1 Abstract

The movement to incorporate datasets into the scholarly record as 'first class' research products (validated, preserved, cited, and credited) has been slowly building momentum for some time, but the pace of developments picked up substantially in the last year. Data publications are beginning to spring up all over the place, but there are still significant debates over formats, processes, and terminology. This article will give an overview of the initiatives underway and the current conversation, highlighting places where consensus seems to have been reached and issues still in contention.

Data publication implementations follow variety of models that differ in, among other things, what kind of documentation is published, where the data resides relative to the documentation, and what validation is performed. Data can be published as a standalone product, as supplemental material to a traditional journal article, or with a descriptive "data paper". Confusing the situation, terms are often used by different initiatives to refer to partially overlapping concepts. The term 'published' always means that the data is public and citable, but it may or may not mean peer reviewed. In turn, data 'peer review' can refer to substantially different processes– although data paper referee guidelines are all fairly similar. There is substantial agreement on the elements of a dataset citation (which closely resembles that of a journal article) but a variety of solutions for citing subsets of datasets or datasets that change over time. Finally, some are already looking past data publication to other metaphors, such as 'data as software', for solutions to unsolved problems.

2 What does "data publication" mean?

The idea (and ideal) of researchers sharing data with one another for the common good is not a new one, but in recent years the conversation has shifted from sharing to "publishing" data. The shift in language reflects a desire to fold datasets into the scholarly record and afford them the same status as traditional research products like journal articles. Although this goal is widely shared, consensus about what "publication" means when applied to data is lacking. Two properties are agreed on: published data is **available** now and for the indefinite future, without gatekeeping by the creator (although access may be limited by subscription or acceptance of a use agreement), and it is formally **citable** in the manner of a journal article. The key aspect of a citation here is that it can be used to identify and locate the dataset from now into the indefinite future; assignment of a persistent identifier like a Digital Object Identifier (DOI) is the usual way to accomplish this. A third property is less agreed on: published data is **trustworthy** based on some community standard, usually involving peer review. Callaghan (2012) draws a distinction between data that has been shared, published, or uppercase-P Published; shared data is available, published data is available and citable, and Published data is available, citable, and trustworthy.

3 Why publish data?

The idea is to increase data sharing by a.) tieing into exisitng mechanisms for awarding credit to reward dataset publishers, and b.) formalize citation and preservation of datasets to combat the significant problem of data loss.

4 What does a data publication look like?

At present, the still-solidifying term “data publication” covers a variety of research objects published via a variety of processes. Given the huge variety of types of data, seems unlikely that any single structure will be ideal for every discipline and every dataset, but we can hope for a managable number of blueprints. Models of data publication can be classified in a variety of ways– for instance, Lawrence (2011) identifies five models based on which organization is reposnisble for what– but for my purposes here, I'll classify data publications based accompanying documentation into three categories: a traditional article, a “data paper”, or nothing.

4.1 Data publication with a traditional journal article

The most familiar model to researchers is data published along with a traditional journal article that uses it as the basis of analysis and conclusions. In the past, the article's publisher typically hosted the datset supplementary material, but that practice is being called into question. The Dryad repository publishes data underlying journal articles. Dryad makes data avialable and citeable, but any assesment of trustworthiness is done as part of the peer review of the article. For some kinds of data, in some fields, this has been the standard for a long time– microarray data, protein structure, nucelotide sequence.

For instance, the Journal of Neuroscience stopped publishing supplemental material in 2010. Journal websites aren't well suited to ensuring data preservation, and they don't provide any means of discovery except through the article. Repositories, whether

4.2 Data publication with a data paper

Data papers are a realatively new type of journal article that describe datasets (collection methods and rationale). Data papers are becoming popular in a variety of formats. What unites them is exclusion of analysis or any attempt to draw conclusions. Data papers are being published in journals dedicated to the format, such as Nature Scientific Data and GeoScience Data Journal, as well as in journals that publish other types of papers, such as F1000 Research, Internet Archaeology, or GigaScience. Most of these journals require the data to be published in a third-party trustworthy repository, although a few are associated with repositories and hande the data themselves. Data papers are peer reviewed (more later); some take the novelty or potential impact of a dataset into consideration, while others only require that the data be scientifically valid.

4.3 Standalone data publication

To be useful for anything, a dataset must have accompanying description, or 'metadata', but this needn't resemble a journal article. Standalone data publications can include rich or relatively thin metadata in structured or freeform flavors. These publications may or may not include an element of peer-review. Figshare, for instance, publishes datasets– providing accessiblilty and citablility– without any form of validation (although a Figshare dataset associated with a datapaper may have been reviewed). On the other hand, Open Context publishes very high quality archeology datasets with optional peer review.

5 How does publication work?

5.1 Availability

The clearly, the essence of publishing anything is to make it public. Science publishing has long had baked into it the notion that access must persist into the future for the use of future researchers. Preserving access to print journals has long been the job of the library, but that's changing like everything else. Likewise, published dataset must be available now and into the future. A dataset that can only be accessed with a paid subscription or after accepting a use agreement can be said to have been published, but one that is provided by its creator over email is not. The status of a dataset located on a researcher's personal website is not entirely clear.

As a practical matter, publishing a datset means depositing it in a trustworthy repository. It is relatively clear what a repository should do– keep the data, unaltered,

5.2 Citability

When a researcher uses a published data set in a paper, they should cite the dataset in the reference list. Data publications have to make this possible. This is generally facilitated by assigning a unique permanent identifier, most commonly a DOI, to the dataset. As long as the DOI is maintained, it can be used by anyone interested to locate the dataset. (It is worth pointing out that assinging a DOI does not, in itself, make something citeable– if the DOI is not maintained, the citation breaks and, conversely a well maintained URL works just as well as a DOI).

5.2.1 Simple Case

In the simplest case, there is substantial agreement that a published dataset should be cited using five elments largely familiar from journal citations: creator(s), title, year, publisher and identifier. The identifier will generally be something, such as a DOI, that can be used to locate the referenced object, as such it can be thought of as replacing the volume and page number used to find an article in a print journal. This format is consistent with the recommendation made by CODATA and with metadata required by DataCite and Thomson Reuters Data Citation Index. However, this article-descended formulation is not adequate to address some of the complications unique to datasets.

5.2.2 Deep Citation

The first major complication that datasets face is the need for deep citation. When supporting an assertion in writing, it is considered sufficiently precise to cite the entirety of the referenced journal article and leave it to the suspicious reader to identify the basis of your assertion. If only part of a dataset is used in a quantitative analysis, you may need to specify exactly the subset in question. Because datasets are so variable in structure, there will probably not be a general solution. The most common approach is to cite the entire dataset and describe the subset in the text of the paper. In some cases, it may be practical to include a date or record number range or a list of variables in the formal citation.

5.2.3 Dynamic Data

A second complication is that datasets are prone to existing in multiple versions or changing over time. In the past, the printed article was a single version of record. Web based publishing and preprint servers such as arXv.org have already complicated the matter. Data publishers are likely to allow or even encourage updating and correction of datasetes. For the results of data analysis to be reproducible, the reader must be able to obtain precisely the version of the data that the researcher used. In the case of dynamic data, that means that previous versions have to be preserved and citable.

As a practical matter, there are two kinds of dynamic data that warrant consideration: growing datasets, to which new data may be added but old data will never be changed or deleted, and reviseable datasets in which data may be added, deleted, or changed over time. Common solutions to add-on data are to include an access date, or a date or record number range in the citation. Revisable datasets are more difficult, but the most common approach is to periodically publish multiple changes as a new version with a version number that can be included in citations.

Controversy persists about dyanmic data and identifiers and different publishers have different policies. DataCite recommends but does not requre that the DOIs that they issue point to immuntable objects. Dataverse, for example, (check up) does not permit changes, but instead recommends that growing datasets be issued a new DOI periodically that refers to the “time-slice” of records added since the last DOI was issued; revisable datasets are to be periodically frozen as a “snapshot” and issued a new DOI.

5.2.4 Just-in-time Identifiers

One potential solution to both deep citation and dynamic data is to turn the identifier-issuing process on its head. Instead of a dataset publisher minting the identifier, the researcher who wants to cite a datset could mint an identifier that refers to precisely the part of the dataset that they wish to cite. The Research Data Alliance (RDA) Data Citation Working Group has put fort a sophisticated proposal suitable for database in which an identifier would wrap together a number of components including specifiying a version of the database and a query over the database that produces the cited dataset. This seems promising, but there are still many technical and policy issues that have to be resolved before this can be widely adopted.

5.3 Trustworthiness

For journal articles, peer-review is the gatekeeper to the scholarly record, meant to ensure some level of trustworthiness. Peer review serves as an initial assessment of quality; the real value and correctness of the work is determined by the relevant community after publication. The same impulse that drives the effort to capture the prestige of the term “publication” and apply it to data drives the effort to apply “peer-review” to data.

Callaghan (2012) [?] draws another useful distinction here: between technical and scientific review. Technical review assures us that the dataset has complete metadata, no missing values that aren’t allowed to be missing, etc. and generally doesn’t require domain expertise. Scientific review evaluates the methods of data collection, the overall plauisbility of the data, and the likely reuse value. Both kinds of review can be done together, or, in the case of a data paper, it’s common for the repository to do the technical review and the data journal to do the scientific review.

Publishers of data papers wrap peer review of the paper and of the datset together. An exception is GigaScience, which assignes a separate data reviewer for technical review of the dataset. Reviewer guidlines are roughly similar accross journals. Nature Scientific Data as a representative example.

While review guidelines are similar, review processes are not. Data paper peer review processes range from traditional (anonymous pre-publication review in NSD) to experimental (open post-publication review in F1000 Research).

More interesting yet are peer reivew processes for standalone datasets. NASA Planetary Data System (PDS) conducts peer review in an in-person meeting with representatives of the repository, the dataset creators, and the reviewers. Open Context goes beyond the simple accept/reject binary of traditional peer review. Instead, each dataset has a rating from 1-5 that indicates how thoroughly it has been reviewed. Essetially, a 3 indicates that the datset has passed technical review, a 4 means that it has passed editorial review, and a 5 means that it has passed external peer review.

5.4 Beyond data publication

Parsons () argues that the metaphor of “publication” is limiting, and only suited to some datasets. He identifies a number of other possible metaphors to apply to differnt kinds of data. One of these seems to be accumulating broad support:

data as software. Under this metaphor, publishing a dataset is analagous to a software release. Any subsequent changes are analagous to new versions.

The open source sofware community has already confronted many of the problems associated with data (managing versions, sharing, collaboration) and developed tools and approaches to address them. Open context came to use Mantis bug tracking software and Git with their data out of purely practical concerns.

Still issues to address: Current VCSs are designed for code (realtively small text files), not large and variegated datasets. Attribution for derived datasets is not clear, but that's likely to be a cultural issue.