Cultural Change in Data Publishing Is Essential

VISHWAS CHAVAN, LYUBOMIR PENEV, AND DONALD HOBERN

Stakeholders across the globe have stressed the urgent need to improve the accessibility and fitness for use of biodiversity data. Huang and colleagues (2013) echoed this concern in *BioScience*, but argued against the data-paper concept as an appropriate mechanism to encourage the publishing of biodiversity data.

We agree wholeheartedly with Huang and colleagues (2013) that “a joint data-publishing and -archiving policy by databases and journals” (p. 5) should be at the heart of the response to these needs. Only approaches that lead to stable, long-term access to well-described, high-quality data sets will be successful. We also believe, however, that data papers are fully compatible with this solution and that Huang and colleagues (2013) misunderstood or left unattended several key characteristics of the data-paper concept.

A data paper is a scholarly journal publication whose primary purpose is to describe a data set or a group of data sets, rather than to report a research investigation (Chavan and Penev 2011). Therefore, it contains facts about data, not hypotheses and arguments in support of those hypotheses based on data, as are found in a conventional research article. Its purposes are threefold: to provide a citable journal publication that brings scholarly credit to data publishers; to describe the data through structured, human-readable extended metadata; and to bring the existence of the data to the attention of the scholarly community.

Data papers and joint data-publishing and -archiving policies by databases and journals in combination address a cluster of related requirements: the persistent, stable, long-term archival of data; a better standardization of data sets and improved confidence about their quality and fitness for use; reliable linkages between scholarly articles and the underlying data; and global acceptance that research data should be readily accessible and reusable.

Publishing data (whether in supplementary files published along with scientific articles or in a separately referenced data paper) should be seen as a necessary step in the publishing process. Every data set published by these means should be documented and included in a stable archival repository.

Such repositories can provide data to researchers in standard views that include the metadata and a data overview. These views can be essentially the same whether the data are published as supplementary material or as a data paper and could be made identifiable through a DOI (digital object identifier) for the data set. In some cases, peer review could occur after the data were published, depending on the availability and interest of the appropriate experts. In any case, the result will be to encourage the reuse of data beyond the original purpose and user group, thereby providing benefits from the investments made in data publishing.

We believe that there are several reasons for which a data paper may be preferred. The first is to address the crucial need to mobilize historical data sets that offer views of the past composition of biodiversity in specific locations or that offer time series for analysis. Mobilizing these data—for example, information about specimens housed in natural history collections—has value quite independent of the possibility of publishing a related research article (Berendsohn et al. 2010). In these contexts, data papers (with peer-reviewed metadata) may help foster data mobilization of and access to currently nondigitized or unpublished data sets. This additional visibility will also serve to multiply the effects of existing public investments and may provide new opportunities to fund digitization. Second, publication of a data paper may be the most effective way to make available a data set with all necessary documentation on methods and other details, while also giving the publishing researcher value in the form of a citable publication. In both of these cases, skilled amateurs may also be able to contribute as coauthors working with professionals to mobilize their data.

We agree with Huang and colleagues (2013) that well-defined quality criteria should be considered for the acceptance of data papers. Examples of such criteria are listed in the Pensoft Data Publishing Guidelines (Penev et al. 2011), and in our view, these criteria will evolve as more and more data papers are submitted and published.

As was noted by Huang and colleagues (2013), the peer-reviewed data paper may, at present, be seen as an extra burden on review processes and resources. However, if the data themselves are not addressed as part of a peer-review process, there are fewer controls on the standards of data and metadata, and their reuse can be problematic (Costello et al. 2012). For this reason, as the data paper becomes a mainstream tool for publication, a new breed of peer reviewers will emerge, with an understanding of data collection, management, and publishing, as well of the potential uses of the data. Building a critical mass of such reviewers will take some time, but it will bring data managers closer to the researchers and will, in turn, strengthen the credibility of both scholarly publications and the peer-review process.

We envisage that the costs of publishing data papers would become an integral part of the research-funding