More Data, Less Process? The Applicability of MPLP to Research Data

by Sophia Lafferty Hess[1], Thu-Mai Christian[2]

Abstract

In their seminal piece, "More Product, Less Process: Revamping Traditional Archival Processing," Greene and Meissner (2005) ask archivists to reconsider the amount of processing devoted to collections and instead commit to the More Product, Less Process (MPLP) 'golden minimum.' However, the article does not specifically consider the application of the MPLP approach to digital data. Data repositories often apply standardized workflows and procedures when ingesting data to ensure that the data are discoverable, accessible, and usable over the long-term; however, such pipeline processes can be time consuming and costly. In this paper, we will apply the principles and concepts outlined in MPLP to the archiving of digital research data. MPLP provides a useful lens to discuss questions related to data quality, usability, preservation, and access: What is the 'golden minimum' for archiving digital data? What unique properties of data affect the ideal level of processing? What level of processing is necessary to serve our patrons most effectively? These queries will contribute to the discussion surrounding how data repositories can develop sustainable service models that support the increasing data management needs of the research community while also ensuring data remain discoverable and useable for the long-term..

Keywords

data curation, data reuse, data quality, data service

Introduction

While Meissner and Greene's (2005) seminal article, More Product, Less Process: Revamping Traditional Archiving Processing, was written with traditional archives in mind, the authors' appeal for a critical assessment and recalibration of archival processing is no less relevant to digital data archives. Soon after the article was published, the new MPLP doctrine became the cause célèbre for much of the archival community, which the authors took to task for its proclivities toward all-or-nothing archival processing that had exacerbated growing backlogs of unprocessed and therefore inaccessible materials. MPLP reinstated user access as the highest of archive priorities, which absolved archivists of the minutiae of item-level arrangement, description, and preservation.

> What constitutes data quality has much to do with users' needs and preferences for discovering, accessing, interpreting, and using data.

For data archives, however, user access is very much tied to the minutiae. The usability of data is inextricable from its specific context: the research question the data were intended to answer, the instruments used to collect the

data, the software programs executed to manipulate and analyze the data, and the methods employed for data collection and analysis (Borgman, 2015). Data are complex, and the archival processing—or what we equate to data curation—required to make them available and usable for researchers has been informed by uncompromising standards of quality. Achieving these quality standards requires data archives to complete a laundry list of skill-intensive, labor-intensive, and time-intensive data curation tasks including normalizing file formats, mitigating confidentiality risks, checking for and correcting data errors, generating and enhancing descriptive metadata, assembling contextual documents, recording checksums, defining undefined variable and value codes, reconciling discrepancies between datasets and codebooks, and so on (Peer, Green, & Stephenson, 2014). Not so different from the backlog situation in traditional archives, compromises in data curation processes are inevitable given the nature of tightly resourced environments in which many data archives operate.

With increasing demand from funding agencies, journal publishers, and research communities for public access to quality data, data archives (and institutional repositories that are becoming ad hoc data archives) must examine the economies of curating archival data collections to the highest degree of quality at scale while keeping user needs at the forefront of curation approaches. The application of MPLP to data curation raises several essential questions about what compromises are allowable, if not inevitable, that will enable data archives to remain solvent while continuing to serve the needs of the user community.

In this paper, we discuss the Odum Institute Data Archive's application of the principal concepts of MPLP to data curation as part of an exercise to assess the scalability of data archive services as demand for them increases. MPLP offers a methodology that enabled us to not only reconsider, but also reaffirm the parameters of the standard data curation processes we use to provide access to quality data.

Data Quality Standards

One of Meissner and Greene's main criticisms of archival processing that obliged them to formulate their MPLP approach is "...the persistent failure of archivists to agree in any broad way on the important components of records processing and the labor inputs necessary to achieve them" (p. 209). This criticism cannot be wholly directed at data archivists, who have achieved some consensus on the important components of data curation as demonstrated in documented best practices that have received wide acceptance among data archives (ICPSR, 2012; Digital Curation Centre, n.d.). These best practices prescribe specific data curation actions that support data quality standards.

What constitutes data quality has much to do with users' needs and preferences for discovering, accessing, interpreting, and using data. Based on results from a study of users' perceptions of data quality, Wang and Strong (1996) identified four dimensions of data quality: 1) intrinsic, referring to the accuracy and credibility of the data; 2) contextual, or the relevancy of the data to the user's goals; 3) representational, relating to the ability to interpret and use the data; and 4) accessibility, or the ability to obtain the data. A more recent study conducted by Faniel, Kriesberg, and Yakel (2015) to determine the factors that elicit social science researchers' satisfaction with data reuse found that users associate data quality with attributes that align with Wang and Strong's quality dimensions. They include completeness (contextual), accessibility (accessibility), ease of operation (representational), and credibility (intrinsic) of the data. These aspects of data quality are often summed up in the notion of data being 'independently understandable' to their intended users (CSSDS, 2012; King, 1995; Lee, 2010; Peer, Green, & Stephenson, 2014).

This is the quality standard to which research data are being held, particularly those subject to data management and sharing mandates that have grown in popularity among funders and journals (e.g., National Endowment of the Humanities, 2012; National Institutes of Health, 2003; National Science Foundation, 2010; Nature, 2015; PLOS, 2014; Science, n.d.). To some degree, this standard also acknowledges the recent scrutiny of published scientific studies, a concerning number of which were reported to have failed to meet the reproducibility benchmark of scientific integrity (Chang & Li, 2015; Freedman et al., 2015; Open Science Collaboration, 2015). In response, the scientific community has called for greater research transparency, which carries the presumption that data underlying published findings are not only shared, but shared in professional data archives that have the expertise and infrastructure to ensure that data are independently understandable to the research community (DA-RT, 2015; Center for Open Science, 2015).

Data archives have long accepted the charge from the scientific community to meet data quality standards to support research transparency. For years, data archives have instituted baseline protocols for acquiring data submissions, preparing data materials for repository ingest, and providing access to usable dataset files based on archival standards for trustworthy repositories. The Reference Model for an Open Archival System (OAIS) is something of a magna carta of archival standards, informing the processing approaches of many data archives (Lee, 2010). OAIS provides a framework of high-level concepts for understanding the requirements for long-term preservation and access of materials. Fundamental to OAIS is the concept of the 'Designated Community,' which is defined as an "identified group of potential Consumers who should be able to understand a particular set of information" (CSSDS, 2012, p. 1-11). In accordance with OAIS, data archives are responsible for giving access to materials that meet the 'independently understandable' criterion for data quality. Meeting this criterion requires that data packages held in data archives include sufficient information for users to apprehend the content, context, and structure of the data, as well as information regarding the unique identity, original source, and allowable uses of the data. What this has meant in practice for the Odum Institute is that, even beyond the various automatic ingest processes executed by archival system technologies, the data archivist is responsible for performing an assortment of critical data curation tasks. Table 1 provides a complete illustration of the Odum Institute data curation pipeline.

This skill-, time-, and resource-intensive data curation is similar to Peer, Green, and Stephenson's (2014) data quality review adopted by Yale University's Institution for Social and Policy Studies (ISPS) and the data curation pipeline employed at the Inter-university Consortium

for Political and Social Research (ICPSR) (Vardigan, 2007). This high-level, or maximal, data curation approach involves an exhaustive list of processing actions that are possible to execute only in small-scale operations as in the case of ISPS, or for well-resourced operations such as ICPSR. For data archives for which neither category applies, maximal data curation may not be feasible and/or sustainable even though our users require it. Here we arrive at an impasse where we need to confront problems of expectation management and resource management as we weigh user requirements for data quality against data archive capabilities.

DATA CURATION PIPELINE

STANDARD CURATION	• Review the dataset file package to ensure all components necessary to describe and interpret the data are present (i.e., codebook, instruments, reports, etc.) • Build the document set (i.e., construct codebooks, locate external documents) • Review data for confidentiality risks • Review data for errors (i.e., wild or out-of-range codes, missing or inconsistent variables, undefined missing values) • Perform data cleaning operations to anonymize data, correct data errors and inconsistencies, and standardize missing values	• Assign a persistent identifier (i.e., DOI) • Apply standard vocabulary • Generate standard DDI metadata to include methodological information and links to associated publications • Add full variable and value label text to dataset	• Normalize files to non-proprietary, software-agnostic preservation formats • Generate derivative files for widely-used software platforms
REPLICATION VERFICATION	• Review the replication data materials for completeness (i.e., README file, code file, etc.) • Review the code for inclusion of commands and comments required for execution • Execute code and compare results to the tables and figures in the manuscript	• Link the replication dataset to the published article	

Table 1. Odum Institute Data Curation Pipeline

More Data, Less Process?

Resolving the problems of expectation management and resource management is at the core of MPLP, which "...can help archivists make decisions about balancing resources so as to accomplish their larger ends and achieve economies in doing so..." (Meissner & Greene, 2010, p. 176). MPLP petitions archivists to pursue the 'good enough,' or 'golden minimum,' in archival processing work, which gives permission to archivists to spend the minimum amount of effort necessary to serve users' needs. Anything beyond the minimum must have "clearly demonstrable business reasons" (p. 240). However, what is 'good enough' for traditional archives may not be 'good enough' for data archives.

To determine what level of processing is considered 'good enough,' MPLP directs archivists to examine three primary task areas: arrangement, description, and preservation. In traditional archives, arrangement refers to the organization of files into physical and intellectual collections in order to preserve the context of the files' creation as well as the order of the files as they were created. Description provides detailed information about the context, characteristics, and content of the materials to allow users to discover them and evaluate their relevance. Preservation deals with the long-term maintenance and protection of materials (Society of American Archivists, n.d.). Though the materials MPLP refers to differ from data objects, these archival processing activities do have their equivalents in data curation.

Arrangement

In MPLP, finding the 'good enough' in arrangement tends towards deliberations over re-labeling and re-foldering archival materials, and whether or not doing so for individual objects is necessary to fulfill the intended purpose of arrangement. According to Meissner and Greene, as well as other foundational texts on archival practices, arrangement is a way to organize materials both physically and intellectually in a way that preserves their context (Society of American Archivist, n.d.). MPLP disputes the meticulousness with which some archivists organize and apply labels to individual objects. Rather than impulsively engaging in such "overzealous housekeeping, writ large" (p. 241), MPLP insists that the archivists discharge themselves of such object-level physical arrangement, which contributes little to users' understanding of the context. Greene and Meissner wrote: "If a user is given an understanding of the whole and the structure and identity of its meaningful parts, then the vagaries that occur within a folder will not prove daunting, and probably not even confusing" (p. 241).

In applying MPLP recommendations to arrangement of data collections, primary focus is on the 'understanding of the whole,' which, for data, is an understanding of their context. This context is contained in codebooks that define each variable and value code; documented data collection instruments such as interview or survey protocols; methodology reports containing comprehensive information on data collection, cleaning and analysis procedures; links to related research products including publications citing the data; and the programming code used to execute data analysis. Arrangement of data is ensuring the presence of these materials and the sufficiency of the information contained in these materials so that the data are 'independently understandable.' Where any of these documents do not exist, we might construct them from scratch, a cumbersome practice of stitching together information from the data producer, related publications, or any other sources that offer useful clues about the context of the data. Data archivists might also insist on performing a meticulous variable-by-variable check of the dataset file to identify and correct errors and inconsistencies.

MPLP questions how much of this attention and diligence to contextual materials is necessary for users' understanding of the whole. Reviewing datasets and correcting coding errors is as, if not more, tedious as shuffling documents among folders. Assembling and copyediting supplemental documents for a dataset is not such a far cry from the meticulous practice of re-labeling folders. Instead, processing approaches for archival arrangement for data should keep focus on the goal of 'understanding of the whole' and determine what the fundamental requirements are for achieving that goal. What is most critical for understanding data is having the information necessary to decipher cryptic variable names and undefined value codes. Data archives may need to reconsider the benefits to users of providing additional and/or enhanced supplementary materials and performing variable-by-variable checks against the amount of resources the archive has to commit to these practices.

Description

As is the case for any type of archival material, the primary purpose of description is to assist users in discovering and accessing materials of interest. MPLP suggests that archivists provide enough information to afford users 'decent access' without expending extra effort on composing lengthy descriptions of an individual object or its context. MPLP discourages verbosity in description, which is considered gratuitous and does not necessarily lend itself to an increase in users' understanding of the materials or their location.

Description as it is performed in the data curation pipeline involves applying standard vocabularies, generating metadata, enhancing variable labels, and assigning a persistent identifier for the data. The generation of standardized Data Documentation Initiative (DDI) metadata for data discovery is extended to include both methodological and contextual details extracted from the document set. While this provides robust metadata for search and discovery, MPLP asks us to consider whether it is perhaps 'good enough' to provide basic discovery metadata without taking the time to incorporate these methodological details such as sample size, weighting procedures, and other contextual information that is available within supplementary documents. Greene and Meissner make the point that as archivists it is not our job to do the research for our patrons, and efficiencies could potentially be gained by minimizing the generation of metadata.

Generating metadata and assigning a persistent identifier also underlie the creation of a stable data citation. Data citation is an essential practice for not only ensuring data producers receive appropriate attribution but also providing persistent access to data, documentation, and code. The Joint Declaration of Data Citation Principles (2014) communicates the importance of data citation as a scholarly practice and provides information on the purpose, function, and attributes of data citations. These principles highlight the role data archives play in the creation of data citations by generating metadata and assigning persistent identifiers and reaffirm this as an essential curation practice.

The other key description task within our pipeline is the enhancement of variable level metadata. For social science survey data, this often takes the form of adding the complete question text to variables. This variable level description allows for much more detailed and comprehensive discovery, examination, and analysis of the data within the repository platform. However, the MPLP model would suggest against this 'item level' description as a processing benchmark and would instead suggest archivists focus on describing the materials as a whole. Understanding how researchers interact with and use repository metadata would help us understand what is 'good enough' for description. Repositories have employed usability testing to inform interface design and the expansion of platform functionalities (Gibbs et al., 2013), an extension of these types of studies could increase our understanding of what metadata fields are most useful to researchers and provide additional evidence for how best to serve users' access needs.

Preservation

Because our focus is on the work of the archivist, a discussion of archive systems technology that are required to effectively preserve data is beyond the scope of this paper. While much of archival preservation actions take place within technological systems, there are

some preservation tasks that archivists perform. Since digital materials are far more fragile than analog materials (Rothenberg, 1999), we concede that MPLP's 'good enough' does not apply as readily to preservation of digital data. However, a consideration of MPLP in our examination of activities in the data curation pipeline--file normalization and optimization--that support long-term preservation allows us to identify potential efficiencies.

Normalizing files into open or preferred file formats allows files to remain accessible and protects against obsolescence. Without normalizing files, data may become unreadable and therefore unusable. A paramount requirement when serving users is ensuring digital material remain accessible into the future; therefore, normalization can be seen as an essential processing practice. In some cases, multiple different derivative copies of a dataset may also be created to allow expanded access to the data. While this increases the dissemination of the data and facilitates reuse, one file normalized into a non-proprietary file format would serve basic user requirements. Although the researcher would then have to read the file into his or her preferred software and variable-level metadata stored within the software package would be lost, as long as that contextual information is available within accompanying documentation then researchers would still be able to fundamentally understand the data. The original file in the proprietary format may also be made available alongside the preservation copy. Perhaps 'good enough' is creating a single non-proprietary version of the data file.

Another possible option includes shifting the burden to the data depositor and only accepting certain file formats for inclusion within the repository. For instance, guideline two of the Data Seal of Approval (2013) states that "the data producer provides the data in formats recommended by the data repository" (p. 12). This guideline shows how a DSA-certified data repository at a minimum must provide recommendations for appropriate file formats but the onus may be placed upon the data producer to comply. Another more automated solution can be seen in certain repository software platforms, such as the Dataverse, that generate a derivative preservation copy for certain data file types upon ingest (Crosas, 2011). An expansion of these types of system functionalities could also lessen the processing burden.

Good Enough" For Data Curation

A reconsideration of primary data curation activities has helped to identify those activities that are essential to ensuring access to quality data. For each of the three processing task areas, there are some activities that, if not performed, will likely make it impossible for users to discover, interpret, and use the data. We offer this 'minimal curation' model as a point of reference for which to engage the data archives community in a discussion on the necessity of intensive data curation processes for supporting data quality and reuse, particularly for tightly resourced environments.

In the MPLP-based 'minimal curation' model (see Table 2), arrangement is reduced to the single task of data file package review, description requires only metadata generation and persistent identification, and preservation is limited to file normalization.

"MINIMAL" DATA CURATION PIPELINE

ARRANGEMENT	DESCRIPTION	PRESERVATION
• Review the dataset file package to ensure all components necessary to describe and interpret the data are present (i.e., codebook, instruments, reports, etc.)	• Assign a persistent identifier (i.e., DOI) • Generate basic descriptive standard DDI metadata for discovery and access	• Normalize files to non-proprietary, software-agnostic preservation formats

Table 2. Proposed Minimal Data Curation Pipeline

Arrangement

Most important to arrangement is ensuring that necessary documentation is included within the data package so that users can understand the context of the data. Whether this documentation takes on the form of a codebook or survey instrument, or some other format, at a minimum documentation should define variables and values and provide some indication of the research methodology and process, for which links to external publications may be sufficient. No longer part of arrangement in the minimal data curation scheme is the variable-by-variable review of the data to identify and remedy errors, discrepancies, and/or sensitive information in the data. By scaling back on the comprehensive data review to this degree, the archive may no longer be able to guarantee the quality of the minutiae of every dataset. In some cases, variables and values may be left undefined, missing values inconsistently or incorrectly coded, and sensitive variables in the dataset might be awaiting unauthorized disclosure.

Certainly, these compromises have the potential to impact overall usability; however, the goal of 'minimal curation' is to ensure that enough information is present for users to understand and interpret the data as a whole. The users still have contextual information available to them so as to assess the overall credibility of the data and to determine whether or not the data are relevant to them. The presence of variable and value definitions in codebooks enables users to make necessary corrections in the data. It is also

not unreasonable to set policies that make data producers responsible for removing sensitive information in their data files and users responsible for reporting the presence of sensitive data. Should the data archiving and research community determine that comprehensive variable level review is the 'golden minimum' for data, then we must also provide "clearly demonstrable business reasons" that dictate this additional task and take into account the additional resources that will be required as a result.

Description

The minimal curation model reduces the amount of metadata generated for a given dataset. Instead of generating extensive DDI metadata and enhancing the variable labels for question-text search queries, the archive would simply generate enough descriptive metadata to allow users to discover and access the data and understand the general scope and topic of the data. This descriptive metadata would also include a persistent identifier and all the information required for a standard data citation.

While this strategy may compromise some of the discovery and online analysis potential for a dataset, it would still serve basic discovery and accessibility requirements. Description to support understanding of the content and context of the data would be left to information contained in supplementary documentation.

Preservation

In regard to preservation, the archive would continue to normalize files into a non-proprietary, system-agnostic file format, as we believe this is necessary to ensure that users are able to properly render the data into the future even as hardware and software systems become obsolete. Rather than producing several different file derivatives, normalization is limited to a single file format that users may convert for use in various software platforms.

Although we did not discuss other preservation activities such as generating and recording checksums, performing fixity checks, and migrating digital content, these are archival processes that are essential for long-term access and reusability and therefore cannot be compromised. However, these preservation activities are performed by archival platform systems and have little effect on archivist-led data curation processes.

What we have identified as a minimal data curation pipeline is neither an endorsement of a new standard of data curation, nor of MPLP itself. 'Minimal curation' is not suggested as an alternative to maximal curation. Maximal curation supports the sharing of the highest quality data that gives greater assurances that data will be reusable into the foreseeable future. 'Minimal curation' is presented to address conflicting priorities in a search for efficiency gains.

Discussion

The outcome of the MPLP exercise of reconsidering data curation processes is a recognition and greater appreciation of our commitment to providing access to quality data for our user community. In our examination of each of our current data curation activities, we were able to reaffirm the value of our practices to our users and their specific needs and expectations for quality data. As MPLP predicted, this exercise reminded us that "choices can be uncomfortable" (p. 233) when attempting to find efficiencies in our current practices, all of which we deem indispensable to our users. But we have little choice but to do so as we anticipate an increase in demand for data curation services. MPLP forced us to think about how each task in our data curation pipeline contributes to our goals. In doing so, we also reaffirmed the necessity and non-negotiable nature of some tasks that must be performed regardless of their intensity.

The search for 'good enough' for data has again left us in a quandary since in many ways meeting the requirements for reuse requires labor-intensive data quality review processes. Several data repositories have implemented a variety of resource management strategies for addressing the challenge of providing high quality data access with limited support. For example, the UK Data Archive has developed different levels of data curation to most effectively respond to users' varying needs (UK Data Archive, 2013). A key aspect of this is clear communication of the data curation tasks that will (and will not) be performed as part of a program of expectation management that distinguishes roles, rights, and responsibilities of the data producer, data user, and data archive.

Shifting responsibility for certain data curation tasks from the data archive to the data producer and data user assumes that data producers and users have an understanding of data quality requirements and the tasks required to meet those requirements, which, unfortunately, is not always the case. To address these challenges, information professionals have produced a proliferation of educational materials and programs to teach researchers strategies for effectively managing their data with eventual data archiving and sharing in mind. While online education programs (such as MANTRA and the Research Data Management and Sharing MOOC) have the potential to reach researchers worldwide, the impact of such education programs is not immediate and does not necessarily guarantee that data meet the standard of being 'independently understandable.'

Tools that facilitate and provide additional functionalities to streamline data curation processes also present opportunities for efficiency gains. Likewise, tools, such as the Open Science Framework, that help moderate and structure research workflows with an end goal of archiving and sharing data have the potential to assist researchers in creating data packages that meet data reuse requirements. However, even with these tools, certain tasks will continue to require manual data curation processes.

Ultimately, determining which data curation processes are essential for archiving and sharing of data that meet certain quality standards requires further research. This research will provide the empirical evidence and rationale for data archives' roles in curating data for

reuse in accordance with the needs of our designated community. Although previous studies have already clearly demonstrated the importance of contextual information (Faniel & Jacobsen 2010; Faniel et al., 2012), additional research is needed to investigate: 1) the designated community's expectations of the archive's role in providing quality data; 2) how variable level reviews affect reuse; 3) how the presentation of contextual information affects use; 4) how users interact with contextual and variable level metadata; and 5) how specific data curation tasks performed by archives directly impact the data quality and satisfaction criteria discussed within the literature.

By expanding our knowledge on the connections between user needs and data curation processes, we will be better equipped to determine what is 'good enough' for data. Likewise, we will be able to substantiate the necessity of data curation, whether it be maximal or minimal, for informed reuse and make clear that simply making data available does not automatically equate to data that are useable. We will then be able to expand and build upon initiatives advocating for the development of sustainable funding models for data archives (ICPSR, 2013).

Conclusion

Performing the conceptual exercise of applying MPLP in many ways raised more questions than it answered. MPLP reaffirmed our belief that a certain amount of processing is necessary to adequately meet users' needs. MPLP also brought to light some gaps in our knowledge about data use that prevents us from truly determining the minimum amount of processing needed. Future research will help us build better understanding of the connection between user needs and data curation processes. In many ways, the exercise suggests that 'good enough' for data still sets the bar pretty high, and building sustainable models to fund data curation will require the data archiving community to articulate the amount of skills, time, and labor that are non-negotiable when a high level of data quality is expected.

Essentially, this exercise boils down to a quotation from Clifford Lynch: "it is clear that an enormous imbalance exists between the resources currently available to fund these efforts and the potentially almost infinite demands of a fully realized data stewardship program; a key strategy in managing this imbalance is the effective use of the specific policy goals, such as data reuse, as shaping and prioritizing mechanisms in shaping an overall stewardship effort" (Lynch, 2013, p. 408). With the growth of data sharing mandates and the increasing focus on research transparency, data archives will play an essential role. However, questions still remain as to how we can best support these needs in a sustainable way that results in data that meet the requirements for reuse.

REFERENCES

Borgman, C.L. (2015) Big data, little data, no data: scholarship in the networked world. Cambridge, Massachusetts: The MIT Press.

Chang, A.C., Li, P. (2015) Is economics research replicable? Sixty published papers from thirteen journals say "usually not." Finance and Economics Discussion Series 2015-083. Washington DC: Board of Governors of the Federal Reserve System. doi:10.17016/FEDS.2015.083

Center for Open Science. (2015) Transparency and Openness Promotion (TOP) Guidelines. Available from: https://cos.io/top/#signatories

Consultative Committee for Space Data Systems. (2012) Reference model for an open archival information system (OAIS) (Magenta Book No. 650.0-M-2). Washington, DC: National Aeronautics Space Agency.

Crosas, M. (2011) The Dataverse Network®: An open-source application for sharing, discovering and preserving data. D-Lib Magazine 17 (1-2). doi:10.1045/january2011-crosas

DA-RT. (2015) The Journal Editors' Transparency Statement (JETS). Available from: http://www.dartstatement.org/#!blank/c22sl

Data Citation Synthesis Group. (2014) Joint Declaration of Data Citation Principles. Martone M. (ed.) FORCE 11, San Diego CA. Available from: /datacitation

Data Seal of Approval. (2013) Data Seal of Approval Guidelines (v.2). Available from: http://datasealofapproval.org/media/filer_public/2013/09/27/guidelines_2014-2015.pdf

Digital Curation Centre (DCC). (n.d) Curation reference manual. Available from: http://www.dcc.ac.uk/resources/curation-reference-manual

Faniel, I.M., Jacobsen, T.E. (2010) Reusing scientific data: How earthquake engineering researchers assess the reusability of colleagues' data. Computer Supported Cooperative Work (CSCW), 19 (3-4), p. 355–375. doi:10.1007/s10606-010-9117-8

Faniel, I.M., Kriesberg, A., Yakel, E. (2012) Data reuse and sensemaking among novice social scientists. Proceedings of the American Society for Information Science and Technology, 49 (1), p. 1–10. doi:10.1002/meet.14504901068

Faniel, I.M., Kriesberg, A., Yakel, E. (2015) Social scientists' satisfaction with data reuse. Journal of the Association for Information Science and Technology. doi:10.1002/asi.23480

Freedman, L.P., Cockburn, I.M., Simcoe, T.S. (2015) The Economics of Reproducibility in Preclinical Research. PLoS Biol 13 (6): e1002165. doi:10.1371/journal.pbio.1002165

Gibbs, E., Lin, L., Quigley, E. (2013) Dataverse usability evaluation: Final report. Available from: http://dataverse.org/files/dataverseorg/files/dataverse_usability_report-participant_omitted.pdf?m=1458571553

Greene, M.A., Meissner, D. (2005) More product, less process: Revamping traditional archival processing. The American Archivist, 68 (2), p. 208-263.

Inter-university Consortium for Political and Social Research (ICPSR). (2012). Guide to social science data preparation and archiving (5th ed.). Ann Arbor, MI: ICPSR. Available from: http://www.icpsr.umich.edu/files/deposit/dataprep.pdf

Inter-university Consortium for Political and Social Research (ICPSR). (2013, June 24-25) Sustaining domain repositories for digital data: A call for change from an interdisciplinary working group of domain repositories. Available from: http://www.icpsr.umich.edu/files/ICPSR/pdf/DomainRepositoriesCTA16Sep2013.pdf

King, G. (1995) Replication, replication. PS: Political Science & Politics, 28 (3), p. 444–452. doi:10.2307/420301

Lee, C.A. (2010) Open Archival Information System (OAIS) reference model. In Encyclopedia of Library and Information Sciences. Taylor & Francis, p. 4020–4030.

Lynch, C. (2014) The next generation of challenges in the curation of scholarly data. In J. M. Ray (Ed.), Research data management: Practical strategies for information professionals. West Lafayette, Indiana: Purdue University Press. Available from: http://www.cni.org/wp-content/uploads/2013/10/Research-Data-Mgt-Ch19-Lynch-Oct-29-2013.pdf

Meissner, D., Greene, M.A. (2010) More application while less appreciation: The adopters and antagonists of MPLP. Journal of Archival Organization, 8 (3-4), p. 174–226. doi:10.1080/15332748.2010.554069

National Endowment for the Humanities (NEH). (2012) Data management plans for NEH Office of Digital Humanitites proposals and awards. Washington DC: National Endowment for the Humanities. Available from: http://www.neh.gov/files/grants/data_management_plans_2015.pdf

National Institutes of Health (NIH). (2003) Final NIH statement on sharing research data (No. NOT-OD-03-032). Bethesda, MD: National Institutes of Health.

National Science Foundation (NSF). (2010) Dissemination and sharing of research results. Arlington, VA: National Science Foundation. Available from: https://www.nsf.gov/bfa/dias/policy/dmpfaqs.jsp#1

Nature. (2013) Availability of data, material and methods policy. Available from: http://www.nature.com/authors/policies/availability.html

Open Science Collaboration. (2015) Estimating the reproducibility of psychological science. Science, 349 (6251), aac4716. doi:10.1126/science.aac4716

Peer, L., Green, A., Stephenson, E. (2014) Committing to data quality review. International Journal of Digital Curation, 9 (1), p. 263–291. doi:10.2218/ijdc.v9i1.317

PLOS. (2014) Data availability policy. Available from: http://journals.plos.org/plosone/s/data-availability

Rothenberg, J. (1999) Ensuring the longevity of digital information. Washington, DC: Council on Library and Information Resources.

Science. (n.d.) Editorial policies: Data deposition. Available from: http://www.sciencemag.org/authors/science-editorial-policies

Society of American Archivists. (n.d.) Glossary of archival and records terminology: Preservation. Available from: http://www2.archivists.org/glossary/terms/p/preservation#.VxKnkfkrJ9M

Society of American Archivists. (n.d.) Glossary of archival and records terminology: Arrangement. Available from: http://www2.archivists.org/glossary/terms/a/arrangement#.VxKn2_krJ9M

UK Data Archive. (2015) Data ingest processing standards. Available from: http://www.data-archive.ac.uk/media/54782/cd079-dataingestprocessingstandards_08_00w.pdf

Wang, R.Y., Strong, D.M. (1996) Beyond accuracy: What data quality means to data consumers. Journal of Management Information Systems, 12, p. 5–33. doi:10.1080/07421222.1996.11518099

Notes

1. Sophia Lafferty-Hess is the Research Data Manager at the Odum Institute for Research in Social Science (228 Davis Library, CB# 3355, University of North Carolina at Chapel Hill), slaffer@email.unc.edu
2. Thu-Mai Christian is the Assistant Director of Archives at the Odum Institute for Research in Social Science (228 Davis Library, CB# 3355, University of North Carolina at Chapel Hill), thumai@email.unc.edu