ESSAY

On the Reuse of Scientific Data

Irene V. Pasquetto, Bernadette M. Randles and Christine L. Borgman

Information Studies, University of California Los Angeles, US

Corresponding author: Irene V. Pasquetto (irenepasquetto@ucla.edu)

While science policy promotes data sharing and open data, these are not ends in themselves. Arguments for data sharing are to reproduce research, to make public assets available to the public, to leverage investments in research, and to advance research and innovation. To achieve these expected benefits of data sharing, data must actually be reused by others. Data sharing practices, especially motivations and incentives, have received far more study than has data reuse, perhaps because of the array of contested concepts on which reuse rests and the disparate contexts in which it occurs. Here we explicate concepts of data, sharing, and open data as a means to examine data reuse. We explore distinctions between use and reuse of data. Lastly we propose six research questions on data reuse worthy of pursuit by the community: How can uses of data be distinguished from reuses? When is reproducibility an essential goal? When is data integration an essential goal? What are the tradeoffs between collecting new data and reusing existing data? How do motivations for data collection influence the ability to reuse data? How do standards and formats for data release influence reuse opportunities? We conclude by summarizing the implications of these questions for science policy and for investments in data reuse.

Keywords: Data Reuse; Open Data; Science Policy; Knowledge Infrastructures

Introduction

Over the last decade or so, a growing number of governments and funding agencies have promoted the sharing of scientific data as a means to make research products more widely available for research, education, business, and other purposes (European Commission High Level Expert Group on Scientific Data 2010; National Institutes of Health 2016; National Science Foundation 2011; Organisation for Economic Co-operation and Development 2007). Similar policies promote open access to data from observational research networks, governments, and other publicly funded agencies. Many private foundations also encourage or require the release of data from research they fund. Whereas policies for sharing, releasing, and making data open have the longest histories in the sciences and medicine, such policies have spread to the social sciences and humanities. Concurrently, they have spread from Europe and the U.S. to all continents.

The specifics of data sharing policies vary widely by research domain, country, and agency, but have many goals in common. Borgman (2015) analyzed many of these policies, and found that the arguments for sharing data could be grouped into four general categories: to reproduce research, to make public assets available to the public, to leverage investments in research, and to advance research and innovation. Implicit in these arguments is the desire to produce new knowledge by reusing shared and open resources (Margolis et al. 2014; Wilkinson et al. 2016). While laudable goals, these arguments are not ends in themselves. Data sharing, in its simplest form, is merely releasing or posting data. The proposed benefits of sharing will be achieved only if the available data are used, or reused by others (Borgman 2015).

Data sharing, particularly the incentives and disincentives to share, is more widely studied than is data reuse (Carlson and Anderson 2007; Faniel and Jacobsen 2010; Tenopir et al. 2011; Treadway et al. 2016; Wallis 2014; Zimmerman 2007). The frequency with which data are shared or reused in the sciences is extremely difficult to assess because of the many meanings of these terms and the diversity of contexts in which sharing and reuse might occur. A number of surveys have asked researchers to report the frequency

with which they share data, reuse data, and circumstances under which they are willing to share or are inclined to reuse data. These surveys vary in the way in which they define "share" or "reuse" – if they define them at all – and use a wide array of methods to build a survey sample, ranging from targeted groups to public postings that acquire a self-selected sample. Results vary accordingly. Most surveys find fairly high percentages of self-reported data sharing and reuse, and even higher rates of intended sharing and reuse (Treadway et al. 2016; Tenopir et al. 2011). In contrast, a study conducted by *Science* of its reviewers, asking where they stored their data, found that most data were stored on lab servers rather than in public repositories (Jasny et al. 2011). Qualitative studies, based on interviews and ethnography, tend to find relatively low levels of data sharing and reuse (Faniel and Jacobsen 2010; Wallis et al. 2013).

Rather than debate how much data sharing and reuse are occurring, here we focus on explicating data reuse as a concept that needs to be understood far more fully. The contested nature of core concepts in data practices has limited progress toward improving the dissemination and reuse of scientific data. We briefly define the concepts of "data," "sharing," and "open" to lay the groundwork for examining the concept of data reuse. The remainder of this short article provides working definitions of these concepts, then presents a set of research questions to be explored by the community.

Conceptual Framework

Terms such as "data," "sharing," "open," and "open data" each encompass many meanings in research publications and in science policy. Key concepts often are conflated or used interchangeably. Here we provide a brief background on each of these terms, any of which is worthy of a long essay.

What are data?

Few data policy documents define "data" explicitly. At most, data are defined by example (e.g., facts, observations, laboratory notebooks), or by context (e.g., "data collected using public funds") (Organisation for Economic Co-operation and Development 2007). Research data sometimes are distinguished from resources such as government statistics or business records (Open Knowledge Foundation 2015). Here we rely on a definition developed earlier, in which data refers to 'entities used as evidence of phenomena for the purposes of research or scholarship' (Borgman 2015, p. 29).

The above definition is useful in determining the point at which some observation, record, or other form of information becomes data. It also helps to explain why data often exist in the eye of the beholder. One researcher's signal – or data – may be someone else's noise (Borgman et al. 2007; Wallis et al. 2007).

However, this phenomenological definition does not address the question of units of data, or the question of degree of processing of data that are shared. Researchers may share a "dataset," which is another contested term. A dataset might consist of a small spreadsheet, a very large file, or some set of files. Determining the criteria for defining a dataset raises various other epistemological questions (Agosti and Ferro 2007; Renear and Dubin 2003). Similarly, the relationship between a research project and a dataset associated with an individual journal article may be one-to-one, many-to-many, or anywhere in between (Borgman 2015; Wallis 2012).

What is data sharing?

"Data sharing" generally refers to the act of releasing data in a form that can be used by other individuals. Data sharing thus encompasses many means of releasing data, and says little about the usability of those data. Examples of sharing include private exchanges between researchers; posting datasets on researchers' or laboratory websites; depositing datasets in archives, repositories, domain-specific collections, or library collections; and attaching data as supplemental materials in journal articles (Wallis et al. 2013). A relatively newer practice in many fields is to disseminate a dataset as a "data paper." Data papers provide descriptions of methods for collecting, processing, and verifying data, as done in astronomy (Ahn et al. 2012), which improves data provenance and gives credit to data producers. Methods of data sharing vary by domain, data type, country, journal, funding agency, and other factors. The ability to discover, retrieve, and interpret shared data varies accordingly (Borgman 2015; Leonelli 2010; Palmer et al. 2011).

What is open data?

"Open data" is perhaps the most problematic term of all, given the array of concepts and conditions to which it may refer (Pasquetto et al. 2016). Baseline conditions for open data usually refer to "fewest restrictions" and "lowest possible costs." Legal and technical availability of data often are mentioned (Open Knowledge

Foundation 2015; Organisation for Economic Co-operation and Development 2007). The OECD specifies 13 conditions for open data, only a few of which are likely to be satisfied in any individual situation (Organisation for Economic Co-operation and Development 2007). Examples of open data initiatives include repositories and archives (e.g., GenBank, Protein Data Bank, Sloan Digital Sky Survey), federated data networks (e.g., World Data Centers, Global Biodiversity Information Facility; NASA Distributed Active Archive Centers), virtual observatories (e.g., International Virtual Observatory Alliance, Digital Earth), domain repositories (e.g., PubMedCentral, arXiv), and institutional repositories (e.g., University of California eScholarship).

Openness varies in many respects. Public data repositories may allow contributors to retain copyright and control over the data they have deposited. Data may be open but interpretable only with proprietary software. Data may be created with open source software but require licensing for data use. Open data repositories may have long term sustainability plans, but many depend on short term grants or on the viability of business models. Keeping data open over the long term often requires continuous investments in curation to adapt to changes in the user community (K. S. Baker et al. 2015).

A promising new development to address the vagaries of open data is the FAIR standards – Findable, Accessible, Interoperable, and Reusable data (National Institutes of Health 2015). These standards apply to the repositories in which data are deposited. The FAIR standards were enacted by a set of stakeholders to enable open science, and they incorporate all parts of the "research object," from code, to data, to tools for interpretation (National Institutes of Health 2015; Wilkinson et al. 2016). For the purposes of this article, open data are those held in repositories or archives that meet the FAIR standards.

Using and reusing data
Even bounding the concepts of data, sharing, and open data, as we have above, data use and reuse are complex constructs. We will constrain the problem even more by focusing on data use and reuse for the purpose of knowledge production, rather than for teaching, presentations, outreach, product development, and so on. We also draw our examples from our own empirical research in the physical and life sciences. Here we identify core questions that we consider essential for understanding data reuse.

Use vs. Reuse of Data
The most fundamental problem in understanding data reuse is to distinguish between a "use" and a "reuse." In the simplest situation, data are collected by one individual, for a specific research project, and the first "use" is by that individual to ask a specific research question. If that same individual returns to that same dataset later, whether for the same or a later project, that usually would be considered a "use." When that dataset is contributed to a repository, retrieved by someone else, and deployed for another project, it usually would be considered a "reuse." In the common parlance of data practices, reuse usually implies the usage of a dataset by someone other than the originator.

When a repository consists entirely of datasets contributed by researchers, available for use by other researchers, then subsequent applications of those datasets would be considered reuse. However, when a research team retrieves its own data from a repository to deploy in a later project, should that be considered a use or a reuse? As scholars begin to receive more credit for data citation, then reuse of deposited data may increase accordingly. Conversely, when researchers obtain data from a repository, they rarely cite the repository, making such uses difficult to track (CODATA-ICSTI Task Group on Data Citation Standards Practices 2013; Uhlir 2012). However, researchers themselves are inconsistent in citing data that they deposit for others to use. Encouraging consistent citation of datasets would increase dissemination.

Some data archives consist of data collected for use by a community, thus any research based on retrieved datasets could be a first "use" of those data. In astronomy, for example, sky surveys collect massive datasets that include images, spectra, and catalogs. The project team prepares the data for scientific use, and then makes the processed datasets available as periodic "data releases" (Szalay et al. 2000). Once released, astronomers use the data for their own scientific objectives (Pasquetto et al. 2015).

Similar large datasets are assembled in the biosciences. For example, computational biologists rely on reference sequence data collected by the Human Genome Project (HGP) for mapping their own new data (Berger et al. 2016). In "next-generation sequencing," DNA molecules are chopped into many small fragments (reads) that bioinformaticians will reassemble in the correct order (Berger et al. 2016; Orelli 2016). Such data collections exist as initial sources of data to ask new questions, rather than assemblages of data collected for myriad purposes by individual researchers and teams.

Reusing data to reproduce research

Reproducibility is the impetus most commonly cited for data sharing. Many fields are claiming a "reproducibility crisis," and demanding more data release for these purposes (M. Baker 2016). Data from a prior study can be reanalyzed to validate, verify, or confirm previous research in a reproducibility study, where the same question is asked again using the same data and analysis methods (Borgman 2015). Slightly different is a replication study, where novel data are used to ask an old question using the same data and analysis methods (Drummond 2009). Reproducibility research is more common in computational sciences, where software pipelines can be recorded (Stodden 2010; Vandewalle et al. 2009). Other mechanisms to reproduce data analyses include Investigation/Study/Assay, Nanopublications, Research Objects, and the Galaxy workflow system (González-Beltrán et al. 2015). Reproducibility has particularly high stakes in the biomedical fields, where the pharmaceutical industry attempts to validate published research for use in developing biomedical products.

However, notions of reproducibility vary widely, due to the many subtle and tacit aspects of research practice (Jasny et al. 2011). Those who study social aspects of scientific practice tend to be highly skeptical of reproducibility claims (Collins 1985; Collins and Evans 2007; Latour and Woolgar 1979).

Independent Reuse vs. Data Integration

Reproducing a study is an example of independent reuse of a dataset. Even so, the dataset is of little value without associated documentation, and often software, code, and associated scientific models. In other cases, a single dataset might be reused for a different purpose, provided the associated contextual information and tools are available.

More complex are the cases where datasets are reused in combination with other data, whether to make comparisons, build new models, or explore new questions altogether. All the datasets involved might be from prior research of others, or available data might be integrated with new observations (Berger et al. 2016; Rung and Brazma 2012).

Datasets can be compared and integrated for a single analysis study, a meta-analysis, parameter modeling, or other purposes. Multiple datasets can be integrated at "raw" or processed levels (Rung and Brazma 2012). Similar datasets or heterogeneous datasets might be combined. For example, multiple raw datasets of gene expression data can be integrated to assess general properties of expression in large sample groups (Lukk et al. 2010). Summary-level gene expression data, such as P values, can be integrated in meta-analysis to compare conditions and diseases (Vilardell Nogales et al. 2011).

In some cases, a primary scientific goal is to integrate heterogeneous datasets into one dataset to allow reuse. An example is the COMPLETE (Coordinated Molecular Probe Line Extinction Thermal Emission Survey of Star Forming Regions) Survey, that integrated new observations with datasets retrieved from public repositories of astronomical observations that cover same regions of the sky (Goodman 2004; Ridge et al. 2006).

Some interdisciplinary fields such as ecology research combine datasets from multiple sources. To understand the impact of an oil spill, ecologists might construct a combined model with data from benthic, planktonic, and pelagic organisms, chemistry, toxicology, oceanography, and atmospheric science, with data on economic, policy, and legal decisions that affect spill response and cleanup (Crone and Tolstoy 2010). Similarly, a combination of social, health and geography data is necessary to develop models to explain the spread and impact of contagious diseases (Groseth et al. 2007).

Reusing a single dataset in its original form is difficult, even if adequate documentation and tools are available, since much must be understood about why the data were collected and why various decisions about data collection, cleaning, and analysis were made. Combining datasets is far more challenging, as extensive information must be known about each dataset if they are to be interpreted and trusted sufficiently to draw conclusions.

Research Questions

The following research questions about data reuse arise from the considerations discussed above. Each question has broad implications for policy and practice. We are pursuing these questions in our own empirical research, and pose them here to the community as a means to stimulate broader discussion.

How can uses of data be distinguished from reuses?

Distinctions between uses and reuse often reflect differences in scientific practice. Some research domains build large data collections that are intended as primary sources for their communities, such as the examples given for astronomy and biosciences. In other domains of the life, physical, and social sciences, repositories

are constructed as secondary sources where researchers can deposit their data. These approaches may be equally worthwhile, but they involve different investments that must be considered in science policy and funding.

When is reproducibility an essential goal?

In fields facing a "reproducibility crisis," substantial investments may be appropriate in releasing datasets associated with individual journal articles or other publications. Such datasets may be encapsulated in a complex container of documentation, code, and software that ensure that procedures can be replicated precisely.

   In other areas, the ability to replicate, verify, or reach the same conclusions by different methods may be more scientifically valuable than reproducibility (Jasny et al. 2011). In these cases, consistent forms of documentation and references to common tools may be more essential than encapsulation of individual datasets.

When is data integration an essential goal?

When scientists need to combine datasets from multiple sources to address a new question, the first challenge is finding ways to integrate the datasets so they can be compared or merged. Considerable data reduction and data loss may occur in the process. These methodological activities can consume project resources. Even in business applications of data integration for reuse, estimates range up to 80% or more of project time spent in "data cleaning" (Mayer-Schonberger and Cukier 2013).

What are the tradeoffs between collecting new data and reusing existing data?

Little is known about the choices scientists make between when to collect data anew and when to seek existing data on a problem. One way to characterize data reuse could be to distinguish between the practices of reusing data collected "inside or outside" the scientist's own research team or project. Scientific teams often use their own data multiple times during a research project. However, scientists also integrate their own datasets with datasets obtained from repositories, colleagues, or other sources.

How do motivations for data collection influence the ability to reuse data?

When do scientists collect data with reuse in mind and when are data sharing and reuse by others an afterthought? How do these choices influence later reuse? Scientists frequently have difficulty imagining how others might use their data, especially others outside their immediate research area (Mayernik 2011). When data are collected with reuse in mind, they probably are more usable in the future by the originating team, as well as to other teams (Goodman et al. 2014). Examples include the HGP reference sequences datasets and the sky surveys described earlier. In other situations, scientists integrate their own data with external data that were originally collected by diverse teams and purposes, as in the case of "dry lab" computational biology or the COMPLETE survey.

How do standards and formats for data release influence reuse opportunities?

Data released in formats that meet community standards are more likely to be analyzed with available tools and to be combined with data in those formats. In astronomy, for example, data usually are disseminated as FITS files, a format established in the 1980s. While the FITS format is a remarkable accomplishment that enables most astronomy data to be analyzed with common software tools, some argue that the standard has reached the limits of its utility (Greisen 2002; Thomas et al. 2014). Even with these standards in place, however, subtle differences in data collection, parameters, instrumentation, and other factors make data integration of FITS files a non-trivial matter (Borgman 2015; Borgman et al. 2016).

   Data integration and reuse are much more difficult in areas where standards are unavailable or premature. Scientists use and reuse their own and others' data for many different kinds of data analyses, such as for single analysis, meta-analysis, or parametric modeling. Exploratory research may be compromised by too much emphasis on data integration and reuse.

Conclusions

While the emphasis of science policy is on data sharing and open data, these are not ends in themselves. Rather, the promised benefits of open access to research data lie in the ability to reuse data. Data reuse is an understudied problem that requires much more attention if scientific investments are to be leveraged effectively. Data use can be difficult to distinguish from reuse. Where that line is drawn will favor some kinds of investments in data collections and archives over others. Data policies that favor reproducibility may

undermine data integration, and vice versa. Similarly, data policies that favor standardization may undermine exploratory research or force premature standardization. Thus, data reuse is not an end in itself either. Rather, data reuse is embedded deeply in scientific practice. Investments in data sharing and reuse made now may have long-term consequences for the policies and practices of science.

Acknowledgements

Competing Interests

The authors have no competing interests to declare.

References

Agosti, M and **Ferro, N** 2007 A formal model of annotations of digital content. *ACM Transactions on Information Systems*, 26(1). DOI: https://doi.org/10.1145/1292591.1292594

Ahn, C P, Alexandroff, R, Allende Prieto, C, Anderson, S F, Anderton, T, Andrews, B H, et al. 2012 The Ninth Data Release of the Sloan Digital Sky Survey: First Spectroscopic Data from the SDSS-III Baryon Oscillation Spectroscopic Survey. *Astrophysical Journal*, 203: 21. DOI: https://doi.org/10.1088/0067-0049/203/2/21

Baker, K S, Duerr, R E and **Parsons, M A** 2015 Scientific Knowledge Mobilization: Co-evolution of Data Products and Designated Communities. *International Journal of Digital Curation*, 10(2). DOI: https://doi.org/10.2218/ijdc.v10i2.346

Baker, M 2016 1,500 scientists lift the lid on reproducibility. *Nature*, 533(7604): 452–454. DOI: https://doi.org/10.1038/533452a

Berger, B, Daniels, N M and **Yu, Y W** 2016 Computational biology in the 21st century: scaling with compressive algorithms. *Communications of the ACM*, 59(8): 72–80. DOI: https://doi.org/10.1145/2957324

Borgman, C L 2015 *Big data, little data, no data: Scholarship in the networked world.* Cambridge, MA: The MIT Press.

Borgman, C L, Darch, P T, Sands, A E and **Golshan, M S** 2016 The durability and fragility of knowledge infrastructures: Lessons learned from astronomy. In: *Proceedings of the Association for Information Science and Technology* (Vol. 53). ASIS&T, pp. 1–10. DOI: https://doi.org/10.1002/pra2.2016.14505301057

Borgman, C L, Wallis, J C, Mayernik, M S and **Pepe, A** 2007 Drowning in Data: Digital library architecture to support scientific use of embedded sensor networks. In *Joint Conference on Digital Libraries.* Vancouver, British Columbia, Canada: Association for Computing Machinery, pp. 269–277. DOI: https://doi.org/10.1145/1255175.1255228

Carlson, S and **Anderson, B** 2007 What Are Data? The Many Kinds of Data and Their Implications for Data Re-Use. *Journal of Computer-Mediated Communication*, 12(2): 635–651. DOI: https://doi.org/10.1111/j.1083-6101.2007.00342.x

CODATA-ICSTI Task Group on Data Citation Standards Practices 2013 Out of Cite, Out of Mind: The Current State of Practice, Policy, and Technology for the Citation of Data. *Data Science Journal*, 12: CIDCR1–CIDCR75. DOI: https://doi.org/10.2481/dsj.OSOM13-043

Collins, H M 1985 *Changing Order: Replication and Induction in Scientific Practice* (Reprint). Chicago: University Of Chicago Press. Retrieved from: http://www.press.uchicago.edu/ucp/books/book/chicago/C/bo3623576.html.

Collins, H M and **Evans, R** 2007 *Rethinking Expertise.* Chicago: University of Chicago Press. DOI: https://doi.org/10.7208/chicago/9780226113623.001.0001

Crone, T J and **Tolstoy, M** 2010 Magnitude of the 2010 Gulf of Mexico oil leak. *Science (New York, N.Y.)*, 330(6004): 634. DOI: https://doi.org/10.1126/science.1195840

Drummond, C 2009 Replicability is not Reproducibility: Nor is it Good Science. In: *ICML '09 Proceedings of the 26th Annual International Conference on Machine Learning (The fourth workshop on evaluation methods for machine learning).* Montreal, Quebec, Canada. Retrieved from: http://cogprints.org/7691/.

European Commission High Level Expert Group on Scientific Data 2010 *Riding the wave: How Europe can gain from the rising tide of scientific data* (Final report of the High Level Expert Group on Scientific Data. A submission to the European Commission). European Union. Retrieved from: https://www.fosteropenscience.eu/content/riding-wave-how-europe-can-gain-rising-tide-scientific-data.

Faniel, I M and **Jacobsen, T E** 2010 Reusing Scientific Data: How Earthquake Engineering Researchers Assess the Reusability of Colleagues' Data. *Journal of Computer Supported Cooperative Work*, 19(3–4): 355–375. DOI: https://doi.org/10.1007/s10606-010-9117-8

González-Beltrán, A, Li, P, Zhao, J, Avila-Garcia, M S, Roos, M, Thompson, M, et al. 2015 From Peer-Reviewed to Peer-Reproduced in Scholarly Publishing: The Complementary Roles of Data Models and Workflows in Bioinformatics. *PLOS ONE*, 10(7): e0127612. DOI: https://doi.org/10.1371/journal.pone.0127612

Goodman, A A 2004 The COMPLETE Survey of Star-Forming Regions on its Second Birthday. *Arxiv preprint astro-ph/0405554*. Retrieved from: https://arxiv.org/abs/astro-ph/0405554.

Goodman, A A, Pepe, A, Blocker, A W, Borgman, C L, Cranmer, K, Crosas, M, et al. 2014 Ten Simple Rules for the Care and Feeding of Scientific Data. *PLoS Computational Biology*, 10(4): e1003542. DOI: https://doi.org/10.1371/journal.pcbi.1003542

Greisen, E W 2002 FITS: A Remarkable Achievement in Information Exchange. In: Heck, A. (Ed.) *Information Handling in Astronomy − Historical Vistas*. Springer: Netherlands, pp. 71–87. Retrieved from: http://link.springer.com/chapter/10.1007/0-306-48080-8_5.

Groseth, A, Feldmann, H and **Strong, J E** 2007 The ecology of Ebola virus. *Trends in Microbiology*, 15(9): 408–416. DOI: https://doi.org/10.1016/j.tim.2007.08.001

Jasny, B R, Chin, G, Chong, L and **Vignieri, S** 2011 Again, and Again, and Again . . . . *Science*, 334(6060): 1225. DOI: https://doi.org/10.1126/science.334.6060.1225

Latour, B and **Woolgar, S** 1979 *Laboratory Life: The Social Construction of Scientific Facts*. Beverly Hills: Sage Publications.

Leonelli, S 2010 Packaging small facts for re-use: databases in model organism biology. *How well do facts travel:* 325–348. Retrieved from: https://www.researchgate.net/profile/Sabina_Leonelli/publication/233438457_Packaging_Small_Facts_for_Re-Use_Databases_in_Model_Organism_Biology/links/0912f50aa0aa49acbe000000.pdf.

Lukk, M, Kapushesky, M, Nikkilä, J, Parkinson, H, Goncalves, A, Huber, W et al. 2010 A global map of human gene expression. *Nature Biotechnology*, 28(4): 322–324. DOI: https://doi.org/10.1038/nbt0410-322

Margolis, R, Derr, L, Dunn, M, Huerta, M, Larkin, J, Sheehan, J, et al. 2014 The National Institutes of Health's Big Data to Knowledge (BD2K) initiative: capitalizing on biomedical big data. *Journal of the American Medical Informatics Association*, 21(6): 957–958. DOI: https://doi.org/10.1136/amiajnl-2014-002974

Mayernik, M S 2011 (June) *Metadata Realities for Cyberinfrastructure: Data Authors as Metadata Creators* (PhD Dissertation). UCLA, Los Angeles, CA. DOI: https://doi.org/10.2139/ssrn.2042653

Mayer-Schonberger, V and **Cukier, K** 2013 *Big Data: A Revolution That Will Transform How We Live, Work, and Think*. Boston: Houghton Mifflin Harcourt.

National Institutes of Health 2015 Commons Home Page | Data Science at NIH. Retrieved from: https://datascience.nih.gov/commons

National Institutes of Health 2016 NIH Sharing Policies and Related Guidance on NIH-Funded Research Resources. Retrieved from: https://grants.nih.gov/policy/sharing.htm

National Science Foundation 2011 Award and Administration Guide: Data Sharing Policy. Retrieved from: https://www.nsf.gov/pubs/policydocs/pappguide/nsf11001/aag_6.jsp#VID4.

Open Knowledge Foundation 2015 Open Definition: Defining Open in Open Data, Open Content and Open Knowledge. Retrieved from: http://opendefinition.org/od/.

Orelli, B 2016 *The Business of Genomic Data*. O'Reilly Media. Retrieved from: http://www.oreilly.com/data/free/the-business-of-genomic-data.csp.

Organisation for Economic Co-operation and Development 2007 *OECD Principles and Guidelines for Access to Research Data from Public Funding*. Paris: Organisation for Economic Co-Operation and Development, p. 24. Retrieved from: http://www.oecd.org/dataoecd/9/61/38500813.pdf.

Palmer, C L, Weber, N M and **Cragin, M H** 2011 The Analytic Potential of Scientific Data: Understanding Re-use Value. *Proceedings of the American Society for Information Science and Technology*, 48(1):

1–10. Retrieved from: http://onlinelibrary.wiley.com/doi/10.1002/meet.2011.14504801174/full. DOI: https://doi.org/10.1002/meet.2011.14504801174

Pasquetto, I V, Sands, A E and **Borgman, C L** 2015 Exploring openness in data and science: What is "open," to whom, when, and why? *Proceedings of the Association for Information Science and Technology*, 52: 1–2. DOI: https://doi.org/10.1002/pra2.2015.1450520100141

Pasquetto, I V, Sands, A E, Darch, P T and **Borgman, C L** 2016 Open Data in Scientific Settings: From Policy to Practice. In: *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. Presented at the 2016 CHI Conference on Human Factors in Computing Systems, New York, NY, USA: ACM, pp. 1585–1596. DOI: https://doi.org/10.1145/2858036.2858543

Renear, A and **Dubin, D** 2003 Towards identity conditions for digital documents. In: *Proceedings of the 2003 international conference on Dublin Core and metadata applications: supporting communities of discourse and practice*. Presented at the International Conference on Dublin Core and Metadata Applications, Seattle, Washington: Dublin Core Metadata Initiative. Retrieved from: http://www.siderean.com/dc2003/503_Paper71.pdf.

Ridge, N A, Francesco, J D, Kirk, H, Li, D, Goodman, A A, Alves, J F, et al. 2006 The COMPLETE Survey of Star-Forming Regions: Phase I Data. *The Astronomical Journal*, 131(6): 2921. DOI: https://doi.org/10.1086/503704

Rung, J and **Brazma, A** 2012 Reuse of public genome-wide gene expression data. *Nature Reviews Genetics*, 14(2): 89–99. DOI: https://doi.org/10.1038/nrg3394

Stodden, V 2010 *The scientific method in practices: Reproducibility in the computational sciences*. Cambridge, MA. Retrieved from: http://ssrn.com/abstract=1550193.

Szalay, A S, Kunszt, P Z, Thakar, A R, Gray, J and **Slutz, D** 2000 The Sloan Digital Sky Survey and its archive. In: *Astronomical Data Analysis Software and Systems IX* (Vol. 216). San Francisco, Calif: Astronomical Society of the Pacific, p. 405. Retrieved from: http://www.adass.org/adass/proceedings/adass99/O1-02/.

Tenopir, C, Allard, S, Douglass, K, Aydinoglu, A U, Wu, L, Read, E, et al. 2011 Data Sharing by Scientists: Practices and Perceptions. *PLoS ONE*, 6(6): e21101. DOI: https://doi.org/10.1371/journal.pone.0021101

Thomas, B, Jenness, T, Economou, F, Greenfield, P, Hirst, P, Berry, D S, et al. 2014 Significant Problems in FITS Limit Its Use in Modern Astronomical Research. *Astronomical Data Analysis Software and Systems XXIII*, 485: 351. Retrieved from: http://www.aspbooks.org/a/volumes/article_details/?paper_id=36249.

Treadway, J, Hahnel, M, Leonelli, S, Penny, D, Groenewegen, D, Miyairi, N, et al. 2016 *The State of Open Data Report*. Figshare. Retrieved from: https://figshare.com/articles/The_State_of_Open_Data_Report/4036398.

Uhlir, P F (ed.) 2012 *For Attribution – Developing Data Attribution and Citation Practices and Standards: Summary of an International Workshop*. Washington, D.C.: The National Academies Press. Retrieved from: http://www.nap.edu/catalog.php?record_id=13564.

Vandewalle, P, Kovacevic, J and **Vetterli, M** 2009 Reproducible Research in Signal Processing. *IEEE Signal Processing Magazine*, 26: 37–47. DOI: https://doi.org/10.1109/MSP.2009.932122

Vilardell Nogales, M, Rasche, A, Thormann, A., Maschke-Dutz, E, Pérez Jurado, L A, Lehrach, H and **Herwig, R** 2011 (May) Meta-analysis of heterogeneous Down Syndrome data reveals consistent genome-wide dosage effects related to neurological processes. *BMC Genomics*, 12: 229. Retrieved from: http://repositori.upf.edu/handle/10230/23032. DOI: https://doi.org/10.1186/1471-2164-12-229

Wallis, J C 2012 *The Distribution of Data Management Responsibility within Scientific Research Groups* (Ph.D. Dissertation). University of California, Los Angeles, United States – California. Retrieved from: http://search.proquest.com/docview/1029942726/abstract?accountid=14512.

Wallis, J C 2014 Data Producers Courting Data Reusers: Two Cases from Modeling Communities. *International Journal of Digital Curation*, 9(1): 98–109. DOI: https://doi.org/10.2218/ijdc.v9i1.304

Wallis, J C, Borgman, C L, Mayernik, M S, Pepe, A, Ramanathan, N and **Hansen, M A** 2007 Know Thy Sensor: Trust, Data Quality, and Data Integrity in Scientific Digital Libraries. In: *Proceedings of the 11th European Conference on Research and Advanced Technology for Digital Libraries* (Vol. LINCS 4675). Budapest, Hungary: Berlin: Springer, pp. 380–391. DOI: https://doi.org/10.1007/978-3-540-74851-9_32

Wallis, J C, Rolando, E and **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: https://doi.org/10.1371/journal.pone.0067332

Wilkinson, M D, Dumontier, M, Aalbersberg, Ij J, Appleton, G, Axton, M, Baak, A, et al. 2016 The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3: 160018. DOI: https://doi.org/10.1038/sdata.2016.18

Zimmerman, A S 2007 Not by metadata alone: the use of diverse forms of knowledge to locate data for reuse. *International Journal on Digital Libraries*, 7(1–2): 5–16. DOI: https://doi.org/10.1007/s00799-007-0015-8