doi.org/10.22498/pages.26.2.52

Open data and the publishing landscape

Alicia J. Newton[1,2]

Every research paper is underlain by data. But, until relatively recently, the accessibility and archiving of this data has been an afterthought to the published paper. Technological advances and efforts to increase reproducibility have pushed data availability to the forefront.

Papers in the paleosciences have always been data rich: Emiliani's (1955) work illustrating glacial-interglacial cycles relied on twelve cores sampled at 10 cm intervals. And from CLIMAP (Climate: Long range Investigation, Mapping, and Prediction) to PAGES 2k Network, paleoclimatologists have also been quick adopters of big-data approaches, combining individual records to generate global maps of temperature change through time. The value of these types of efforts is immediately recognizable by the wider paleo community. However, the open data practices that support these efforts have grown more slowly.

Today, the data that underlie the CLIMAP reconstruction are available from a variety of repositories found by a simple internet search. However, at the time of the compilation in 1981, files would have been shared peer to peer, with some smaller data tables contained within publications. Surprisingly, peer-to-peer sharing remains a prominent mode of data sharing, with 31% of Earth scientists opting not to archive data in a repository or include data in supplementary materials of publications (Stuart et al. 2018).

Peer-to-peer sharing is quick, but has a number of downsides. On a practical level, data that isn't archived may be unprotected. Many scientists still store data on personal or external hard drives, where it is vulnerable to theft, format or program obsolescence, or simply an errant cup of coffee (Baynes 2017). On a broader level, requiring personal outreach to obtain data can hinder scientists with fewer connections or who face a language barrier. And data stored in this manner may be lost when scientists retire or leave academia.

In the paleosciences, and geosciences more broadly, data archiving in open repositories takes on an additional importance: it can be exceedingly expensive to obtain samples through means such as ocean or ice-core drilling, and materials such as meteorites or certain fossils can be extremely rare. And some samples may prove irreplaceable as material is lost through erosion, land-use changes, and as glaciers melt. As signatories to the Coalition on Publishing Data in the Earth and Space Sciences (COPDESS) Statement of Commitment (copdess.org/statement-of-commitment), publishers have recognized this importance.

Figure 1: Respondents to a survey of 1,500 scientists raised substantial concerns about the reproducibility of the published literature, and reported their own experiences with failure to reproduce results (Baker 2016). Open data is one avenue being explored to help increase confidence in the scientific record. Image credit: Edwyn Mayhew.

Why open data?

In 2016, 90% of researchers surveyed by *Nature* raised major concerns about the reproducibility of the scientific record, with few people convinced that all of the published literature would be reproducible (Fig. 1; Baker 2016). In the Earth and environmental sciences, about 40% of respondents were unable to reproduce even their own work in at least one instance; over 60% were unable to reproduce the findings of others. Increased openness of data, methods, and code can help improve confidence in the scientific record.

Geoscientists certainly recognize the importance of data sharing, with 69% of Earth scientists making their data available in a repository or supplementary materials (Stuart et al. 2018). This movement towards data availability is driven by a growing recognition that making supporting data open offers benefits for both data producers and the broader scientific endeavor (Schmidt et al. 2016). Specifically, data sharers are motivated by the desire to help accelerate scientific research, and also to increase the visibility and dissemination of their research output (Stuart et al. 2018). Intriguingly, the survey found that funder and publisher requirements were not as strong of an incentive to release data.

But is available data always open data? In the geosciences, 28% of respondents only made data available in the electronic supplementary materials (Fig. 2). Whether or not this material sits behind a paywall varies by publisher: *Nature Geoscience* and the Nature Research journals make this material free to read, but other journals require a subscription for access. The format and content of the supplementary-data

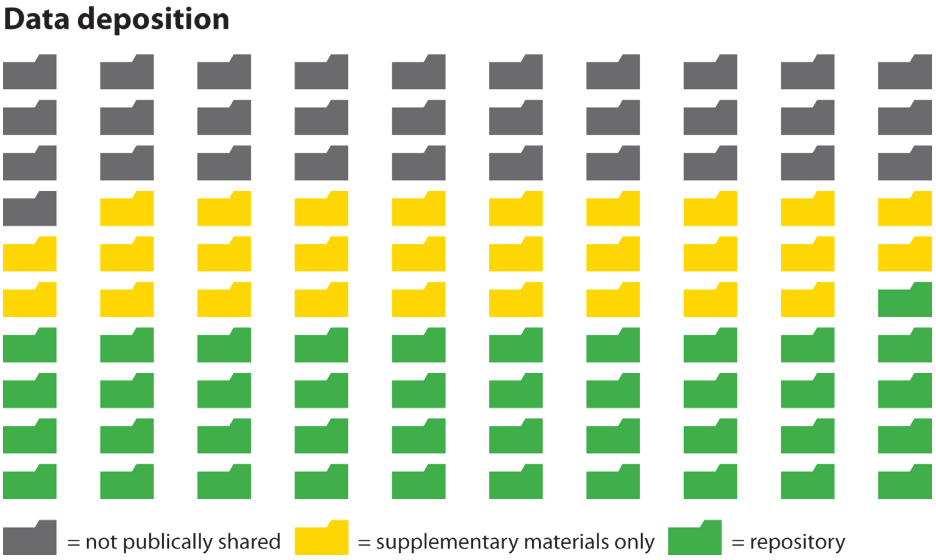

Figure 2: How discoverable is the data behind a paper? Stuart et al. (2018) surveyed 365 Earth scientists about their experiences and if and how they made the data associated with their work available. Each folder represents 1% of the survey response. Image credit: Edwyn Mayhew.

tables may also be less than ideal, with pdf tables not always easy to import into other software.

Springer Nature has started a trial in which electronic supplementary materials from articles published in BioMed Central and Springer Open journals is hosted on Figshare. These files are freely accessible and uniquely identifiable with a separate DOI, helping the data behind a paper to find its own audience (Hyndman 2016).

Recognition and reward

Beyond altruism and a desire to contribute to scientific advances, there are other benefits for researchers who make their data widely available. In *Paleoceanography*, articles that were published alongside publicly-available datasets saw a 35% greater citation rate than the journal average (Sears 2011). Across all disciplines, data availability provides a citation boost between 9 and 50% (Baynes 2017).

The rise of peer-reviewed data journals helps to provide credit for data generators, beyond a traditional scientific publication. Journals like *Scientific Data* and *Earth System Science Data* publish "data descriptors". These articles describe the collection and processing of a dataset that has been released through a public repository. The descriptors provide sufficient metadata and related information to allow for easy use of the data, but refrain from interpretation and extensive analysis. Data descriptors also can accompany a traditional scientific publication, and can allow for an expanded dataset to be released: for instance, $\delta^{13}C$ data that was collected alongside oxygen isotopes but not featured in the interpretation or additional parts of a record that were generated but not the focus on the paper. In these instances, the data descriptor can have a different lead author than the main paper, perhaps giving due credit to a student researcher who led the data collection but played a smaller role in the interpretation.

Data-descriptor papers can also serve as a way to release and promote the reuse of datasets that might otherwise live in a proverbial desk drawer: data from student summer projects, null results, or the never-written up thesis chapter can all be released for others to work from and build upon. In these cases, the data generators can receive appropriate recognition for their work – and potentially the reward of citations of the data descriptor and data set – even if the interpretation of the data might not be sufficient to warrant a traditional publication.

Into the future

In 2015, COPDESS released a statement of commitment, which was signed by most Earth and environmental science publishers and data repositories. Signatories from the publishing side agreed to promote the use of appropriate community repositories to their authors, and direct authors to relevant resources, for instance through lists maintained at the COPDESS website. The statement also encouraged publishers to develop clear statements about requirements for data availability. The Nature Research journals have long required authors to make materials, data, and code available without undue qualification. Nature Research also encourages authors to freely release data through repositories (nature.com/authors/policies/availability.html). Data-availability statements, which are now available to readers without a subscription, tell readers how to access the data reported in the manuscript, as well as any previously published data used in the analysis (Nature 2016; Hrynaszkiewicz et al. 2016). Code-availability statements require authors to report whether any code associated with the work is accessible.

Of course, much of this data still remains in supplementary information (Fig. 2), and may be only partially accessible, or lacks the essential metadata and standardization that would be provided by curators at a repository. Led by AGU, some signatories to the original COPDESS statement are addressing this concern through the Enabling FAIR Data Project. This project, which is supported by Nature Research and other publishers, will support authors to make sure that the data behind their publications are Findable, Accessible, Interoperable, and Reusable (FAIR; Wilkinson et al. 2016). Importantly, the National Computational Infrastructure of Australia is also supporting the project, providing the expertise required to start to tackle the terabyte-sized elephant in the room that is model output.

Although these and other challenges remain, the combined efforts of funders, publishers, repositories, and open-data advocates are ushering in a new era of data openness. Open data helps ensure the integrity of the scientific record, while new metrics and venues ensure that data generators are recognized and rewarded for their work. And the community stands to benefit as well, as increasingly easy data access facilitates powerful big-data approaches to understanding past environments.

AFFILIATIONS

[1]Nature Geoscience, Nature Research, London, UK
[2]Now at: The Geological Society of London, London, UK

CONTACT

Alicia J. Newton: aliciajillnewton@gmail.com

REFERENCES

Baker M (2016) Nature 533: 454-454

Baynes G (2017) In: The state of open data 2017. Holtzbrinck Publishing Group, 17-19

Emiliani C (1955) J Geol 63: 538-578

Hrynaszkiewicz I et al. (2016) Standardising and harmonising research data policy in scholarly publishing. bioRxiv, 7 pp

Hyndman A (2016) New partnership with Springer Nature to make research more accessible. Figshare blog

Nature editorial staff (2016) Nature 537: 138

Schmidt B et al. (2016) PLOS one 11: e0146695

Sears JRL (2011) Data sharing effect on article citation rate in paleoceanography, AGU Abstract IN53B-1628

Stuart D et al. (2018) Practical challenges for researchers in data sharing. Figshare, paper

Wilkinson MD et al. (2016) Sci Data 3: 160018