Open Data and the Social Contract of Scientific Publishing

TODD J. VISION

We owe the effectiveness of the scientific enterprise in large part to the social contract under which scientists publish their findings in such a way that they may be confirmed or refuted and receive credit for their work in return. Because of the limitations of the printed page, data have been largely left out of this arrangement. We have grown accustomed to reading papers in which tables, figures, and statistics summarize the underlying data, but the data themselves are unavailable. There are exceptions, such as DNA sequences, for which there exist specialized public repositories that authors are required to use. But the vast majority of data types do not have such repositories.

Occasionally, these "orphan data" are provided by authors as online supplements. More often, they are left for authors to share only upon request. Unfortunately, such requests are often left unfulfilled. Wicherts and colleagues (2006) requested data for 141 recent empirical articles in psychology. Though the corresponding authors had signed a certification of compliance with the American Psychological Association's Ethical Principles, which includes the principle of sharing data for independent verification, only one-quarter of the authors furnished their data after six months of repeated requests. Such practices undermine scientific credibility; in a survey of 1240 geneticists, 28 percent reported that they had been unable at some point in the previous three years to confirm published findings because of data withholding (Campbell et al. 2002).

Data are a classic example of a public good, in that shared data do not diminish in value. To the contrary, shared data can serve as a benchmark that allows others to study and refine methods of analysis, and once collected, they can be creatively repurposed by many hands and in many ways, indefinitely. For this reason, many voices in recent years have advocated for the removal of barriers to the availability and reusability of scientific data (e.g., Schofield et al. 2009), and specifically for the release of scientific data into the public domain (*http://pantonprinciples.org*).

One way to achieve this would be to require data archiving at the time of publication, when authors are motivated and able to provide their data in a reusable form. Among the geneticists surveyed by Campbell and colleagues (2002), 12 percent admitted to not honoring at least one request for published data in the preceding three years; the most commonly given reason for denying requests was the amount of effort required for compliance (80 percent of respondents). Unarchived data files are often misplaced, corrupted, or the software in which they were produced becomes obsolete. Memories fade.

Journals are in the best position to promote the practice of archiving "small-science" data upon publication, and some are now stepping up to the challenge. A consortium of journals in ecology and evolutionary biology has signed on to a joint data archiving policy (JDAP) that requires, as a condition for publication, that the original data reported in an article be archived in an "appropriate public repository" prior to publication (e.g., Whitlock et al. 2010). The policy has some flexibility; journals may offer the option of a one-year embargo on public availability in order to protect the authors' first-mover advantage. In special circumstances, longer embargoes may be granted at the discretion of an editor, and sensitive information is exempted altogether.

What should constitute an "appropriate public repository?" We could rely on authors to archive their data on laboratory and institutional Web sites. However, relatively few authors and institutions have the means or resources to maintain their own data archives, such sites are not stable, and this approach does little to promote discoverability or long-term preservation.

A better approach would be to expand the role of the online supplementary materials sections of journals. However, this possibility has some major weaknesses. Supplemental data can seldom be discovered except by manual examination of individual articles. A paywall often limits access. Publishers put few resources into maintaining supplemental data and may even fail to migrate data when journals change hands. In an eye-opening study, Anderson and colleagues (2006) reported that of the links to supplemental materials ostensibly hosted on the Web site for the journal *Genetics*, 40 percent were unavailable just one year after publication.

A third and, I would argue, superior approach is a disciplinary repository that has data as its primary focus and is shared by a scientific community larger than a single journal or publisher. The benefits of this model can best be illustrated by describing the workings of Dryad (*http://datadryad.org*), a digital repository designed specifically to enable authors to archive data upon publication and to promote the reuse of that data. The governing board of the repository is composed of representatives from a consortium of partner journals. The consortium has grown out of the original core of ecology and evolutionary biology journals that signed on to the JDAP. It currently includes more than a dozen journals, both society-owned and commercial.

One requirement for Dryad is that it be able to host any kind of orphan data.

Therefore, the format and contents of the data files cannot practically be standardized, though journals are free to require minimal content standards or format conventions should they so choose, and the articles themselves provide important context for understanding the data.

A second critical requirement for Dryad is that it minimize the burden of submission for the author. To achieve this, partner journals provide Dryad with the bibliographic information for each article in advance of publication. Then, at the time of deposition, authors follow a link to a preexisting record in the Web submission system, log in, and upload their electronic files with some optional descriptive metadata and a "read me" file. To further minimize deposition burden, Dryad is developing interfaces to enable one-stop data submission for cases where some of the data belong in more specialized repositories.

Dryad promotes data citations by assigning a unique, persistent, and resolvable digital object identifier (DOI) for inclusion in the published article. This takes the form of a DataCite DOI (*www.datacite.org*). Data are dedicated to the public domain through a Creative Commons Zero waiver (*http://creativecommons.org/publicdomain/zero/1.0*), which makes the terms of reuse both clear and nonrestrictive. A statement of community norms advises scientists who reuse the data to cite both the paper and the data as separate research products. Thus, Dryad provides a positive incentive for data archiving without erecting unnecessary barriers to data reuse.

A shared data repository such as Dryad provides additional value to its holdings. To enhance discoverability and reusability, Dryad's curators enrich each record with keywords from controlled vocabularies. They also convert or migrate file formats when needed and manage version control should there be updates. The contents are exposed to the Web using a variety of standard and emerging search and retrieval technologies.

Ensuring long-term preservation of research data requires financial sustainability. With few exceptions, funding agencies are willing to provide only ephemeral funding for data repository research and development, and end users cannot be charged if the primary goal is to make the data open and accessible. Support must come instead from the stakeholders that already support the business of scholarly publication: authors, scientific societies, research libraries, and academic institutions. Publishers, too, have an interest in supporting a repository that saves them the greater expense of hosting supplemental materials themselves.

Estimates of the combined online and print publication costs of a single scientific article range from $2000 to $10,000 (King 2007). On the basis of projections for Dryad, the marginal cost of data publication would be only a small fraction (< 2 percent) of this sum, provided that the repository has sufficient volume (on the order of 10^4 new submissions annually). Many more journals would need to participate in the consortium to achieve this economy of scale, but the potential for consortium growth is huge. Thomson Reuters indexes more than a half-million abstracts annually in its BIOSIS database (*http://thomsonreuters.com*). Sufficient volume could potentially be achieved within the largest publishing houses, but repositories managed by major commercial publishers would leave many independent scientific journals out in the cold, and would have a worryingly strong incentive to sell access to the data. In short, if we are serious about seeing research data preserved and reused, we must be willing to support the enterprise.

Permanent archives for published research data would allow us to write an amendment to the centuries-old social contract governing scientific publishing and give data their due.

References cited

Anderson NR, Tarczy-Hornoch P, Bumgarner RE. 2006. On the persistence of supplementary resources in biomedical publications. BMC Bioinformatics 7: 260.

Campbell EG, Clarridge BR, Gokhale M, Birenbaum L, Hilgartner S, Holtzman NA, Blumenthal D. 2002. Data withholding in academic genetics: Evidence from a national survey. Journal of the American Medical Association 287: 473–480.

King DW. 2007. The cost of journal publishing: A literature review and commentary. Learned Publishing 20: 85–106.

Schofield PN, Bubela T, Weaver T, Portilla L, Brown SD, Hancock JM, Einhorn D, Tocchini-Valentini G, Hrabe de Angelis M, Rosenthal N. 2009. Post-publication sharing of data and tools. Nature 461: 171–173.

Whitlock MC, McPeek MA, Rausher MD, Rieseberg L, Moore AJ. 2010. Data archiving. American Naturalist 175: 145–146.

Wicherts J, Borsboom D, Kats J, Molenaar D. 2006. The poor availability of psychological research data for reanalysis. American Psychologist 61: 726–728.

Todd J. Vision (tjv@bio.unc.edu) is with the Department of Biology at the University of North Carolina in Chapel Hill, and the National Evolutionary Synthesis Center in Durham, North Carolina.

doi:10.1525/bio.2010.60.5.2