171 Second Street, Suite 300
San Francisco, CA 94105
http://creativecommons.org

Recommendations for independent scholarly publication of data sets

Jonathan Rees

Creative Commons Working Paper
March 2010

In an ideal world, any data collected by a research study would be available to anyone interested in validating or building on that data, just as is the documentation describing the study itself. Some data has value that goes beyond the study for which it is generated, and getting the data to those who can use it for reanalysis, meta-analysis, and other applications unimagined by the study authors is to everyone's benefit.

Data reuse failure is receiving growing recognition as a problem for the research community and the general public. The road to reuse is perilous, involving as it does a series of difficult steps:

1. The author must be professionally motivated to publish the data
2. The effort and economic burden of publication must be acceptable
3. The data must become accessible to potential users
4. The data must *remain* accessible over time
5. The data must be discoverable by potential users
6. The user's use of the data must be permitted
7. The user must be able to understand what was measured and how (materials and methods)
8. The user must be able to understand all computations that were applied and their inputs
9. The user must be able to apply standard tools to all file formats
10. The user must be able to understand the data in detail (units, symbols)

This report considers how the genre of the *data paper,* suitably construed, might be used to help a data set survive these trials.

Cited resources together with others of interest are listed at the end.

Data Papers

A data paper is a publication whose primary purpose is to expose and describe data, as opposed to analyze and draw conclusions from it. The data paper enables a division of labor in which those possessing the resources and skills can perform the experiments and observations needed to collect potentially interesting data sets, so that many parties, each with a unique background and ability to analyze the data, may make use of it as they see fit.

An important aspect of a data paper is that those who generate the data are independently recognized for having done so, both immediately when the data paper is published and over time as others build on and cite the work.

The data paper genre is at present obscure and poorly exploited. Our search yielded just a few journals announcing that they publish data papers. Some notable examples:

- *Ecological Archives* has been publishing data papers since 2000. *EA* "provides a reward mechanism (in the form of peer-reviewed, citable objects) for the substantial effort required to compile and adequately document large data sets of ecological interest."
- *Earth System Science Data* began in 2009 and publishes data papers exclusively. It gives its purpose as "the publication of articles on original research data(sets), furthering the reuse of high (reference) quality data of benefit to Earth System Sciences."
- The *International Journal of Robotics Research* has begun to publish data papers, and has given an eloquent motivation for the practice.

Recommendations

The data paper genre can be expanded and enriched so that it takes on the role of filling all gaps in the data reuse pipeline. Here we present advice for publishers who would use the data paper genre to address the problems of data set accessibility and reuse.

Set a standard. There won't be investment in data set reusability unless granting agencies and tenure review boards see it as a legitimate activity. A journal that shows itself credible in the role of enabling reuse will be rewarded with submissions and citations, and will in turn reward authors by helping them obtain recognition for their service to the research community.

Aggressively implement a clean separation of concerns. To encourage submissions and reduce the burden on authors and publishers, avoid the imposition of criteria not related to data reuse. These include importance (this will not be known until after others work with the data) and statistical strength (new methods and/or meta-analysis may provide it). The primary peer review criterion should be adequacy of experimental and computational methods description in the service of reuse.

Complement previously published papers that present the data. Data sets usually warrant independent treatment even if the materials and methods and/or the data set itself is already published in a traditional paper. The traditional data paper has presentation of materials and methods independent of analysis as a primary purpose, and publishing data in this way achieves the most principled separation of concerns. However, recognizing the conventional practice of combining materials and methods with analysis in one paper, we propose as an alternative a second kind of data paper that provides materials and methods by citation. Such a paper can provide value by focusing on the data instead of the results, filling in missing elements such as computational materials and methods, abstract of data set, curated metadata, or data archiving, any of which may be neglected by the traditional publishing process. Following Callaghan et al. we'll call this an *overlay paper*.

Review archival status, or provide an archiving solution. To provide data many publications refer to web sites such as laboratory-based servers that lack credible archival status. The publisher of a data paper should ensure that the data set itself (and not just the paper) is archived in one or, preferably, more than one independent document or data repository so that it will be available and retain its integrity beyond the retirement of the principal investigator.

Insist on adequate documentation. Data sets that are not ready for integration with other data are not reuseable. A downstream user must be able to understand how quantities in this data set compare to quantities found in other data sets.

- Data elements such as column headings and entries must be comprehensible both by someone competent in the field and by their informatics-oriented colleagues.
- Units must be clear.
- The manner of measurement and normalization must be specified.
- Data is often the result of computational processing. The provenance of the inputs (e.g. external databases consulted) and the formulas or software used to do the processing should be provided.

This aspect of review should be carried out by someone knowledgeable in informatics. Researchers may be unaware of what is involved in using someone else's data.

Encourage use of standard file formats, schemas, and ontologies. It is impossible to know what file formats will be around in ten years, much less a hundred, and this problem worries digital archivists. Open standards such as XML, RDF/XML, and PNG should be encouraged. Plain text is generally transparent but risky due to character encoding ambiguity. File formats that are obviously new or exotic, that lack readily available documentation, or that do not have non-proprietary parsers should not be accepted. Ontologies and schemas should enjoy community acceptance.

Items that are indicated as conforming to a format, schema, or ontology should be validated as such.

Dispel legal uncertainties. Facts and ideas are not protected by copyright under traditional copyright law, but European sui generis database rights may protect a database consisting of unoriginal facts, and a data set may include copyrightable elements, such as commentary or other original expression. A user of the data may want to be able to combine the data with others for meta-analysis, for contribution to a federated knowledge base, for federated search, for annotation of other data sets, or innumerable other purposes. For widest latitude of use and best scalability, and therefore greatest return to the research community, the entirety of the data set, including any incidental copyrightable elements, should dedicated to the public domain. Note that public domain does not preclude making a request for attribution or other terms of use following community norms. Such a request may be as effective - or more effective - at getting users to follow desired practices based on norms as any attempted legal restrictions.

Conclusion

We encourage everyone involved in data sharing and reuse to take a holistic view of the data reuse problem. The attention that the various pieces of the problem are receiving is welcome, but it's not just about review, or publication, or deposit guidelines, or archiving. All parts of the system must work together if we are to create the incentives needed for adequate publication and therefore to maximize return on research investment. The data paper in both its traditional and overlay forms can pick up where conventional publications leave off and provide the advocacy for reuse that the research cycle demands.

Prepared with help from MacKenzie Smith, Tim Danford, Alan Ruttenberg, Kaitlin Thaney, John Wilbanks, and Thinh Nguyen.

Resources

- Data sharing and reuse
  - Sustainable Digital Data Preservation and Access Network Partners (DataNet). Program announcement, U.S. National Science Foundation, 2007. http://www.nsf.gov/pubs/2007/nsf07601/nsf07601.htm
  - Jeremy P. Birnholtz and Matthew J. Bietz. Data at work: supporting sharing in science and engineering. http://portal.acm.org/citation.cfm?id=958160.958215&type=series
  - Campbell EG et al. Data withholding in academic genetics: evidence from a national survey. *JAMA* 287(4):473-80, 2002. http://www.ncbi.nlm.nih.gov/pubmed/11798369
  - Samuelle Carlson and Ben Anderson. What Are Data? The Many Kinds of Data and Their Implications for Data Re-Use. *Journal of Computer-Mediated Communication* 12(2), article 15. http://jcmc.indiana.edu/vol12/issue2/carlson.html
- Data papers
  - Cornell Southeast Asia Program Data Papers Series. http://seapdatapapers.library.cornell.edu/s/seap/
  - *Ecological Archives*. http://esapubs.org/archive/default.htm
  - *Earth System Science Data*. http://earth-system-science-data.net/
  - *International Journal of Robotics Research*. http://ijr.sagepub.com/.
  - Paul Newman and Peter Corke. Editorial: Data Papers — Peer Reviewed Publication of High Quality Data Sets. *International Journal of Robotics Research* 28:587, 2009. http://ijr.sagepub.com/cgi/reprint/28/5/587
  - S. Callaghan et al. Overlay Journals and Data Publishing in the Meteorological Sciences. *Ariadne* issue 60, 2009. http://www.ariadne.ac.uk/issue60/callaghan-et-al/
- Deposit and archiving guidelines
  - National Science Board. NSB-05-40, Long-Lived Digital Data Collections Enabling Research and Education in the 21st Century. US National Science Foundation, 2005. http://www.nsf.gov/pubs/2005/nsb0540/
  - Deposit Data & Findings. Inter-University Consortium for Political and Social Research. http://www.icpsr.umich.edu/icpsrweb/ICPSR/access/deposit/index.jsp
  - The Dataverse Network Project. http://thedata.org/citation/standard
  - The Wellcome Trust. Sharing Data from Large-scale Biological Research Projects: A System of Tripartite Responsibility. http://www.genome.gov/Pages/Research/WellcomeReport0303.pdf
  - Green, T. We Need Publishing Standards for Datasets and Data Tables. OECD Publishing White Paper, OECD Publishing, 2009. http://dx.doi.org/10.1787/603233448430
- Public domain
  - Kaitlin Thaney. Sharing data on the Web. Nodalities blog, February 2010. http://blogs.talis.com/nodalities/2010/02/sharing-data-on-the-web.php
  - Thinh Nguyen. Freedom to Research: Keeping Scientific Data Open, Accessible, and Interoperable. Creative Commons, 2008. http://sciencecommons.org/wp-content/uploads/freedom-to-research.pdf