HHS Public Access

Author manuscript

Am Psychol. Author manuscript; available in PMC 2019 February 01.

Published in final edited form as:

Am Psychol. 2018 ; 73(2): 111–125. doi:10.1037/amp0000242.

Data Sharing in Psychology

Maryann E. Martone,
Department of Neurosciences at the University of California, San Diego

Alexander Garcia-Castro, and
Ontology Engineering Group at the Polytechnic University of Madrid

Gary R. VandenBos
Former Publisher of American Psychological Association and has been involved in data sharing initiatives for over 25 years

Abstract

Routine data sharing, defined here as the publication of the primary data and any supporting materials required to interpret the data acquired as part of a research study, is still in its infancy in psychology, as in many domains. Nevertheless, with increased scrutiny on reproducibility and more funder mandates requiring sharing of data, the issues surrounding data sharing are moving beyond whether data sharing is a benefit or a bane to science, to what data should be shared and how. Here, we present an overview of these issues, specifically focusing on the sharing of so-called "long tail" data, that is, data generated by individual laboratories as part of largely hypothesis-driven research. We draw on experiences in other domains to discuss attitudes towards data sharing, cost-benefits, best practices and infrastructure. We argue that the publishing of data sets is an integral component of 21st century scholarship. Moreover, although not all issues around how and what to share have been resolved, a consensus on principles and best practices for effective data sharing and the infrastructure for sharing many types of data are largely in place.

Why are we talking about data sharing?

The simple answer is, because we can (and we should). Prior to computers and the internet, sharing data-routinely was really not possible beyond what could be published in journals or books. Consequently, a culture grew up around scholarly publishing where data were considered disposable after some specified regulatory period, and certainly where the production of a data set on its own was not considered a work of scholarship. Rather it was the hypotheses proposed, the experimental design, the analysis, and the insights gained from collected data that were valued and preserved through our system of journals and books.

Few would argue that the full potential of a data set can be captured in prose, but the scientific article has held such a privileged place in scientific communication for so long that the current push to treat scientific data differently than in the past has encountered much

Correspondence: Maryann E. Martone; mmartone@ucsd.edu.

Editor's note. This article is part of a special section titled "Data Sharing in Psychology: An Idea Whose Time Has Come," published in the February–March 2018 issue of American Psychologist. Jennifer Crocker and Leah Light served as editors of the special section.

resistance across fields. The resistance is particularly acute in fields like psychology characterized by relatively small data sets produced by small teams of scientists through often complex experimental designs, so called “long tail” data (Ferguson, Nielson, Cragin, Bandrowski, & Martone, 2014).

First, some definitions. For the purposes of this article, we define data as the measurements, observations or facts taken or assembled *for* analysis as part of a study and upon which the results and conclusions of the study are based. Primary data are defined as raw or minimally processed data that are collected for a study. Metadata are attributes of data or a data set. Here we mostly focus on what are usually called descriptive metadata, such as the subjects used or the experimental paradigm. Our definition of data is adapted from Borgman (Borgman, 2015), who notes that as almost anything can be data, they don’t become data until they are used as evidence (pg 27). We believe that this definition is sufficiently broad to cover the diversity of data across psychological fields. We refer readers to Borgman (Borgman, 2015) for a thorough discussion of data and their place in scholarship, and specifically to chapter 2 for additional discussions about data definitions.

Many fields have recognized the value of collecting large, prospective data sets that serve as a public resource for the field (e.g., the Human Genome, Sloan Digital Sky Survey (http://www.sdss.org/), Human Connectome Project (http://www.humanconnectomeproject.org/), Allen Brain Atlas (http://brain-map.org/), Psychiatric Genomics Consortium (https://www.med.unc.edu/pgc), to name a few. These latter cases, often called “big science,” can amass much larger sample sizes compared to smaller studies. Such a scale is often required for statistically powering certain types of studies such as probing genotype-phenotype relationships. For example, the Psychiatric Genomics Consortium recently reported over 100 regions across the genome associated with schizophrenia by analyzing genomic data from over 150,000 subjects (Ripke et al., 2014).

Some research domains in psychology (e.g., developmental psychology and educational psychology), have a history of public longitudinal studies (e.g., National Longitudinal Study of Youth (https://www.nlsinfo.org/content/cohorts/nlsy79) or are beginning to assemble large data sets (e.g., psychotherapy; (Owen & Imel, 2016)). Neuropsychology/cognitive psychology is represented in the large neuroimaging data sets available (e.g., the Alzheimer’s Disease Neuroimaging Initiative, and the National Database for Autism Research, NDAR). Other examples are available from the APA website: http://www.apa.org/research/responsible/data-links.aspx

In this article, however, we focus specifically on the sharing of long tail data sets in psychology, where the rationale, benefits and mechanisms for sharing data are less clear to many. Psychology has long recognized the importance of making this type of data available. As far back as 1983, Ceci and Walker offered a proposal in the American Psychologist, the official journal of the American Psychological Association (APA), for a mandate of data sharing (Ceci, & Walker, 1983). This tradition largely covers making the data available “upon request.” That is, authors were expected to provide the primary data if requested. But current funder mandates and open science advocate calls for data sharing go well beyond this type of 1:1 transaction.

Today, data sharing has taken on a broader range of meanings, from “upon request” to making data available in supplemental files associated with a journal article to “publishing” the data through deposition into a public repository, along with any supporting materials required to interpret them, independent of consulting with the original authors (Alter & Vardigan, 2015). Publication of data, as opposed to one-on-one sharing, requires more effort on the part of the researchers and open infrastructures, including standards, to support. The most successful examples of this type of data in biomedicine are sequence and structure databases, where deposition of a gene sequence or a protein structure is a condition of publishing in many journals, e.g., Proceedings of the National Academy of Science (http://www.pnas.org/site/authors/editorialpolicies.xhtml). This type of data sharing is still in early days in psychology (although it is interesting that Ceci and Walker (1983) called for the creation of a public data bank). As more data are shared and more funders and journals are starting to push for or require data sharing, both mechanisms and best practices for sharing long tail data are starting to take shape. In the following, we consider the practice and promise of data sharing in psychology, drawing on experiences in other fields.[1]

Impetus for data sharing: Open science and e-science movements

Almost all major funding agencies in the US and abroad are developing policies around open sharing of research data and other research products (Gewin, 2016; McKiernan et al., 2016). The impetus for the push towards open science, which at a minimum encompasses the open sharing of the products of research (such as research articles, data and code) is driven by both human- and machine-based considerations. It reflects both the need to modernize our scholarly communications system in light of new technologies (the internet, cloud storage and computing, and mobile devices) and the promises of new insights to be gained from increased human- and machine-based access to research outputs.

Human considerations focus on long-standing and emerging values, such as the integrity of scientific research, which calls for greater transparency in light of recent concerns about fake data (McNutt, 2015; Yong, 2012) and scientific reproducibility (here defined as the ability to carry out the same analyses as in the original study and reach similar conclusions) in psychology and other fields (Gewin, 2016; Open Science Collaboration et al., 2015). Those concerned about animal welfare see sharing of primary data as a means of both reducing the number of animals used and ensuring that any experiments performed are more effective (http://3rs.ccac.ca/en/research/reduction/data-sharing.html). Other human values focus on the rights of the public to access research results generated from taxpayer funding. Some have criticized the field of psychology for restricting too much of its content behind a locked paywall, although that is slowly changing as APA and other publishers collaborate with PubMed and Wellcome Trust (Fischman, 2008; Guterman, 2008).

The real excitement, particularly in the age of big data and data science, is generated by machine-based access, where exposing data to new types of algorithms and the ability to combine data across studies can lead to new research questions and insights that simply

[1]Material in this article reflects issues discussed in the Dagstuhl Perspectives Workshop on Digital Scholarship and Open Science in Psychology and the Behavioral Sciences, July 19-24, 2015 (DIO: 10.4230/DagRep.5.7.42). All of the authors were participants at the Dagstuhl Workshop.

cannot be gained from analysis of individual data sets (Perrino et al., 2013). While some might argue that meta-analysis delivers some of the same benefits, our current practices of studying effect sizes from pooled data across studies historically arose because we didn't have access to raw data. Gene Glass, one of the pioneers of meta-analysis, notes: "Meta-analysis was created out of the need to extract useful information from the cryptic records of inferential data analyses in the abbreviated reports of research in journals and other printed sources…" (Glass, 2000). In fact, sharing of data sets with individual participant data can be seen as the next evolution of meta-analysis, as it allows for more powerful analyses and, particularly if negative data are shared, fewer biases than current methods (Cooper & Patell., 2009; Perrino et al., 2013). Glass himself (2000) calls forcefully for data sharing, stating "Meta-analysis needs to be replaced by archives of raw data that permit the construction of complex data landscapes that depict the relationships among independent, dependent and mediating variables."

The APA's Data Sharing Task Force report summarized the benefits of data sharing, including promoting progress, enhancing transparency, accountability and reproducibility and increasing the power and breadth of psychological studies (BSA, 2015b). Towards this end, we are seeing more calls for creating a basic culture not only for a culture of human-centric sharing (i.e., routine availability of the data underlying a paper for others to reuse), but a movement towards an e-Science vision, where researchers conduct their research digitally and where data standards and programmatic interfaces make it easy for machines to access and ingest large amounts of data (Figure 1).

eScience, like Open Science, has many flavors and does not presume a single platform or infrastructure, but does require that the basic elements of the domain (i.e., its concepts, researchers, and instruments) have a digital presence that is designed for both humans and machines. It requires that the process of translating ideas and resources into digital products through methods and protocols be built on a firm digital foundation within the originating laboratory and that this valuable provenance not be lost as data are transferred to repositories. It requires that the products of research, including the data, be easily coupled to computational resources that can operate on them at scale, and that the provenance of these research products can be tracked as they transition between users. It requires that data be dependable (see Lishner, 2012) for a concise set of core recommendations to improve the dependability of psychological research) and that it be hosted by trustworthy repositories for long term stability. It requires the ability to find, access and reuse digital artifacts using computational interfaces (e.g., API's-application programmatic interface) with minimal restrictions. As will be discussed later, these requirements have recently been succinctly articulated as the FAIR (findable, accessible, interoperable and reusable) data principles (Wilkinson et al., 2016).

Data sharing success stories

Do we have any examples that the premises and promises of open, eScience are true? That is, if researchers have access to lots of independently collected data, can they do anything significant with them? The experiences of spinal cord and traumatic brain injury in biomedicine offer one of the most compelling examples of how sharing data and new data

science approaches can lead to new findings. Translational research in both domains had a poor track record of reproducibility and in leading to clinically meaningful outcomes, despite countless promising studies in the laboratory (Steward, Popovich, Dietrich, & Kleitman, 2012). The lack of reproducibility led the US National Institute of Neurological Disease and Stroke to fund several centers dedicated to reproducing the major findings in spinal cord injury (SCI), the Facilities of Research Excellence—Spinal Cord Injury" (FORE—SCI). The results were dismal, and the complexities of reproducing paradigms and protocols across laboratories were starkly revealed. Rehabilitation psychologists and neuropsychologists are particularly familiar with these efforts, but the issue of reproducibility is at the forefront of psychology these days in many domains.

In light of these findings, the scientific community responded by making their individual data sets available (Visualized Syndromic Information and Outcomes for Neurotrauma-SCI; VISION-SCI)[2], allowing aggregation across dozens of laboratories and thousands of animals with spinal cord injury (Nielson et al., 2014). Researchers contributed not only their primary data, but also animal care and laboratory records, so called "file drawer" and "background" data (Wallis, Rolando, Borgman, ackson, & Brinkley, 2013). This approach yielded benefits arguably even beyond large, well-controlled prospective studies, because it allowed a fuller sampling of what Ferguson and colleagues call the "syndromic space" (Nielson et al., 2014), that is, the full presentation and course of a disease or syndrome given many varying initial and subsequent conditions (e.g., extent and location of injury, physiological state, level of care). Each study contributed a slice of a large multi-dimensional space representing the sequelae of spinal cord injury. The results of this effort and similar efforts in TBI led to much better predictive models, pointed towards new therapeutic areas and provided more robust cross-species biomarkers of functional recovery. It was sharing the data, including *both* positive and negative results, primary and background data, and not just hypotheses, protocols, and results that led to a greater understanding of the phenomenon. Gathering all of the heterogeneous, non-standardized data also led to efforts by NIH and the community to define common data elements (CDE's) and minimal information models that should be routinely collected by researchers in these fields (reviewed in (Ferguson et al., 2014)).

Nielson and colleagues (Nielson et al., 2015) recently published a remarkable follow up. Applying a new data algorithm, they reanalyzed the data from a large multi-center preclinical drug trial from the 1990's that failed to show efficacy due to small effect sizes. When they did a re-analysis, including the file drawer data collected from these trials, using novel machine learning technologies agnostic of hypotheses, they found large effect sizes related to perioperative neuro-critical care that masked any potential treatment effects. This new "data lens" showed that reproducible patterns did emerge from these data, despite disappointing findings in the original analyses. The re-analyses revealed new information about injury outcomes and drug effects and a completely unexpected but very robust correlation between blood pressure at time of injury and functional outcomes.

[2]These data sets were shared with Dr. Ferguson and colleagues but have not yet been made publicly available. The Spinal Cord Injury community is, however, participating in the creation of a Data Commons for the sharing of this type of data (https://scicrunch.org/odc-sci).

How this finding translates into the clinic is still being tested, and biological systems always seem to confound our best attempts at understanding them. But this example illustrates that applying new algorithms even to old data can point towards avenues for explorations that could not have been predicted reading the research reports alone. They also suggest that psychologists should be aggressive in data preservation, ensuring that valuable data are removed from file drawers and made potentially available for reuse. Such valuable data include often overlooked records such as animal care logs. As Ferguson and colleagues note, the data analyzed collectively represented a multi-million dollar investment by the NIH in what was perceived largely as a failed study (Lindsay, 2015).

Similar collaborative data sharing and synthesis activities are underway in psychology through the Collaborative Data Synthesis Study on Adolescent Depression Trials (CDSADT), an effort to test interventions to prevent or treat adolescent depression. As described in Perrino et al., (2013), CDSADT is assembling a large data set from existing randomized trials with longitudinal outcomes in order to investigate which interventions are most efficacious for different populations. As of 2015 (Perrino et al., 2015), the study had gathered de-identified data from 19 trials that had examined intervention effects in youth depression. Beyond the benefits of having a larger sample size, the authors note that combining data across studies has the potential to increase under-represented groups within the data set (Perrino et al., 2015).

Point and counterpoint: Arguments against data sharing

Counterbalancing the arguments in support of data sharing is a list of arguments against it. We focus in this section on negative attitudes towards data sharing rather than impediments that practically or legally prevent sharing data. We recognize that some sensitive data, most no-tably human subject data, cannot be shared openly without some restrictions, although there are accepted standards for making de-identified human subject data available. Similarly, often the necessary permissions have not been obtained to allow sharing, e.g., the IRB approval for collection of the data does not include explicit permission to share (Perrino et al., 2013; Lindsay, 2017). We address these issues under “Best practices for data sharing”.

Arguments against data sharing are remarkably similar across fields, including psychology (Alter & Vardigan, 2015; Eisenberg, 2015; Tenopir et al., 2011). The key objections can be summarized as follows:

1. Fear for reputation: Someone will use my data against me by finding errors in my data or statistics.
2. Fear of scooping, aka, the “research parasite” (Longo & Drazen, 2016): Someone will do an analysis that I was planning to do, and then they will claim the scientific credit for my work;
3. Fear of harassment: Release of primary data on certain subjects may open the data provider to abuse or prosecution.

4. Too much effort: Who will pay for my time and other expenses for preparing a code book and preparing the data for storage and retrieval/re-use (Baldwin & Del Re, 2016)?

5. No one will understand them: My data are too complicated to understand and making them available may lead to bad science (Longo & Drazen, 2016).

6. No one needs to understand them: My data really don't have any use beyond this study and I've already extracted any meaning from them and published the results

7. The field will stagnate, because no one will collect new data, just re-analyze the old (Roche et al., 2014).

We break these arguments down into two basic categories – (a) utility and (b) negative conse-quences to the data producer.

Utility (or the lack thereof) arguments generally posit that beyond large prospective data sets released to the public, data gathered during most intricate and complex experiments have little value because no one else would understand them. Many researchers cannot fathom how sharing their data could advance science and, indeed, fear the opposite: data in the hands of naive or malicious analysts could lead to poor science. Of course, one might argue that given the current concerns about reproducibility within the behavioral sciences (Open Science Collaboration et al., 2015), lack of sharing leads to its share of poor or at least non-reproducible science. Data science is still in its infancy, and its further development depends on a supply of data from which to learn. But as we outline in our examples above, and as the new data science shows, people find uses for data that we could not foresee.

Perhaps the most solid utilitarian argument in favor of data sharing concerns transparency and reproducibility. Regardless of whether data can be re-used, upping the standards for making such primary data available for inspection along with article publication is hard to argue against. Full access to data might be considered the ultimate evidence, as it allows an interested party to inspect the measurements made. Indeed, inspection of data or suspicions arising because the data were not available led to the uncovering of research irregularities, at best, and outright fraud, at worst, in several high profile studies (McNutt, 2015; Yong, 2012). In the days when data were stored in boxes or filing cabinets, data typically went un-inspected, but even then, graduate students in psychology were taught to ensure that all records, including data records, were preserved for five years after a study was published (Association, 2010). In fact, perusal of the average laboratory or office of older researchers reveals that we hold onto these records a lot longer, suggesting we are loathe to part with our old data. Having the original data also allows an independent investigator to replicate the results of analyses or try new analysis approaches on the same data to see if results can be reproduced.

As for the argument that "no one will ever collect new data," many data, even very large, expensive data sets, eventually become obsolete as they are supplanted by better data generated by newer and better measurement instruments and techniques. Sometimes new data cannot be generated (e.g. those generated from patient HM and other rare subjects), so

preservation of existing data is paramount. And finally, reducing the amount of new data collected is an explicit goal of the 3R (Replacement, Refinement, Reduction; (http://3rs.ccac.ca/en/research/reduction/data-sharing.html).) movement for animal studies.

The second set of arguments is largely based on possible harm to the data producer. In the most extreme cases, politically charged or controversial research, e.g., animal experiments involving companion species and primates, fully open sharing of primary data along with animal care records may make data producers vulnerable to targeting by animal rights activists, leading to physical or legal harm. Even the most ardent supporters of open science generally recognize that there are limits to openness when either researchers or subjects' safety may be compromised. But what about professional harm? Is sharing of primary data a detriment to one's career?

Researchers sharing data might be harmed if others ("data parasites" or "freeloaders") benefit from the data before the original researchers can fully mine their data. Such concerns may be particularly acute for early career researchers, who in addition may be penalized for the time and effort taken to prepare the data for publication at the expense of producing more research articles. In surveys of data sharing attitudes across scientists, Tenopir and colleagues (Tenopir et al., 2011; Tenopir et al., 2015) indeed found that younger scientists were much less likely to want to share their data than older researchers.

Researchers also fear that they will be attacked through their data, particularly in the case of replication studies, ("hostile replications," "weaponizing data"), which sometimes happens when someone finds mistakes, or when someone deliberately tries to undermine a study through multiple re-analyses of the data and selective reporting of results (Lewandowsky & Bishop, 2016). Rouder (Rouder, 2015) refers to a feeling of professional vulnerability engendered by open data, and "the self-appointed replication police who are viewed in some quarters as trying to shame otherwise good researchers" (p. 1068). Some whose work was part of the psychology replication project expressed negative feelings about the experience (Bohannon, 2014). While such concerns are very human, at least one study on the effect of retraction on scientists' reputation indicated that provided that rigorous standards were adhered to, errors and mistakes do not damage scientific reputations (Ebersole et al., 2016).

The issue of who is harmed by sharing data needs to be balanced against who is harmed by not sharing data. To the extent that making the data on which a study rests available helps root out mistakes and allows more rapid discounting of spurious effects, younger scientists and students are spared wasting valuable time and effort on unproductive avenues. In some cases, lack of transparency may lead to entire branches of a field being discredited, causing harm to all that are associated with it (Kahneman, 2012). But, science also has an obligation to those who are funding it and on whose behalf studies are performed. If new discoveries from data can occur more rapidly when many have access to the data more quickly, those who may benefit from such results, e.g., SCI patients, adolescents with depression, are harmed by any delays in the scientific process. In a well-publicized study in the British Medical Journal in 2015, "Restoring Study 329", a re-analysis of a major clinical trial testing the efficacy of two antidepressants in adolescents disputed the conclusions of the original paper (Le Noury et al., 2015). Not only did the two drugs in question have no

beneficial effect in adolescents, but their use was associated with significant adverse events. These authors concluded: “Access to primary data from trials has important implications for both clinical practice and research, including that published conclusions about efficacy and safety should not be read as authoritative. The reanalysis of Study 329 illustrates the necessity of making primary trial data and protocols available to increase the rigor of the evidence base” (p. 1).

Research communities are trying to come to grips with new regulatory policies on making data available and the professional vulnerability that comes with them by developing practices and norms that attempt to balance these needs (BSA, 2015b). For example, many journals and repositories permit embargo periods on published data, where data are held in private for a specified period of time before being released to the public. Embargo periods can range in length from a few months to two years, depending on the type of data and the repository (Assante, Candela, Castelli, & Tani, 2016). During this time, the producers of the data control access to them, so they can continue to mine them or only share with selected collaborators. Embargo periods are seen by some as a necessary and reasonable approach to making data available (see NEMJ editorial (Sharing, 2016)), while others see them as an unacceptable barrier to the open sharing of data and therefore an impediment to the progress of science.

CDSADT is establishing a collaborative model that works with researchers willing to share their data to mitigate concerns they might have and to engender a sense of shared owner-ship of the collective data (Perrino et al., 2013; Perrino et al., 2015). Investigators were invited to participate in all stages of the research, including identifying the research questions to be addressed. Towards this end, they established a facilitating team to work with and fully involve the different stakeholders in all aspects of the study. The team was careful to balance the need for researchers to come together to answer important questions about adolescent depression through data sharing with concerns expressed by individual researchers about the potential drawbacks to their individual research programs.

Although both sides of the arguments on data sharing make many claims, most of these appear in editorials, perspective pieces or the blogosphere. We have some case studies, but relatively few long-term studies at this point to buttress either side. For example, are careers significantly harmed by scooping? Piwowar & Vision (Piwowar & Vision, 2013) have studied citation rates in research papers in molecular biology that did or did not make data available. They found that most use of published microarray data sets by third parties occurred five to six years after they were published. In contrast, the original authors tended to publish within two years on these data sets. Such data might support the idea of two-year embargos, but they also suggest that the reuse was not motivated by the desire to discredit the study or act as a parasite, but rather the need for the data. The lack of studies on which to base data policies is partly due to the difficulty in our current system of tracking down papers that have re-used data (Read et al., 2015), something that those involved in data citation and resource identification are trying to rectify (Bandrowski et al., 2015; Starr et al., 2015). The APA Publications & Communication Board recommends citation of data sets reused in studies, and they may make such a practice mandatory in the future.

The CDASDT study and model for collaboration should provide an important source of data, both about researchers' experiences with sharing data and in the outcomes of creating large, synthesized pools of data (Perrino et al., 2013). As of this writing, the study is still underway, but we hope that the outcomes and the experiences of participants are well documented. An ambitious plan to measure the impact of open practices on careers has just been launched through funding by the NSF as a collaboration between UC Irvine and the Open Science Framework, a platform for open science with roots in Psychology (reported in McKiernan et al., 2016). In this study, groups of students will be trained in the use of the OSF and open science and their career outcomes will be tracked relative to those receiving more traditional training.

Whose data are they?

At the heart of many arguments around data sharing is the relationship between data, scientists and a scientific work. Are data truly disposable, or merely "a supplement to the written record of science"(Wallis et al., 2013 pg 2)? Are they private property (i.e., part of the research assets assembled over time by individual laboratories)? Or, are data an integral part of a research paper, along with the introduction, materials and methods, results and discussion? Can a data set be something more – a work of scholarship in its own right, reflecting the ingenuity and skills of those who designed the study and were able to carry out the sometimes arduous and lengthy process of data collection? Data sets should, in that case, be treated on par with narrative works and be viewed as publications in their own right. When analysis of these published data leads to discovery, they would be cited, acknowledged and rewarded alongside with making the discovery. Re-use of data is therefore celebrated rather than decried (Longo & Drazen, 2016).

Depending on whether one views data as proprietary research resources or as part of the integral output of a research study, data policies and arguments for and against sharing take on a different flavor. If data are just another research resource accrued and assembled by the researcher at great expense and effort, then they are owned by the researcher, who may be required to share them if requested. Indeed, the APA ethical guidelines explicitly state that the psychologist should "… not withhold the data on which their conclusions are based from other competent professionals who seek to verify the substantive claims through reanalysis…" (American Psychological Association, 2010), although few mechanisms exist to track compliance with these requests (Vanpaemel, Vermorgen, Deriemaecker, & Storms, 2015). Within a publication, such sharing is generally acknowledged along the lines of "We are grateful to Professor X for supplying the data used to produce this study", but not formally cited. Those who receive the data are not free to redistribute them or post modified versions without explicit permission. The producers may even expect compensation for the direct expense of generating a useable copy of their data (Vanpaemel et al., 2015). If researchers give up sole access to the data they collected, compromises such as embargo periods enable the original researchers time to fully exploit their data before others make free use of it. It also makes sense to release data under a license that restricts use of the data without special permission (i.e., no commercial use).

If, on the other hand, data are viewed as not a resource but as an integral part of a research study, “a manifest of our intellectual output” (Brembs, 2014), then do such practices make sense? If authors asked to publish only an abstract, introduction, and discussion and embargo the results for two years so that they could conduct the follow up studies, would that be acceptable? Would we require anyone who reads the paper to contact the author if they intended to try to replicate it? Would we forbid anyone to use the results contained within the paper for other studies or let them use the results only if they don’t publish on them? Would we require approval from an editorial review board to use the results and then only for 12 months? Of course not. Scientists are expected to publish their intellectual output. We expect that scientists will include the results and provide the evidence on which these results rest. We would expect others to point out flaws in our data, just as they do our results and arguments. We would expect that people would try to build from what we had published and derive new insights. And, we would celebrate the citations of our data and code just as we prize the citations of our papers as a measure of our impact. Our published data become part of the “shoulders of giants” on which others stand.

Incentives for data publishing

In discussions of data publishing, the benefits are often couched in terms of benefits to science and society, while the downsides are often framed in terms of negative impacts on the data producer. Thus, there is increasing recognition that data publishing must be rewarded or at least recognized. The National Institute of Health is investing in the creation of a data discovery system akin to Pub Med for discovering and re-using data sets (datamed.org). The NSF now allows data sets and code to be listed alongside articles in a CV. The Association for Psychological Science has adopted the use of Open Science Foundation badges in its journal *Psychological Science* to recognize authors who share their data and materials (Eich, 2014) According to analyses, display of these badges is correlated with significantly increased sharing of data (Lindsay, 2017; Kidwell et al., 2016). For an example of badges, see (Farmer, Baron-Cohen, & Skylark, 2017)). Piwowar (Piwowar & Vision, 2013) first identified a significant increase in citations to papers that made data available for reuse, which has since been replicated in other domains (see McKiernan et al., 2016).

Efforts are also underway to encourage authors to share their data by tapping into the current paper-based reward system. The APA open journal, *Archives of Scientific Psychology*, has as a policy that re-users of data offer the collectors co-authorship on a paper (Cooper, & VandenBos, G. R., 2013). New types of journals and article types specifically for making data available are coming into existence. So called data journals and data papers differ from regular research reports in that they are based on the rich description and curation of a data set rather than analysis of it, e.g. Yu et al., (2017). Data papers require that the data be deposited in an approved repository and offered under an open license and sometimes that additional data standards be enforced. Examples of data journals include the *Journal of Open Psychology Data, Nature Scientific Data*, and *GigaScience*, but many traditional journals also accept data papers (e.g., *Brain and Behavior*). Both the data and the publication are indexed by citation indexes.

Many publishers are adopting more formal systems for data citation data sets so that they are cited the same way as articles (Altman & Crosas, 2014). In our current publishing system, authors adopt a variety of styles for referencing data when they are re-used, from accession numbers, to URL's, to citing a paper associated with the data set. Some journals set aside a special section on data use which contain lists of data sets and other resources. However, these practices make it difficult to track when and where data are cited (Read et al., 2015). A formal citation system assigns credit for the re-use of data and establishes links to the evidence on which claims are based while providing the means for tracking and therefore measuring impact of data re-use. Citations to data sets would look like citations to articles, with a standard set of metadata, and would appear in the reference list. With the ability to list published datasets on a scientific CV, cite them within published articles, and search for them via tools like DataMed, data would finally take their place as a primary product of scholarship.

But besides funder mandates and credit for published data, are there other more immediate benefits that accrue to those who publish their data?

Better data management

Studies on the availability of research data in psychology and other fields note that many labs were unable to comply with requests for data simply because they no longer could access the data themselves (Vanpaemel et al., 2015). Whether the data were inadvertently lost with system failures, in an outdated format, could not be located, or were impenetrable to human understanding because the graduate student or post-doc left the laboratory, many labs are plagued by poor data management practices. The difference between searching across file systems, floppy disks or previous students and lab employees for data vs a simple query to an online data repository represents an enormous cost savings. Platforms like the Open Science Framework, Zenodo, GitHub and Figshare, provide tools to help researchers manage their data from project inception, which makes it easier to release data to the public when it is ready. Textbooks, such as Cooper, (2016), directly demonstrate and train graduate students and early career professionals how to efficiently and effectively input, code, and manage data as part of the ethical component to the research enterprise.

Better data

Taking the premise that no scientist wants to be sloppy, but errors do occur, both research (Wicherts, Bakker, Molenaar, Rouder, & Iverson, 2011) and anecdotal evidence from scientists (e.g., Rouder, 2015), indicate that knowing that data will become available tends to make researchers more careful with their data. Mistakes and errors can be costly, and so being forced to pay attention to the integrity and documentation of the data from the outset can potentially lead to significant returns later on.

Establishing trust

The experience of the spinal cord injury community highlights the importance of shared data in establishing trustworthiness in domains where reproducibility of results is, or is perceived to be, a problem. With the recent well publicized results of the Reproducibility Project, psychology got a lot of bad press about the low level of reproducibility and questionable

research practices (Vanpaemel et al., 2015). Such bad press can lead to a crisis in confidence in the integrity and reliability of a field, and potentially harm all those who are associated with a discipline, particularly young scientists starting out (Kahneman, 2012). As experienced by the SCI community, making the data available signals transparency and trust in the process of self-correction in science.

Are psychologists publishing their data?

In response to the promise of open data in other fields and funder mandates, the APA Publications & Communications Board convened two task forces in 2011–2012 that recommended that the APA journals program adopt more open data policies that emphasize proper data management in the laboratory and the publishing of data in suitable repositories[3]. To move these recommendations into practice, the APA launched the *Archives of Scientific Psychology*, "an open methodology, collaborative data sharing, open access journal" (Cooper, 2013). Data associated with these publications is deposited in the Interuniversity Consortium for Political and Social Research (ICPSR) data repository. However, reuse of these data currently requires approval from the Archives Review Committee for the APA, and is granted for 12 months only (Data Sharing Policies, Archives of Scientific Psychology). Perhaps because of the strict policies on data sharing, uptake of the journal has been rather modest, with only 30 articles published to date (http://psycnet.apa.org/PsycARTICLES/journal/arc/5/2, accessed Aug 27, 2017). More recently, the APA Board of Scientific Affairs convened a Working Group on Data Sharing, which produced a report supporting such initiatives (BSA, 2015a), although no formal governance-wide policy on data sharing has been approved as overall APA policy. *Psychological Science* has also recently introduced policies to make it easier for reviewers to access data and is strongly encouraging authors to make their data available (Lindsay, 2017).

Tenopir and colleagues as part of the NSF-funded Data One project surveyed scientists' attitudes and practices regarding data sharing in 2011 and again in 2015 (Tenopir et al., 2011; Tenopir et al., 2015). Their samples included social scientists, including psychologists. In their 2015 follow-up, psychologists were among the most negative about data sharing, with 30% of respondents indicating that data should not be shared, although the sample was small (21 respondents in the follow up study). Nevertheless, evidence that voluntary sharing of data is occurring is available. The APA, which publishes 92 journals, began in 2003 to allow authors to voluntarily add supplemental files of stimulus material, datasets, and program to the digital version of their journal articles. By mid-2016, there were a total of 3,363 articles in the PsycARTI-CLES database with attached supplemental files. The use of supplemental files has increased each year since the introduction of this feature, which reached its peak in 2016 with 808 supplemental files being added. Among the total pool, one out of 10 supplemental files is a dataset. The APA abstract database, PsycINFO,

[3]Harris Cooper, Mark Applebaum, and Gary VandenBos were actively involved in urging the P&C Board to establish the first and second P&C Board Data Sharing Task Forces. They also played a key role in encouraging the APA Board of Scientific Affairs to undertake an independent analysis of the role and need for data sharing initiatives in psychology. Harris Cooper and Gary VandenBos proposed the concept of the Archives of Scientific Psychology to the P&C Board, and they were later invited to become the Founding Editors.

contains another 1,622 additional records that link to articles in non-APA journals which include datasets.

Both the studies of Tenopir (Tenopir et al., 2015) and Vanpaemel and colleagues (Vanpaemel et al., 2015) indicate that routine deposition of data within a public data repository is currently far from the norm in psychology. In the Vanpaemel study, which requested data from over 300 recent papers in psychology, only four authors pointed them to data in public repositories. However, a search across public databases for “psychology” returns several thousand data sets (Table 1). One psychologist, Rouder (Rouder, 2015), makes his data “born open” through a computational pipeline that automatically uploads all data generated in the lab to a public repository. Thus, both through supplemental materials and deposition in repositories, these statistics reflect a small but growing willingness for some psychologists to release their data publicly.

Best practices in data sharing

Whether or not psychology as a field fully embraces data sharing at this time, scholarship continues to transition to digital and the trend with funder and journal mandates is towards greater transparency and sharing (Gewin, 2016). Communities’ attitudes towards data sharing change over time (Tenopir et al., 2015). For example, the neuroimaging community objected vociferously to making MRI data available in 2002, but now is characterized by many public data sets and specialized repositories (Poline et al., 2012). Therefore, for the remainder of this article, we focus less on the why and more on the how. What are the best practices for publishing research data and what are practical steps to achieve them? It is interesting to note that the astronomy community, which is characterized by large, shared pools of data via several large repositories and a very positive attitude towards data sharing (Tenopir et al., 2015), also is unaware of best practices for sharing their personal data sets (Pepe, Goodman, Muench, Crosas, & Erdmann, 2014). Many of the recommendations and issues for sharing data effectively will be the same whether data are viewed as a research resource or a research product, but data as a research product entails additional levels of attention, much as publishing a full research paper requires more attention than writing up notes.

Recently, a lot of attention has been paid in biomedicine to the FAIR principles for data (Wilkinson et al., 2016). FAIR stands for Findable, Accessible, Interoperable and Re-usable. FAIR principles lay out a set of attributes for data to be maximally re-usable, for both humans and machines (elaborated in Table 2). The FAIR principles give broad guidance as to how to make prepare and maintain data for sharing, but leave the precise interpretations and implementations up to individual communities (Mons et al., 2017).

An example in the behavioral sciences of making data FAIR is the Research Domain Criteria (RDOC), an National Institute of Mental Health initiative that illustrates an important effort in data sharing, storing, and standardization in psychopathology (Sanislow et al., 2010). RDoC is a research framework that integrates many levels of information (from genomics to self-report) to better understand basic dimensions of functioning underlying the full range of human behavior from normal to abnormal. The initiative’s goal is to “devise a system for

identifying and integrating constructs for disordered cognitive, neural, genetic, or epigenetic processes that hold particular promise to [explain] psychiatric symptoms" (Sanislow et al., 2010) p. 633).

The RDoCdb provides the research community a data repository for the harmonization and sharing of research data related to this initiative, and also accepts and shares human subjects data related to all fields of mental health research. The RDOC initiative delivers FAIR data; the infrastructure allows data sets to be ***Findable*** as it assigns unique identifiers to the datasets, data and metadata are indexed in a searchable database and, throughout the design of the RDOC infrastructure there is a rich metadata and data management layer. It is ***Accessible*** as data and metadata are retrievable by the identifier over a standard communication protocol and the standards themselves are open and available. It is ***Interoperable*** as data and metadata are represented in a formal, accessible, and shared language and, it includes references to other data. RDOC data are also ***Reusable*** as they have a clear data usage policy addressing both, technical and legal aspects; data and metadata elements are the result of a community agreement and data have well defined provenance.

Practical steps towards FAIR

As is evident from the elements involved in RDoC, creating a data ecosystem and making data fully FAIR requires significant effort and components such as open instruments, formats, and vocabularies. Components like common data elements (CDE's) and minimal information models also play a role. It requires sufficient resources-funds, personnel and tools- for researchers to prepare FAIR data, even at the most basic level. And, as outlined in Figure 1, it requires that we not just devote attention to the data set itself, but to the processes that led to the creation of the data, specifically, the methods and protocols.

Achieving the envisioned level of interoperability and reusability is currently beyond reach for many fields at this stage (Wilkinson et al., 2016). But many reasonable steps that can be taken by researchers right now to improve access to data are outlined in a recent editorial by Lindsey (2017). The simplest and most effective mechanism for publishing data is to deposit data in a qualified data repository (Gewin, 2016; Roche et al., 2014; White et al., 2013; Nosek et al., 2015). Qualified data repositories ensure long term maintenance of the data, generally enforce community standards, and handle things like obtaining an identifier, maintaining a landing page with appropriate descriptive metadata, and providing programmatic access. Analysis of "link rot", that is URL's that are broken, within published papers indicate that links to data in repositories are more stable than to those on individual's servers (Pepe et al., 2014). A variety of types of repositories are available, from specialized repositories developed around a specific domain (e.g., ICPSR), or data type (e.g., NITRC-IR, openfMR)I, but also general repositories that will take all domains and most data types (e.g., Figshare, Dryad, OSF, DataVerse, Zenodo – see Table 3). Many research institutions are maintaining data repositories for their researchers as well (e.g., University of California DASH).

Compared to bioinformatics, earth sciences and neuroscience, psychology and other social and behavioral sciences have invested in relatively few open domain-specific repositories. The APA is using the ICPSR as the repository for their open journal, the *Archives of Scientific Psychology*. PsycTESTS and the Mental Measurements Yearbook are the go-to resource for psychology instruments, protocols and surveys, but these are not open.

But even without the advantages of specialist repositories, depositing data in any competent repository ensures that data are recoverable. In the SCI example given earlier, the data had to be meticulously gathered from laboratories and in some cases, valuable data such as the animal care records were in danger of being destroyed. Significant time and expense had to be expended just obtaining and digitizing the data, even before the laborious process of cleaning and harmonizing the data could begin. Had the data been deposited in a repository, even as a compressed archive with minimal to no curation, the effort that went into gathering the data and amount of data lost would have been significantly less. In other words, at the most basic level, deposition in a repository makes data potentially re-usable. Not without effort and cost, but if the data are gone, no amount of effort or cost will suffice. And in the age of Big Data, machine-based learning and AI, we cannot know how much curation will be necessary in the future to make data fully interoperable and re-usable.

Several helpful reviews and textbooks are now available to train researchers in the proper handling and preparation of data for sharing (e.g., Cooper, 2016; Gewin, 2016; White et al., 2013). But the essence of effective data publishing was summarized succinctly by White and colleagues (White et al., 2013):

1. Well documented data are easier to understand (to which we add: and generally involve a code book and well described methods and protocols);
2. Properly formatted data are easier to use in a variety of software,
3. Data that are shared in established repositories with no or minimally restrictive licenses are easier for others to find and use.

To which we would add a fourth: Prepare to share. It is essential to plan from the inception of the study to make the data as open as possible by securing all necessary permissions to publish the data and associated materials like protocols, and that the necessary resources are available to do any preparation work. Such plans should include ensuring appropriate IRB approvals for sharing human subjects data or other background materials like animal care records which may belong to the university, and resolving any intellectual property concerns with your institution and colleagues ***before*** the data are collected (Carroll et al., 2015; Steneck, 2007).

Conclusions

Whether we are talking about big data, small data, open data, or dark data, we are still very much in the early days of sharing, publishing and re-using data. But thanks to significant investments in infrastructure for managing, hosting and distributing research data over the past few decades, the barriers to wholesale and routine data sharing are perceived as largely cultural and no longer purely technological (Sablonniere, Auger, Sabourin, & Newton,

2012). Best practices and guidance on how to manage data in psychology and other social and behavioral sciences, from inception to publishing, are starting to become available (Cooper, 2016).

Yes, psychology and most of the other social and behavioral sciences have a long way to go before they fully transition to an eScience. Psychology must commit to and invest in, the necessary components of this ecosystem, including open formats, data standards, tools and ontologies. But, eScience runs on data. For most long tail data, both the means and the mechanisms for publishing of data so that they are findable and accessible are in reach. The field of psychology must make the commitment to mandate it. More journals should start implementing the Transparency and Openness Guidelines issued by the Center for Open Science (Nosek et al., 2015), which lay out a series of requirements for moving towards data publishing.

There are still many challenges facing science as it transitions more and more of its assets to digital: long term sustainability and preservation of data, dealing with extremely large and complicated data sets, citing and publishing slices of longitudinal and dynamic data sets. Many domains are working on solutions for these (e.g., Research Data Alliance (https://www.rd-alliance.org/data-citation-making-dynamic-data-citable-wg-update.html); (Honor, Haselgrove, Frazier, & Kennedy, 2016). But some questions about the how and why of data sharing will only be answered when we have a lot of data with which to answer them. So, we believe the time is ripe to ramp up efforts in psychology to make data open, FAIR and citable.

Acknowledgments

The authors wish to thank Drs. Jessica Nielson and Adam Ferguson for their helpful comments. AGC acknowledges the KOPAR, H2020-MSCA-IF-2014, Grant Agreement Number 655009.

Biographies

Maryann E. Martone

Alexander Garcia-Castro

Gary R. VandenBos

References

Alter GC, Vardigan M. Addressing Global Data Sharing Challenges. Journal of Empirical Research on Human Research Ethics. 2015; 10:317–323. DOI: 10.1177/1556264615591561 [PubMed: 26297753]

Altman, M., Crosas, M. The Evolution of Data Citation: From Principles to Implementation. IASSIST Quarterly. 2014. Retrieved from http://informatics.mit.edu/publications/evolution-data-citation-principles-implementation

Association, A.P. Ethical Principles of Psychologists and Code of Conduct. American Psychological Association. 2010. Retrieved from http://www.apa.org/ethics/code/index.aspx

Assante M, Candela L, Castelli D, Tani A. Are Scientific Data Repositories Coping with Research Data Publishing? Data Science Journal. 2016; 15doi: 10.5334/dsj-2016-006

Baldwin SA, Del Re AC. Open access meta-analysis for psychotherapy research. Journal of Counseling Psychology. 2016; 63:249–260. DOI: 10.1037/cou0000091 [PubMed: 27078196]

Bandrowski A, Brush M, Grethe JS, Haendel MA, Kennedy DN, Hill S, Vasilevsky N. The Resource Identification Initiative: A cultural shift in publishing. F1000Research. 2015; 4doi: 10.12688/f1000research.6555.2

Bohannon J. Psychology. Replication effort provokes praise–and 'bullying' charges. Science (New York, NY). 2014; 344:788–789. DOI: 10.1126/science.344.6186.788

Borgman, CL. Big data, little data, no data: scholarship in the networked world. Cambridge, MA: MIT Press; 2015.

Brembs, B. What is the difference between text, data and code?. Blog. 2014. Retrieved from http://bjoern.brembs.net/2014/03/what-is-the-difference-between-text-data-and-code/#annotations:Czg4floAEeaBzgNu7V4RAA

BSA. 2015 Annual Report. 2015a. Retrieved from http://www.apa.org/science/leader-ship/bsa/2015-report.aspx

BSA. Data Sharing: Principles and Considerations for Policy Development. American Psychological Association's Board of Scientific Affairs; 2015b. Retrieved from http://www.apa.org/science/leadership/bsa/data-sharing-report.pdf

Carroll MW, Chylinski K, Fonfara I, Hauer M, Doudna JA, Charpentier E. Sharing Research Data and Intellectual Property Law: A Primer. PLOS Biology. 2015; 13:e1002235.doi: 10.1371/journal.pbio.1002235 [PubMed: 26313685]

Ceci SJ, Walker E. Private Archives and Public Needs. American Psychologist. 1983; 38:414–423.

Cooper H, VandenBos GR. Archives of scientific psychology: A new journal for a new era. Archives of Scientific Psychology. 2013; 1(1):1–6.

Cooper, HM. Ethical choices in research: managing data, writing reports, and publishing results in the social sciences. American Psychological Association; 2016.

Cooper HP, Erika A. The Relative Benefits of Meta-Analysis Conducted With Individual Participant Data Versus Aggregated Data. Psychological Methods. 2009; 14(2):165–176. DOI: 10.1037/a0015565 [PubMed: 19485627]

Ebersole CR, Axt JR, Nosek BA, Borsboom D, Bowman S, Breckler S. Scientists' Reputations Are Based on Getting It Right, Not Being Right. PLOS Biology. 2016; 14:e1002460.doi: 10.1371/journal.pbio.1002460 [PubMed: 27171138]

Eich E. Business Not as Usual. Psychological Science. 2014; 25:3–6. DOI: 10.1177/0956797613512465 [PubMed: 24285431]

Eisenberg N. Thoughts on the Future of Data Sharing - Association for Psychological Science. APS Observer. 2015; 28 Retrieved from http://www.psychologicalscience.org/index.php/publications/observer/2015/may-june-15/thoughts-on-the-future-of-data-sharing.html#annotations:41mr2mMKEea7tlNoNJ_84A.

Farmer GD, Baron-Cohen S, Skylark WJ. People With Autism Spectrum Conditions Make More Consistent Decisions. Psychological Science. 2017; 28:1067–1076. DOI: 10.1177/0956797617694867 [PubMed: 28635378]

Ferguson AR, Nielson JL, Cragin MH, Bandrowski AE, Martone ME. Big data from small data: data-sharing in the'long tail'of neuroscience. Nature neuroscience. 2014; 17(11):1442–1447. [PubMed: 25349910]

Fischman, J. About-Face: Psychological Association Will Not Charge for Open Access. Chronicle of Higher Education. 2008. Retrieved from http://www.chronicle.com/article/About-Face-Psychological/41329

Gewin V. Data sharing: An open mind on open data. Nature. 2016; 529:117–119. DOI: 10.1038/nj7584-117a [PubMed: 26744755]

Glass, GV. Meta-Analysis at 25. 2000. Retrieved from http://www.gvglass.info/papers/meta25.html

Guterman, L. Psychological Association Will Charge Authors for Open-Access Archiving - The Chronicle of Higher Education. Chronicle of Higher Education. 2008. Retrieved from http://www.chronicle.com/article/Psychological-Association-Will/41311

Honor LB, Haselgrove C, Frazier JA, Kennedy DN. Data Citation in Neuroimaging: Proposed Best Practices for Data Identification and Attribution. Frontiers in Neuroinformatics. 2016; 10:34.doi: 10.3389/fninf.2016.00034 [PubMed: 27570508]

Kahneman, D. A proposal to deal with questions about priming effects. 2012. Retrieved from http://www.nature.com/polopoly_fs/7.6716.1349271308!/suppinfoFile/KahnemanLetter.pdf

Kidwell MC, Lazarević LB, Baranski E, Hardwicke TE, Piechowski S, Falkenberg LS, Nosek BA. Badges to Acknowledge Open Practices: A Simple, Low-Cost, Effective Method for Increasing Transparency. PLOS Biology. 2016; 14:e1002456.doi: 10.1371/journal.pbio.1002456 [PubMed: 27171007]

Le Noury J, Nardo JM, Healy D, Jureidini J, Raven M, Tufanaru C, Abi-Jaoude E. Restoring Study 329: efficacy and harms of paroxetine and imipramine in treatment of major depression in adolescence. BMJ (Clinical research ed). 2015; 351:h4320.doi: 10.1136/bmj.h4320

Lewandowsky S, Bishop D. Research integrity: Don't let transparency damage science. Nature. 2016; 529:459–461. DOI: 10.1038/529459a [PubMed: 26819029]

Lindsay DS. Sharing Data and Materials in Psychological Science. Psychological Science. 2017; 28:699–702. DOI: 10.1177/0956797617704015 [PubMed: 28414920]

Lindsay, G. The Latest Medical Breakthrough In Spinal Cord Injuries Was Made By A Computer Program. CoExist. 2015. Retrieved from http://www.fastcoexist.com/3052282/the-latest-medical-breakthrough-in-spinal-cord-injuries-was-made-by-a-computer-program

Lishner DA. A concise set of core recommendations to improve the dependability of psychological research. Review of General Psychology. 2012; 19:52–68. DOI: 10.1037/gpr0000028

Longo DL, Drazen JM. Data Sharing. New England Journal of Medicine. 2016; 374:276–277. DOI: 10.1056/NEJMe1516564 [PubMed: 26789876]

McKiernan EC, Bourne PE, Brown CT, Buck S, Kenall A, Lin J, Yarkoni T. How open science helps researchers succeed. eLife. 2016; 5doi: 10.7554/eLife.16800

McNutt, M. Editorial retraction. Science. 2015. Retrieved from http://science.science-mag.org/content/early/2015/05/27/science.aac6638

Mons B, Neylon C, Velterop J, Dumontier M, da Silva Santos LOB, Wilkinson MD. Cloudy, increasingly FAIR; revisiting the FAIR Data guiding principles for the European Open Science Cloud. Information Services & Use. 2017; 37:49–56. DOI: 10.3233/ISU-170824

Nielson JL, Paquette J, Liu AW, Guandique CF, Tovar CA, Inoue T, Ferguson AR. Topological data analysis for discovery in preclinical spinal cord injury and traumatic brain injury. Nature communications. 2015; 6:8581.doi: 10.1038/ncomms9581

Nielson JL, Guandique CF, Liu AW, Burke DA, Lash AT, Moseanko R, Ferguson AR. Development of a Database for Translational Spinal Cord Injury Research. Journal of Neurotrauma. 2014; 31:1789–1799. DOI: 10.1089/neu.2014.3399 [PubMed: 25077610]

Nosek BA, Alter G, Banks GC, Borsboom D, Bowman SD, Breckler SJ, Chambers CD. Promoting an open research culture. Science. 2015; 348:1422–1425. DOI: 10.1126/science.aab2374 [PubMed: 26113702]

Open Science Collaboration, O. S. Hempel C, Hempel C, Oppenheim P, Meehl PE, Platt JR, Cumming G. Estimating the reproducibility of psychological science. Science (New York, NY). 2015; 349:aac4716.doi: 10.1126/science.aac4716

Owen J, Imel ZE. Introduction to the special section “Big‘er’ Data”: Scaling up psychotherapy research in counseling psychology. Journal of Counseling Psychology. 2016; 63:247–248. DOI: 10.1037/cou0000149 [PubMed: 27078195]

Pepe A, Goodman A, Muench A, Crosas M, Erdmann C. How do astronomers share data? Reliability and persistence of datasets linked in AAS publications and a qualitative study of data practices among US astronomers. PloS one. 2014; 9:e104798.doi: 10.1371/journal.pone.0104798 [PubMed: 25165807]

Perrino T, Howe G, Sperling A, Beardslee W, Sandler I, Shern D, Brown CH. Advancing Science Through Collaborative Data Sharing and Synthesis. Perspectives on Psychological Science. 2013; 8:433–444. DOI: 10.1177/1745691613491579 [PubMed: 24244216]

Perrino T, Beardslee W, Bernal G, Brincks A, Cruden G, Howe G, Brown CH. Toward scientific equity for the prevention of depression and depressive symptoms in vulnerable youth. Prevention science: the official journal of the Society for Prevention Research. 2015; 16:642–651. DOI: 10.1007/s11121-014-0518-7 [PubMed: 25349137]

Piwowar HA, Vision TJ. Data reuse and the open data citation advantage. PeerJ. 2013; 1:e175.doi: 10.7717/peerj.175 [PubMed: 24109559]

Poline JB, Breeze JL, Ghosh S, Gorgolewski K, Halchenko YO, Hanke M, Kennedy DN. Data sharing in neuroimaging research. Frontiers in Neuroinformatics. 2012; 6:9.doi: 10.3389/fninf.2012.00009 [PubMed: 22493576]

Read KB, Sheehan JR, Huerta MF, Knecht LS, Mork JG, Humphreys BL, Sicilia MA. Sizing the Problem of Improving Discovery and Access to NIH-Funded Data: A Preliminary Study. PLOS ONE. 2015; 10:e0132735.doi: 10.1371/journal.pone.0132735 [PubMed: 26207759]

Ripke S, Neale BM, Corvin A, Walters JTR, Farh KH, Holmans PA, O’Donovan MC. Biological insights from 108 schizophrenia-associated genetic loci. Nature. 2014; 511:421–427. DOI: 10.1038/nature13595 [PubMed: 25056061]

Roche DG, Lanfear R, Binning SA, Haff TM, Schwanz LE, Cain KE, Viney I. Troubleshooting Public Data Archiving: Suggestions to Increase Participation. PLoS Biology. 2014; 12:e1001779.doi: 10.1371/journal.pbio.1001779 [PubMed: 24492920]

Rouder JN. The what, why, and how of born-open data. Behavior Research Methods. 2015; :1–8. DOI: 10.3758/s13428-015-0630-z [PubMed: 24683129]

Sablonniere RDL, Auger E, Sabourin M, Newton G. Facilitating Data Sharing in the Behavioural Sciences. Data Science Journal. 2012; 11:DS29–DS43. DOI: 10.2481/dsj.11-DS4

Sanislow CA, Pine DS, Quinn KJ, Kozak MJ, Garvey MA, Heinssen RK, Cuthbert BN. Developing constructs for psychopathology research: research domain criteria. Journal of abnormal psychology. 2010; 119:631–639. DOI: 10.1037/a0020909 [PubMed: 20939653]

Sharing. Toward Fairness in Data Sharing. New England Journal of Medicine. 2016; 375:405–407. DOI: 10.1056/NEJMp1605654 [PubMed: 27518658]

Starr J, Castro E, Crosas M, Dumontier M, Downs RR, Duerr R, Clark T. Achieving human and machine accessibility of cited data in scholarly publications. PeerJ Computer Science. 2015; 1:e1.doi: 10.7717/peerj-cs.1

Steneck, NH. ORI Introduction to the Responsible Conduct of Research. ORI - The Office of Research Integrity; 2007.

Steward O, Popovich PG, Dietrich WD, Kleitman N. Replication and reproducibility in spinal cord injury research. Experimental neurology. 2012; 233:597–605. DOI: 10.1016/j.expneurol.2011.06.017 [PubMed: 22078756]

Tenopir C, Allard S, Douglass K, Aydinoglu AU, Wu L, Read E, Frame M. Data Sharing by Scientists: Practices and Perceptions. PLoS ONE. 2011; 6:e21101.doi: 10.1371/journal.pone.0021101 [PubMed: 21738610]

Tenopir C, Dalton ED, Allard S, Frame M, Pjesivac I, Birch B, Dillman DA. Changes in Data Sharing and Data Reuse Practices and Perceptions among Scientists Worldwide. PLOS ONE. 2015; 10:e0134826.doi: 10.1371/journal.pone.0134826 [PubMed: 26308551]

Vanpaemel W, Vermorgen M, Deriemaecker L, Storms G. Are We Wasting a Good Crisis? The Availability of Psychological Research Data after the Storm. Collabra. 2015; :1.doi: 10.1525/collabra.13

Wallis JC, Rolando E, Borgman CL, Ackson TB, Brinkley JF. If We Share Data, Will Anyone Use Them? Data Sharing and Reuse in the Long Tail of Science and Technology. PLoS ONE. 2013; 8:e67332.doi: 10.1371/journal.pone.0067332 [PubMed: 23935830]

White EP, Baldridge E, Brym ZT, Locey KJ, McGlinn DJ, Supp SR. Nine simple ways to make it easier to (re)use your data Ideas in Ecology and Evolution. 2013:6.

Wicherts JM, Bakker M, Molenaar D, Rouder JN, Iverson GJ. Willingness to Share Research Data Is Related to the Strength of the Evidence and the Quality of Reporting of Statistical Results. PLoS ONE. 2011; 6:e26828.doi: 10.1371/journal.pone.0026828 [PubMed: 22073203]

Wilkinson MD, Dumontier M, Aalbersberg IJJ, Appleton G, Axton M, Baak A, Musen MA. The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data. 2016; 3:160018.doi: 10.1038/sdata.2016.18 [PubMed: 26978244]

Yong E. The data detective. Nature. 2012; 487:18–19. DOI: 10.1038/487018a [PubMed: 22763527]

Yu H, Shen Z, Miao C, Leung C, Chen Y, Fauvel S, Salmon CT. A dataset of human decision-making in teamwork management. Scientific Data. 2017; 4:160127.doi: 10.1038/sdata.2016.127 [PubMed: 28094787]

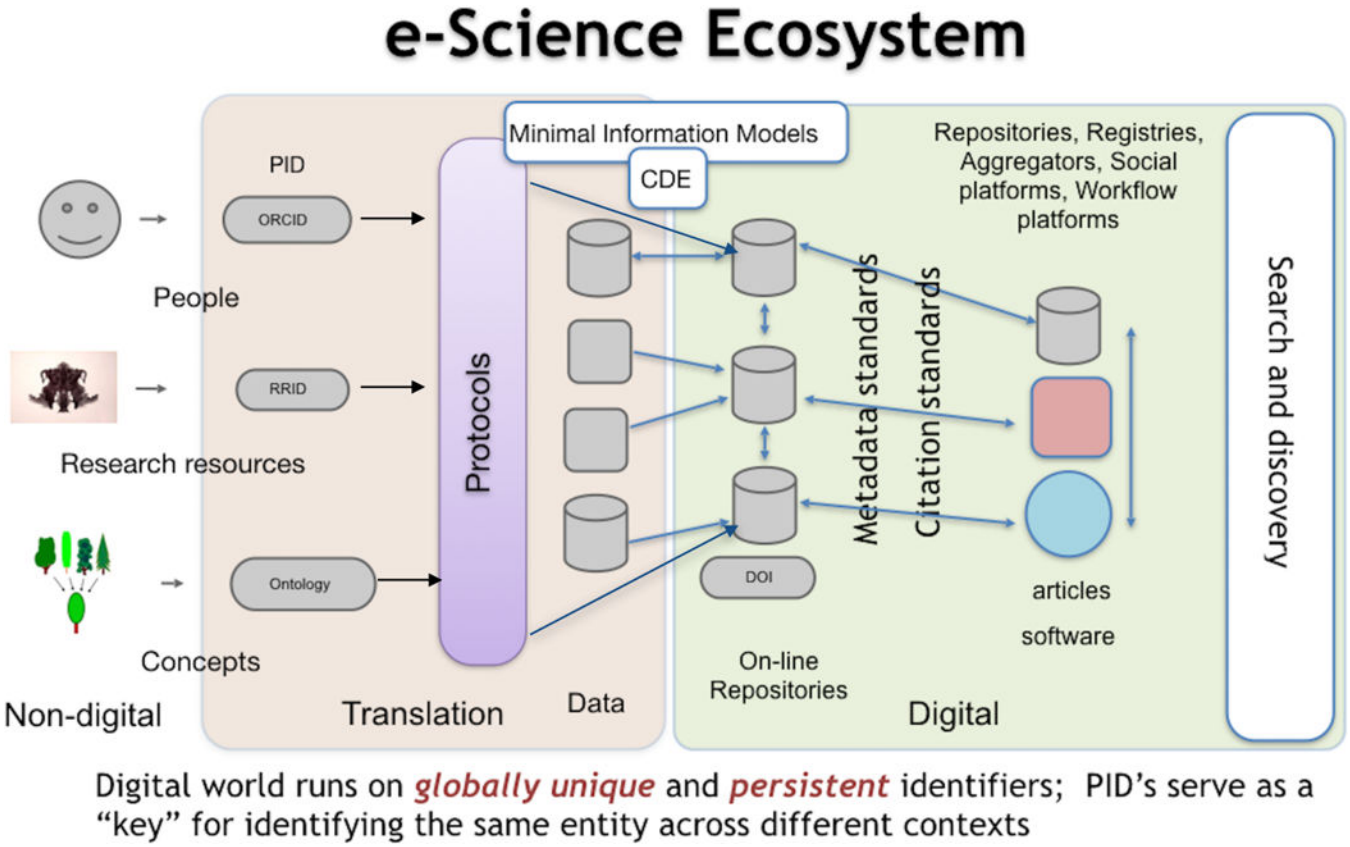

Figure 1.

Components of an e-Science ecosystem and the transition of entities from the non-digital to the digital environment. Each of these external entities, e.g., people, concepts is ideally accompanied by a persistent identifier (PID) that uniquely identifies the entity in a machine friendly format. For example, the identifier system for people is the ORCID; for research resources, the Research Resource Identifiers (RRIDs; Bandrowski et al., 2015). Experiments performed within the laboratory result in digital data, which may be held locally, but which should transition into the e-Science ecosystem by deposition in a reputable repository. Deposition should also include a complete accounting of the protocols and methods used to produce the data. Repositories assign a persistent identifier like a DOI (Digital Object Identifier). Additional standards, e.g., minimal information models, common data elements (CDE's) can be implemented within the laboratory to make this transition more seamless. CDE's provide standard ways of collecting certain types of data, e.g., demographic data. Once within the eScience ecosystem, these objects are able to be re-used. Provenance is tracked programmatically as data are combined with other data, transformed by software or referenced within an article. Data as well as other research objects are designed so that they are amenable to search and discovery (see discussion of FAIR principles later in the article).

Table 1

Psychology representation in major on-line repositories. Unless indicated, the repositories were searched through their on-line user interface for the term “psychology”. For mixed repositories that house more than just data sets, e.g., OSF, GitHub, we used the “data set” filter when available. For OSF and GitHub, it was not possible to filter for data sets specifically.

Repository	Records returned	Data accessed + comments
Dryad	52	8/20/2016
ICPSR	1198	8/20/2016
DataVerse	1219	8/20/2016; searched the Harvard Dataverse, which is open for data deposition from all institutions
Psychology Dataverse	3	8/20/2016
Open Science Framework	540 Registrations; 731 projects	8/20/2016; Cannot search for datasets directly
Zenodo	10	8/20/2016; Searched psychology + dataset filter
GitHub	44	Search: Psychology data
Figshare	19,803	8/22/2016; Obtained from FigureShare personnel (personal communication) as the number of hits is not listed on website
Data Cite	1, 203	8/22/2016; Searched psychology + dataset filter

Table 2

The FAIR principles and their their details (column 2) from Wilkinson et al., (2016). Some implications are presented in column 3.

Principle	Details	Some implications
Findable	F1. (meta)data are assigned a globally unique and persistent identifier F2. data are described with rich metadata F3. metadata clearly and explicitly include the identifier of the data it describes F4. (meta)data are registered or indexed in a searchable resource	"Findable" means both that data are annotated in a form that makes them easy to discover and that their location is reliably found. Almost everyone has experienced frustration with hyperlinks in papers that are broken. Thus, the principles recommend that data be given a persistent identifier (e.g., a DOI), that resolves to the data independently of where the data are stored.
Accessible	A1. (meta)data are retrievable by their identifier using a standardized communications protocol A1.1 the protocol is open, free, and universally implementable A1.2 the protocol allows for an authentication and authorization procedure, where necessary A2. metadata are accessible, even when the data are no longer available	For humans, accessible also means providing information on where and how to obtain the data. ***Accessible does not mean that all data are available to everyone, all the time***. But even for protected data, it is often possible to share basic descriptive metadata that can let researchers know such data is available if they have the necessary qualification to access them.
Interoperable	I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation. I2. (meta)data use vocabularies that follow FAIR principles I3. (meta)data include qualified references to other (meta)data	The requirements for interoperability depend largely on the implementation of standards for both representing the data (e.g., common data formats and elements, but also the metadata, in the form of controlled vocabularies or ontologies). However, if these standards are not free to use or are in proprietary formats, then interoperability is similarly impacted.
Reusable	R1. meta(data) are richly described with a plurality of accurate and relevant attributes R1.1. (meta)data are released with a clear and accessible data usage license R1.2. (meta)data are associated with detailed provenance R1.3. (meta)data meet domain-relevant community standards	Re-usability encompasses both the provision of sufficient metadata to make the data understandable to a human, and the licensing conditions under which the data are made available. Failure to provide clear licenses is a major impediment to re-use of the data because the legal rights to re-use data are not made explicit (Carrel, 2015).

Table 3

List of repositories referenced in this article

Repository	URL	Full name	Domain
ADNI	http://adni.loni.usc.edu/	Alzheimer's Disease Neuroimaging Initiative	Alzheimer's disease neuroimaging
NDAR	https://ndar.nih.gov/	National Database for Autism Research	Autism and related disorders neuroimaging
OSF	https://osf.io/	Open Science Framework	General
NITRC-IR	http://nitric.org	Neuroimaging Tool and Research Clearinghouse Image repository	Neuroimaging; includes a tool and image repository
OpenfMRI	https://openfmri.org/		fMRI data
NeuroVault	http://neurovault.org/		fMRI statistical maps
ICPSR	https://www.icpsr.umich.edu/icpsrweb/	Interuniversity Consortium for Political and Social Research	Social science and behavioral data
FigureShare	http://Figureshare.com		General
Dryad	http://datadryad.org		General
DataVerse	http://dataverse.org		General, but Dataverse repositories for specific types of data have been established
DataCite	http://datacite.org		Registry of data sets