

CHAPTER 11

Peering Outward: Data Curation Services in Academic Libraries and Scientific Data Publishing

Patricia Hswe

This work is licensed under the Creative Commons Attribution License 4.0 (CC BY 4.0).

Introduction

In the sciences, data are so pivotal they can be considered a chief currency the domain deals in—in more than one sense. First, data are central to the reproducibility of research results, which aid in verifying science. As such, they signify an asset, and dissemination of them enhances their value. Second, data enable existing research to be expanded upon and transformed. The recycling and reuse of data can lead to new types of experimentation, adding, as well, to the value of data. Third, data help keep science itself current, making the timely sharing of them all the more essential. It is little wonder that, embodying the dependency for the way science is done, data are viewed—in particular, of late, by federal grant-funding agencies and the US Office of Science and Technology Policy (OSTP)—as worth sharing and thus worth making broadly discoverable and accessible. Freely available data are public goods.

As this “democratization of data” continues apace, spurred by federal mandates, data publishing offers a new frontier for various entities having a vested interest in their currency and in their longevity. Besides governments and grant agencies, these stakeholders include—but are not limited to—researchers, scholarly publishers, archives, data repositories, and academic libraries. For many of them, the publication of data offers both challenges and opportunities, though for a range of different reasons (Kratz and Strasser 2014). For example, while a majority of researchers may support broader dissemination of quality data that publishing standards would likely foster, the reward structure for most promotion and tenure cycles in the humanities, social sciences, and sciences continues to favor the publication of articles and monographs over that of data sets (Griffiths 2009). For many science journal publishers, how to publish data remains a puzzle to be solved, although there are communities of interest, such as in biodiversity research, making sizable strides toward the creation of frameworks for data publishing (Chavan and Ingwersen 2009) and advocating for journal-based data publishing policies (De Wever et al. 2012). There is also debate about the use of the term *publication* in this arena and how it functions as a metaphor, though not necessarily the most felicitous one. Because scientific data

are a special case, consisting of many variables in the sense of formats, types, metadata, versions, and standards, and because *publication* is innately connected, in the production of research literature, with articles and monographs, it may not be appropriate to apply *publication* to describe the wide dissemination of data (Parsons and Fox 2013).

For academic libraries, traditionally viewed as keepers of data and content, the concept of data publishing—while still a new frontier—is less strange than it may initially appear, largely because important groundwork has already been laid. First, efforts to develop and promote standards for data citation, which is closely tied to data publication, have long involved libraries, as reflected in the library membership of DataCite, the organization that has led work in this area;¹ libraries are in the business of facilitating discovery and access—it makes sense that data about data, such as citations, matter to them. Second, many academic libraries, whether on their own or in collaboration with a university press, run scholarly publishing operations and are familiar with the processes thereof, including the implementation and customization of software applications for publishing, the establishment of criteria for publishing, and the design and development of production workflows. Third, a great number of libraries also manage institutional repositories (IRs), a primary purpose of which is to preserve and ensure persistent access to the scholarly record; as part of this mission, some libraries are sharing data sets via their IRs.² Finally, even before the data management plan (DMP) requirement from the National Science Foundation (NSF) came into effect, libraries had started building out services for e-science data support and anticipating needs for data publishing (Soehner, Steeves, and Ward 2010). The NSF DMP mandate mobilized many libraries to respond with new services and tools, such as the DMPTool (<https://dmptool.org/>), and to revamp existing services for enhanced relevance to faculty and students in light of the requirement. Because it is early days yet for library-based data curation services, professional development opportunities for librarians to become more informed about data curation infrastructure, practices, and services have also been on the rise. Examples are the E-Science Institute (offered by DuraSpace and the Digital Library Federation), the Data Scientist Training for Librarians course at the Harvard-Smithsonian Center for Astrophysics, and the New England Science Boot Camp for Librarians, which has an e-science focus. Academic libraries are also creating new roles, such as data management services librarian positions, to concentrate on this area more strategically. The Council on Library and Information Resources (CLIR), which has sponsored a library-based postdoctoral fellowship program for the humanities since 2004, received funding in 2013 from the Alfred P. Sloan Foundation to make possible new postdoctoral opportunities, also centered in academic libraries, for data curation in the sciences and social sciences. All of these related activities and tactics are preparing academic libraries well for collaboration with other stakeholders, such as researchers

¹ DataCite members consist of data services and centers, information science and technology institutes and councils, research institutes, and libraries. Library members include the British Library, California Digital Library, the German National Library of Medicine, Purdue University Libraries, and Harvard Library.

² Examples of IRs that accept and preserve data sets are Penn State's ScholarSphere (<https://scholarsphere.psu.edu>), Purdue University Research Repository (PURR, <https://purr.purdue.edu>), University of Minnesota's Digital Conservancy (<http://conservancy.umn.edu>), and University of California at San Diego's Digital Collections (<http://library.ucsd.edu/dc>).

and publishers, in advancing data publishing for the sciences (CLIR 2013).

Data curation services in libraries are poised to help make strides in science data publishing. A chief objective in data curation is ensuring that data are shareable. To provide curation services for research data is, in part, to foster channels of access to those data, such as through citation and publication. This chapter teases out the synergy between publishing services and data curation services in libraries. It reports on the current status of each type of service, providing context and drawing out comparisons between library publishing and data publishing. The complications surrounding peer review of data sets are also examined. Such background sets the stage for assessing the state of data publishing in the sciences by looking in brief at data policies currently enacted by journal publishers for associating articles with data sets, data repositories that publishers and researchers use for linking data with publications, and the genres of the data journal and the data paper. It also captures briefly what some programs and services in publishing and data curation at academic libraries are currently accomplishing in data publishing. As the chapter suggests, the paradigm for publication of data in the sciences seems always to be shifting. The goal of the chapter, however, is to lay a foundation for understanding scientific data publishing, as well as the role that data curation services in libraries can play in it.

But First, the Basics: Data Curation and Sharing versus Publication

In concept, data curation is the “active and ongoing management of data through its lifecycle of interest and usefulness” to the scholarly and scientific research enterprise (Cragin et al. 2007). In practice, it encompasses a range of activities: collection of data sets, often including a selection and appraisal process; documentation and description of them in accordance with a community’s best practices and standards to optimize for sharing, discovery, and retrieval; assurances for dissemination, access, use, and reuse so that analysis, integration, and visualization of data may take place; and storage, preservation, and migration for persistent access. (Higgins 2008; Michener et al. 2012) Important to mention, as well is that curation tracks usage not only for repurposing possibilities but also to inform future deaccessioning measures and decisions. Library services for data curation can address both external needs, such as those of researchers, and internal needs, such as those of library professionals who work with researchers. They also address the needs of library collections, such as digital collections, including those that are IR-based. As a conduit for access and sharing, the publication of data may be considered an integral activity in the curation of data.

To afford an appreciation of the data publishing landscape, it helps to know what is generally meant by *data publication*. Read (2013) offers a well-conceived definition:

A data publication takes data that has been used for research and expands on the “why, when and how” of its collection and processing, leaving an account of the analysis and conclusions to a conventional article. A data publication should include metadata describing the data in detail such as who created the data, the description of the type of data, the versioning of the data, and most importantly where the data can be accessed (if it can be accessed at all). The main purpose of a data publication is to provide adequate information about the data so that it can be reused by another researcher in the future, as

well as provide a way to attribute data to its respective creator. Knowing who creates data provides an added layer of transparency, as researchers will have to be held accountable for how they collect and present their data. Ideally, a data publication would be linked with its associated journal article to provide more information about the research.

Read's précis covers chief attributes of published data, including information about data generation and processing and detailed metadata that captures versioning and leads to access. Publication of data enables adequate context for encouraging reuse and crediting the producer of the data. A tacit yet critical requirement for ongoing access to data is the preservation of them. In effect, much of this definition embodies the central activities of data curation in practice. As proposed by Read and the leaders in the field from whom he derived the above synthesis, the publication of research data hinges largely on the curation of them.

It is also worthwhile distinguishing between data sharing and data publishing. Sharing data could mean making data sets available at the website for one's research laboratory, e-mailing data at a colleague's request, depositing them into a repository, or, as Read observes in his definition, linking data to an article publication. The DMP mandates issued by the NSF and the National Endowment for the Humanities (NEH), as well as a similar policy long in effect at the National Institutes of Health (NIH), stress the sharing and availability of data, rather than the publishing of them (Costello et al. 2013)—although the publication of research results is strongly encouraged, if not required. No doubt, the emphasis on sharing is owed to a combination of at least the following factors: (1) the word *sharing* bespeaks the broadest possible sense of distributing or circulating content; (2) a progressively accepted practice, among communities of interest with stakes in NSF or NIH funding, is the sharing of data via disciplinary or data-specific repositories, such as GenBank (<https://www.ncbi.nlm.nih.gov/genbank>), Ocean Biogeographic Information System (www.iobis.org), and the Predicted Crystallography Open Database (www.crystallography.net/pcod); and (3) without a clear understanding of what the publishing of data means across these communities of interest, funding agencies can hardly require that data resulting from a supported project be *published* rather than simply made publicly available for sharing.

A related nuance to consider is that, just as the deposit of a research paper into an IR is often held up as an example of open-access *publishing*, albeit with a “lowercase p,” so may data sets that are deposited into a disciplinary repository, or even an IR, be understood to be similarly published. Cornell University's Research Data Management Service Group (RDMSG) codifies data publishing in a comparable manner. It takes *data publication* to mean “all strategies by which an investigator might make their [*sic*] data available to a broader audience” (Cornell University 2014, under “Data Publication”). The RDMSG categorizes or classes these strategies accordingly: (1) deposit to a discipline-specific data repository; (2) submission to a journal publisher in conjunction with a related publication; (3) deposit to an IR; and (4) publication via an independently developed infrastructure for data distribution. Whereas the first three strategies are straightforward, the last one is less clear. The RDMSG does not elaborate on the final strategy listed, but it could be taken to mean publishing frameworks tailored for distinct types of

data, such as marine science data (Brauer and Hasselbring 2013), polar scientific data (Wenfang et al. 2013), and biodiversity data (Chavan and Ingwersen 2009), or to mean the infrastructure that publishers of data journals have created—again based on the needs and expectations of distinct research communities or subfields.

Publication is a mode of sharing, nonetheless, but not vice versa. More and more, scientists and data curation experts are pushing for a formalized notion of data publishing. Callaghan et al. (2013, 194) view formal data publication as “a service over and above the simple act of posting a dataset on a website, in that it includes a series of checks on the dataset of either a technical (format, metadata) or a more scientific (is the data scientifically meaningful?) nature.” Furthermore, in formal publication of data, the persistence of the data is assured, and discovery and open evaluation of the data are facilitated. Lawrence et al. (2011, 7) promote a similar definition: “In this paper we define to Publish (with a capital P) data, as: ‘To make data as permanently available as possible on the Internet.’ This Published data has been through a process which means it can appear along with easily digestible information as to its trustworthiness, reliability, format and content.” These interpretations go several specific steps beyond the definitions from Read and from Cornell’s RDMSG, particularly in terms of reviewing data for quality assurance purposes (e.g., “checks,” “evaluation,” “trustworthiness,” “reliability”) and in terms of prioritizing enduring access (“persistence of the data,” “make data as permanently available as possible”). It is interesting that neither definition integrates the notion of discovery, although Lawrence et al. (2011, 10) consider “discovery metadata” essential to include in a taxonomy of metadata for “the Publication process.”

Key to the notion of data persistence is the mechanism of the digital object identifier (DOI) and data citation standards in general. Per Lawrence et al. (2011), persistence also relates to repeated findability of data—that is, whether one is able to find the data set again, which constitutes an identifier issue. Data citation renders many of the benefits that scientific publication brings about: attribution and credit; data reuse and a means to verify the data; evidentiary information; and, as already mentioned, access and persistence. A critical feature for data citation is unique identification, or the DOI, which allows for machine readability of digital data. Repository services that accept and curate data are adopting data citation standards increasingly, assigning DOIs to their data sets. In February 2014, Force 11, an organization that advocates for improvement of research communication and frameworks for e-scholarship, issued a “Joint Declaration of Data Citation Principles,” in which additional fundamentals of data citation are expressed: citation of data should be considered as important as citations of publications; they should lead to the data being cited as well as to any related metadata, code, and other documentation; and approaches to data citation should be responsive to diverse community needs and practices but not at the risk of jeopardizing interoperability. While data citation is not equivalent to data publication, there is evidence of concern that data published with a DOI may be considered a previous publication—as if assignment of a DOI, which enacts persistence, is akin to publication. For example, the data policies for *F1000Research*, an open science journal that relies on a post-publication, open peer-review process, list journals that

“have confirmed that they would not view publication of datasets with a DOI and associated protocol information as prior publication, if a more standard (analysis/conclusions) article based on the data was subsequently submitted to them” (F1000Research 2014).

The foregoing assertions about data publishing capture key tenets of data curation. Data that are curated should be persistently accessible, too, and give credit and attribution. If curated data are quality data, and curation adds value to data, then are these data not equivalent to published data? With respect to repository services and the question of whether they qualify as publishers of data or not, the name changes of the data repository Pangaea tell a revealing story. From 1996, when it was launched, until 2003, the repository service was called Pangaea: Network for Geological and Environmental Data. In 2004, it became known as Publishing Network for Geoscientific and Environmental Data. Since 2011, Pangaea has billed itself as Data Publisher for Earth and Environmental Science, explicitly signaling its key purpose for data. Pangaea offers rich metadata for its data sets, which themselves are rigorously reviewed by an editorial board of earth and environmental scientists for “completeness and consistency” (Pangaea 2014). It also assigns each data set a DOI for persistence (particularly in terms of the ability to be located and accessed) and to render the data citable and discoverable. In addition, the journal *Earth System Science Data* includes Pangaea among its recommended repositories. Few may dispute data sets in Pangaea as being of published quality. It is not clear, however, how Pangaea is facilitating discovery and interoperability, which—as Muñoz (2013, citing Parsons and Fox 2013) rightly points out—should be prominent concerns in data publishing, as should exposing the issues of big data through “latency, rapid versioning, reprocessing, and computational demands” (Muñoz 2013, citing Parsons and Fox 2013, WDS37). Indeed, how to publish “big data” currently stands as an intractable issue. For this reason, and for other reasons mentioned later in this chapter, the terminology of *publishing* and *publication* may not be appropriate for data. Yet, *publishing* and *publication* have such cultural weight in academia that not to try to accord data sets the legitimacy and resources they deserve as tenure-worthy products through some form of publication may be ill-advised, both for data and for scientists.

If data curation in practice is intended to ensure quality data, as well as access to and use of that data, then as library-based services in data curation evolve, they present an immense opportunity for shaping how data publication may be done. Services in data curation may also be ripe for collaboration with publishing services within academic libraries and beyond.

Data Curation Services and Library Publishing Services: Context and Comparison

Data curation services in US academic libraries are still in their infancy; likewise with library-based data publishing. *Academic Libraries and Research Data Services*, an ACRL white paper, notes that only a very few libraries have such services, which it defines as focusing on the demands of the complete data lifecycle (Tenopir, Birch, and Allard 2012). Based on survey responses from libraries at associate’s colleges, baccalaureate colleges, and research/doctoral universities, Tenopir, Birch, and Allard report that approximately 25 percent to 30 percent of

respondents intend to begin services in this area in the next couple of years. Similarly, in *Research Data Management Services: SPEC Kit 334*, issued a year after the ACRL white paper, Fearon et al. (2013, 11) refer to the “growing pains” and “early stages” of developing these services in libraries. They differentiate between services for broad data support and services for research data management (RDM). A key finding is that for most libraries (93 percent of respondents, or sixty-eight libraries out of a total of seventy-three that responded), broad data support services primarily constitute assistance to faculty and students in search of data sets for their own use. Tenopir, Birch, and Allard (2012) also present a comparable takeaway in their white paper. Fearon et al. (2013, 12) describe RDM services as “providing information, consulting, training or active involvement” for areas such as data management planning and guidance, metadata, and sharing and curation of research data. In this sense, however, fewer libraries—though still almost three-quarters of respondents (or fifty-four libraries)—can claim active participation. The SPEC Kit authors encouraged respondents to document that they were providing RDM services, even if the services amounted to only Web-based resources for data management planning guidance and reference. Another significant survey response, because of its implications for infrastructure and data sharing, is the number of libraries running IRs—sixty-four, or 88 percent of respondents. Neither the ACRL white paper nor the ARL SPEC Kit mentions data publishing per se, although the latter delves into data archiving as a mechanism for data sharing.

By contrast, library publishing services seem relatively mature. Far more established operations exist for this area than for data curation. Data curation services are probably at the stage where library publishing services stood in 2007, when ARL surveyed its member libraries about publishing services. The survey found that 44 percent of the responding libraries (which numbered eighty in total) already had some form of such services, and 21 percent disclosed plans to develop them (Hahn 2008). If the current edition of the *Library Publishing Directory* (Lippincott 2013) gives any indication, libraries have gained further traction in scholarly publishing since then. According to the directory, there are ninety-eight academic libraries in North America engaged in publishing services. (The directory includes listings for seven additional libraries outside the United States and Canada.) Most of these libraries publish faculty-led, peer-reviewed journals, and more than half support the publication of journals started by students; many also publish electronic theses and dissertations (ETDs). Monographs, conference proceedings, and technical/research reports make up most of the remaining types of publications covered by these services. Publications that are open-access are the norm for most services.

The directory also hints at a kind of “data publishing readiness” among library-based publishers. Many of the libraries—more than half—note in the Formats field of the directory that they publish data (albeit in what sense is not clear). Also relevant to data publishing are some of the additional services the directory highlights in its introduction: metadata, analytics, outreach, and DOI assignment/allocation of identifiers. Another worthwhile statistic, which the directory does not surface explicitly, is the number of libraries specifying “dataset management” among

their additional services—just under one-third, or roughly 27 percent. In addition, most library publishing services are at institutions with IRs; they acknowledge a digital preservation strategy; and roughly one-third are considering the integration of digital preservation services. As digital preservation and data reflect burgeoning concerns, these statistics imply that library-based publishing, though long operational at many institutions, is still evolving and may have more in common with data curation services as these services, too, mature.

These statistics convey that building capacity to support researchers in data publishing is not an unreasonable ambition. Hahn (2008, 10) veritably indicated such in her summary of the 2007 ARL survey, with particular reference to the repository service component of library publishing: “Evolving repository services, which house and disseminate institutional records, theses and dissertations, pre-prints, post-prints, learning objects, and research data, can inspire a range of inquiries about potential publishing services.” As further evidence, Mullins et al. (2012), in their report on a series of library publishing workshops held in spring 2011, note the attendees’ openness to developing such services to address research life cycle issues, including data management, and to test out different modes of disseminating scholarly content. Indeed, the publication of science data qualifies as an example of both: a research life cycle endeavor—one consonant with data curation concerns—as well as an experimental service. In addition, library publishing and data publishing have comparable missions. For library publishing, the overarching mission is to provide unfettered access to peer-reviewed scholarship and allow authors to retain their copyright. Data publishing, as a formalized mode of data sharing, echoes such an aim, especially in terms of availability, discovery, quality, standards, and attribution. Like the publishing done by libraries, data publishing contributes to a public good. There is more to the ethos of scientific data publishing in comparison with typical library publishing, however, which tends to favor humanities content (although social sciences are also represented): the sciences depend on access to reliable data and other research results for purposes of verifiability and accountability (Borgman 2008). Moreover, applying data curation practices to scientific research aids in ensuring overall reproducibility. In this sense, the publication of data may be viewed as a curation tactic and thus about more than access.

Conceptually and operationally, scholarly publishing that is library-based is better defined than data publishing vis-à-vis data curation services in libraries—and not only because library publishing has been around longer. Another pivotal reason is the homogeneity of the content and the format that academic libraries commonly publish: the subject matter stems mainly from the humanities and social sciences, and the publication format is overwhelmingly text, though in a variety of genres. Furthermore, as the *Library Publishing Directory* (Lippincott 2013) implicitly confirms, with few exceptions libraries are publishing what they have always *collected*. The containers—for example, the monograph and the journal—have also not changed for most library-based publishers, even in online environments. It is true that experimentation in this area, such as CommentPress, a MediaCommons Press product (<http://futureofthebook.org/commentpress/>), for example, has enabled innovative leaps in recent years, particularly in the practice of open peer review. Data publishing may be more likely to

occur in an academic library if its collection policies and mission statements were formally articulated to include, and thus promote, data set collection, particularly as produced by the researchers of the library's institution. If publishing services and data curation services in libraries had adequate infrastructure and other support for experimental, even risk-taking ventures, then perhaps collaborative pilot projects for data publication would be the norm rather than the exception.

Data Publishing and Peer Review: Peas in a Pod, or Strange Bedfellows?

Speculation about what is possible for data publishing as part of a suite of data curation services in a library ultimately raises questions about peer review and quality assurance standards for data. Namely, what determines these standards for data? In the text-based scholarly publishing performed by library publishing services, the standards for quality work are well understood. Library publishing typically observes peer review and other quality assurance processes through practices long in place in academic communities. Scholars who review article manuscripts for journal publications know whereof they evaluate; they write, as well as review, in genres familiar to them—genres that also count toward tenure. Criteria for peer evaluation of data for publication are currently less concrete (Parsons and Fox 2013; Parsons, Duerr, and Minster 2010; Griffiths 2009). Scientists frequently advocate for progress toward such criteria and argue that data sets, like journal articles, should be treated as first-rate research products and thus inform promotion and tenure decision making (Gorgolewski, Margulies, and Milhan 2013; Reilly et al. 2011; Callaghan et al. 2013; Lawrence et al. 2011; Parsons, Duerr, and Minster 2010). In the case of articles linked to data files, the assignment to peer review a data set, in addition to the published article, can also prove burdensome, as the experience of reviewers for the *Journal of Neuroscience* ultimately conveyed (Socha 2013, citing Maunsell 2010). Reviewers found the task of refereeing article manuscripts along with supplemental materials, which could include data sets, increasingly insurmountable. The extent of the additional files (which in time were equaling the length of the articles), the growing tendency of them to reflect content that actually belonged in the main article, and, thus, the challenge that referees faced in assessing supplemental materials with any depth, drove editors of the journal to cease their acceptance (Maunsell 2010). Parsons and Fox (2013, WDS39) have noted how slow the review process can be for substantial sets of data; current approaches for refereeing “will not scale to handle the growing deluge of data.” The lack of scalability affects time to publication, a serious impediment in the peer-review process (Bornmann 2011), and was arguably a factor for the *Journal of Neuroscience* in its management of review practices for supplemental materials.

Deciding upon principles of review of scientific data for publication is a daunting endeavor. Data are heterogeneous and dynamic in nature. Data formats and types vary considerably across science disciplines, making “standardization” a slippery, if not also hollow, concept. As Tenenbaum, Sansone, and Haendel (2014, 2) caution, “even data standards experts do not agree on what constitutes a data standard.” Such disparity makes metadata for data sets an especially

intractable problem. Communities of interest have their own data models and metadata schemas, eroding the likelihood of interoperability and data that are shareable and reusable (Willis, Greenberg, and White 2012). Or scientists may understand the value that metadata brings to their data but are not well informed about standards for it (Cragin et al. 2010). Even when metadata standards are established, they are infrequently used or incorrectly implemented, making them “almost standards” (Edwards et al. 2011, 683). Unsurprisingly, the lack of metadata use is also a persistent issue. A DataONE survey conducted by Tenopir et al. (2011) found that of the 1,329 scientists who responded, almost 60 percent disclosed that they do not apply any metadata standards; another roughly 20 percent said they “use their own lab metadata standard” (9). In addition to these metadata hurdles is the complication of data versioning, a significant aspect of managing data. Data versions must be tracked, thereby raising the question of which version of a data set to subject to review for publication or how to capture for publication a changing data set, especially given the general understanding of publication as an act that finalizes and fixes for perpetuity a research investigation and its findings. Because data sets can evolve over time, their use and value may not be fully realized until well into the future, which renders current practices for peer review inadequate for them (Parsons and Fox 2013). The validity of a data set also is neither uniformly nor easily determined within science disciplines, let alone across subsets of them (Parsons, Duerr, and Minster 2010). An ever-moving target, data—and thus their publication—resist a “one size fits all” solution. And, as Parsons and Fox (2013, citing de Waard et al. 2006, 2008; Kuhn 1996; and Latour 1987) also caution, the tendency to model the peer-review publishing of data on that of scientific articles is itself problematic. The substance and intent of each differ radically: based on investigative findings, articles are “designed to persuade” (Parsons and Fox 2013, WDS38), while data constitute fact.

Although it may seem that data publishing and peer review make strange bedfellows, there are signs hinting at an enhanced perception and treatment of data as meriting peer review and thus being tenure-worthy. One sign is the Peer REview for Publication and Accreditation of Research Data in the Earth Sciences (PREPARDE) project (PREPARDE 2014). Based in the United Kingdom, PREPARDE has been working on deliverables for five aspects of data publishing: journal and data repository workflows, scientific review of data sets, cross-linking between repositories and data publishers, data repository accreditation, and stakeholder engagement and dissemination. The project has partnered with Wiley-Blackwell and its new *Geosciences Data Journal*, an open-access, online-only, peer-reviewed publication, on the workflows piece and on the creation of procedures and policies for scientific review of data sets to guide scientists refereeing for the data journal. Formal publication of just the data—that is, minus the preparation of an article manuscript—also holds the promise of swifter dissemination and thus access. Some in the sciences and the social sciences, as well as in academic libraries, have argued for publication outputs that are solely devoted to accounts of research data, such as data articles, or data papers, and data journals (Callaghan et al. 2013; Guy and Duke 2013; Kansa and Kansa 2013; Chavan and Penev 2011; Kunze et al. 2011). The rise in prominence of these types of data publication may, in time, lend them the cachet they need to be considered as

scholarship that counts toward promotion and tenure. Other encouraging signs come from the NSF. Since 2012 it has permitted researchers applying for grant funding to cite data sets and software code in their biosketches; for this purpose, the agency changed the heading for Publications to Products in its biosketch format. In early 2014 it issued a “Dear Colleague” letter to solicit collaborative workshops and exploratory research proposals in the areas of data citation and attribution and of metrics reflecting the impact of these practices: “Unlike generally accepted citation-based metrics for papers, software and data citations are not systematically collected or reported. NSF seeks to explore new norms and practices in the research community for software and data citation and attribution, so that data producers, software and tool developers, and data curators are credited for their contributions” (Tornow and Farnam 2014). If the NSF is going to require DMPs and thus expect scientists to care for their data more systematically, then shifting policy and supporting research efforts to prioritize citation, attribution, and metrics for data mark logical developments. Yet, just as proper attribution via use of data citation standards can serve as an impetus for sharing data (McCallum et al. 2013; Socha 2013), more incentives for researchers to publish data must also be created if researcher culture, attitudes, and behaviors are to change—an awareness that cuts across the sciences and social sciences (Costello et al. 2013; Gorgolewski, Margulies, and Milhan 2013; AGU 2012; Lawrence et al. 2011; Barton, Smith, and Weaver 2010; Griffiths 2009; Swan and Brown 2008).

Toward Publication: Data Policies, Data Repositories, Data Journals

The paradigm for scientific data publishing is in flux, but there are policies, systems, standards, and publication genres being engaged and interconnected for dissemination of data that are germane to this paradigm. These include publishers’ data policies, data repositories, and data journals and data papers, as well as new genres for data publishing, such as the Data Descriptor, formulated by the journal *Scientific Data* and intended to present a detailed, peer-reviewed data set, complete with the methods and analyses associated with producing it and understanding it. The examples of these genres discussed below arise mainly from nonlibrary publishing enterprises. The act of peering outward at these models for data publishing not only reveals possibilities for similar or complementary operations in academic libraries, especially in the context of data curation services, but is also aspirational. Libraries should strive to be peers with those outside that are meeting and anticipating the needs of researchers effectively. In this sense, too, libraries should be “peering outward.”

With the growing expectation for science journals to link published articles to the relevant data sets, publishers’ data policies have also risen in importance. The trend toward open data has impelled many publishers to require researchers to ensure that data from their articles are available for others to access on publication. The Public Library of Science (PLOS) is one of the latest publishers to revise their data policies: as of March 2014 it requires researchers to make their data sets for articles completely available upon submission, thus before publication, and to include a “Data Availability Statement” asserting PLOS policy compliance (Bloom 2013). The practice of associating published articles to data sets also works well for scientists who would

rather not share data until their research has been published (Tenopir et al. 2011). Some journals host the data themselves, like *CODATA Data Science Journal* and the journals published by the Ecological Society of America (ESA), which also maintains a data registry. The *CODATA Data Science Journal* accepts data in any format, including proprietary ones, while ESA journals take in data only for data papers that are in open formats. The ESA also demands fairly detailed metadata for the data sets, and it charges a one-time fee for publishing data papers.

Rather than hosting data sets themselves, many journals offer recommendations for repositories where researchers may deposit their data. In its list of data-sharing options (Bloom 2013), PLOS recommends deposit of data to public repositories, preferably ones that are certified as trusted and with open licensing policies, such as Creative Commons Attribution (CC BY). There are also data repositories that were created expressly for deposit of data sets associated with scientific article publications, such as Pangaea and Dryad (datadryad.org). An advantage for authors submitting to journals collaborating with Dryad is the repository's Submission Integration service, detailed on the Dryad website, which is essentially a workflow that couples processes for article submission with those for data deposit. Through automated notice, journals let Dryad know when a manuscript is about to be processed; Dryad establishes a placeholder for the data set record; journals encourage authors to archive their data in Dryad when they submit, giving them access to the link for the placeholder record, where they may upload the files for their data; the author deposits data files into Dryad, which approves the data set and generates a DOI for it; and the DOI is then passed onto the relevant journal and applied to the article so the data set can be accessed, tracked, and cited. The goal of Dryad's Submission Integration service is to make deposit of data into the repository as seamless an experience as possible for researchers. It also activates two-way access, or linking, between the journal article and the data set, which allows each to gain more visibility. Dryad is able to furnish such a service largely because of its membership-based business model, a development that occurred after its NSF funding ended in 2012. Each journal that integrates with Dryad does so for a fee, or the integration occurs as a benefit of organizational membership in Dryad.

Close partnerships between data repositories and publishers, such as that which Dryad enjoys and from which it is able to create a valuable service, are not as common as they should be, however. For the most part, publishers provide authors with a list of recommended repositories but little more. The onus is still on the researcher, who must figure out which repository is best for the data (or what to do in the event that no suggested repository appears suitable), learn the guidelines of that repository, and prepare the data for submission—in addition to the other work that is required to finalize an article manuscript for publication. Resources have surfaced in the last few years that offer some guidance on data repositories. One of these is DataBib (<http://databib.org/>), a well-curated registry of information about data repositories that researchers can use to locate repository services relevant for their data types. Data journal publishers in particular, such as Ubiquity Press, are selective in the repositories they advise authors to use; the ones they list adhere to the publisher's standards for peer review of data. (Although data journals are devoted to publishing data sets, they also must advise their authors

on where to deposit the data being featured and discussed in the data paper or data article to allow other researchers to access and use the data sets.) The website for the *Earth System Science Data (ESSD) Journal* (<http://www.earth-system-science-data.net/>) explicitly displays the requirements that repositories accepting authors' data sets must fulfill: the repository has to mint a DOI for the data set; it must make the data set freely available (i.e., charge no fees for access); it accords the data set the equivalence of a Creative Commons Attribution license; and the repository satisfies the topmost criteria for ensuring ongoing access. While *ESSD* displays a short list of repositories at its site, cautioning that the list is not exhaustive, it also urges authors to see whether data repositories they are familiar with meet the journal's criteria.

In 2014 a new online, open-access data publication, *Scientific Data*, emerged that introduced an inventive genre for data publishing—the peer-reviewed Data Descriptor: “a combination of traditional scientific publication content and structured information curated in-house . . . designed to maximize reuse and enable searching, linking and data mining” (NPG 2013). A main motivation for creating *Scientific Data* is to assist researchers in complying with data management requirements. Six principles lie at the core of what *Scientific Data* is trying to achieve: credit, reuse, quality, discovery, openness, and service. As part of its focus on service, *Scientific Data* strives to lower barriers for researchers submitting their data sets, such as automating deposit of them into Dryad, or the figshare repository service (in the event that there is not a community-driven repository available for the data); provide professional services in data curation to make certain standards are observed so that content is discoverable; facilitate visual interpretations of the content, as well as pathways to the content via robust linking and searching; and rapid evaluation and decision processes, resulting in prompt publication of the data. *Scientific Data* recommends as part of its data policies that data be submitted in the “rawest” form that will benefit the scientific community and bring out the broadest possible repurposing of the data. It urges authors to use data repositories that are discipline-specific for their data and community-driven. Its data policies include criteria for trusted data repositories that scientists are expected to consult when deciding which one to use. The criteria call for expertise in curation; implementation of “relevant, community-endorsed reporting requirements”; provision “for confidential review” of the data; application of “stable identifiers” for the data; and “public access to data without unnecessary restrictions” (NPG 2014). *Scientific Data* provides a template for the Data Descriptor manuscript, which incorporates sections for, among other things, background and summary of the data, methods, data records, technical validation, and usage notes (which are optional).

With multiple senses of “data publishing” at play—is it sharing, dissemination, or publication?—and without conducting a formal environmental scan via a survey, it is difficult to know who is doing what in data publishing and at which US academic libraries. A few examples do come to the fore, though, that, because of the high level of curation involved, could be called data publishing. Cornell has repurposed its Datastar repository, originally for staging data, as a metadata-rich data registry, taking advantage of Semantic Web technologies such linked open data and VIVO, a networking tool (Wright et al. 2013). Another is the data publishing

investigation on which the California Digital Library (CDL) has partnered with the PREPARDE project. CDL has been a leader in the United States promoting data citation through its EZID identifier service, as well as data preservation and access via Merritt, its repository service. CDL hired a CLIR postdoctoral fellow to take the lead in exploring data publishing possibilities. It is creating specifications, toward implementation, for a data paper, at minimal cost and minimal effort for the researcher, that would be formed from a record for an EZID citation; it would “identify a publishable dataset, complete with author, title, date, abstract, and links to stored data” (PREPARDE 2014). Also within the PREPARDE project, CDL is partnering with UC Berkeley in curation of medium-to-large data sets. As part of developing a service model for data publishing, CDL issued a survey to determine how researchers think about, and engage in, data publication. Another academic library that is testing the waters with data publishing is Purdue. By coordinating workflows among its data repository, IR, and university press, Purdue has been able to automate linking of technical and project reports shared through its IR with their related data sets, which are deposited separately (Scherer, Zilinski, and Matthews 2013). Finally, as active collaborators with faculty in research activities, librarians at Johns Hopkins are paving new paths for how data, as a result of these partnerships, can be accessed—such as through interactive visualizations of data, projected on a wall, that are transformed via hand and body motions, making the wall “a new form of publishing” (Monastersky 2013). Johns Hopkins has also developed a program in which it will contract, for a fee, with researchers who have been awarded project funding: the libraries will commit to curation and storage of the research data produced by projects for a period of five years, renewable thereafter (Monastersky 2013). A key benefit of such a commitment is the rich, contextual information that curation will engender for the data sets, making the sharing, if not also the publication, of data uncomplicated.

Conclusion

Data curation services, especially those that leverage expertise across departments and subject libraries, have many roles to play in this area in support of researchers managing, and perhaps publishing, their data. Topics for instruction in academic libraries tend to focus on vended database resources and tools for management of citations, generally helping researchers look for and find materials in support of their research and organize those materials. However, there could be regular instruction offerings that complement creation and collection of data sets, such as sessions on data repositories, data publishing and citation principles, data publication genres, tools like the Open Data Commons toolkit for providing and using open data, and what scientific publishing entails overall, as well as on the best practices for maintaining data files locally. This instruction could be geared toward postdocs, beginning graduate students, and advanced undergraduates, as well as early-career faculty—particularly those who play the role of “data keeper” in their research labs. It could also complement instruction already being done on data management planning. Since the Data Descriptor template for *Scientific Data* effectively helps researchers tell the story of a data set, the template could serve as an “inreach” tool for librarians wishing to learn more about what data sharing and publishing are about. Another

resource, the Data Curation Profiles Toolkit (<http://datacurationprofiles.org/>), developed by Purdue and the University of Illinois at Urbana-Champaign, could be applied for similar purposes.

In a slight reversal of collection development responsibilities, librarians could familiarize themselves with the types of data that faculty and students are collecting and consider approaches for strategically developing collections of their institution's data sets—to assist in curating them with an eye toward their dissemination and reuse. Based on overall knowledge about their constituents' data, and with the assistance of a resource like DataBib, librarians could also become familiar with the repository services that are applicable to the data being generated by their researchers. In the event that an IR is not appropriate, then familiarity with DataBib would help them have suggestions at the ready when meeting with faculty and students about data management planning. Metadata librarians, working with their libraries' IR managers and liaison librarians, could engage with researchers on various outreach efforts, such as metadata education and training that would include best practices for file naming and management, data normalization and cleanup, information on data standards for specific disciplines, and approaches to making researchers' data discoverable and accessible.

Since many library publishing services are well versed in copyright and fair use, they could collaborate with data curation services to establish guidance on intellectual property rights and data. A hopeful trend is that of open data, particularly given the data-sharing policy put into effect by PLOS. Along this line, libraries could repurpose position responsibilities of appropriate staff to include monitoring of external developments regarding policies for data sharing, at the national level as well as at the publisher level. Such monitoring, itself an example of peering outward, is useful not only for reasons of internal apprising. It can also inform possible discussion in the context of university governance—for example, how an institution needs to respond to decisions made by the OSTP so that it acts as a whole in compliance. Keeping track of the pulse of initiatives at funding agencies, such as the NSF, opens up possibilities for direct participation from academic libraries, particularly when the initiative addresses areas in which many libraries already have strengths, such as in citation standards and bibliometrics. The best practices for data citation, attribution, and metrics tracking, supported by the NSF's "Dear Colleague" letter mentioned earlier, constitute an area in which library-based services in data curation and in publishing could partner with information schools and with relevant departments in the sciences and social sciences on workshops and research proposals appropriate for this call.

Libraries should also partner formally with the research institutes on their campuses, as well as with the Office of the Vice President for Research (OVPR), in developing more centralized, scalable, and programmatic efforts and services toward improved data management practices for faculty, students, and staff. An institution's OVPR is often the campus entity that provides guidelines for the responsible conduct of research, under which best practices for data management would fall. There are opportunity costs too dear for libraries, information technology services, and the OVPR to afford if an institutional approach to data curation programs is not realized. Perhaps the steepest costs are data loss and lack of access to data,

which are ultimately tied to an institution's ability to foster and gain more research funding and more research partnerships. Vines et al. (2014) note the adverse effects of "article age" on availability of data sets. In summary, the older the publication, the harder it was for Vines et al. to contact researchers for access to the relevant data sets, primarily because of obsolete e-mail information, loss of data, and barriers to data due to "inaccessible storage media" (95). As Vines et al. suggest, support on an institutional scale for author identity services, such as the Open Researcher and Contributor ID (ORCID), as well as guidance on ORCID and on researcher networking opportunities evident in Google Scholar and ResearchGate, could ultimately help increase researcher access and data set availability, regardless of the age of the article. Libraries could expand their instruction offerings to include such guidance, advising researchers on identity and research reputation management tools, which offer additional channels for discovery of data set citation and attribution. Programs in the responsible conduct of research should encompass an understanding of these and other issues related to proper management of data sets produced by an institution's researchers.

As early as 2002, Gray et al. connected curation of scientific data with publication and archiving, stating, "Librarians would describe documenting the metadata as *curating* the data. They have thought deeply about these issues, and we would do well to learn from their experiences." (104, emphasis in the original). In the same article, summarizing, they say, "Data publication is really data curation," thus binding together the library's central role in publishing and curation of data, if not as data publisher. Others have made similar parallels between publishing and curation (Muñoz 2013; Ray, Choudhury, and Furlough 2009), suggesting new models for library organizational structures and collection development and management practices. As data policies, particularly at the level of publishers, funding agencies, and the federal government, move toward more openness and transparency, opportunities are opening up in tandem for libraries to participate and partner in data sharing and publishing efforts. There is also much that services in data curation and in library publishing can learn from each other, perhaps to the extent that peer review and data sets might make sense instead of seeming at odds—as long as libraries keep peering outward.

Works Cited

- AGU (American Geophysical Union). 2012. "Earth and Space Science Data Should Be Widely Accessible in Multiple Formats and Long-Term Preservation of Data Is an Integral Responsibility of Scientists and Sponsoring Institutions." Adopted May 29, 1997; revised May 2009, February 2012. http://sciencepolicy.agu.org/files/2013/07/AGU-Data-Position-Statement_March-2012.pdf.
- Barton, C., R. Smith, and Weaver 2010. "Data Policies, Practices, and Rewards in the Information Era Demand a New Paradigm." *Data Science Journal* 9: IGY95-IFY99. doi: 10.2481/dsj.SS_IGY-003.
- Bloom, Theo. 2013. "Data Access for the Open Access Literature: PLOS's Data Policy." The Public Library of Science (PLOS) website, December 12 <http://www.plos.org/data-access-for-the-open-access-literature-ploss-data-policy/>.
- Borgman, Christine L. 2008. "Data, Disciplines, and Scholarly Publishing." *Learned Publishing* 21, no. 1 (January): 29–38. doi:10.1087/095315108X254476.
- Bornmann, Lutz. 2011. "Scientific Peer Review." *Annual Review of Information Science and Technology* 45, no. 1: 197–245. doi:10.1002/aris.2011.1440450112.
- Brauer, Peter and Wilhelm Hasselbring. "PubFlow: A Scientific Data Publication Framework for Marine Science" (paper presented at PubMan Days 2013 conference, Munich, Germany, October 23-24, 2013). Accessed January 24, http://eprints.uni-kiel.de/22400/1/vortrag_pubmanDays.pdf.
- Callaghan, Sarah, Fiona Murphy, Jonathan Tedds, Rob Allan, John Kunze, Rebecca Lawrence, Matthew S. Mayernik, and Angus Whyte. 2013. "Processes and Procedures for Data Publication: A Case Study in the Geosciences." *International Journal of Digital Curation* 8, no. 1: 193–203. doi:10.2218/ijdc.v8i1.253.
- Chavan, Vishwas S., and Peter Ingwersen. 2009. "Towards a Data Publishing Framework for Primary Biodiversity Data: Challenges and Potentials for the Biodiversity Informatics Community." *BMC Bioinformatics* 10, Supp. 14: S2. doi:10.1186/1471-2105-10-S14-S2.
- Chavan, Vishwas, and Lyubomir Penev. 2011. "The Data Paper: A Mechanism to Incentivize Data Publishing in Biodiversity Science." *BMC Bioinformatics* 12, Supp. 15: S2. doi:10.1186/1471-2105-12-S15-S2.
- CLIR (Council on Library and Information Resources). 2013. "CLIR Receives Sloan Foundation Grant for Data Curation Fellows." News release, April 1. www.clir.org/about/news/pressrelease/sloan-data-curation-award.
- Cornell University. 2014. "Sharing Data." Research Data Management Service Group website, accessed February 24. <http://data.research.cornell.edu/content/sharing-data>.
- Costello, Mark J., William K. Michener, Mark Gahegan, Zhi-Qiang Zhang, and Philip E. Bourne. 2013. "Biodiversity Data Should Be Published, Cited, and Peer Reviewed." *Trends in Ecology and Evolution* 28, no. 8 (August): 454–61.

doi:10.1016/j.tree.2013.05.002.

- Cragin, Melissa H., P. Bryan Heidorn, Carole L. Palmer, and Linda C. Smith. An Educational Program in Data Curation. Poster presented at the American Library Association Science & Technology Section Conference, Washington, D.C. June 2007.
https://www.ideals.illinois.edu/bitstream/handle/2142/3493/ALA_STS_poster_2007.pdf?sequence=2.
- Cragin, Melissa H., Carole L. Palmer, Jacob R. Carlson, and Michael Witt. 2010. "Data Sharing, Small Science and Institutional Repositories." *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 368, no. 1926 (September): 4023–38. doi:10.1098/rsta.2010.0165.
- de Waard, Anita, Leen Breure, Joost G Kircz, and Herre Van Oostendorp. 2006. "Modeling Rhetoric in Scientific Publications." Paper presented at the International Conference on Multidisciplinary Information Sciences and Technologies, Mérida, Spain, October 25-28, 2006.
- de Waard, Anita, and Joost Kircz. 2008. "Modeling Scientific Research Articles – Shifting Perspectives and Persistent Issues." In *Open Scholarship: Authority, Community, and Sustainability in the Age of Web 2.0 - Proceedings of the 12th International Conference on Electronic Publishing*, edited by Leslie Chan and Susanna Mornatti. Toronto: ELPUB, 21: 234-245.
- De Wever, Aaike, Astrid Schmidt-Kloiber, Mark O. Gessner, and Klement Tockner. 2012. "Freshwater Journals Unite to Boost Primary Biodiversity Data Publication." *BioScience* 62, no. 6: 529–30. doi:10.1525/bio.2012.62.6.2.
- Edwards, Paul N., Matthew S. Mayernik, Archer L. Batcheller, Geoffrey C. Bowker, and Christine L. Borgman. 2011. "Science Friction: Data, Metadata, and Collaboration." *Social Studies of Science* 41, no. 5 (October): 667–90. doi:10.1177/0306312711413314.
- Fearon, David, Jr., Betsy Gunia, Barabara E. Pralle, Sherry Lake, and Andrew L. Sallans. 2013. *Research Data Management Services: SPEC Kit 334*. Washington, DC: Association of Reserach Libraries, July.
- F1000Research. 2014. "Can I Publish an Analysis of My Published Dataset in Other Journals?" Accessed February 26. <http://f1000research.com/data-policies>.
- Force 11. 2014. "Joint Declaration of Data Citation Principles: Final." February. www.force11.org/datacitation.
- Gorgolewski, Krzysztof J., Daniel S. Margulies, and Michael P. Milham. 2013. "Making Data Sharing Count: A Publication-Based Solution." *Frontiers in Neuroscience* 7, no. 9. doi:10.3389/fnins.2013.00009.
- Gray, Jim, Alexander S. Szalay, Ani R. Thakar, Christopher Stoughton, and Jan vandenBerg. 2002. "Online Scientific Data Curation, Publication, and Archiving." In *Proceedings of SPIE* 4846, "Virtual Observatories," edited by Alexander S. Szalay, 103–7. Bellingham,

- WA: SPIE. doi:10.1117/12.461524.
- Griffiths, Aaron. 2009. "The Publication of Research Data: Researcher Attitudes and Behaviour." *International Journal of Digital Curation* 4, no. 1: 46–56. doi:10.2218/ijdc.v4i1.77.
- Guy, Marieke, and Monica Duke. 2013. "The Rise of the Data Journal." Presentation at IASSIST, Cologne, Germany, May 31. www.slideshare.net/MariekeGuy/the-rise-of-the-data-journal.
- Haendel, Melissa A., Nicole A. Vasilevsky, and Jacqueline A. Wirz. 2012. "Dealing with Data: A Case Study on Information and Data Management Literacy." *PLOS Biology* 10, no. 5 (May 29): e1001339. doi:10.1371/journal.pbio.1001339.
- Hahn, Karla L. 2008. *Research Library Publishing Services: New Options for University Publishing*. Washington, DC: Association of Research Libraries.
- Higgins, Sarah. 2008. "The DCC Lifecycle Curation Model." *The International Journal of Digital Curation* 3, no. 1: 134-140. doi:10.2218/ijdc.v3i1.48.
- Kansa, Eric C., and Sarah Witcher Kansa. 2013. "We All Know That a 14 Is a Sheep: Data Publication and Professionalism in Archaeological Communication." *Journal of Eastern Mediterranean Archaeology and Heritage Studies* 1, no. 1: 88–97. http://muse.jhu.edu/journals/journal_of_eastern_mediterranean_archaeology_and_heritage_studies/v001/1.1.kansa01.html.
- Kratz, John, and Carly Strasser. 2014. "Data Publication: Consensus and Controversies" [v2; ref status: indexed, <http://f1000r.es/3hi>] *F1000Research* 3:94. doi: 10.12688/f1000research.3979.2
- Kuhn, Thomas S. 1996. *The Structure of Scientific Revolutions*. 3rd edition. Chicago: University of Chicago Press.
- Kunze, John, Trisha Cruse, Rachael Hu, Stephen Abrams, Kirk Hastings, Catherine Mitchell, and Lisa Schiff. 2011. *Practices, Trends, and Recommendations in Technical Appendix Usage for Selected Data-Intensive Disciplines*, version 2011.01.18. Oakland: California Digital Library. <http://www.cdlib.org/services/uc3/docs/dax.pdf>.
- Latour, Bruno. 1987. *Science in Action: How to Follow Scientists and Engineers through Society*. Cambridge, MA: Harvard University Press.
- Lawrence, Bryan, Catherine Jones, Brian Matthews, Sam Pepler, and Sarah Callaghan. 2011. "Citation and Peer Review of Data: Moving Towards Formal Data Publication." *International Journal of Digital Curation* 6, no. 2: 4–37. doi:10.2218/ijdc.v6i2.205.
- Lippincott, Sarah K., ed. 2013. *Library Publishing Directory*. Atlanta: Library Publishing Coalition. www.librarypublishing.org/sites/librarypublishing.org/files/documents/LPC_LPDirectory2014.pdf.
- Maunsell, John. 2010. "Announcement Regarding Supplemental Material." *Journal of Neuroscience* 30, no. 32: 10599–600. www.jneurosci.org/content/30/32/10599.

- McCallum, I., H.-P. Plag, S. Fritz, and S. Nativi. 2013. "Data Citation Standard: A Means to Support Data Sharing, Attribution, and Traceability." In *Proceedings of the 16th International Conference on Heavy Metals in the Environment*, edited by Nicola Pirrone. *E3S Web of Conferences* 1: 28002. doi:10.1051/e3sconf/20130128002.
- Michener, William K., Suzie Allard, Amber Budden, Robert B. Cook, Kimberly Douglass, Mike Frame, Steve Kelling, Rebecca Koskela, Carol Tenopir, and David A. Vieglais. 2012. "Participatory Design of DataONE—Enabling Cyberinfrastructure for the Biological and Environmental Sciences." *Ecological Informatics* 11 (September): 5–15. doi:10.1016/j.ecoinf.2011.08.007.
- Monastersky, Richard. 2013. "Publishing Frontiers: The Library Reboot." *Nature* 495, no. 7442 (March 27): 430–32. doi:10.1038/495430a.
- Mullins, James L., Catherine Murray Rust, Joyce L. Ogburn, Raym Crow, October Ivins, Allyson Mower, Daureen Nesdill, Mark P. Newton, Julie Speer, and Charles Watkinson. 2012. *Library Publishing Services: Strategies for Success*. Final Research Report. Washington, DC: SPARC, March. http://docs.lib.purdue.edu/cgi/viewcontent.cgi?article=1023&context=purduepress_ebooks.
- Muñoz, Trevor. 2013. "Data Curation as Publishing for Digital Humanists." *Trevor Muñoz* (blog). May 30. <http://trevormunoz.com/notebook/2013/05/30/data-curation-as-publishing-for-dh.html>.
- NPG (Nature Publishing Group). 2013. "NPG to launch *Scientific Data* to Help Scientists Publish and Reuse Research Data." News release, April 4. http://www.nature.com/press_releases/scientificdata.html.
- _____. 2014. "*Scientific Data*: Data Policies," under "Data Deposition Policy." Accessed February 24. www.nature.com/scientificdata/for-authors/data-deposition-policies.
- Pangaea. 2014. Website. Accessed January 6. <http://pangaea.de>.
- Parsons, Mark A., Ruth Duerr, and Jean-Bernard Minster. 2010. "Data Citation and Peer Review." *Eos, Transactions American Geophysical Union* 91, no. 34 (August): 297–98. doi:10.1029/2010EO340001.
- Parsons, Mark A., and Peter A. Fox. 2013. "Is Data Publication the Right Metaphor?" In "Proceedings of the 1st WDS Conference in Kyoto 2011," special issue, *Data Science Journal* 12 (February 10): WDS32–WDS46. doi:10.2481/dsj.WDS-042.
- PREPARDE (Peer REview for Publication & Accreditation of Research data in the Earth Sciences). 2014. Project website, accessed October 13. <http://proj.badc.rl.ac.uk/preparde>.
- Ray, Joyce, Sayeed Choudhury, and Michael J. Furlough. 2009. "Digital Curation and E-Publishing: Libraries Make the Connection." Presentation, 29th Annual Charleston Library Conference, Charleston, SC, November 6. <http://docs.lib.purdue.edu/cgi/viewcontent.cgi?article=1065&context=charleston>

- Read, Kevin. 2013. "Data Publishing: Who Is Meeting This Need?" *Kevin the Librarian* (blog), June 11 <http://kevinthelibrarian.wordpress.com/2013/06/11/data-publishing-who-is-meeting-this-need>.
- Reilly, Susan, Wouter Schallier, Sabine Schrimpf, Eefke Smit, and Max Wilkinson. 2011. "Report on Integration of Data and Publications." Miscellaneous. *EPIC387 P.*, October 17, 2011. http://www.alliancepermanentaccess.org/wp-content/uploads/downloads/2011/11/ODE-ReportOnIntegrationOfDataAndPublications-1_1.pdf.
- Scherer, David, Lisa Zilinski, and Courtney Matthews. 2013. "Opportunities and Challenges of Data Publication: A Case Study from Purdue." Presentation, 33rd Annual Charleston Library Conference, Charleston, SC, November 8. http://docs.lib.purdue.edu/lib_fspress/38.
- Socha, Yvonne, ed. 2013. "Out of Cite, Out of Mind: The Current State of Practice, Policy, and Technology for the Citation of Data." *Data Science Journal* 12 (September): CIDCR1–CIDCR75. doi:10.2481/dsj.OSOM13-043.
- Soehner, Catherine, Catherine Steeves, and Jennifer Ward. 2010. *E-Science and Data Support Services: A Study of ARL Member Institutions*. Washington, DC: Association of Research Libraries. www.arl.org/storage/documents/publications/escience-report-2010.pdf.
- Swan, Alma, and Sheridan Brown. 2008. *To Share or Not to Share: Publication and Quality Assurance of Research Data Outputs*. London: Research Information Network, June. www.rin.ac.uk/our-work/data-management-and-curation/share-or-not-share-research-data-outputs.
- Tenenbaum, Jessica D., Susanna-Assunta Sansone, and Melissa Haendel. 2014. "A Sea of Standards for Omics Data: Sink or Swim?" *Journal of the American Medical Informatics Association* 21, no. 2: 200–3. doi:10.1136/amiajnl-2013-002066.
- Tenopir, Carol, Suzie Allard, Kimberly Douglass, Arsev Umur Aydinoglu, Lei Wu, Eleanor Read, Maribeth Manoff, and Mike Frame. 2011. "Data Sharing by Scientists: Practices and Perceptions." *PLOS ONE* 6, no. 6 (June 29): e21101. doi:10.1371/journal.pone.0021101.
- Tenopir, Carol, Ben Birch, and Suzie Allard. 2012. *Academic Libraries and Research Data Services: Current Practices and Plans for the Future*. An ACRL white paper. Chicago: Association of College and Research Libraries, June.
- Tornow, Joanne, and Farnam Jahanian. 2014. "Dear Colleague Letter: Supporting Scientific Discovery through Norms and Practices for Software and Data Citation and Attribution." NSF 14-059. Washington, DC: National Science Foundation, April 11. www.nsf.gov/pubs/2014/nsf14059/nsf14059.jsp.
- Vines, Timothy H., Arianne Y. K. Albert, Rose L. Andrew, Florence Débarre, Dan G. Bock, Michelle T. Franklin, Kimberly J. Gilbert, Jean-Sébastien Moore, Sébastien Renaut, and Diana J. Rennison. 2014. "The Availability of Research Data Declines Rapidly with Article Age." *Current Biology* 24, no. 1 (94–97). doi:10.1016/j.cub.2013.11.014.

- Wenfang, Cheng, Zhang Jie, Zhang Beichen, and Yang Rui. 2013. "A Multidisciplinary Scientific Data Sharing System for the Polar Region." In *Proceedings: 12th International Symposium on Distributed Computing and Applications to Business, Engineering Science (DCABES)*, 167–70, 2013, edited by Souheil Khaddaj. doi:10.1109/DCABES.2013.54.
- Willis, Craig, Jane Greenberg, and Hollie White. 2012. "Analysis and Synthesis of Metadata Goals for Scientific Data." *Journal of the American Society for Information Science and Technology* 63, no. 8 (August): 1505–20. doi:10.1002/asi.22683.
- Wright, Sarah J., Wendy A. Kozlowski, Dianne Dietrich, Huda J. Khan, Gail S. Steinhart, and Leslie McIntosh. "Using Data Curation Profiles to Design the Datastar Dataset Registry." *D-Lib Magazine* 19, no. 7/8 (July 2013). doi:10.1045/july2013-wright.