doi.org/10.22498/pages.26.2.50

Building open data: Data stewards and community-curated data resources

John W. Williams[1,2], D.S. Kaufman[3], A. Newton[4,5] and L. von Gunten[6]

Open data advance the pace of discovery in the paleogeosciences. Community-curated data resources and data stewards, together, offer a solution for jointly maximizing the volume and quality of open data. All can assist, at both individual and institutional levels.

Open data, long a good idea, are now mission-critical to advancing and accelerating the pace and breadth of discovery in the paleogeosciences. We seek to understand the past dynamics of the Earth system and its interacting subsystems, across a wide range of timescales, and to use this knowledge to inform society in a new era of global change. However, the scale of the system is too vast, and the volume and variety of data too large, for any single investigator or team to be able to integrate it. Open scientific data, gathered into curated data resources, are essential to integrating this information at scales beyond the capacity of any single team. Such data can then support big-data applications, where inferential power is proportional to data size and richness, such as machine learning, proxy system modeling (Dee et al. 2016), and data-model assimilation (Hakim et al. 2016). Ultimately, the goal is to form an open architecture of scientific data as complex, deep, and interlinked as the Earth system itself.

The benefits of open data extend beyond scientific objectives. For individual investigators, open-data resources provide services of data archival and increasing data visibility. In the genetics literature, papers with published data have a 9% higher citation rate than similar studies without published data (Piwowar and Vision 2013). Open data enable interdisciplinary research and knowledge exchange across disciplines. Open data also empower early-career scientists and scientists from the Global South, enable transparency and reproducibility, and return the fruits of publicly and privately funded research to the public domain (Soranno et al. 2014).

Multiple initiatives are underway to support and encourage best practices in open data. Publishers have launched the FAIR initiative: data must be findable, accessible, interoperable, and reusable (Wilkinson et al. 2016). Funding agencies are setting firmer standards for publicly funded data (National Science Foundation 2018). Multiple authors have called for open data (Soranno et al. 2014; Schimel 2017; Kaufman and PAGES 2k special-issue editorial team 2018). Open code and software are becoming the norm, facilitated by open-source languages (e.g. R, Python), platforms for sharing code (e.g. GitHub, BitBucket), and notebooks for sharing scientific workflows (e.g. RMarkdown, Jupyter).

Nonetheless, both cultural and technical barriers remain (Heidorn 2008), with only 25% of geoscientific data submitted to open-data repositories (Stuart et al. 2018). Most scientists are willing to share data once published, but many lack the time to prepare datasets and metadata for open publication, or the training and tools to do so efficiently. Some communities lack established data standards and repositories, with particular difficulties in finding an appropriate home for terabyte-scale datasets. Systems for data citation and provenancing remain underdeveloped, so it is hard for scientists to receive the credit due for data publication. Data curation adds value to open data, thereby navigating the big-data challenge of maximizing both data volume and veracity (Price et al. 2018), but effective data curation requires dedicated time by experts, which needs to be recognized and rewarded.

These challenges to open data are real but tractable and can be resolved through a combination of cultural and technological solutions.

One key emerging solution is the combined rise of community-curated data resources and linked networks of data stewards (CCDRs; Figs. 1, 2). CCDRs serve as loci where experts can contribute and refine data, establish data standards and norms, and ensure data quality. If open data are a commons, then CCDRs provide a governance framework for managing the commons. In this framework, data stewards (or data editors, see Diepenbroeck, this issue) are positions of service and leadership that are equivalent in function and prestige to journal editors, dedicating a portion of their time and expertise to ensure that published data are of high quality and meet community standards. The broader cultural goal is to establish norms of data openness – in which we commit to contributing our data to community data resources – and data stewardship, in which

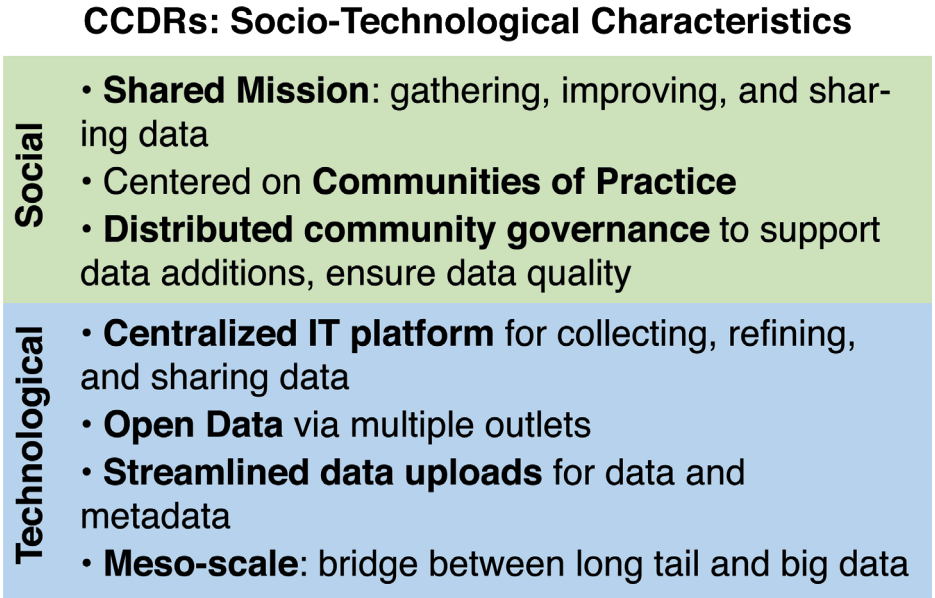

Figure 1: Community-curated data resources (CCDRs) as both social and technological solutions for supporting open data. Social characteristics include a shared scientific mission, communities of practice centered on domain experts, and governance mechanisms that facilitate participation and leadership by a broad and diverse base of experts. Technological characteristics include a central platform with support for uploading, curating, and providing data; and systems that facilitate open data access and data uploads. Because CCDRs are closely tied to their expert communities, they tend to be meso-scale intermediaries between individual data generators and big-data initiatives.

we commit to adding value to community data resources on an ongoing basis.

Multiple related initiatives are underway to build open and high-quality community data resources, stewarded by experts. Publishers have created journals specifically devoted to data publication (Newton, this issue). In paleoclimatology, PAGES 2K has established pilot examples of open data and data stewardship for global-scale data syntheses (PAGES 2k Consortium 2017). The LiPD and LinkedEarth ontologies provide flexible data standards for paleoclimatic data, with editors able to approve ontology extensions (McKay and Emile-Geay, this issue). The Neotoma Paleoecology Database has established a system of member virtual constituent databases, each with data stewards charged with prioritizing data uploads and defining variable names and taxonomies (Williams et al. 2018). The Paleobiology Database uses data authorizers to ensure quality data uploads (Uhen et al. 2013 and this issue). Some efforts focus on curating primary measurements and others on higher-level derived inferences (McKay and Emile-Geay, this issue).

Technologically, the broad need is to move open-data resources from systems of record to systems of engagement (Moore 2011), in which we move beyond models of submitting datasets to static data repositories to systems that support crowdsourcing and ongoing efforts to publish and improve data. Such infrastructure must support data discovery, archival, citation, tracking, annotation, and linking. Flexible and extensible data models are needed to support both existing and new proxies (McKay and Emile-Geay, this issue). Controlled vocabularies and common semantic frameworks are needed to tame the heterogeneity of proxy measurements. Systems for data annotation are needed to flag and correct data errors. Systems for microattribution and provenancing are needed to track data usage from initial publication to subsequent incorporation into broad-scale data syntheses. Assigning DOIs to datasets is a first step; subsequent steps are to include these DOIs in all future publications to appropriately credit data generators. Journals and citation indices will need to adopt linked data systems, tracking data usage, with ability to link to thousands of individual records, so as to avoid arbitrary limits caused by fixed limits to the number of references. New tools are needed that streamline the collection and passing of data from point of collection to data resource. Because effort is the main barrier to open data, good data management should be maximally automated.

For open data to power the next generation of scientific discovery, we must all pitch in. Scientists must commit to making their data available in open public repositories, join governance, and serve as data stewards. Publishers, as they adopt FAIR data standards, should endorse and support open community data resources that meet these standards. Funding agencies should support development of open-data standards for data types where none yet exist and provide modest but sustained support for open-data resources, under the logic that costs of supporting CCDRs are cheap relative to costs of regenerating primary data. We must launch data-mobilization campaigns that are science driven (e.g. PAGES 2k Consortium 2017), using these campaigns to prioritize rescues of dark data. Professional societies should establish mechanisms to endorse community data standards and open platforms and, where possible, provide support via a portion of membership dues. Just as professional journals were the mainstay of communicating scientific knowledge in the 19th and 20th centuries, open, high-quality community data resources will be a mainstay of communicating and advancing knowledge in the coming decades.

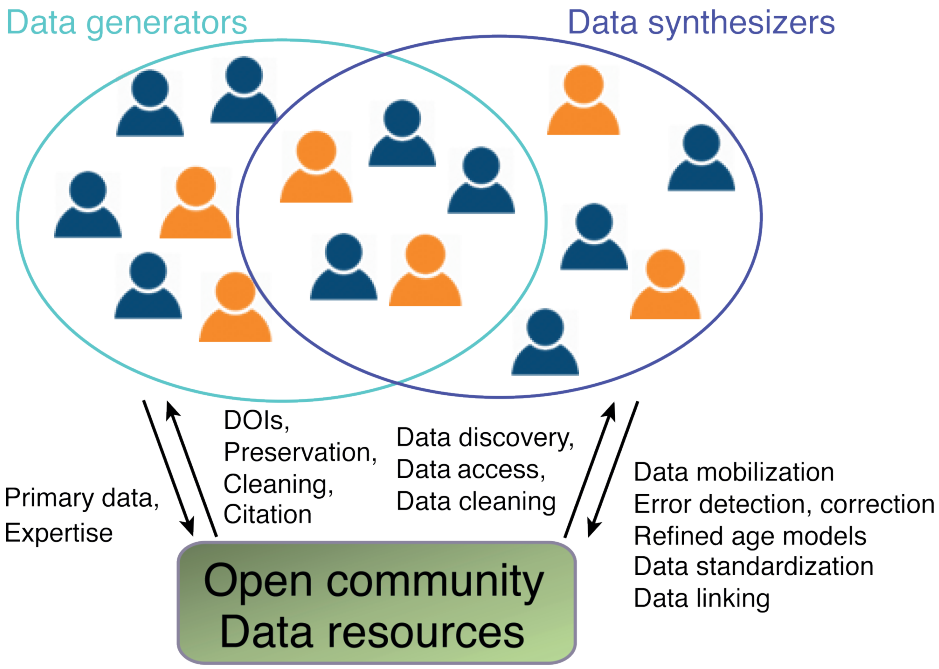

Figure 2: Paleodata CCDRs and their relationships of engagement with their overlapping research communities of data generators, stewards, and synthesizers. Data generators provide the primary data to CCDRs and receive in return DOIs for data citation and tracking and assistance in meeting community data standards. Synthesizers benefit from CCDRs through the services of improved data discovery, access, and cleaning, while returning to CCDRs the services of data mobilization for dark data, detection and correction of errors in CCDRs, updated and improved age models, and assistance in linking CCDRs with other data resources. Data stewards (orange), drawn from both communities, support data curation and ensure that community data norms are met, akin to the role of editors in peer-reviewed journals.

AFFILIATIONS
[1]Department of Geography, University of Wisconsin-Madison, USA
[2]Neotoma Paleoecology Database
[3]School of Earth and Sustainability, Northern Arizona University, Flagstaff, USA
[4]Nature Geosciences Editorial Office, London, UK
[5]Geological Society of London, UK
[6]PAGES International Project Office, Bern, Switzerland

CONTACT
John (Jack) W. Williams: jww@geography.wisc.edu

REFERENCES

Dee SG et al. (2016) J Ad Model Earth Sy 8: 1164-1179
Hakim GJ et al. (2016) J Geophys Res Atmos 121: 6745-6764
Heidorn PB (2008) Libr Trends 57: 280-299
Kaufman DS, PAGES 2k special-issue editorial team (2018) Clim Past 14: 593-600
Moore G (2011) Systems of engagement and the future of enterprise IT. A sea change in enterprise IT. AIIM, 14 pp
National Science Foundation (2018) Data and Sample Policy. nsf.gov/geo/geo-data-policies/ear/ear-data-policy-apr2018.pdf
PAGES 2k Consortium (2017) Sci Data 4: 170088
Piwowar HA, Vision TJ (2013) PeerJ 1: e175
Price GJ et al. (2018) Nature 558: 23-25
Schimel D (2017) Front Ecol Environ 15: 175
Soranno PA et al. (2014) BioScience 65: 69-73
Stuart D et al. (2018) Practical challenges for researchers in data sharing. SpringerNature, 17 pp
Uhen MD et al. (2013) J Vert Paleontol 33: 13-28
Wilkinson MD et al. (2016) Sci Data 3: 160018
Williams JW et al. (2018) Quat Res 89: 156-177