REVIEW

REVISED

Data publication consensus and controversies [v3; ref status: indexed, http://f1000r.es/4ja]

John Kratz, Carly Strasser

California Digital Library, University of California Office of the President, Oakland, CA, 94612, USA

v3 **First published:** 23 Apr 2014, **3**:94 (doi: 10.12688/f1000research.3979.1)
Second version: 16 May 2014, **3**:94 (doi: 10.12688/f1000research.3979.2)
Latest published: 16 Oct 2014, **3**:94 (doi: 10.12688/f1000research.3979.3)

Abstract

The movement to bring datasets into the scholarly record as first class research products (validated, preserved, cited, and credited) has been inching forward for some time, but now the pace is quickening. As data publication venues proliferate, significant debate continues over formats, processes, and terminology. Here, we present an overview of data publication initiatives underway and the current conversation, highlighting points of consensus and issues still in contention. Data publication implementations differ in a variety of factors, including the kind of documentation, the location of the documentation relative to the data, and how the data is validated. Publishers may present data as supplemental material to a journal article, with a descriptive “data paper,” or independently. Complicating the situation, different initiatives and communities use the same terms to refer to distinct but overlapping concepts. For instance, the term *published* means that the data is publicly available and citable to virtually everyone, but it may or may not imply that the data has been peer-reviewed. In turn, what is meant by data peer review is far from defined; standards and processes encompass the full range employed in reviewing the literature, plus some novel variations. Basic data citation is a point of consensus, but the general agreement on the core elements of a dataset citation frays if the data is dynamic or part of a larger set. Even as data publication is being defined, some are looking past publication to other metaphors, notably “data as software,” for solutions to the more stubborn problems.

Open Peer Review

Referee Status: ✓ ✓ ✓

	Invited Referees 1	2	3
REVISED version 3 published 16 Oct 2014	✓ report		
REVISED version 2 published 16 May 2014		✓ report	✓ report
version 1 published 23 Apr 2014	? report		

1 **Mark Parsons**, Research Data Alliance USA, **Peter Fox**, Rensselaer Polytechnic Institute USA

2 **Mark Costello**, University of Auckland New Zealand

3 **Ingrid Dillo**, Data Archiving and Networking Services (DANS) Netherlands

Discuss this article

Comments (3)

Corresponding author: John Kratz (John.Kratz@ucop.edu)

How to cite this article: Kratz J and Strasser C. **Data publication consensus and controversies [v3; ref status: indexed, http://f1000r.es/4ja]** *F1000Research* 2014, **3**:94 (doi: 10.12688/f1000research.3979.3)

Grant information: JK is supported by a Council on Library and Information Resources/Digital Library Foundation Postdoctoral Fellowship in Data Curation for the Sciences and Social Sciences funded by the California Digital Library and the Alfred P. Sloan Foundation.
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: No competing interests were disclosed.

First published: 23 Apr 2014, **3**:94 (doi: 10.12688/f1000research.3979.1)
First indexed: 10 Jun 2014, **3**:94 (doi: 10.12688/f1000research.3979.2)

REVISED **Amendments from Version 2**

This version no longer presents three models for data publication based on documentation. Instead, we treat documentation as an essential feature and discuss three forms of documentation in parallel with forms of availability, citation, and validation. The figure has been updated to reflect this reorganization.

Numerous minor additions, corrections and clarifications were made throughout in response to referee and reader comments. Most significantly, the discussions of paper-independent documentation and validation have been expanded, as has the concluding "beyond data publication".

See referee reports

What does data publication mean?

The idea that researchers should share data to advance knowledge and promote the common good is an old one, but in recent years the conversation has shifted from sharing data to *publishing* data[1–3]. This shift in language stems from the conviction that datasets should join the scholarly record and be afforded the same first-class status as traditional research products like journal articles[4,5]. While many in the scholarly communication community share this goal, different people and organizations often refer to different things with the phrase *data publication*.

Lawrence *et al.* (2011) define formal data Publication (upper-case "P") as making data as permanently available as possible following "a process which means it can appear along with easily digestible information as to its trustworthiness, reliability, format and content"[3]. Callaghan *et al.* (2012) draw an explicit distinction between Published and published data: published data is at least available, while Published data is persistent, documented, and peer-reviewed[5]. Publication refers to the scholarly literature, while publication is used in the sense of any kind of printed and distributed material. Actual usage is considerably more complicated. *Data publication* overlaps with terms like *data sharing*, *data release*, and *open data*. A data publication might be a spreadsheet on a website, a set of images in an institutional archive, a stream of readings from a weather station transmitted over the internet, or a peer-reviewed article describing a dataset; a data publisher might be a data journal publisher, archive, database, or repository.

Despite uncertainty over precisely what qualifies, the scholarly communication community largely agrees on three essential properties of a data publication (Figure 1)[2,5]. First, published data is publicly **available** now and for the indefinite future; access might demand payment of fees or acceptance of a legal agreement, but is not subject to the whims of the author. Second, published data must be adequately **documented** such that, at a minimum, a researcher

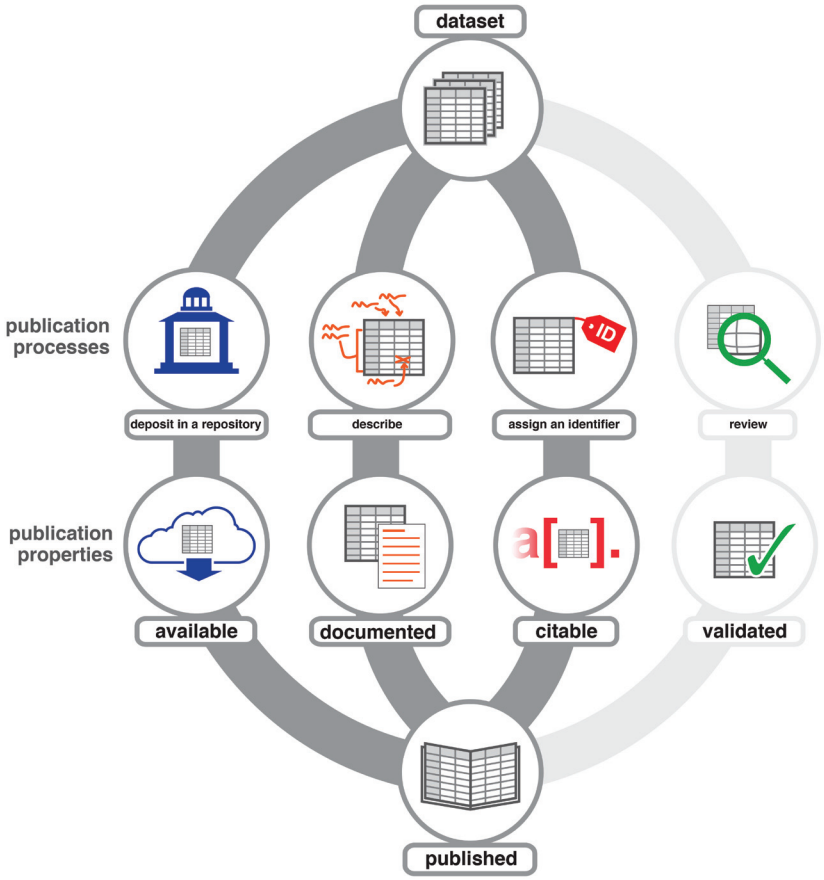

Figure 1. To be published, datasets are typically deposited in a repository to make them available, documented to support reproduction and reuse, and assigned an identifier to facilitate citation. Some, but not all, publishers review datasets to validate them.

in the same field could reproduce or reuse it. Third, like a book or journal article, a data publication can be formally **cited**. Data citation maintains the integrity of the expanded scholarly record and offers a reward– in the currency of academic prestige– to encourage researchers to publish data. Open questions flock around a fourth property: how and to what extent a published dataset must be **validated**. Here, we will consider data that is persistently available, documented, and citable to be published, whatever the level of validation.

Why publish data?

The underlying goals of data publication are to enable research to be **reproduced** and data to be **reused**. Hidden primary data exacerbates science's very public "reproducibility crisis"[6–10], recently illustrated by the collapse of a pair of irreproducible *Nature* articles describing a simple method to transform any cell into a stem cell[11,12]. Psychology's "closed data culture"[13] enabled Diederik Stapel to invent data for an astonishing 55 papers, prompting calls for routine psychology data publication[13–15]. Widespread publication of the data underlying research papers could help expose honest errors as well as fraud[16]. The leaders of the US National Institutes of Health (NIH) recently suggested "greater transparency of the data that are the basis of published manuscripts" as one way to improve scientific reproducibility[17].

Journals already frequently require authors to supply underlying data on request. In 2011, Alsheikh-Ali *et al.* found that 88% of high-impact journals required a statement regarding the availability of underlying data, and half of those made willingness to provide data a condition of publication[18]. However, authors of 59% of the papers examined in the study failed to adhere to the availability instructions. Vines *et al.* (2014) could only obtain underlying data from 101 of 516 papers published from 1991 to 2011[19]. Availability dropped off sharply with time; out of the 62 oldest papers, data was available from only two. Now, some journals require that underlying data be published simultaneously with the article. In 2010, a coalition of Ecology and Evolutionary Biology journals began to require that the data underlying articles be archived with a maximum embargo of one year[20,21]. *F1000Research* has had a similar policy (without an embargo period) since its inception, and the *Public Library of Science (PLOS)* journals followed suit earlier this year[22].

Although there can be no substitute for funding new experiments and data collection, appropriate data reuse lowers costs and accelerates research. Documenting, publishing, and archiving data is time consuming and costly, but usually far less so than repeating the data collection. Open Context published archaeological data from a site in eastern Turkey at the substantial cost of $10,000–15,000, but this expense is minor compared to $800,000 spent to collect the data[23]. Piwowar (2011) contrasted the impact of $100,000 in National Science Foundation (NSF) grants, which generates an average of three to four papers, with an estimate that the same investment in curating, archiving, and publishing data could contribute to over 1,000 publications[24]. Furthermore, while some data is merely expensive to replace, time-dependent or ephemeral data, (e.g., climate records or observations of unique astronomical events) can never be recreated for any price[25].

Availability

Fundamentally, to publish is to make public, and to publish data is to make data publicly available. Present availability requires mechanisms for access; future availability also requires preservation (e.g., long-term storage, format migration)[25–27]. As in print publication, published data need not be free or legally unencumbered, and data use agreements constrain many published datasets. If access is limited, it should be contingent on clear and objective criteria; writing a request to the creator for permission should not be part of the process. For example, before granting access to restricted data, The interuniversity Consortium for Political and Social Research (ICPSR) evaluates the applicant's ability to handle the data securely, but not the merit of the research. The most common source of access restrictions is the need to protect the privacy of human research subjects. In the United States, the Health Insurance Portability and Accountability Act of 1996 (HIPAA) Privacy Rule severely limits the disclosure of medical information[28].

As a practical matter, publishing a dataset usually includes depositing it in a trustworthy repository. What constitutes "trustworthy" is somewhat subjective and there are a handful of certification schemes to choose from. In 2007, The Center for Research Libraries (CRL) published the most extensive scheme: the Trusted Repository Audit Checklist (TRAC)[29]. Many repositories consult TRAC for self-assessment, but only four (listed by the CRL) have completed the lengthy and rigorous process to be officially certified. That same year, DANS released the Data Seal of Approval (DSA) guidelines; 31 repositories have been stamped with the DSA since then. The Trusted Digital Repository framework incorporates the DSA, a TRAC-derived standard, and a third standard from the German Institute for Standardization (DIN) to give repositories flexibility in the processes and standards by which they are to be certified. Repositories seeking to join the World Data System (WDS) are certified to perform particular role (i.e., data publisher) based on a self-description and possibly a site visit; the WDS currently boasts 56 regular members.

Even taken together, these standards certify only a fraction of the hundreds of repositories in operation (e.g., the 973 now listed Databib or the 609 at re3data.org). In practice, the perceived trustworthiness of a repository often derives from the reputation of its managing organization. For instance, repositories run by governments or large universities are likely to be considered trustworthy (although the effects of the 2013 US government shutdown on the PubMed biomedical article database[30] might give one pause).

Documentation

To be useful or reproducible, a dataset must be accompanied by descriptive information (i.e., metadata)[25]. Preparing documentation is frequently the most laborious step for researchers in taking data from useable within the lab to useable by others, and rewarding this effort is a major impetus for data publication. Dataset documentation– which might resemble a paper– is a natural hook for bringing data into the scholarly record. The Opportunities for Data Exchange (ODE) project elaborated Jim Gray's pyramidal model of online scientific data[31] into five classes of relationship between data and the literature: 'desk-drawer' data and four forms of publication[4].

Similarly, five classes of data publication described by Lawrence *et al.* (2011) have recognizably different kinds of documentation[3]. Note, however, that a single dataset may have relationships with multiple articles or other documentation, and an article may use or describe multiple datasets. Here, we will discuss three non-mutually-exclusive relationships with the literature: a dataset may **supplement** a traditional research paper, be the **subject** of a "data paper", or be **independently documented** by its publisher.

Data that supplements a paper

The most familiar kind of data publication is a traditional journal article accompanied by underlying data. That data can be hosted by the journal as supplementary material or deposited in a third-party repository. The trend is away from supplemental material because repositories are considered to be better suited to ensure long-term preservation and access to the data. For instance, *The Journal of Neuroscience* stopped publishing supplemental material in 2010; the announcement promotes disciplinary repositories as "vastly superior to supplemental material as a mechanism for disseminating data"[32]. Data underlying any peer-reviewed or otherwise "reputable" publication can be deposited in the Dryad repository. Dryad makes data available and citable, but the publisher of the article must manage any assessment of scientific validity. Research Compendia compiles published articles together with all the underlying code and data. Beyond repositories like these specifically for paper-related data, many more publishers that do not require such a relationship are nevertheless pleased to publish data underlying or described by a paper.

This kind of data publication supports reproduction of an analysis, but not necessarily reuse. For example, the *PLOS* data policy requires publication of only the data needed to reproduce the article's finding. Consequently, not all of the data collected must be published, and the documentation need not support reuse for an unrelated purpose.

Data as the subject of a paper

A **data paper** describes a dataset with thoroughly detailed rationale and collection methods, but lacks any analysis or conclusions[33]. Data papers are flourishing as a new article type in journals such as *F1000Research*, *Internet Archaeology*, and *GigaScience*, as well as in dedicated journals like *Earth System Science Data*[34], *Geoscience Data Journal*, Nature Publishing Group's *Scientific Data*, and a trio of "metajournals" from Ubiquity Press. The strength of a data paper is in providing rich documentation, which is especially useful for unique and heterogeneous "long-tail"[35] research data.

Data paper length and structure varies between journals, but the tendency is toward a short, tightly structured format. All journals require an abstract, collection methods, and a description of the dataset; a few encourage authors to suggest potential uses for the data (e.g., *Internet Archaeology*, and *Open Health Data*). Some journals supplement this general framework with field-specific sections. (e.g., *Internet Archaeology* and the *Journal of Open Archaeology Data* each include a section for temporal and geographic scope). Data papers are most sharply defined not by the presence of any particular information, but by the absence of analysis or conclusions. A crisp distinction from other article types is important because many journals do not consider a data paper to be prior publication if the authors seek to publish an analysis of the same dataset (e.g., *Nature*-titled journals, *Science*, and others listed by *F1000Research*).

Data journals generally limit themselves to publishing the description of the dataset; a trusted repository publishes the data itself. For instance, *Scientific Data* and *Geoscience Data Journal* each direct authors to a list of approved repositories. One exception, *GigaScience* hosts data in an integrated repository named *GigaDB*. Another, *The International Journal of Robotics Research*[33] permits authors to host datasets on their own websites.

Data papers are predated by an approach that Lawrence *et al.* (2011) call *data publication by proxy*, in which a paper providing a general description of a database or dataset serves as a citable proxy[3]. Proxy publications are distinguished from data papers in that they may contain analysis or conclusions drawn from the dataset and they may not contain all of the information needed to use the data. For example, the Climatic Research Unit of the University of East Anglia asks dataset users cite to papers associated with a dataset instead of the data itself. In the biosciences, *Nucleic Acids Research (NAR)* annually publishes a massive issue devoted to such articles; the 2014 database issue featured 58 papers describing new databases and 123 updates on existing resources[36]. Participating databases typically ask users to cite the most recent *NAR* paper. Proxy publication or data paper citation serves to award scholarly credit, but fails at other functions of citation and should be supplemented with direct citation of the data.

Independent documentation

Dataset documentation need not take the form of a journal article. Together with data, repositories and databases publish documentation– minimal or rich, structured or freeform– that sometimes fulfills the needs of reproducibility and reuse without reference to the literature. Even so, an independently documented dataset might also be described by a data paper or support any number of traditional articles. Academic, governmental, and commercial repositories publish data from diverse place- and interest-based research communities through varying processes with or without linkage to the literature.

Institutional repositories preserve and publish any kind of data generated by the research communities they serve, e.g., University of California researchers deposit data in Merritt, while Purdue University researchers use the Purdue Research Repository (PURR). At the national level, the Dutch Data Archiving and Networked Services (DANS) accepts a broad range of data from researchers in the Netherlands. Figshare and Zenodo publish data from any researcher in any field. These broad-topic publishers are well suited to handle heterogeneous or long-tail data that does not fit comfortably in a specialized repository. But, because repositories typically cannot assemble domain expertise across such a broad range of disciplines, this inclusiveness imposes limits on documentation requirements and validation. While Figshare and Zenodo do accommodate rich documentation, they require very little.

Interest-based research communities are served by a thriving ecosystem of specialized data publishers. The broadest of these publishers serve entire disciplines, e.g., the Digital Archaeological Record

(tDAR). A narrower example from the life sciences is the group of databases centered around model organisms, such as WormBase[37] or FlyBase[38]; these databases aggregate diverse, but finite, data types and benefit from extensive domain expertise. Along similar lines, a data publisher may deal with a particular data-type, such as gene expression data in the Gene Expression Omnibus (GEO) or seismological data in SeismicPortal. Focus on a particular type of data facilitates rigorous technical validation and development of specialized metadata requirements to ensure the data is useable. For instance, GEO data ingest meshes with Minimum Information About a Microarray Experiment (MIAME)[39] documentation guidelines[40]. As a final example, a publisher might be devoted to a particular scientific instrument or facility, such as the One Degree Imager Portal, Pipeline, and Archive (ODI-PPA) or the Worldwide LHC Computing Grid), the massive infrastructure built to handle the output of the Large Hadron Collider (LHC). Unlike most other publishers, these emphasize real time access to data coming off the instruments.

Because researchers know the databases that serve their community, data in disciplinary repositories is easy to discover and because it is relatively standardized, it is easy to reuse. A disadvantage is that the data from a single research program can be distributed across many repositories (e.g., gene expression data in one, sequence data in another), whereas an institutional or broad-scope repository can publish the whole research story.

Citability

Data citation is the element of publication that has come the farthest toward consensus. In early 2014, a coalition of organizations brought together by Future Of Research Communication and E-Scholarship (FORCE11)[41] released a Joint Declaration of Data Citation Principles. The first of the eight principles states, in part, that "[d]ata citations should be accorded the same importance in the scholarly record as citations of other research objects, such as publications". Most of the time, this means that when a published dataset contributes to a paper, it should be cited formally in the reference list.

Unfortunately, actual practice lags far behind this consensus. Not all article publishers allow data citations in the references and, even when permitted, most authors refer to data in the text without a formal citation[42]. Many data publishers provide no guidance on citation; others ask users to cite a proxy publication (e.g., from the *NAR* database issue). However, a growing number of data publishers do supply users with explicit citation instructions; Dryad, Figshare, and Zenodo dataset landing pages all display a formatted citation and links for import into reference managers.

Many data publishers facilitate formal citation by assigning unique permanent identifiers, most commonly the same ones used for journal articles: Digital Object Identifiers (DOIs). In addition to precisely specifying what resource is being cited, a DOI can be resolved to locate the referenced dataset. Note, however, that a DOI is neither sufficient nor necessary for citability, which demands that the referenced object be persistent and locatable via the citation. If a dataset moves and the DOI is not updated with the new location, the citation breaks. Conversely, a well-maintained web-address works as well as a DOI in theory– although a DOI is more likely to be maintained in practice.

Simple case

The present consensus is that a dataset should be cited using, at a minimum, five elements largely familiar from article citations: creator(s), title, year, publisher and identifier. This format agrees with Committee on Data for Science and Technology (CODATA) recommendations[43] and conveys all the information required to obtain a DataCite DOI[44] or be listed in the Thomson-Reuters Data Citation Index. The basic format works well when a dataset can be cited like an article, but that is not always the case.

Deep citation

One major complication data citation faces is the need for **deep citation**. When supporting an assertion in writing, it usually suffices to cite the entirety of an article or the page of a book and leave it to the inquisitive reader to find the relevant passage. But, to reproduce an analysis performed on a subset of a larger dataset, the reader needs to know exactly what subset was used (e.g., a limited range of dates, only the adult subjects, wind speed but not direction). Datasets vary so widely in structure that there may not be a good general solution for describing subsets. The most common suggestion is to cite the entire dataset in the reference list and describe the subset in the text of the paper[45]. The Federation of Earth Science Information Partners (ESIP) and the National Snow and Ice Data Center (NSIDC) both recommend defining the subset in the citation itself, using a format suited to the dataset's internal structure (e.g., a temporal or spatial range, a list of variables, or an internal identifier).

Dynamic datasets

A second major complication arises when datasets change. In the past, the printing process cemented one version of an article as the version of record. Even for traditional scholarly literature, web-based publishing and preprint servers (e.g., arXiv.org) are complicating the situation, but datasets are especially prone to be **dynamic**. Two kinds of dynamic datasets warrant consideration: **growing** datasets that add new data while never changing or deleting existing data, and **revisable** datasets where data may by added, deleted, or changed.

Consider USC00046336, a weather station at the Oakland Museum of California. Each day, the high temperature, low temperature and amount of precipitation recorded at the Museum[46] flow, together with data from more than 20,000 other stations, into the swelling Global Historical Climate Network (GHCN)-Daily[47] dataset. Or, consider WormBase, the genome database used by the *Caenorhabditis elegans* research community. WormBase encompasses genomic sequences of *C. elegans* and 20 related species massively annotated with gene structures, protein sequences, expression patterns, and a host of other information from empirical data and computational predictions. Every two months, WormBase administrators respond to new data and better computational models by issuing a revised version with new material added and inaccurate material deleted or corrected.

Additions and updates to published datasets are extremely valuable, but a researcher seeking to reproduce an analysis of a dynamic dataset needs access to a particular version. To enable that access, previous versions must be preserved and citable. Growing datasets can be cited with an access date or a date range in the citation, as recommended by ESIP and NSIDC. Revisable datasets are more difficult; the most common approach is to accumulate revisions and periodically publish a new version with a citable version number. For example, WormBase identifies each release with a version number and makes all of the previous versions available.

Controversy persists around the specific issue of identifiers for dynamic datasets. DataCite recommends, but does not insist, that their DOIs refer to immutable digital objects. NSIDC and ESIP instruct researchers to use a single identifier for growing datasets and include the access date in the citation; each major version of a revisable datasets gets a new identifier, but minor versions do not. In contrast, the Digital Curation Centre (DCC), Dataverse, and the UK Natural Environment Research Council (NERC) insist that any change to a dataset should trigger a new identifier[5,45,48]. To handle the difficulties with dynamic data that this policy creates, the DCC recommends periodically issuing growing datasets a new identifier that refers to the *time-slice* of new records and freezing versions of revisable datasets as individually-identified *snapshots*.

Just-in-time identifiers

The difficulties surrounding deep citation and dynamic data could potentially be solved by turning the identifier-issuing process on its head. Instead of the dataset publisher issuing identifiers for data at the level that researchers seem likely to cite, researchers could issue identifiers for only the part of the dataset that they want to cite.

The Research Data Alliance (RDA) Data Citation Working Group recently put forth a sophisticated proposal applicable to data in (or convertible to) databases. Identifiers created under this scheme would wrap together identification of a database, a query to return the cited dataset, the version of the database queried for this analysis, and a number of other useful components. The ultimate promise is to provide a simple yet precise citation for any selection of data, at the cost of technical complexity "under the hood".

Validation

Data validation is the least resolved aspect of data publication, and fundamental questions are still unanswered: What minimum level of quality should a published dataset guarantee? How and by what criteria can datasets be evaluated against that guarantee? How should dynamic datasets be handled? Is literature peer-review an appropriate model?

Callaghan *et al.* (2012)[5] draw a useful distinction between **technical** and **scientific** review. Technical review verifies that a dataset is complete, its description is complete, and that the two match up. Domain expertise is generally not required, and many repositories provide at least some level of technical review. Scientific review evaluates the methods of data collection, the overall plausibility of the data, and the likely reuse value. Scientific review does require domain expertise, making this level of validation more difficult to organize[13]. When data is published with a data paper, review may be split between the repository for technical review and the data journal for scientific review.

Data paper peer review

Peer review guarantees that journal articles entering the scholarly record reach some level of validity (although the aforementioned reproducibility crisis calls into question exactly what that level is). In many fields, peer-reviewed publications enjoy a much higher status than any other literature. Any effort to apply the prestige of "publication" to datasets cascades naturally into an effort to apply the prestige of "peer review". But as data validation seeks to model itself on literature peer review, literature peer review itself is in flux[49–51]. Open peer review at *F1000Research* and post-publication commenting at PubMed Commons are just two of many ongoing web-enabled experiments in article evaluation.

Journal article reviewers traditionally consider whether the methods used are appropriate for the questions asked and the data collected support the conclusions drawn. In the absence of particular questions and conclusions, it is not obvious what peer review of data should certify. A dataset may serve for some purposes, but not for others and a reviewer may anticipate many potential uses for the data, but surely not all[52]. Researchers are already over-whelmed by peer review of articles[53] and could find any increased workload unreasonable. Despite all these difficulties, venues for peer-reviewed data papers are opening rapidly.

Data paper journals wrap scientific peer review of the paper and the dataset together into a single process. *GigaScience*, an exception, assigns technical review of the dataset to a separate data reviewer. The guidelines that various data journals provide to reviewers are fairly uniform, except that about half consider novelty or potential impact, while the rest only require the dataset to be scientifically sound. Although the guidelines are similar, review processes differ widely.

As an example, compare *Biodiversity Journal* and *Scientific Data*. Both journals divide reviewer guidelines into three sections along similar lines, which *Biodiversity Journal* calls "quality of the data", "quality of the description", and "consistency between manuscript and data". *Scientific Data* follows a traditional peer-review process: an editor appoints reviewers who are encouraged to remain anonymous. In contrast, review at *Biodiversity Journal* follows a flexible and open process featuring entirely optional anonymity and multiple types of reviewer. There, an editor appoints two or three "nominated" reviewers who must report back and several "panel" reviewers who read the paper and only comment at their discretion. Additionally, the authors may choose to open the paper to public comment during the review process.

Independent data validation

Data journals all model their data validation more or less faithfully on literature peer review, but independent data validation practices and proposals are considerably more varied. Lawrence *et al.* (2011) propose a set of independent data peer review guidelines similar to the ones used by data journals[3]. Each of The National Aeronautics and Space Administration (NASA) Distributed Active Archive Centers (DAACs) draws on an affiliated User Working Group for

domain expertise. The NSIDC combines an internal assessment of the effort that will be required to publish a dataset at a desired level of service (roughly corresponding to technical review) with an external assessment of scientific quality. The Planetary Data System (PDS) peer-reviews datasets via an in-person meeting with representatives of the repository, the dataset creators, and the reviewers.

Pre-publication validation can be supplemented or replaced by post-publication feedback from successful or unsuccessful reusers. Parsons *et al.* (2010) suggest that "data use in its own right provides a form of review", and go on to point out that the context of reuse demonstrates that the data is not simply "good", but fit for some particular purpose[52]. The DANS repository solicits feedback from researchers who use its datasets: users are asked to rate the dataset on a one to five scale in each of six criteria (e.g., data quality, quality of the documentation, structure of the dataset)[54,55]. Researchers trust peer review in part because they understand the process and its limitations; if researchers come to understand them, alternate pre- or post-publication validation processes could potentially provide the same level of assurance.

Two examples from archaeology, Open Context and the Digital Archaeological Record (tDAR), illustrate the diversity of approaches to data validation. Open Context provides multiple validation processes that incorporate peer review beyond a simple accept/reject binary[23]. Each Open Context dataset is rated from one to five based not on quality *per se*, but on the thoroughness of the validation; a one comes with no guarantees, a three has passed a technical review, and a five has passed external peer review. Whereas Open Context is a boutique publisher, focusing on data presentation and reuse, tDAR is a large repository primarily concerned with collecting and preserving archaeology data for future use. tDAR is able to operate at scale by performing only technical validation and streamlining data deposition with a minimum of mandatory description. However, tDAR also serves as a platform for high-quality data publication. The repository accommodates contributors who provide more information, and much of the content is deposited by digital curators who can be relied on to supply rich descriptions. Furthermore, two data paper journals, *Internet Archaeology* and *Journal of Open Archaeological Data*, recommend both tDAR and Open Context as repositories for their peer-reviewed data. Thus, data validation depends not only on discipline and data type, but on a host of external factors, including the goals of the organizations and researchers involved.

Beyond data publication

Consensus abides wherever traditional scholarly publication offers a clear model for data; controversy churns wherever the literature offers only murky guidance. Static datasets of manageable size and simple structure can be made available, identified, and cited like the literature. Dynamic and complex datasets raise questions that attract multiple and sometimes conflicting answers. Where the guidance of the print metaphor threatens to give out, it must be extended creatively— or abandoned for another approach entirely.

Parsons and Fox (2013)[56] argue that thinking about data through the metaphor of print publication is often misleading. They advocate treating publication as only one metaphor in a larger ecosystem of metaphors for sharing data. For example, they associate the uniform, high-volume output from instruments like the Large Hadron Collider with industrial production and suggest "Big Iron" as an alternative metaphor for this kind of data.

Another alternative metaphor that seems to be gaining particular traction is "data as software"[57]. Here, one thinks of releasing a dataset like a piece of software and regards subsequent changes as analogous to updated versions. The open-source software community has already developed many potentially relevant tools for working collaboratively, managing multiple versions, and tracking attribution. Ram (2013)[58] catalogs a multitude of scientific uses for the software version control system Git, including data management. Open Context uses Git and Mantis Bug Tracker to track and correct dataset errors. The Dat project "aim[s] to bring to data a style of collaboration similar to what Git brings to source code". Furthermore, projects such as IPython Notebook integrate data, processing, and analysis into a single package. Unfortunately, scientific software struggles for recognition[59] just as data does, so that metaphor offers little guidance for navigating the academic reward system. On the other hand, the publication metaphor targets this system explicitly, but leaves numerous other gaps.

Although some aspects of data publication have matured to a firm and useful consensus– exemplified most powerfully by the Joint Declaration of Data Citation Principles– the field as a whole is still burgeoning. Controversial issues, such as validation, may be best addressed by presenting an array of options rather than converging on a single solution. In the ongoing conversation, *data publication* may come to refer to only those means of dissemination most directly drawn from the scholarly literature, or it may open as a canopy over a range of approaches. Whichever the case, it is our hope and expectation that for the foreseeable future, mixing of metaphors and contemplation of the unique properties of research data will continue to yield novel forms of data-centered scholarly production.

Author contributions

JK collected information and prepared the first draft of the manuscript. JK and CS designed the scope and direction of the study. Both authors contributed to the writing and editing of the manuscript.

Competing interests

No competing interests were disclosed.

Grant information

JK is supported by a Council on Library and Information Resources/ Digital Library Foundation Postdoctoral Fellowship in Data Curation for the Sciences and Social Sciences funded by the California Digital Library and the Alfred P. Sloan Foundation.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgements

The authors would like to thank colleagues at the CDL and Jodi Reeves Flores for productive discussions. Margaret Smith and Christina Doyle contributed invaluable suggestions for editing the manuscript.

References

1. Costello MJ: **Motivating online publication of data.** *BioScience.* 2009; **59**(5): 418–427.
Publisher Full Text
2. Smith VS: **Data publication: towards a database of everything.** *BMC Res Notes.* 2009; **2**: 113.
PubMed Abstract | Publisher Full Text | Free Full Text
3. Lawrence B, Jones C, Matthews B, *et al.*: **Citation and peer review of data: Moving towards formal data publication.** *Int J Digit Curation.* 2011; **6**(2): 4–37.
Publisher Full Text
4. Reilly S, Wouter S, Schrimpf S, *et al.*: **Report on integration of data and publications.** *Zenodo.* 2011.
Publisher Full Text
5. Callaghan S, Donegan S, Pepler S, *et al.*: **Making data a first class scientific output: Data citation and publication by NERC's environmental data centres.** *Int J Digit Curation.* 2012; **7**(1): 107–113.
Publisher Full Text
6. Mobley A, Linder SK, Braeuer R, *et al.*: **A survey on data reproducibility in cancer research provides insights into our limited ability to translate findings from the laboratory to the clinic.** *PLoS One.* 2013; **8**(5): e63221.
PubMed Abstract | Publisher Full Text | Free Full Text
7. Pashler H, Harris CR: **Is the replicability crisis overblown? three arguments examined.** *Perspect Psychol Sci.* 2012; **7**(6): 531–536.
Publisher Full Text
8. Zimmer C: **Rise in scientific journal retractions prompts calls for reform**. *The New York Times.* 2012.
Reference Source
9. Hiltzik M: **Science has lost its way, at a big cost to humanity.** *Los Angeles Times.* 2013.
Reference Source
10. Begley CG, Ellis LM: **Drug development: Raise standards for preclinical cancer research.** *Nature.* 2012; **483**(7391): 531–533.
PubMed Abstract | Publisher Full Text
11. Cyranoski D: **Acid-bath stem-cell study under investigation.** *Nature.* 2014.
Publisher Full Text
12. Tabuchi H: **One author of a startling stem cell study calls for its retraction**. *The New York Times.* 2014.
Reference Source
13. Doorn P, Dillo I, Van Horik R: **Lies, damned lies and research data: Can data sharing prevent data fraud?** *Int J Digit Curation.* 2013; **8**(1): 229–243.
Publisher Full Text
14. Committee L, Committee N, Committee D: **Flawed science: The fraudulent research practices of social psychologist diederik stapel.** *Tech Rep.* 2012.
Reference Source
15. Wicherts JM: **Psychology must learn a lesson from fraud case.** *Nature.* 2011; **480**(7375): 7.
PubMed Abstract | Publisher Full Text
16. Drew BT, Gazis R, Cabezas P, *et al.*: **Lost branches on the tree of life.** *PLoS Biol.* 2013; **11**(9): e1001636.
PubMed Abstract | Publisher Full Text | Free Full Text
17. Collins FS, Tabak LA: **Policy: NIH plans to enhance reproducibility.** *Nature.* 2014; **505**(7485): 612–613.
PubMed Abstract | Publisher Full Text | Free Full Text
18. Alsheikh-Ali AA, Qureshi W, Al Mallah MH: **Public availability of published research data in high-impact journals.** *PLoS One.* 2011; **6**(9): e24357.
PubMed Abstract | Publisher Full Text | Free Full Text
19. Vines TH, Albert AYK, Andrew RL: **The availability of research data declines rapidly with article age.** *Curr Biol.* 2014; **24**(1): 94–7.
PubMed Abstract | Publisher Full Text
20. Whitlock MC, McPeek MA, Rausher MD: **Data archiving.** *Am Nat.* 2010; **175**(2): 145–146.
PubMed Abstract | Publisher Full Text
21. Fairbairn DJ: **The advent of mandatory data archiving.** *Evolution.* 2011; **65**(1): 1–2.
PubMed Abstract | Publisher Full Text
22. Bloom T, Ganley E, Winker M: **Data access for the open access literature: PLOS's data policy.** *PLoS Biol.* 2014; **12**(2): e1001797.
Publisher Full Text | Free Full Text
23. Kansa EC, Kansa SW: **We all know that a 14 is a sheep: Data publication and professionalism in archaeological communication.** *J Endocrinol Metab Arch Heritage Studies.* 2013; **1**(1): 88–97.
Reference Source
24. Piwowar HA, Vision TJ, Whitlock MC: **Data archiving is a good investment.** *Nature.* 2011; **473**(7347): 285–285.
PubMed Abstract | Publisher Full Text
25. Gray J, Szalay AS, Thakar AR, *et al.*: **Online scientific data curation, publication, and archiving**. 2002; 103.
Publisher Full Text
26. Waters D, Garrett J: **Preserving Digital Information. Report of the Task Force on Archiving of Digital Information.** *ERIC.* 1996.
Reference Source
27. Beagrie N: **Digital curation for science, digital libraries, and individuals.** *Int J Digit Curation.* 2008; **1**(1): 3–16.
Publisher Full Text
28. **Office for Civil Rights. Renal resource guide**. 2003.
Reference Source
29. **Center for Research Libraries (U.S.) and OCLC. Trustworthy repositories audit & certification (TRAC) criteria and checklist**. Center for Research Libraries; OCLC Online Computer Library Center, Inc Chicago: Dublin, Ohio. 2007.
Reference Source
30. Hayden EC: **NIH shutdown effects multiply.** *Nature.* 2013.
Publisher Full Text
31. Gray J: **Jim gray on eScience: A transformed scientific method**. In Tony Hey, STewarT Tansley, Stewart, and Kristin Tolle, editors, *The fourth paradigm: data-intensive scientific discovery*, pages xvii–xxxi. USA. Microsoft Research. 2009.
Reference Source
32. Maunsell J: **Announcement regarding supplemental material.** *J Neurosci.* 2010; **30**(32): 10599–10600.
Reference Source
33. Newman P, Corke P: **Data papers — peer reviewed publication of high quality data sets.** *Int J Rob Res.* 2009; **28**(5): 587–587.
Publisher Full Text
34. Pfeiffenberger H, Carlson D: **"Earth system science data" (ESSD) — a peer reviewed journal for publication of data.** *D-Lib Magazine.* 2011; **17**(1/2).
Publisher Full Text
35. Bryan Heidorn P: **Shedding light on the dark data in the long tail of science.** *Libr Trends.* 2008; **57**(2): 280–299.
Publisher Full Text
36. Fernández-Suárez XM, Rigden DJ, Galperin MY: **The 2014 nucleic acids research database issue and an updated NAR online molecular biology database collection.** *Nucleic Acids Res.* 2014; **42**(Database issue): D1–D6.
PubMed Abstract | Publisher Full Text | Free Full Text
37. Harris TW, Baran J, Bieri T, *et al.*: **WormBase 2014: new views of curated biology.** *Nucleic Acids Res.* 2014; **42**(Database issue): D789–793.
PubMed Abstract | Publisher Full Text | Free Full Text
38. St Pierre SE, Ponting L, Stefancsik R, *et al.*: **FlyBase 102--advanced approaches to interrogating FlyBase.** *Nucleic Acids Res.* 2014; **42**(Database issue): D780–788.
PubMed Abstract | Publisher Full Text | Free Full Text
39. Brazma A, Hingamp P, Quackenbush J, *et al.*: **Minimum information about a microarray experiment (MIAME)-toward standards for microarray data.** *Nat Genet.* 2001; **29**(4): 365–371.
PubMed Abstract | Publisher Full Text
40. Barrett T, Edgar R: **NCBI GEO standards and services for microarray data.** *Nat Biotechnol.* 2006; **24**(12): 1471–1472.
PubMed Abstract | Publisher Full Text | Free Full Text
41. **FORCE11. Improving future research communication and e-scholarship**. 2012.
Reference Source
42. Mooney H, Newton M: **The anatomy of a data citation: Discovery, reuse, and credit.** *J Libr schol commun.* 2012; **1**(1): eP1035.
Publisher Full Text
43. CODATA-ICSTI Task Group on Data Citation Standards and Practices. **Out of cite, out of mind: The current state of practice, policy, and technology for the citation of data.** *Data Sci J.* 2013; **12**: 1–75.
Publisher Full Text
44. Starr J, Gastl A: **isCitedBy: a metadata scheme for DataCite.** *D-Lib Magazine.* 2011; **17**(1).
Publisher Full Text
45. Altman M, King G: **A proposed standard for the scholarly citation of quantitative data.** *D-Lib Magazine.* 2007; **13**(3/4).
Publisher Full Text
46. **Global Historical Climate Data Network.** Daily summaries station details: OAKLAND MUSEUM, CA US, GHCND:USC00046336.
Reference Source
47. Menne MJ, Durre I, Vose RS, *et al.*: **An overview of the global historical climatology network-daily database.** *J Atmos Ocean Technol.* 2012; **29**(7): 897–910.
Publisher Full Text
48. Ball A, Duke M: **How to cite datasets and link to publications**. 2012.
Reference Source
49. Pulverer B: **A transparent black box.** *EMBO J.* 2010; **29**(23): 3891–3892.
PubMed Abstract | Publisher Full Text | Free Full Text
50. Herron DM: **Is expert peer review obsolete? A model suggests that post-publication reader review may exceed the accuracy of traditional peer review.** *Surg Endosc.* 2012; **26**(8): 2275–2280.
PubMed Abstract | Publisher Full Text

51. Kriegeskorte N, Walther A, Deca D: **An emerging consensus for open evaluation: 18 visions for the future of scientific publishing.** *Front Comput Neurosci.* 2012; **6**: 94.
PubMed Abstract | Publisher Full Text | Free Full Text

52. Parsons MA, Duerr R, Minster JB: **Data citation and peer review.** *Eos, Transactions American Geophysical Union.* 2010; **91**(34): 297–298.
Publisher Full Text

53. Diederich F: **Are we refereeing ourselves to death? The peer-review system at its limit.** *Angew Chem Int Ed Engl.* 2013; **52**(52): 13828–9.
PubMed Abstract | Publisher Full Text

54. Grootveld M, Van Egmond J: editors. **Data Reviews, peer-reviewed research data**. Number 5 in DANS Studies in Digital Archiving. Data Archiving and Networked Services. 2011.
Reference Source

55. Grootveld M, Van Egmond J: **Peer-reviewed open research data: Results of a pilot.** *Int J Digital Curation.* 2012; **7**(2): 81–91.
Publisher Full Text

56. Parsons MA, Fox PA: **Is data publication the right metaphor?** *Data Sci J.* 2013; **12**: WDS32–WDS46.
Publisher Full Text

57. Schopf JM: **Treating data like software: a case for production quality data**. In *Proceedings of the 12th ACM/IEEE-CS joint conference on Digital Libraries*, JCDL '12, New York, NY USA, 2012; 153–156. ACM.
Publisher Full Text

58. Ram K: **Git can facilitate greater reproducibility and increased transparency in science.** *Source Code Biol Med.* 2013; **8**(1): 7.
PubMed Abstract | Publisher Full Text | Free Full Text

59. Pradal C, Varoquaux G, Langtangen HP: **Publishing scientific software matters.** *J Comput Sci.* 2013; **4**(5): 311–312.
Publisher Full Text

Open Peer Review

Current Referee Status: ✓ ✓ ✓

Version 3

Referee Report 06 November 2014

doi:10.5256/f1000research.5878.r6447

Mark Parsons
Research Data Alliance, Troy, NY, USA

A nice rewrite and a very much improved paper, especially since it is now properly classified as a review article. Just a few small corrections listed below.

Page 4, Para 5:
"Writing a request to the creator should [very rarely] be part of the process" As noted before creator permission is sometimes legitimate.

Page 5, para 2:
"The most familiar kind of data publication is a traditional journal article accompanied by underlying data. " [citation needed] or change the sentence.

Page 5, para 7
"Data papers are predated by an approach that Lawrence et al. (2011) call data publication by proxy..." This is not true. As Hans Pfeiffenberger notes, ESSD dates back to 2008. Indeed I wouldn't be surprised if Lawrence chatted a bit with ESSD editors, Pfeiffenberger and Carlson, in preparing his paper.

Page 6, para 1
NSIDC does do external scientific reviews but they only go outside when they don't have expertise in house. So it's sort of a combined approach. A quibble, but they pride themselves on in-house scientific expertise and engagement.

Page 6 para 2:
This paragraph is a little confused. Domain repositories don't usually serve interdisciplinary use very well, but I don't see how it's necessarily bad to be distributed. What do you mean by publishing the whole research story? No one entity can do that.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Competing Interests: No competing interests were disclosed.

Version 2

Referee Report 10 June 2014

doi:10.5256/f1000research.4518.r4540

Ingrid Dillo
Data Archiving and Networking Services (DANS), The Hague, Netherlands

The article focuses on a topic that receives a lot of interest these days. Therefore it is very timely. The article provides a useful and valuable overview of the current state of affairs and the ongoing debate. The title does justice to the content of the article. So does the abstract.

This is the second version of the article. I do not see many changes in the text based on the earlier critical comments made by Mark Parsons and Peter Fox.

The overview is very informative for everyone who needs a quick introduction into the subject. I do miss the opinion of the authors themselves on the issues at hand and on the quoted suggestions by others. This would have been appropriate in a concluding paragraph.

Detailed comments:

1. In the Introduction I miss a clear link between data publishing and data citation and creating the possibility for researchers to receive academic credits for their work on data. This academic credit is crucial as an incentive for researchers to put valuable time and effort in sharing their data.

2. In the paragraph *Why publish data?* a reference to the Dutch fraud cases might be useful, as these cases got a lot of international attention and more or less triggered the discussion in the Netherlands with respect to research data management, long term preservation of data and data publishing and citation. A reference could be:

   Doorn P, Dillo I, van Horik R: Lies, Damned Lies and Research Data: Can Data Sharing Prevent Data Fraud? *International Journal of Digital Curation.* 2013; **8**(1): 229-243
   http://dx.doi.org/10.2218/ijdc.v8i1.256

3. In the paragraph *Types of data publication* a threefold model is introduced to categorize data publications. It is not clear how this model relates to the terminology and categorisation presented in the *introduction.* This could be somewhat confusing for the reader.

   Furthermore, there are of course many other models available, e.g. that of the The Data Publication Pyramid, developed on the basis of the Jim Gray pyramid, to express the different manifestation forms that research data can have in the publication process:

   Reilly S, Schallier W, Schrimpf S, *et al.*: Report on integration of data and publications. October 2011. Located at:
   http://www.stm-assoc.org/2011_12_5_ODE_Report_On_Integration_of_Data_and_Publications.pc

   Or the model presented in the report: Costas, R., Meijer, I., Zahedi, Z. and Wouters, P. (2013). The Value of Research Data - Metrics for datasets from a cultural and technical point of view. A Knowledge Exchange Report, available from www.knowledge-exchange.info/datametrics

4. With respect to trustworthy digital repositories, I would like to add a few comments. First of all, in Europe a European Framework for Audit and Certification of Digital Repositories is emerging. It contains three certification standards (DSA, DIN31644/NESTOR seal and ISO13636) and three levels of certification (basic, extended and formal) see: http://www.trusteddigitalrepository.eu/Site/Trusted%20Digital%20Repository.html

   Of these three standards, only DSA has been up and running for some time now ,with 31 seals awarded and 30 ongoing self-assessments at this moment. The NESTOR seal has become available only very recently and the ISO standard is not yet officially available. The accompanying ISO 16919 standard: Requirements for bodies providing audit and certification of candidate trustworthy digital repositories, has been published very recently and now the ISO organization needs to be set up in the different countries, including the training of national auditors. The audits done by CRL are not fully official, since CRL is no formal ISO accreditation body.

   In Europe we see a growing interest in TDRs, coming from funders who want to push open data and data sharing and demand the deposit of publicly funded data in long term TDRs. Furthermore European research infrastructures and projects are also looking more and more into the issue of trust hen it comes to data sharing and a groeing number of them is incorporating (parts of) the DSA guidelines into there repositories and policies (e.g. CESSDA, CLARIN, EUDAT).

   Yet another certification procedure is offered by the ICSU/WDS to repositories that aim to become a member of the World Data System. See: https://www.icsu-wds.org/community/membership/certification

   The certification of TDRs could also help publishers/editorial boards with Data Availability Policies to point their authors to the right repositories for the long-term storage of their data.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Competing Interests: No competing interests were disclosed.

Referee Report 28 May 2014

doi:10.5256/f1000research.4518.r4543

Mark Costello
Institute of Marine Science, University of Auckland, Auckland, New Zealand

Having made some similar comments myself I must agree with this review. But a few points may merit amendment:

Abstract:
Note that what 'publication' means is the same as in print as in digital form. It does not imply peer-review or editorial oversight in either format.

Citability:
Yes, data citations are important but some datasets and web-based resources do not show how they should be cited, some journals do not allow citations to web resources in the Reference list, and even

were both possible, too many authors neglect to cite actual datasets and instead cite a web site (which may have many datasets) or a related print paper.

I agree that a DOI is not enough and only permanent if it is updated when documents are moved. A full author-tile-publisher citation as you suggest is more informative and human readable.

I do not think it is problematic to cite parts of datasets. Pages and chapters in books are already cited for example. In most datasets it is also possible to identify individual dat records. Also, the actual data used could be provided in an Appendix so the reader is left in no doubt. Neither do I think 'versioning' is a problem. Where new data are added (e.g. to a time-series) then they comprise a new dataset, as they would if published in print. Where many corrections are made they a dataset can be treated like a paper; i.e. the original can be 'retracted' and replaced, or the new version be published with the metadata stating that it is more accurate.

You mention venues for peer-reviewed data papers. For this article to advance previous articles, perhaps it could expand on these venues and how they manage the details of the peer-review process?

I have published a few papers you may find of interest:

1. Costello MJ, Wieczorek J. 2014. Best practice for biodiversity data management and publication. *Biological Conservation,* 173, 68-73. http://www.vliz.be/en/imis?module=ref&refid=234968

2. Costello MJ, Appeltans W, Bailly N, Berendsohn WG, de Jong Y, Edwards M, Froese R, Huettmann F, Los W, Mees J, Segers H, Bisby FA. 2014. Strategies for the sustainability of online open-access biodiversity databases. *Biological Conservation* 173, 155-165. http://www.marinebiology.ugent.be/component/imis/?module=ref&refid=230520

3. Costello MJ, Michener WK, Gahegan M, Zhang Z-Q, Bourne P. 2013. Data should be published, cited and peer-reviewed. *Trends in Ecology and Evolution* 28 (8), 454-461. http://dx.doi.org/10.1016/j.tree.2013.05.002

4. Costello, M.J., Vanden Berghe E. 2006. “Ocean Biodiversity Informatics” enabling a new era in marine biology research and management. *Marine Ecology Progress Series* 316, 203-214. http://www.int-res.com/abstracts/meps/v316/

5. Costello MJ, Michener WK, Gahegan M, Zhang Z-Q, Bourne P, Chavan V. 2012. Quality assurance and intellectual property rights in advancing biodiversity data publications. ver. 1.0, Copenhagen: Global Biodiversity Information Facility, Pp. 33, ISBN: 8792020496. Accessible at http://links.gbif.org/qa_ipr_advancing_biodiversity_data_publishing_en_v1.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Competing Interests: No competing interests were disclosed.

Version 1

Referee Report 06 May 2014

doi:10.5256/f1000research.4264.r4541

Mark Parsons[1], **Peter Fox**[2]
[1] Research Data Alliance, Troy, NY, USA
[2] Rensselaer Polytechnic Institute, Troy, NY, USA

General Comments:
Note: This review was written by Parsons and accepted (with some modification) by Fox. Insights likely come from conversations between Fox and Parsons, errors from Parsons.

I am very glad the authors wrote this essay. It is a well-written, needed, and useful summary of the current status of "data publication" from a certain perspective. The authors, however, need to be bolder and more analytical. This is an opinion piece, yet I see little opinion. A certain view is implied by the organization of the paper and the references chosen, but they could be more explicit. The paper would be both more compelling and useful to a broad readership if the authors moved beyond providing a simple summary of the landscape and examined *why* there is controversy in some areas and then use the evidence they have compiled to suggest a path forward. They need to be more forthright in saying what data publication means to them, or what parts of it they do not deal with. Are they satisfied with the Lawrence *et al.* definition? Do they accept the critique of Parsons and Fox? What is the scope of their essay?

The authors take a rather narrow view of data publication, which I think hinders their analyses. They describe three types of (digital) data publication: Data as a supplement to an article; data as the subject of a paper; and data independent of a paper. The first two types are relatively new and they represent very little of the data actually being published or released today. The last category, which is essentially an "other" category, is rich in its complexity and encompasses the vast majority of data released. I was disappointed that the examples of this type were only the most bare-bones (Zenodo and Figshare). I think a deeper examination of this third category and its complexity would help the authors better characterize the current landscape and suggest paths forward.

Some questions the authors might consider: Are these really the only three models in consideration or does the publication model overstate a consensus around a certain type of data publication? Why are there different models and which approach is better for different situations? Do they have different business models or imply different social contracts? Might it also be worthy of typing "publishers" instead of "publications"? For example, do domain repositories vs. institutional repositories vs. publishers address the issues differently? Are these models sustaining models or just something to get us through the next 5-10 years while we really figure it out?

I think this oversimplification inhibited some deeper analysis in other areas as well. I would like to see more examination of the validation requirement beyond the lens of peer review, and I would like a deeper examination of incentives and credit beyond citation.

I thought the validation section of the paper was very relevant, but somewhat light. I like the choice of the term validation as more accurate than "quality" and it fits quite well with Callaghan's useful distinction between technical and scientific review, but I think the authors overemphasize the peer-review style approach. The authors rightly argue that "peer-review" is where the publication metaphor leads us, but it may be a false path. They overstate some difficulties of peer-review (No-one looks at every data value? No, they use statistics, visualization, and other techniques.) while not fully considering who is responsible for what. We need a closer examination of different roles and who are appropriate validators (not necessarily conventional peers). The narrowly defined models of data publication may easily allow for a

conventional peer-review process, but it is much more complex in the real-world "other" category. The authors discuss some of this in what they call "independent data validation," but they don't draw any conclusions.

Only the simplest of research data collections are validated only by the original creators. More often there are teams working together to develop experiments, sampling protocols, algorithms, etc. There are additional teams who assess, calibrate, and revise the data as they are collected and assembled. The authors discuss some of this in their examples like the PDS and tDAR, but I wish they were more analytical and offered an opinion on the way forward. Are there emerging practices or consensus in these team-based schemes? The level of service concept illustrated by Open Context may be one such area. Would formalizing or codifying some of these processes accomplish the same as peer-review or more? What is the role of the curator or data scientist in all of this? Given the authors's backgrounds, I was surprised this role was not emphasized more. Finally, I think it is a mistake for science review to be the main way to assess reuse value. It has been shown time and again that data end up being used effectively (and valued) in ways that original experts never envisioned or even thought valid.

The discussion of data citation was good and captured the state of the art well, but again I would have liked to see some views on a way forward. Have we solved the basic problem and are now just dealing with edge cases? Is the "just-in-time identifier" the way to go? What are the implications? Will the more basic solutions work in the interim? More critically, are we overemphasizing the role of citation to provide academic credit? I was gratified that the authors referenced the Parsons and Fox paper which questions the whole data publication metaphor, but I was surprised that they only discussed the "data as software" alternative metaphor. That is a useful metaphor, but I think the ecosystem metaphor has broader acceptance. I mention this because the authors critique the software metaphor because "using it to alter or affect the academic reward system is a tricky prospect". Yet there is little to suggest that data publication and corresponding citation alters that system either. Indeed there is little if any evidence that data publication and citation incentivize data sharing or stewardship. As Christine Borgman suggests, *we need to look more closely at who we are trying to incentivize to do what*. There is no reason to assume it follows the same model as research literature publication. It may be beyond the scope of this paper to fully examine incentive structures, but it at least needs to be acknowledged that building on the current model doesn't seem to be working.

Finally, what is the takeaway message from this essay? It ends rather abruptly with no summary, no suggested directions or immediate challenges to overcome, no call to action, no indications of things we should stop trying, and only brief mention of alternative perspectives. What do the authors want us to take away from this paper?

Overall though, this is a timely and needed essay. It is well researched and nicely written with rich metaphor. With modifications addressing the detailed comments below and better recognizing the complexity of the current data publication landscape, this will be a worthwhile review paper. With more significant modification where the authors dig deeper into the complexities and controversies and truly grapple with their implications to suggest a way forward, this could be a very influential paper. It is possible that the definitions of "publication" and "peer-review" need not be just stretched but changed or even rejected.

Detailed comments:

- The whole paper needs a quick copy edit. There are a few typos, missing words, and wrong verb tenses. Note the word “data” is a plural noun. E.g., Data are not software, nor are they literature. (NSICD, instead of NSIDC)

- **Page 2, para 2:** “citability is addressed by assigning a PID.” This is not true, as the authors discuss on page 4, para 4. Indeed, page 4, para 4 seems to contradict itself. Citation is more than a locator/identifier

- In the discussion of “Data independent of any paper” it is worth noting that there may often be linkages between these data and myriad papers. Indeed a looser concept of a data paper has existed for some time, where researchers request a citation to a paper even though it is not the data nor fully describes the data (e.g the CRU temp records)

- **Page 4, para 1:** I’m not sure it’s entirely true that published data cannot involve requesting permission. In past work with Indigenous knowledge holders, they were willing to publish summary data and then provide the details when satisfied the use was appropriate and not exploitive. I think those data were “published” as best they could be. A nit, perhaps, but it highlights that there are few if any hard and fast rules about data publication.

- **Page 4, para 2:** You may also want to mention the WDS certification effort, which is combining with the DSA via an RDA Working Group:

- **Page 4, para 2:** The joint declaration of data citation principles involved many more organizations than Force11, CODATA, and DCC. Please credit them all (maybe in a footnote). The glory of the effort was that it was truly a *joint* effort across many groups. There is no leader. Force11 was primarily a convener.

- **Page 4, para 6:** The deep citation approach recommended by ESIP is not to just to list variables or a range of data. It is to identify a “structural index” for the data and to use this to reference subsets. In Earth science this structural index is often space and time, but many other indices are possible--location in a gene sequence, file type, variable, bandwidth, viewing angle, etc. It is not just for “straightforward” data sets.

- **Page 5, para 5:** I take issue with the statement that few repositories provide scientific review. I can think of a couple dozen that do just off the top of my head, and I bet most domain repositories have some level of science review. The “scientists” may not always be in house, but the repository is a team facilitator. See my general comments.

- **Page 5, para 10:** The PDS system is only unusual in that it is well documented and advertised. As mentioned, this team style approach is actually fairly common

- **Page 6, para 3:** Parsons and Fox don’t just argue that the data publication metaphor is limiting. They also say it is misleading. That should be acknowledged at least, if not actively grappled with.

We have read this submission. We believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however we have significant reservations, as outlined above.

Competing Interests: No competing interests were disclosed.

Author Response 12 May 2014

John Kratz, California Digital Library, USA

Thank you for refereeing our paper and thank you especially for delivering your report so quickly.

We submitted the paper as a review article, not an opinion piece, and it was reclassified somewhere along the way. I contacted an editor at F1000 about the issue, and I believe it will be switched back shortly. While there is undoubtedly a viewpoint inherent in the way we have organized the manuscript, it was our intention to deliver a timely summary of the current landscape as a foundation for future thinking, not to offer prescriptions or to endorse particular approaches. We have no shortage of opinions about data publication, and a true opinion piece may follow at some point, but our aim here was to remain fairly neutral. I think the paper you are asking for would also be valuable, but it's an entirely different paper from the one we have written.

That said, your report is full of suggestions for expansion of analysis and clarification of scope that would absolutely improve the paper (e.g. the question of why some issues resist consensus more than others is an excellent one), and we will certainly address them in the next version.

Competing Interests: I am an author of the selected paper.

Discuss this Article

Version 2

Reader Comment 22 Aug 2014

Leonardo Candela, ISTI-CNR, Italy

Rather than a comment, I highlight here a potential issue in Reference 3. If I'm not mistaking it should be: Lawrence, B.; Jones, C.; Matthews, B.; Pepler, S. & Callaghan, S. ***Citation and Peer Review of Data: Moving Towards Formal Data Publication*** *International Journal of Digital Curation,* **2011***, 6*, 4-37 doi:10.2218/ijdc.v6i2.205

Competing Interests: No competing interests were disclosed.

Version 1

Reader Comment 06 May 2014

Eric Kansa, Open Context (http://opencontext.org), USA

This is an excellent, and a tremendously useful overview of the issues involved in data publishing. From my perspective in archaeology, the discussion of tDAR and Open Context is useful, since these different systems try to serve different needs. You may find this poster by Beth Sheehan comparing these different

systems useful as
well: http://www.slideshare.net/asist_org/rdap14-comparing-disciplinary-repositories-tdar-vs-open-context

One small point of clarification on a minor factual point. The Journal of Open Archaeological Data (JOAD) also lists Open Context (http://opencontext.org) as a repository for data, see:
http://openarchaeologydata.metajnl.com/about/editorialPolicies#opencontext

Similarly, Internet Archaeology also lists Open Context in the same vein:
http://intarch.ac.uk/authors/data-papers.html

Competing Interests: I direct Open Context (see: http://opencontext.org/about/people), so I have a professional interest in discussions of this project.

Reader Comment 02 May 2014

Konrad Hinsen, Centre de Biophysique Moléculaire (CNRS), France

First of all, thanks for this article, which is a good introduction to the problems surrounding data publication.

One aspect which deserves more attention is the question "What is data?" Or, more precisely, which categories of data should be distinguished with respect to publication? This is related to the last paragraph of this article that starts with "Ultimately, while “data as software” is promising, data is not software." Data is indeed not software - but software is data.

I would like to propose the following categories of scientific data:

1. Observational data. This is the "raw input" of science: data from experiments, observations, polls, etc.

2. Machine-readable information generated by humans. This category includes software, input files, workflows, etc. Information for human consumption but also stored electronically could be included as well: articles, drawings, software documentation, etc.

3. Data resulting from a computation: processed observational data, output of simulations, etc.

Data in category 1 is not reproducible in any way, and thus needs to be archived and published. Data in category 2 cannot be reproduced exactly by anyone else, but could be regenerated approximately from less complete/precise data by a domain expert. Nevertheless, it should be archived and published as well in order to produce a complete and accurate record of scientific activities. Data in category 3 can be reproduced by computation if the data in categories 1 and 2 is available. It may be convenient to share it nevertheless, in particular if recomputation is expensive, but it's less fundamental than categories 1 and 2.

I believe that these categories are more useful than the traditional separation into data, software, and writeup, in particular for questions such as archiving, citing, and updating. In particular, the vague term "dataset" does not distinguish clearly between categories 1 and 3.

Competing Interests: none

Reader Comment 01 May 2014

Hans Pfeiffenberger, Alfred Wegener Institut, Germany

Dear authors,
your article is a very noteworthy and valuable, broad overview of many of the issues surrounding "data publication". I would like to offer this as recommended reading to anybody unfamiliar with the field. However, there is one omission and one erroneous/misleading statement which I strongly suggest to correct:

- In the first paragraph of "Data as the subject of a paper" you list a number of quite representative examples of data journals, but manage to omit the probably first example of a "pure" data journal (with peer review of data), ESSD, founded in 2008.
  A brief summary of ESSD's rationale and approach was published 2011 in D-Lib Magazine, doi:10.1045/january2011-pfeiffenberger

- In the second paragraph of "Citability" you write "DOI is neither sufficient nor necessary for citability- if a dataset moves and the DOI is not updated, the citation breaks and, conversely a well-maintained web-address works as well as a DOI."

  I regard this as strongly misleading, at least for a novice to the domain of publishing or identifiers/DOIs: What typically breaks, sooner or later, is a bookmark with a "normal" URL. The DOI system - which I would characterize as "handle system with a policy" - was set up to work around that fact of life. The contracts data centers (DC) have to sign with "their" (DataCite) DOI registration agency typically contain wording such as: "DC has to ensure that registered content will be available for the entire duration of the agreement." (See "contractual form" , linked to from TIB's "DOI registration" page.) Admittedly, this and other such agreements are difficult to find.

  By the way, this agreement also adresses the issue of fixity: "Once an item is registered, it may not be altered. If an item is changed, it has to be registered with a new DOI name."

Beyond those corrections, I suggest you provide the reader with some pointers about the venues where the ongoing discussions about data publication issues are actually being led. E.g., there are a number of working and interest groups at the Research Data Alliance (not just the one on Data Citation)

best regards,
Hans Pfeiffenberger

Competing Interests: I happen to be the founder and chief editor of ESSD

Reader Comment 30 Apr 2014

Chris Hartgerink, Tilburg University, Netherlands

Possibly of interest to your paper is dat, a program in development to provide version control of datasets (more so than git is able to). It has received funding recently from the Knight Foundation (see here) and is something worth looking out for in terms of data sharing, but more importantly, preservation and logging.

Thank you for writing this — it provides a succinct introduction to an important issue.

Competing Interests: No competing interests were disclosed.