

Data: figshare is a repository for "figures, datasets, media, papers, posters, presentations and filesets"[36] that offers unlimited storage space for data that is made publicly available on the site. Researchers can upload any file type to figshare, and attempts are made to display all file types in a web browser. The repository's publically available content is replicated in "CLOCKSS's geographically and geopolitically distributed network of redundant archive nodes."[37] figshare hosts the supplemental data for all seven PLOS journals.

Data review: Curatorial review is not part of the figshare model. In an article by Ned Stafford (2013), "Peter Murray-Rust, a chemist at the University of Cambridge, UK, says he likes the figshare model, allowing researchers to 'publish first and sort out the problems of formats quality, et cetera later." In other words, data review is not part of the figshare model and is left to the researcher to sort out.

Metadata: Upon submitting data to figshare, a metadata record is created, based upon Dublin Core. It is not reviewed.

Other stakeholders and data quality review

So far, our examination of data quality review has focused on data repositories and archives, as increasingly that is where data can be found. Three other important

[34] http://figshare.com
[35] http://www.digital-science.com/products/figshare
[36] http://figshare.com/about
[37] http://figshare.com/faqs

stakeholders have an interest in data quality and may hold the keys to data quality review: the researchers themselves, academic libraries, and scholarly journals.

Researchers

There is agreement that researchers are best positioned to do a lot to ensure the quality of the data they share for future reuse. For example, Donald J. Treiman (2009), in his book on data analysis in Stata, recommends archiving .do and .log files as a professional practice, using Stata codebook commands to document a data file, and including them with papers submitted for publishing. In a recent white paper, Ember and Hanisch (2013) lament that "[t]he broader community of data creators and users does not fully appreciate what it takes to preserve data for future use. This leads to assumptions that online storage using systems like Dropbox are adequate, ignoring the needs of curation, preservation, interoperability, and metadata." Research culture and habit seem to play a significant role (Sandve et al., 2013). Could research teams themselves take on more of the curatorial tasks similar to those done by data archives? Part of the solution might be to incorporate the right training, guidance and tools that support data quality into the habits of researchers as part of the efforts to make their data independently understandable over time.

The production of metadata is often cited as one of the most significant barriers to researchers sharing data. (Tenopir, C. et al, 2011) Edwards et al (2011) state that "just as with data themselves, creating, handling, and managing metadata products always exacts a cost in time, energy, and attention: metadata friction." Tools that make the creation and capture of metadata during the research process are essential. Many of these tools would have to be domain specific, but a suite of curatorial tools for capturing contextual and descriptive information needs to be developed.

Other data quality issues arise during the research process in addition to the challenges of metadata production and capture including, inconsistent labeling, coding errors, version confusion, and lack of awareness about problems with proprietary formats that might not be usable by others and are difficult to migrate or emulate over time. These aspects of data quality would also benefit from having the right tools in the research space. Data and code review, for example, could take place in collaborative research spaces, allowing researchers to do quality review work while actively engaged in the research process. In their "10 Simple Rules for the Care and Feeding of Scientific Data," Goodman et al (2014) recommend "publish[ing] workflow as context." Similarly, Tyler Walters (2014) discusses how researchers are using repositories to deposit "data generated in the first stages of a research project," and points out the work by the Sustainable Environment - Actionable Data (SEAD) initiative to "support coauthorship, shared tagging, microcitation, threaded discussions, and reviewing and commenting on data and research projects." [38] These collaborative research environments do not provide long-term preservation, and ideally could develop seamless integration with long-lived repositories. The advantage of considering virtual research environments as essential components for data quality is that many of the data quality review tasks are performed before files are deposited in repositories. Capturing the 'workflow' of the research team could go a long way in addressing the challenges of producing data that is independently usable, especially if guidelines are followed in regard to documentation, file formats, persistent identifiers, and the inclusion of methodology statements and documents explaining research methods or decisions about sampling.

[38] http://sead-data.net

Some examples of collaborative research platforms[39] that could capture data, metadata, and workflow needed for informed reuse include the following: GitHub,[40] a hosted Git repository popular among open source developers. As a collaborative workflow, it allows one to "take part in collaboration by forking projects, sending and pulling requests, and monitoring development."[41] Increasingly, it is used for other collaborative projects, including research.[42] The emerging Open Science Framework[43] is "part network of research materials, part version control system, and part collaboration software." It has many potential uses, and is so far mostly recognized as the site of a project on reproducibility in psychology research.[44] Zenodo[45] enables uploading files into its system directly from Dropbox. Finally, the newly announced figshare Projects[46] system provides collaborative spaces for private, secure file management based upon the figshare platform.

Data and code review could also take place after publication; once materials are released, the scientific community could review them. In the future, there may be incentives for researchers to do so, and post-publication crowd-sourced peer review may prove to be a successful model. Services supporting these efforts include RunMyCode[47] which enables easy dissemination of the necessary pieces required to submit the research to scrutiny by fellow scientists and ResearchCompendia,[48] a "web service allowing people to share the research software and data associated with a scientific publication."[49] Other tools such as Active Papers[50] which consists of "a file combining datasets and programs in a single package, which also contains a detailed history of which data was produced when, by running which code, and on which machine," may prove to contribute to data quality as well. These services and tools are important

[39] Note that these platforms are often referred to as "repositories," but they are in effect locations for managing changes to code, often collaboratively and openly. This is in distinction to the standard definition of repository as place to store and maintain things (http://www2.archivists.org/glossary/terms/r/repository).In fact, GitHub, for example, states that it does not provide archiving (https://help.github.com/articles/can-i-archive-a-repository)

[40] https://github.com

[41] http://www.crunchbase.com/company/github

[42] Zach Jones' describes GitHub's appeal: It offers a hosting environment for a "complete research project (that) is reproducible and transparent by default in a more comprehensive manner than a typical journal mandated replication archive. With a public Git repository the data, any manipulation code, and the associated models are available at any time that a change was "committed" to a file tracked in said Git repository. Keeping data, data manipulation code, model code, code for visualizations (tables and graphs), along with the manuscript in a Git repository on GitHub (or a similar site such as Bitbucket) thus subsumes and extends the advantages of journal maintained replication archives." (Jones, 2013) See also http://dat-data.com.

[43] https://osf.io

[44] "Scientists can use OSF for free to archive, share, find, register research materials and data. Journals, funders and scientific societies can use the OSF as back-end infrastructure for preregistration, data and materials archiving, and other administrative functions. Labs can organize, share, and archive study materials among team members. . Manage scientific workflow and increase transparency. OSF provides versioning of files, and projects can be copied by using forking."

[45] https://zenodo.org/

[46] Using figshare Projects, any academic, can at no cost store and share their research outputs privately, to "create collaborative spaces and control who else has access to these spaces. The activity stream allows researchers to keep track of who has viewed, commented, added notes to, or uploaded files to a collaborative space - adding a layer of transparency to collaborations."
http://figshare.com/blog/Upgrade_to/110

[47] http://www.runmycode.org

[48] http://researchcompendia.org

[49] http://researchcompendia.org/faq

[50] http://www.activepapers.org

facilitators for people who wish to have their data and code validated via a peer review process.

Academic Libraries

Academic institutions, and their libraries, increasingly desire to be involved in the life-cycle data management process. (Burnett, 2013) Some libraries have a history of including data files in their collection policies, and they support tools for data reuse and analysis, but most have only partnered with individual local researchers to provide guidance on data acquisitions or research data management. Exceptions are data libraries and data archives that have taken on stewardship of datasets, sometimes going back to the early 1970's. Institutional repositories are making progress in taking on the role of stewardship of data outputs by their affiliated researchers. "In libraries, we see a similar trend of assisting researchers with the creation of metadata and its ingest along with research data into a repository for preservation and access." (Riley, 2014) In biomedical libraries, "informationists" work with research teams to advise on "data management and curation, including metadata standards and preservation and preparation of data for sharing" (Federer, 2013) And finally, Kimpton and Minton Morris (2014) point out that "(t)hough some libraries are accepting deposits without intervention, most try to review data as it is added to make sure that it includes appropriate bibliographic information."

In most cases, however, institutional repository services are not able to take on the responsibility of reviewing data beyond basic bibliographic-level information, and they rely upon data being properly prepared for sharing prior to submission. In a 2013 survey of members, the Association of Research Libraries found that none of the responding institutions offered or carried out a clearly defined data quality review; instead libraries addressed "data management best practices (both online resources and workshops), helping researchers identify (and apply) appropriate metadata standards, research file organization and naming, data citation, data sharing and access, and data storage and backup." (Fearon, 2013).

It has been proposed that partnerships between data archives and institutional repositories be established so that the services and expertise of high end curatorial institutions can be shared by those who are not able to take on those tasks. (Green and Gutmann, 2007) The ARL survey previously mentioned also encouraged "collaboration within the library, across a campus, and sometimes across institutions… A common theme throughout the survey is the recognition that, in order to provide comprehensive RDM services and to support scientists throughout the data lifecycle, libraries need to collaborate, either formally or informally, with other units at the institution." (Fearon, 2013) For example, at the UCLA SSDA, a pilot project has been initiated to study the data quality review and curation processes workflow. One objective is to determine the possibility of developing a cooperative data curation infrastructure where some tasks are carried out by the archive and some by the institutional repository.

Scholarly Journals

It is too often the case that, "the amount of real data and data description in modern publications is almost never sufficient to repeat or even statistically verify a study being presented." (Goodman et al, forthcoming) Some scholarly journals have begun requiring that data are published with articles, and must meet a minimal set of requirements. Others take it further: it is the policy of the *American Economic Review* to "publish papers only if the data used in the analysis are clearly and precisely documented and are readily available to any researcher for purposes of replication."

(AEA, 2014) However, there is no quality review of the submissions. Allan Dafoe calls for better replication practices, particularly in political science. He places responsibility on authors to provide quality replication files, but also suggests that journals encourage high standards for replication files and that they conduct a "replication audit" which will "evaluate the replicability and robustness of a random subset of publications from the journal." (Dafoe, 2013) A document produced at a workshop held at the British Library on peer review recently recommended that, "publishers should provide simple and, where appropriate, discipline-specific data review (technical and scientific) checklists as basic guidance for reviewers." (Tedds et al, 2013)

An example of a journal that takes an active role in data review is the *Journal of Open Psychology Data*[51] (JOPD) that requires open peer review of data descriptions and data deposit. Its peer review process has been developed "to ensure that each paper correctly describes the data, and that it has been openly archived in accordance with best practices. The datasets themselves are not reviewed in terms of validity or importance." All JOPD data papers are peer reviewed according to the following criteria: "The methods section of the paper must provide sufficient detail that a reader can understand how the dataset was created, and would within reason be able to recreate it."[52]

The F1000 group identifies the "complexity of the relationship between the data/article peer review conducted by our journal and the varying levels of data curation conducted by different data repositories." (Lawrence, 2013) The group provides detailed guidelines for authors on what is expected of them to submit and ensures that everything is submitted and all checklists are completed. (F1000, 2014) It is not clear, however, if they themselves review the data to make sure it replicates results.

Scientific Data[53] is "a new open-access, online-only publication for descriptions of scientifically valuable datasets." The journal uses "a new type of content called the Data Descriptor, which combines traditional narrative content with curated, structured descriptions of research data" including detailed methods and technical analyses supporting data quality. Data are not contained in the journal, but can be accessed via references and links to both related journal articles and data files stored at data repositories (particularly in figshare or Dryad). Professional in-house curation of the data descriptions "helps to ensure standardized and uniformly discoverable content." *Scientific Data*'s aims align with the measures of quality to which we refer in this paper in these areas: "Offer transparency in experimental methodology, observation and collection of data… Ensure all interested parties -- scientists, policy makers, NGOs, companies, funders and the public -- can find, access, understand and reuse the data they need."[54]

[51] http://openpsychologydata.metajnl.com
[52] Detailed guidelines specify that, "(t)he deposited data must include a version that is in an open, non-proprietary format. The deposited data must have been labeled in such a way that a 3rd party can make sense of it (e.g. sensible column headers, descriptions in a readme text file). The deposited data must be actionable – i.e. if a specific script or software is needed to interpret it, this should also be archived and accessible. Participant data should be sufficiently anonymized and appropriate consent forms should be signed." (JOPD, 2014)
[53] http://www.nature.com/scientificdata
[54] http://www.nature.com/scientificdata/principles

Committing to Data Quality Review

Our review of various players in the research data ecosystem reveals that data quality review is not uniformly practiced. At this time, we see little evidence that researchers, academic libraries and scholarly journals are committed to fully reviewing data to ensure quality, and we explained the ways in which general data repositories fall short of full data quality review. The data quality review practices at data archives can go a long way toward ensuring that data are accurate, complete, well documented, and that they are delivered in a way that maximizes their use and reuse. We acknowledge that our perspective has been focused on the social sciences, and that conversations with other disciplines are productive. Exciting developments in biology, for example, include investments by organizations such as ENCODE[55] and EMBL-EBI[56] in data quality. In addition, we acknowledge variation in practice not only among disciplines, but among individual researchers. This paper does not intend to cover all of the mime types, varieties of research habits and workflows, or technologies and tools. Still, as evident in our research for this paper, some domain specific data archives currently offer the most comprehensive data quality review.

While data archives may currently be best positioned to carry out such review, we believe that reviewing the quality of the data is the responsibility of *any* entity that assumes responsibility over the data (Peer, 2011). We think that the stakeholders and caretakers of scientific materials such as data and code must share the responsibility of meeting the challenges of data quality review in order to ensure that data, documentation, and code are of the highest quality so as to be independently understandable for informed reuse, in the long term. The commitment to data quality review, however, has to involve the entire research community for two reasons.

First, domain specific data archives have limitations. The models described at ICPSR, ISPS, and SSDA at UCLA may not be applicable to other contexts, and indeed may not always be employed by other domain-specific archives. Quality review requires significant investment in staffing, relationships, and resources. The ISPS Data Archive and the UCLA SSDA staff have data management and archival skills, as well as domain and statistical expertise. Both invest in relationships with researchers and learn about their research interests and methods to facilitate communication and trust. Further, the reproducibility imperative at ISPS does not neatly apply to more generalized data, or to data that is not tied to publications. In other instances, a larger lab, greater volume of research, or simply more data will require greater resources and may prove the level of review we endorse challenging. All of this requires the right combination of domain, technical and interpersonal skills as well as time, which translates into higher costs. A recent white paper on "Sustaining Domain Repositories for Digital Data" has articulated the financial impact of the demands of data stewardship and "aims to start a conversation with funding agencies about how secure and sustainable funding can be provided for domain repositories." (ICPSR, 2013) With regard to ICPSR, quality review practices are done within the context of a large consortium of paying members, and the level of review ICPSR offers has come to be expected from the 'gold standard' data archive in the United States. Still, ICPSR's staff and financial resources are finite, it is specific in selection and scope, and access is sometimes limited only to members. New initiatives, such as ICPSR's service, openICPSR (Lyle, 2013a), which facilitates

[55] (ENCODE) Encyclopedia of DNA Elements: http://encodeproject.org/ENCODE/qualityMetrics.html
[56] (EMBL-EBI) European Molecular Biology Laboratory's European Bioinformatics Institute: http://www.ebi.ac.uk/services

data deposit into an open repository and provides a review by professional data curators who are experts in developing metadata for the social and behavioral sciences, might sidestep some of these limitations. This landscape is constantly changing; as Margaret Hedstrom (2013) points out, data archives and repositories still need to work out exactly what role they want to play in the data supply chain.

Second, in many situations it is imperative that quality review occurs outside repositories because data are being disseminated in a variety of ways. Obviously, if there are no curatorial services in place, the full burden of quality review falls to the researchers and whatever support they have available prior to publishing data, and they need to locate a trusted place to put and get data. Yet, even as Guédon and Stodden make a compelling argument that open repositories hold the key to the future credibility of the scientific enterprise, Christine Borgman reminds us that most repositories and archives follow the letter, not the spirit, of the law: They take steps to share data, but they do not review the data. "Who certifies the data? Gives it some sort of imprimatur?" she asks (Borgman, 2013). Even when review steps are taken – normalizing data to one format such as SPSS, for example – how can we be sure that there was no loss of precision (e.g., formats, missing values, labels)? As Stodden pointed out at Open Repositories 2013, it is not clear "who, if anyone, checks replication pre-publication." (Stodden, 2013b) She suggested that this activity is community-dependent, often done by students or other researchers continuing a project, and that community can adjust norms by rewarding high integrity, verifiable research.

If researchers are not familiar with the repository and archive options in their subject area, it can be difficult for them to determine what type of curatorial review of data, documentation, and code various repositories and archives really do. One way to locate repositories for sharing and storing research data by subject discipline is to search a digital repository register (e.g., OpenAIRE, Databib, and re3data.) But it is difficult to assess what curatorial practices each of the repositories offer. There is no question that these can be very useful tools, but we suggest that it would be helpful if they would include information about the level of curatorial review, if any, that has been given to data after submission. There have been efforts to develop criteria for ensuring a level of data quality as it relates to repository operations. For example, the registry re3data.com uses a quality standard icon to indicate that a repository is either "certified or supports a repository standard."[57] Certification of repositories commonly focuses upon the important aspects of a repository's implementation, sustainability, and technical adequacy.[58] However, we find that repository certification metrics do not include explicit information about how much, and what types, of data quality review are done by the archive or repository itself. The Data Seal of Approval[59] differs from the other certification methods in that it has clear requirements that are assigned specifically to the data producer. The data producer is required to deposit the data with sufficient information for others to assess the quality of the data and compliance with disciplinary and ethical norms; provide the data in formats recommended by the data repository, and

[57] http://www.re3data.org/faq

[58] For example, criteria have been developed by NESTOR (http://www.dcc.ac.uk/resources/repository-audit-and-assessment/nestor#sthash.l5I1RcPs.dpuf), the Digital Curation Center (http://www.dcc.ac.uk), and Digital Preservation Europe (http://www.digitalpreservationeurope.eu/announcements/drambora) to evaluate long-term digital repositories. The most extensive process for evaluating repository functions, TRAC (http://www.crl.edu/archiving-preservation/digital-archives/metrics-assessing-and-certifying-0), was developed by a task force under the auspices of OCLC's Research Libraries Group (RLG) and the National Archives and Records Administration (NARA); this has evolved into the Trustworthy Digital Repository Checklist (TDR), now the ISO16363 standard, largely based upon the TRAC metrics.

[59] http://datasealofapproval.org/en/information/about

provide the data together with the metadata requested by the data repository. There are no explicit requirements for the repository to complete a data review or undergo the curatorial actions we describe in this paper.

We strongly believe that, as more entities take on various roles in the review, curation, or dissemination of data – especially entities removed from the original data producers – that strong controls should be put in place to ensure that there is no potential for unintentional (and even intentional?) changes that can significantly alter the data. For example, cleaning files could unknowingly reduce decimal precision due to imprecise format specifications to revising codes. As Lyle (2013b) cautions, cleaning and enhancing files should always be weighed against potential distortion. As should be clear in this paper, review is intended to contribute to long term independently understandable informed use of the data, and in no way jeopardize any other aspects of data quality (e.g., accuracy, authenticity, verity, etc.). A serious conversation about ways to ensure "zero harm" to data and code needs to take place in the scientific community.

In spite of these challenges, we believe that stewardship of data requires this type of quality review because it leads to better science (Peer, 2013). Usable "[d]ata-rich research environments can promote new fields of study, improve understanding of complex systems such as the Earth's climate, and lead to new products such as pharmaceutical drugs." (Wallis, Rolando & Borgman, 2013) This endeavor requires more and better tools, as well as smart, effective partnership among the various stakeholders. "The social nature of science and the network of interested stakeholders in the future of access to scientific data," says Gold (2007), "make it essential to develop social and policy tools to support this future." As Jones et al (2006) observe, the key is "to find the balance of responsibility for documenting data between individual researchers and trained data stewards who have advanced expertise with appropriate metadata standards and technologies." And, the National Digital Stewardship Alliance recently urged the scientific community to, "work together to raise the profile of digital preservation and campaign for more resources and higher priority given to digital preservation, and to highlight the importance of digital curation and the real costs of ensuring long term access (NDSA, 2014).

Conclusion

Independently understandable informed reuse of data in the long term is in jeopardy: Data are being lost at an alarming rate (Gibney & Van Noorden, 2013, Vines et al., 2014). At the same time, more data than ever are being released publicly. Unfortunately, in both scenarios, there is still significant misunderstanding about what is necessary to archive data for long term usability. Digital preservation practices go beyond storing and managing the bits. (Owens 2012) We can think of a continuum of data curation that progresses from a basic level where data are accepted "as is" for the purpose of storage and discovery, to a higher level of curation which includes processing for preservation, improved usability, and compliance, to an even higher level of curation which also undertakes the verification of published results.

Data archives have traditionally taken on 'gold standards' of data processing as described above, but repositories vary widely in the curatorial processing they offer for incoming data, and in the preservation services they can provide over the long term. Researchers sometimes believe that assigning a persistent link, e.g., a DOI, and maintaining redundant backups will be enough to make data accessible and understandable for decades. Certainly repository systems offer more secure homes for

research data than researchers may have had in the past, but we suggest that threats remain. Among the pitfalls of this approach is the lack of quality review when data are submitted to digital repositories. Those researchers who follow the guidelines we describe have better odds that their data will be independently usable over time, but what if those guidelines have not been followed? Wouldn't it be better to catch and correct the problems with formats, metadata, missing data, mismatches between data and code, disclosure review, etc. when the data are submitted and reviewed by a research team, a repository, or an archive, rather than waiting for those problems to prevent long term understandability and use of the data? The lack of quality review as a curatorial practice can have severe consequences and can contribute to the loss of data over time.

A conversation about reviewing the data we put in repositories is a sign of maturity in the scholarly community, across all scientific domains, and recognition that simply sharing data is necessary, but not sufficient. We call on the community as a whole to commit to data review by practicing it and by demanding to know when it has been done. Our hope is that it becomes a cornerstone in standard approaches to data curation and will become common practice once appropriate tools and frameworks are in place.

References

[online guide] American Economics Association. (AEA) (2014). The American Economics Review: Data Availability Policy. Retrieved from: http://www.aeaweb.org/aer/data.php

[unpublished report] Altman, M. (2012). Mitigating Threats to Data Quality Throughout the Curation Lifecycle. Draft. Retrieved from https://docs.google.com/document/d/1LXtK9-P6F65tLZTy533y1Q62ePDYlGgFElWHw0-Ouk4/edit

[presentation] Ashley, K. (2013). Data Quality and Data Curation - a personal view. Presentation from OAI Geneva June 20, 2013. Retrieved from http://www.slideshare.net/kevinashley/oai8-ashleyplain

[presentation] Borgman, C. L (2013). ADS, Astronomy, and Scholarly Infrastructure. Retrieved from http://conf.adsabs.harvard.edu/ADSXX

[journal article] Borgman, C. L. (2012). The conundrum of sharing research data. *Journal of the American Society for Information Science and Technology*. 63(6), p. 1059–1078. Retrieved from http://works.bepress.com/borgman/258

[blog post] Bruce, T. R. & Hillmann, D. (2013). Metadata quality in a linked data context. Voxpopulii. Legal Information Institute. Retrieved from http://blog.law.cornell.edu/voxpop/2013/01/24/metadata-quality-in-a-linked-data-context

[blog post] Burnett, P. (2013). What is the role of a librarian in Research Data Management? Practicing Development. Retrieved from http://blog.inasp.info/research-data-management-role-librarians

[journal article] CODATA/ITSCI Task Force on Data Citation. (2013). Out of cite, out of mind: The Current State of Practice, Policy and Technology for Data Citation. *Data Science Journal 12*, 1-75. http://dx.doi.org/10.2481/dsj.OSOM13-043

[presentation] Crosas, M. (2013). Promoting Transparency in Social Science Research. Panel on Open Data and Protection of Research Subjects at the workshop on Promoting Transparency in Social Science Research: Innovation, Collaboration, Impact at BITTS, UC Berkeley. Retrieved from http://thedata.org/files/thedata_new2/files/opendata-datatags-mercecrosas.pdf

[report] Consultative Committee for Space Data Systems (CCSDS). (2012). *Reference model for an Open Archival Information System (OAIS)* (Magenta Book CCSDS 650.0-B-1). p. 3-1. Retrieved from http://public.ccsds.org/publications/archive/650x0m2.pdf

[working paper] Dafoe, A. (2013). Science Deserves Better: The Imperative to Share Complete Replication Files. Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2318223

[online guide] DataOne. (2014). Best Practices. Retrieved from http://www.dataone.org/all-best-practices

[online guide] The Digital Archaeological Record. (tDAR). (2013). Retrieved from https://dev.tdar.org/confluence/display/TDAR/Planning+Your+Data+Contribution

[journal article] Edwards, P. N., Mayernik, M. S., Batcheller, A. L., Bowker, G. C., & Borgman, C. L.. (2011). Science friction: Data, metadata, and collaboration. *Social Studies of Science* 41, p. 667-690. doi:10.1177/0306312711413314. Retrieved from http://pne.people.si.umich.edu/PDF/EdwardsEtAl2011ScienceFriction.pdf

[white paper] Ember, C. & Hanisch, R. (2013). Domain Repositories for Digital Data: A White Paper. Retrieved from http://datacommunity.icpsr.umich.edu/sites/default/files/WhitePaper_ICPSR_SDRDD_121113.pdf

[online guide] F1000 Research (2014). F1000Research Author Guidelines. Retrieved from http://f1000research.com/author-guidelines

[report] Fearon, D., Gunla, B., Pralle, B. I., Lake, S. & Sallans, A. (2013). Research Data Management Services. ARL Spec Kit 334. Washington, D.C. p.14. Retrieved from http://publications.arl.org/Research-Data-Management-Services-SPEC-Kit-334

[journal article] Federer, L. (2013). The librarian as informationist: A case study. *Journal of the Medical Library Association* 101 (4), p. 298-302. doi: 10.3163/1536-5050.101.4.011. Retrieved from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3794685

[journal article] Gibney, E., & Van Noorden, R. (2013). Scientists losing data at a rapid rate. *Nature*. doi:10.1038/nature.2013.14416 Retrieved from http://www.nature.com/news/scientists-losing-data-at-a-rapid-rate-1.14416

[online journal article] Gold, A. (2007). Cyberinfrastructure, Data, and Libraries, Part 1: A Cyberinfrastructure Primer for Librarians. *D-Lib Magazine 13* (9-10). doi:10.1045/september20september-gold-pt1 Retrieved from http://www.dlib.org/dlib/september07/gold/09gold-pt1.html

[journal article] Goodman, A., Pepe, A., Blocker, A.W., Borgman, C.L., Cranmer, K., Crosas, M., Di Stefano, R., Gil, Y., Groth, P., Hedstrom, M., Hogg, D.W., Kashyap, V., Mahabal, A., Siemiginowska, A., and Slavkovic, A. (2014). 10 Simple Rules for the Care and Feeding of Scientific Data. *PLOS Computational Biology*. Preprint retrieved from: http://arxiv.org/abs/1401.2134

[journal article] Green, A. & Gutmann, M.. (2007). Building Partnerships Among Social Science Researchers, Institution-based Repositories, and Domain Specific Data Archives. *OCLC Systems and Services: International Digital Library Perspectives 23*, p. 35-53. Retrieved from http://hdl.handle.net/2027.42/41214

[presentation] Hedstrom, M. (2013). Certification and Trust in the Data-Supply Chain. Keynote address at the Data Seal of Approval Conference, Ann Arbor, Michigan, Retrieved from http://datasealofapproval.org/en/news-and-events/news/2013/10/15/dsa-conference-2013-great-succes/

[presentation] Hedstrom, M., Niu, J., and Marz, K. (2006). Producing Archive Ready Data Sets. Presented at the 2006 Annual Meeting of the International Association for Social Science Technology and Information. Retrieved from http://www.iassistdata.org/downloads/2006/f2_hedstrometal.pdf

[white paper] ICPSR (2013). Sustaining Domain Repositories for Digital Data: A White Paper. http://www.icpsr.umich.edu/files/ICPSR/pdf/DomainRepositoriesCTA16Sep2013.pdf

[online guide] Inter-university Consortium for Political and Social Research (ICPSR). (2012). Guide to Social Science Data Preparation and Archiving: Best Practice Throughout the Data Life Cycle (5th ed.). Retrieved from http://www.icpsr.umich.edu/icpsrweb/content/deposit/guide

[blog post] Jones, Z. (2013). Git/GitHub, Transparency, and Legitimacy in Quantitative Research. The Political Methodologist. Retrieved from http://thepoliticalmethodologist.com/2013/11/18/gitgithub-transparency-and-legitimacy-in-quantitative-research

[journal article] Jones, M. B., Schildhauer, M. P., Reichman, O. J. and Bowers, S. (2006). The New Bioinformatics: Integrating Ecological Data from the Gene to the Biosphere *Annual Review of Ecology, Evolution, and Systematics* 37. p. 519-544. doi:10.1146/annurev.ecolsys.37.091305.110031. Retrieved from http://www.annualreviews.org/doi/pdf/10.1146/annurev.ecolsys.37.091305.110031

[online guide] Journal of Open Psychology Data (2014). Peer review process. Retrieved from http://openpsychologydata.metajnl.com/about/editorialPolicies#peerReviewProcess

[book chapter] Kimpton, M. & C. Minton Morris. (2014). Managing and Archiving Research Data: Local Repository and Cloud-Based Practices. In: Ray, J. M., ed. *Research Data Management: Practical Strategies for Information Professionals.* p. 225.

[journal article] King, G. (1995). Replication, Replication. *PS: Political Science and Politics 28*. p.443–499. Retrieved from http://gking.harvard.edu/files/gking/files/replication.pdf

[blog post] Lawrence, R. (2013). Data review: Making it happen! F1000Research blog. Retrieved from http://blog.f1000research.com/2013/03/25/data-review-making-it-happen

[presentation] Lyle, J. (2013a). Public Data Access Perspectives (and a Product) from a Membership-based Organization. Presented at the 2013 Biennial ICPSR Meeting. Retrieved from http://www.icpsr.umich.edu/files/membership/or/ormeet/PublicDataAccess2013.pptx

[presentation] Lyle, J. (2013b). Domain Repositories and Institutional Repositories Partnering to Curate: Opportunities and Examples. Presented at the 2013 Annual Research Data and Preservation Meeting (RDAP). Retrieved from http://www.slideshare.net/asist_org/2-rdap13-lyle

[book chapter] Lyle, J., Alter, G. & Green, A. (2014). Partnering to Curate and Archive Social Science Data. In: Ray, J. M., ed. *Research Data Management: Practical Strategies for Information Professionals.* P. 203-222.

[report] National Digital Stewardship Alliance (2014). National Agenda for Digital Stewardship 2014. Retrieved from http://www.digitalpreservation.gov/ndsa/documents/2014NationalAgenda.pdf

[blog post] Owens, T. (2012). NDSA Levels of Digital Preservation: Release Candidate One. *The Signal: Digital Preservation*. US Library of Congress. Retrieved from http://blogs.loc.gov/digitalpreservation/2012/11/ndsa-levels-of-digital-preservation-release-candidate-one

[journal article] Peng, R. D. (2011). Reproducible research in computational science. *Science*. 334(6060), p. 1226-1227. Retrieved from: http://europepmc.org/articles/PMC3383002/reload=0;jsessionid=7dJktUNN5xsRbVeGxWc0.0

[blog post] Peer, L. (2013). The Role of Data Repositories in Reproducible Research. Institution for Social and Policy Studies, Yale University blog. Retrieved from http://isps.yale.edu/news/blog/2013/07/the-role-of-data-repositories-in-reproducible-research#.UtHmtv2dkds

[working paper] Peer, L. (2011). Building an Open Data Repository: Lessons and Challenges. Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1931048

[journal article] Peer, L. & Green, A. (2012). Building an Open Data Repository for a Specialized Research Community: Process, Challenges and Lessons. *International Journal of Digital Curation* 7(1). P. 151-162. doi:10.2218/ijdc.v7i1.222 Retrieved from http://www.ijdc.net/index.php/ijdc/article/download/212/281

[book chapter] Ray, J. M. (2014). Introduction to *Research Data Management*. In: Ray, J. M., ed. *Research Data Management: Practical Strategies for Information Professionals.* P. 2.

[book chapter] Riley, J. (2014). Metadata Services. In: Ray, J. M., ed. *Research Data Management: Practical Strategies for Information Professionals.* P. 151.

[journal article] Sandve G. K., Nekrutenko, A., Taylor, J., Hovig, E. (2013). Ten Simple Rules for Reproducible Computational Research. *PLoS Computational Biology* 9(10): e1003285. doi:10.1371/journal.pcbi.1003285 Retrieved from http://www.ploscompbiol.org/article/info%3Adoi%2F10.1371%2Fjournal.pcbi.1003285

[online guide] Scientific Data (2014). Guide to referees. Retrieved from http://www.nature.com/scientificdata/guide-to-referees/#review-process.

[online journal article] Stafford, N. (2013). Figshare to offer institutional data platform. *Chemistry World*. Retreived from http://www.rsc.org/chemistryworld/2013/09/figshare-offer-institutional-university-data-platform.

[blog post] Stodden, V. (2013a). Changes in the Research Process Must Come From the Scientific Community, not Federal Regulation. Retrieved from http://blog.stodden.net/2013/09/24/changes-in-the-research-process-must-come-from-the-scientific-community-not-federal-regulation/#more-231

[presentation] Stodden, V. (2013b). Re-use and Reproducibility: Opportunities and Challenges. Open Repositories 2013 Keynote. Retrieved from http://or2013.net/sites/or2013.net/files/OR2013-July92013-STODDEN.pdf

[presentation] Tedds, J., Lawrence, R., Stoneham, G. C., Newbold, E., Kotarski, R., Leigh, R., Balzter, H., Wilson, B., Callaghan, S., Murphy, F., Mayernik, M., Kunze, J., Whyte, A., Hodson, S. (2013). Research Data Peer Review & the PREPARDE project. Retrieved from http://proj.badc.rl.ac.uk/preparde/attachment/blog/FinalMeeting/Tedds-PREPARDE-Data-Peer-Review-23Aug2013.pdf (slide 12)

[online journal article] Tenopir, C. A., Allard, S., Douglass, K., Aydinoglu, A.U., Wu, L., Read, E., Manoff, M., Frame, M. (2011). Data sharing by scientists: Practices and perceptions. *PLoS ONE* 6(6). doi:10.1371/journal.pone.0021101 Retrieved from http://www.plosone.org/article/info:doi/10.1371/journal.pone.0021101

[book] Treiman, D. J. (2009). *Quantitative Data Analysis: Doing Social Research to Test Ideas*. New York: Jossey-Bass/John Wiley & Sons. P. 68-69, 404.

[online guide] UK Data Archive (UKDA). (2014). Create and Manage Data. Retrieved from http://www.data-archive.ac.uk/create-manage

[report] UK Data Archive (UKDA). (2011). Managing and Sharing Data: Best Practices for Researchers. Retrieved from http://www.data-archive.ac.uk/media/2894/managingsharing.pdf

[online journal article] Vines, T.H., Albert, A. Y. K., Andrew, R. L., Débarre, F., Bock, D. G., Franklin, M. T., Gilbert, K. J., Moore, J., Renaut, S. & Rennison, D. J. (2014). The Availability of Research Data Declines Rapidly with Article Age. *Current Biology 24(1).* p. 94-97. Retrieved from http://download.cell.com/current-biology/pdf/PIIS0960982213014000.pdf

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

[book chapter] Walters, T. (2014). Assimilating Digital Repositories in the Active Research Process. In: Ray, J. M., ed. *Research Data Management: Practical Strategies for Information Professionals.* p. 189 – 201.

[journal article] Wang, R. Y. & Strong, D. M. (1996). Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems.* 12(4). Retrieved from http://www.thespatiallab.org/resources/data%20quality(JMIS).pdf

[journal article] White, E P., Baldridge, E., Brym, ,Z. T., Locey, K. J., McGlinn, D. J., & Supp, S. R. (2013). Nine simple ways to make it easier to (re)use your data. *Ideas in Ecology and Evolution* 6(2). 1–10. doi:10.4033/iee.2013.6b.6.f Retrieved from http://library.queensu.ca/ojs/index.php/IEE/article/download/4608/4898

Acknowledgments

We thank Jared Lyle of ICPSR for comments on an earlier version of this paper. We also recognize the members of the International Association of Social Science Information Services and Technology (IASSIST), upon the 40th anniversary of the organization, for their dedication to data quality and data related services.

Appendix: Quality Measures in Practice: A Comparison

	ICPSR	ISPS	UCLA SSDA	Dataverse	Dryad	figshare
REVIEW FILES						
Create persistent	Y; DOI	Y; handles for	Y; via	Y; Handles	Y; DOI	Y; DOI

ID		files	Dataverse; handle for study; UNF v5 for data files	and DOI for dataset		
Record file sizes and formats	Y	Y	Y	Y	Y	Y
Create checksums	Y	N; Pending	Y	UNF v5 for tabular data files; MD5 for all other files	Y	Y
Check for completeness, confirm all files are present (data, and required documentation and code if available)	Y	Y	Y	N	Y	N
Create study-level metadata record including file information	Y	Y	Y	Y	Y	Y
Create citation	Y	Y	Y	Y	Y	Y
Create non-proprietary file formats for preservation	Y	Y: ASCII data file; PDF and DDI XML documentation files; R statistical code files	Y; ASCII data file; PDF and DDI XML documentation files	Y; Tabular data files (SPSS, Stata, and R), files converted to tab delimited files for use in Dataverse tools	Y; Limited to specific file formats	N; Pending
REVIEW DOCUMENTATION						
Confirm comprehensive descriptive information	Y	Y	Y	N	Completeness and correctness of the metadata (e.g. information about the associated publication, indexing keywords) are checked	N
Confirm methodology and sampling information	Y	Y	Y	N	N	N
Create documentation compliant with community standards, e.g., DDI XML	Y	Y	Y	Y; DDI XML, Dublin Core	N	N
REVIEW DATA						
Run frequencies and check for undocumented or out of range codes	Y	Y	Y	N; but tools are available	N	N

Standardize missing values; check for consistency and skip patterns	Y	N	Y; review is done, but changes made by depositor	N	N	N
Check and edit variable and value labels	Y	Y	Y	Tools available for viewing labels in data files	N	N
Check and add question wording (surveys)	Y	N	Y	Tools available for viewing question wording if present in data file	N	N
Review data for confidentiality issues; Recode variables to address confidentiality concerns	Y; in consultation with data depositor	Y; in consultation with data depositor	Y; review is done, but changes made by depositor	N; but DataTags.org to be offered	Y; but no changes to data are made	N
Generate multiple data formats for dissemination	Y	Y; ASCII and R	Y; ASCII data with Stata and SPSS set up files	Y data outputs in Text, R Data, S plus, or Stata	Y; some files are converted upon submission	N
REVIEW CODE						
Check and verify replication code	N	Y	N	N	N	N
PUBLISH & LINK						
Publish to access system	Y	Y	Y	Y	Y	Y
Link to other research products (e.g., publications, registries, grants)	Y; See Bibliography of Data-Related Literature	Y; links to related ISPS Publications and Projects; links to other repositories	Y; Data Citation Index	Y; links can be put into metadata record	Y; Data are linked to and from corresponding publication and, to and from select specialized data repositories.	Y; PLOS
PRESERVE						
Migration strategy for file formats	Y	N; Pending	Y	Convert tabular data files to preservation formats upon ingest	N; Pending	N; Pending
Monitor bits	Y	N; Pending	LOCKSS via DataPass consortium	LOCKSS	CLOCKSS	CLOCKSS

NOTES:
All changes to data, documentation, and code are reviewed and recorded during processing.
Preservation actions are more involved than the two discussed here.

SOURCES:
ICPSR Data Management and Curation, Data Enhancement:
http://www.icpsr.umich.edu/icpsrweb/content/datamanagement/lifecycle/ingest/enhance.html

ICPSR Data Preparation Guide, Deposit Data:
http://www.icpsr.umich.edu/icpsrweb/content/deposit/guide/index.html

Dataverse information is for the Harvard Dataverse Network which offers services that might not be available if Dataverse is installed locally.
Dataverse FAQ: http://thedata.org/book/faq-using-harvard-dataverse-network#q5
Dataverse Network Guides: http://thedata.harvard.edu/guides/
Dataverse Replication Guidelines: http://thedata.org/book/replication-guidelines

Dryad Curation Manual: http://wiki.datadryad.org/wg/dryad/images/2/25/Curation_man_2013-12-12.pdf
Dryad FAQ: http://datadryad.org/pages/faq
Dryad Preservation Policy (in development): http://wiki.datadryad.org/Preservation_Policy
Dryad Terms of Service: http://datadryad.org/themes/Mirage/docs/TermsOfService-Letter-2013.08.22.pdf

figshare FAQ: http://figshare.com/faqs