ISSUE BRIEF

Data Communities

A New Model for Supporting STEM Data Sharing

May 13, 2019

Danielle Cooper
Rebecca Springer

ITHAKA S+R

Ithaka S+R provides research and strategic guidance to help the academic and cultural communities serve the public good and navigate economic, demographic, and technological change. Ithaka S+R is part of ITHAKA, a not-for-profit organization that works to advance and preserve knowledge and to improve teaching and learning through the use of digital technologies. Artstor, JSTOR, and Portico are also part of ITHAKA.

ITHAKA is interested in disseminating this brief as widely as possible. Please contact us with any questions about using the report: research@ithaka.org.

Introduction

There is a growing perception that science can progress more quickly, more innovatively, and more rigorously when researchers share data with each other.[1] Policies and supports for data sharing within the STEM (science, technology, engineering, and mathematics) academic community are being put in place by stakeholders such as research funders, publishers, and universities, with overlapping effects. Additionally, many data sharing advocates have embraced the FAIR data principles – holding that data must be findable, accessible, interoperable, and reusable, by both humans and machines – as the standard benchmark for data sharing success.[2] There is also an emerging scholarly literature evaluating the efficacies of some of these policies, although this literature tends to either focus on discrete disciplines[3] or particular journal or funder initiatives.[4]

[1] The authors are deeply grateful to Sayeed Choudhury, Alastair Dunning, Themba Flowers, Maggie Levenstein, Cameron Neylon, Oya Rieger, and Elizabeth Yakel for their comments on a draft of this issue brief. Any errors or omissions remain our own.

[2] Mark D. Wilkinson et al., "Comment: The FAIR Guiding Principles for Scientific Data Management and Stewardship," *Scientific Data* 3:160018 (Mar. 15, 2016), DOI: 10.1038/sdata.2016.18.

[3] A limited selection: Katherine G. Akers and Jennifer Doty, "Disciplinary Differences in Faculty Research Data Management Practices and Perspectives," *The International Journal of Digital Curation* 8.2 (2013): 5-26, DOI: 10.2218/ijdc.v8i2.263; Dominique G. Roche, "Public Data Archiving in Ecology and Evolution: How Well Are We Doing?" *PLOS Biology* 13:11, e1002295 (Nov. 10, 2015), DOI: 10.1371/journal.pbio.1002295; Philip Herold, "Data Sharing among Ecology, Evolution, and Natural Resource Scientists: an Analysis of Selected Publications," *Journal of Librarianship and Scholarly Communication* 3:2, eP1244 (2015), DOI: 10.7710/2162-3309.1244; Yi Shen, "Data Sharing Practices, Information Exchange Behaviors, and Knowledge Discovery Dynamics: A Study of Natural Resources and Environmental Scientists," *Environmental Systems Research* 6:9 (2017), DOI: 10.1186/s40068-017-0086-5; Alessandro Blasimme et al., "Data Sharing for Precision Medicine: Policy Lessons and Future Directions," *Health Affairs* 37:5 (2018): 702-9, DOI: 10.1377/hlthaff.2017.1558; Joshua D. Wallach, Kevin W. Boyack and John P.A. Ioannidis, "Reproducible Research Practices, Transparency, and Open Access Data in the Biomedical Literature, 2015-2017," *PLOS Biology* 16:11, e2006930, DOI: 10.1371/journal.pbio.2006930; Christie Wiley, "Data Sharing and Engineering Faculty: An Analysis of Selected Publications," *Science & Technology Libraries* 37:4 (2018), DOI: 10.1080/0194262X.2018.1516596; Dan Sholler et al., "Enforcing Public Data Archiving Policies in Academic Publishing: A Study of Ecology Journals," *Big Data & Society* (Mar. 25, 2019), DOI: 10.1177/2053951719836259.

[4] Victoria Stodden, Peixuan Guo, and Zhaokun Ma, "Toward Reproducible Computational Research: An Empirical Analysis of Data and Code Policy Adoption by Journals," *PLOS ONE* 8:6 (June 21, 2013), DOI: 10.1371/journal.pone.0067111; Nicole A. Vasilevsky et al., "Reproducible and Reusable Research: Are Journal Data Sharing Policies Meeting the Mark?" *PeerJ* e3208 (Apr. 25, 2017), DOI: 10.7717/peerj.3208; Florian Naudet et al., "Data Sharing and Reanalysis of Randomized Controlled Trials in Leading Biomedical Journals with a Full Data Sharing Policy: Survey of Studies Published in *The BMJ* and *PLOS Medicine*," *BMJ* 360:k400 (Feb. 13, 2018), DOI: 10.1136/bmj.k400; Lisa M. Federer et al., "Data Sharing in PLOS ONE: An Analysis of Data Availability Statements," *PLOS ONE* 13:5, e0194768, DOI: 10.1371/journal.pone.0194768; Jessica L. Couture et al., "A Funder-Imposed Data Publication Requirement Seldom Inspired Data Sharing," *PLOS ONE* 13:7, e0199780 (July 6, 2018), DOI: 10.1371/journal.pone.0199789.

By contrast, many scientists are not engaging in data sharing and remain skeptical of its relevance to their work. Through a series of studies on scholarly research practices at Ithaka S+R, we have found that scientists in a variety of fields, including chemists, agricultural scientists, and civil and environmental engineers, tend not to make their data widely available.[5] This reticence stands in contrast to the fact that over 40 percent of scientists reported that analyzing pre-existing quantitative data was highly important to their research in the Ithaka S+R US Faculty Survey 2018.[6] Barriers to sharing include the fear of being "scooped," wariness of data being misused, or the belief that the benefits of sharing data do not outweigh the effort required to format, contextualize, and upload research data in a way that is suitable for reuse. There is growing awareness of these challenges in academic support communities, and much has been written about possible solutions, ranging in scale from domain-specific technical solutions to systemic interventions like facilitating data citation and publication.[7]

As organizations and initiatives designed to promote STEM data sharing multiply – within, across, and outside academic institutions – there is a pressing need to decide

[5] Matthew P. Long and Roger C. Schonfeld, "Supporting the Changing Research Practices of Chemists," *Ithaka S+R*, Feb. 26, 2013, DOI: 10.18665/sr.22561, 29-31; Danielle Cooper et al., "Supporting the Changing Research Practices of Agriculture Scholars," *Ithaka S+R*, June 7, 2017, DOI: 10.18665/sr.303663, 24-26; Danielle Cooper, Rebecca Springer, et al., "Supporting the Changing Research Practices of Civil and Environmental Engineering Scholars," *Ithaka S+R*, Jan. 16, 2018, DOI: 10.18665/sr.310885, 22-28.

[6] Melissa Blankstein and Christine Wolff-Eisenberg, "Ithaka S+R US Faculty Survey 2018," *Ithaka S+R*, Apr. 12, 2019, DOI: 10.18665/sr.311199, 22.

[7] Domain-specific technical solutions: Dave A. Chokhi, Michael Parker, and Dominic Kwiatkowski, "Data Sharing and Intellectual Property in a Genomic Epidemiology Network: Policies for Large-Scale Research Collaboration," *Bulletin of the World Health Organization* 84:5 (May 2006): 382-87; John B. Freymann et al., "Image Data Sharing for Biomedical Research--Meeting HIPAA Requirements for De-identification," *Journal of Digital Imaging* 25 (2012): 14-24, DOI: 10.1007/s10278-011-9422-x. Systemic interventions: Mark J. Costello, "Motivating Online Publication of Data," *Bioscience* 59 (2009): 418-27, DOI: 10.1525/bio.2009.59.5.9; Vishwas Chavan and Lyubomir Penev, "The Data Paper: A Mechanism to Incentivize Data Publishing in Biodiversity Science," *BMC Bioinformatics* 12:S2 (2011), https://doi.org/10.1186/1471-2105-12-S15-S2.

Sarah Callaghan et al., "Making Data a First Class Scientific Output: Data Citation and Publication by NERC's Environmental Data Centres," *International Journal of Digital Curation* 7:1 (2012): 107-113, DOI: 10.2218/ijdc.v7i1.218; Paul E. Uhlir, *For Attribution: Developing Data Attribution and Citation Practices and Standards: Summary of an International Workshop* (National Academies Press: Washington, DC, 2012), DOI: 10.17226/13564; Yvonne M. Socha, ed., "Out of Cite, out of Mind: The Current State of Practice, Policy, and Technology for the Citation of Data," *Data Science Journal* 12 (Sept. 13, 2013), DOI: 10.2481/dsj.OSOM13-043; Benjamin S. Arbucklke et al., "Data Sharing Reveals Complexity In the Westward Spread of Domestic Animals across Neolithic Turkey," *PLOS One* 9:6, e99845 (June 13, 2014), DOI: 10.1371/journal.pone.0099845; "Recommended Practices to Promote Scholarly Data Citation and Tracking: The Role of the Data Citation Index," *Clairvate Analytics*, 2007, file:///U:/Useful%20articles/Data%20Sharing/Recommended%20Practices%20Promote.pdf.

Angela Cochran, "Data Transparency and Civil Engineers," *The Scholarly Kitchen*, Feb. 6, 2019, https://scholarlykitchen.sspnet.org/2019/02/06/data-transparency-and-civil-engineers/.

strategically on the best ways to move forward. Central to this decision is the issue of scale. Is data sharing best assessed and supported on an international or national scale? By broad academic sector (engineering, biomedical)? By discipline? On a university-by-university basis? Or using another unit of analysis altogether? To the extent that there are existing initiatives on each of these scales, how should they relate to one another? How do we design support for data sharing in order to align as closely as possible with the practices and interests of scholars, in order to maximize buy-in?

In this issue brief, we build on our ongoing research into scholarly practices to propose a new mechanism for conceptualizing and supporting STEM research data sharing.[8] Successful data sharing happens within **data communities**, formal or informal groups of scholars who share a certain type of data with each other, regardless of disciplinary boundaries. Drawing on Ithaka S+R findings and the scholarly literature, we identify what constitutes a data community and outline its most important features by studying three success stories, investigating the circumstances under which intensive data sharing is already happening. We contend that stakeholders who wish to promote data sharing – librarians, information technologists, scholarly communications professionals, and research funders, to name a few – should work to identify and support **emergent data communities**. These are groups of scholars for whom a relatively straightforward technological intervention, usually the establishment of a data repository, could kickstart the growth of a more active data sharing culture. We conclude by responding to some potential counterarguments to this call for bottom-up intervention and offering recommendations for ways forward.

[8] We have chosen to focus this issue brief on quantitative STEM data sharing in order to maintain a manageable scope, and because this is where the majority of research and support efforts have concentrated thus far. There is a growing body of research into how quantitative and qualitative data sharing in the social sciences and humanities can be supported. See recently, for example: Renata Gonçalves Curty, "Factors Influencing Research Data Reuse in the Social Sciences: An Exploratory Study," *International Journal of Digital Curation* 11:1 (2016): 96-117, DOI: 10.2218/ijdc.v11i1.401; Libby Bishop and Arja Kuula-Luumi, "Revisiting Qualitative Data Reuse: A Decade On," *SAGE Open* (Jan.-Mar. 2017), DOI: 10.1177/2158244016685136; Sara Mannheimer et al., "Qualitative Data Sharing: Data Repositories and Academic Libraries as Key Partners in Addressing Challenges," *American Behavioral Scientist* (June 28, 2018), DOI: 10.1177/0002764218784991; Ben Marwick and Suzanne E. Pilaar Birch, "A Standard for the Scholarly Citation of Archaeological Data as an Incentive to Data Sharing," *Advances in Archaeological Practice* 6:2 (2018): 125-43, DOI: 10.1017/aap.2018.3; Nicole Ruggiano and Tam E. Perry, "Conducting Secondary Analysis of Qualitative Data: Should We, Can We, and How?" *Qualitative Social Work* 18:1 (2019): 81-97, DOI: 10.1177/1473325017700701. There are also a few data sharing success stories outside STEM fields, most notably the Inter-University Consortium for Political and Social Research (ICPSR), which hosts a repository of quantitative social and behavioral science data: https://www.icpsr.umich.edu/icpsrweb/.

The Data Community

This issue brief focuses on understanding what makes scholars willing to share their data – and on applying that understanding strategically in order to improve and increase sharing going forward. We recognize that this is only one aspect of the work that is needed in this area. Numerous professional organizations (CODATA, DCC, FORCE11, GO FAIR, RDA, and RDAP, to name just a few),[9] in addition to a panoply of smaller projects and working groups, are making significant strides in defining standards and best practices in important technical areas such as metadata creation, discoverability, machine readability, and long-term preservation. These efforts hold out the promise that in the future scholars will share their data in ways that maximize its usefulness for research innovation. However, there is also a need to address scholarly practice on a much more fundamental level. The best policies and standards achieve little without buy-in from the research community, and our research shows that in most fields that buy-in has not yet been achieved.[10] Put simply, STEM researchers must be convinced to share their data in the first place before they can be taught how to share it well.

One problem is that relatively little effort has gone toward establishing conceptual models that accurately describe the ways in which scholars are already sharing their data – and therefore the ways in which they might be best supported in moving toward increased sharing. For example, there has been little critical thought given to the question of whether the discipline is the correct unit of analysis by which to study scholars' attitudes toward and practices of data sharing. In what follows, we argue that it is not, and propose a new conceptual model for understanding how researchers share data.

Data Sharing Success Stories: Three Examples

Existing research points to some preliminary ways forward. For instance, when Ithaka S+R studied the research practices of chemists, agricultural scientists, public health scholars, and civil and environmental engineers, we found that these scholars are much more comfortable with sharing data in a personal and ad-hoc manner, either with other

[9] CODATA (Committee on Data of the International Council for Science): http://www.codata.org/; DCC (Digital Curation Centre): http://www.dcc.ac.uk/; FORCE11 (The Future of Research Communications and e-Scholarship): https://www.force11.org/; GO FAIR: https://www.go-fair.org/; RDA (Research Data Alliance): https://rd-alliance.org/; RDAP (Research Data Access and Preservation): https://rdapassociation.org/.

[10] Long and Schonfeld, "Chemists," 29-30; Cooper, "Agriculture Scholars," 25-26, Cooper and Daniels, "Public Health Scholars," 23; Cooper and Springer, "Engineering Scholars," 25-26.

scholars they know and trust or, most commonly, with collaborators.[11] That is, many scholars implicitly conceive of data sharing as a social activity, and when they share data they rely on networks of professional relationships. This suggests that studies of data sharing should take note of informal links among researchers in addition to formal departmental affiliations. Another promising area of analysis has been the creation of data curation profiles.[12] These evidence-based analyses of data lifecycles help to focus our attention on the technical processes for sharing specific types of data, rather than on disciplinary generalizations.

There is also a need to look carefully at the circumstances under which data sharing on a wider, technology-enabled scale is already happening. A number of data repositories can already be considered success stories; some of these, like the Cambridge Crystallographic Data Centre's Cambridge Structural Database (CSD), are well known, whereas others, such as FlyBase, are relatively unheard of. What do these success stories have in common, and what can they teach us about the possibilities for strategically facilitating data sharing in the sciences? In order to answer these questions, we highlight examples of three successful data sharing initiatives. This selection is by no means meant to be exhaustive; rather, it shows how data sharing can happen in a range of fields and scales.

> *Cambridge Structural Database.* This initiative, run by the Cambridge Crystallographic Data Center, catalogs information about the crystal structures, or arrangements of atoms, of known crystalline materials.[13] It began in 1965 as a computer file and companion series of print volumes which included bibliographic information and numerical data extracted from published articles. In the 1990s, the .CIF file format was developed to organize this information digitally; this format is now standard.[14] Most journals require that new structure determinations associated with articles they publish be submitted to CSD. As a result, virtually all published crystal structures are included in the database.[15] Some researchers also submit crystal structures which have not yet been peer reviewed or published.[16] CSD is important in crystallography, a subfield of chemistry, but is also used by researchers working on drug discovery and materials science.

[11] Long and Schonfeld, "Chemists," 29; Danielle Cooper, Katherine Daniel et al., "Supporting the Changing Research Practices of Public Health Scholars," *Ithaka S+R*, Dec. 14, 2017, 10.18665/sr.305867, 23; Cooper and Springer, "Engineering Scholars," 23-24.

[12] Michael Witt, "Constructing Data Curation Profiles," *The International Journal of Digital Curation* 3:4 (2009), DOI: 10.2218/ijdc.v4i3.117; Melissa H. Cragin et al., "Data Sharing, Small Science and Institutional Repositories," *Philosophical Transactions of the Royal Society A* 368 (2010): 4023-38, DOI: 10.2218/ijdc.v4i3.117.

[13] See https://www.ccdc.cam.ac.uk/solutions/csd-system/components/csd/.

[14] Colin R. Groom et al., "The Cambridge Structural Database," *Acta Crystallographica* B72 (2016): 171-79, DOI: 10.1107/S2052520616003954, 173-74.

[15] Ibid., 177.

[16] Long and Schonfeld, "Chemists," 31.

FlyBase. This database records gene and genome sequences, as well as species information, for the insect family *drosophilidae*, or the fruit fly.[17] This information is useful to scientists in a variety of fields because fruit flies are one of several "model organisms" that reproduce easily and have genetic similarities to humans, making them ideal research subjects.[18] FlyBase was established in 1992 as a dataset which combined existing *drosophilia* records with several overlapping researcher mailing lists. It is funded by the NIH's National Human Genome Research Institute, and its website incorporates not only data access and submission capabilities, but sophisticated navigation tools, a bibliography, a researcher directory, and a discussion forum. It is paralleled by several other NIH-funded databases dedicated to the genetics of other model species, including yeast, the nematode worm, the western clawed frog, the mouse, and the zebrafish. There is significant overlap between the genetic sequences contained in FlyBase and the far larger GenBank, a database containing all publicly available DNA sequences regardless of species.[19]

DesignSafe-CI. "CI" stands for "cyber infrastructure." This data repository, run by the NSF-funded Natural Hazards Engineering Research Infrastructure, allows researchers to store, access, and analyze data related to natural disasters in the cloud.[20] Its predecessor was the NEEShub (Network for Earthquake Engineering Simulation), which traced its origins to a California-based research network established in 1988.[21] Although the database accepts any type of data related to natural disasters – including sensor readings, point clouds from lidar scans, and images – the file formats are generally standard ones and are therefore accessible to users with sufficient domain expertise. A number of civil and environmental engineering scholars spoke positively about DesignSafe-CI because they received staff support in formatting the data and metadata they uploaded.[22] Another strength of the database is its integration with multiple stages of the scholarly workflow. With permission, scholars can upload 100 terabytes or more of raw data and can use built-in tools to analyze and eventually provide wider access to their data.

[17] See https://flybase.org/. Madeline A. Crosby et al., "FlyBase: Genomes by the Dozen," *Nucleic Acids Research* 35:1 (Jan. 2007): D486-91, DOI: 10.1093/nar/gkl827. For the database's history see https://wiki.flybase.org/wiki/FlyBase:About.

[18] For a discussion of how model organisms achieved their status as standard research subjects precisely because advocates recognized their potential to anchor research communities, see Ankeny and Leonelli, "Repertories," 21-22.

[19] See https://www.ncbi.nlm.nih.gov/genbank/.

[20] See https://www.designsafe-ci.org/.

[21] Ixchel M. Faniel and Trond E. Jacobsen, "Reusing Scientific Data: How Earthquake Engineering Researchers Assess the Reusability of Colleagues' Data," *Computer Supported Cooperative Work* 19 (2010): 355-75, DOI: 10.1007/s10606-010-9117-8.

Rathje et al., "DesignSafe: New Cyberinfrastructure for Natural Hazards Engineering," *Natural Hazards Review* 18.3 (2017): 1-7, 10.1061/(ASCE)NH.1527-6996.0000246.

[22] Cooper and Springer, "Engineering Scholars," 27.

Defining the Data Community

All three of these examples involve the creation, or growth, of what we call a "data community." A data community is a fluid and informal network of researchers who share and use a certain type of data, such as crystallographic structures, DNA sequences, or measurements relating to natural disasters. In coining the term "data community," we are building upon a few scattered – yet important – observations in the existing literature that point in this direction, but we are unaware of any prior systematic effort to define an equivalent concept in relation to research data sharing.[23] The concept of a "data community" is also grounded in sociological theories which relate the formation of communities of practice to social relationships and learned identities.[24] Most notably, data communities can be considered one type of the *repertoire*-based scientific research communities described by Sabina Leonelli and Rachel Ankeny.[25]

A data community is a fluid and informal network of researchers who share and use a certain type of data.

In order to understand what a data community *is*, we need to recognize what it is *not*. A data community is not the same thing as a discipline. Indeed, the members of a single data community will often belong to a number of different disciplines. For example, a

[23] Christine L. Borgman observes that in 2010 the National Science Foundation defined data management in relation to "communities of interest," which she infers to mean something close to the "data communities" described here: "The Conundrum of Sharing Research Data," *Journal of the American Society for Information Science and Technology* 63:6 (2012): 1059-78, DOI: 10.1002/asi.22634. Alison Callahan et al. write about the need for a "data sharing community" in spinal cord injury research, and identify current data sharing activities that could lead to one, but do not describe this concept systematically: "Developing a Data Sharing Community for Spinal Cord Injury Research," *Experimental Neurology* 295 (2017): 135-43, DOI: 10.1016/j.expneurol.2017.05.012. See also the research and initiatives cited in notes 12 and 61. For a comparable effort to re-conceptualize an aspect of scholarly communications as community-based, see John Hartley et al., "Do We Need to Move from Communication Technology to User Community? A New Economic Model of the Journal as a Club," *Learned Publishing* 32:1 (2019): 27-35, DOI: 10.1002/leap.1228. We also drew inspiration from the history of arXiv (https://arxiv.org/), a scholarly community based on sharing preprints as opposed to datasets: Paul Ginsparg, "It Was Twenty Years Ago Today…" arXiv:1108.2700v2 [cs.DL], Sept. 13, 2011.

[24] Jean Lave and Etienne Wenger, *Situated Learning: Legitimate Peripheral Participation* (Cambridge, 1991).

[25] Sabina Leonelli and Rachel A. Ankeny, "Repertoires: How to Transform a Project into a Research Community," *BioScience* 65:7 (July 2015): 701-08, DOI: 10.1093/biosci/biv061; Rachel A. Ankeny and Sabina Leonelli, "Repertoires: A Post-Kuhnian Perspective on Scientific Change and Collaborative Research," *Studies in History and Philosophy of Science* 60 (2016): 18-28, DOI: 10.1016/j.shpsa.2016.08.003.

search of the name "John" in FlyBase's directory of "Fly People," or directory members, returns researchers affiliated with twelve different department types.[26] Although biologists predominate, a range of other departmental affiliations, from physics to entomology, are also represented. Additionally, not all the researchers working in any one discipline will belong to the same data community – not all biologists are interested in fly genetics. A researcher can belong to several data communities or no data communities, and can move in and out of data communities as their research topics and practices change.

Table 1

Department	"Fly People" Named John
Biology/Biological Sciences	19
Cell Biology/Molecular Biology/Biochemistry	9
Ecology and Evolution	1
Entomology	1
Genetics	4
Life Sciences	1
Medicine	2
Natural Sciences	1
Neurobiology	3
Physics	1
Statistics	1
Total	43

This is important because studies of data sharing tend to either lump all science researchers together or speak of "disciplinary" cultures and standards.[27] It is true that disciplinary affiliation is important in some respects: university resources are often allocated at the department level, and some disciplines, such as medicine, are more

[26] The directory is at https://flybase.org/community/find.

[27] Exceptions to this trend include Borgman, "Conundrum," and Cragin, "Data Sharing."

dedicated to the notion of “reproducibility” than others.[28] But our research shows that attitudes toward and practices of data sharing vary just as much *within* disciplines as *between* disciplines, if not more. For example, in all four of the STEM-adjacent fields studies by Ithaka S+R, many of the researchers interviewed expressed either skepticism or ambivalence toward sharing data on a wide scale.[29] However, in both agriculture and civil and environmental engineering, interviewees who do not themselves work with genetic data commented on how colleagues whose work has a biological component routinely use and share gene sequences through databases.[30] “I mean, the molecular people are always trying to put the gene sequences in the gene banks and do those kind of things,” commented one agriculture researcher.[31]

Thinking about data sharing in terms of data communities rather than disciplines can allow us to represent scholarly activities more accurately. Scientific research is becoming increasingly interdisciplinary, as grant-funded projects bring together scholars from diverse backgrounds to tackle complex issues. A recent research project which investigated the data sharing habits of scientists at Rutgers University, Temple University, and Pennsylvania State University discovered that some graduate students identify more readily with the subject of their research than with a formal discipline: for instance, several identified themselves as cancer researchers even though one was a mathematician, another a physicist, and so on.[32] This suggests that it would be more intuitive for these researchers to upload their data to a platform designed to facilitate cancer research than to a physics data repository or mathematics data repository.[33]

[28] For examples of studies of data sharing in reproducibility-focused disciplines see: Adam R. Ferguson et al., “Big Data from Small Data: Data-Sharing in the ‘Long Tail’ of Neuroscience,” *Nature Neuroscience* 17:11 (Nov. 2014): 1442-48, DOI: 10.1038/nn.3838; Florian Naudet et al., “Data Sharing and Reanalysis of Randomized Controlled Trials in Leading Biomedical Journals with a Full Data Sharing Policy: Survey of Studies Published in The BMJ and PLOS Medicine,” *BMJ* 360:k400 (2018): DOI: 10.1136/bmj.k400; Stodden, Guo and Ma, “Toward Reproducible Computational Research”; Vasilevsky, “Reproducible and Reusable Research”; Wallach, Boyack and Ioannidis, “Reproducible Research Practices.” For an overview of current issues see National Academies of Sciences, Engineering and Medicine, *Reproducibility and Replicability in Science* (Washington, DC, 2019), DOI: 10.17226/25303.

[29] Long and Schonfeld, “Chemists,” 29-30; Cooper, “Agriculture Scholars,” 25-26, Cooper and Daniels, “Public Health Scholars,” 23; Cooper and Springer, “Engineering Scholars,” 25-26.

[30] Cooper and Springer, “Engineering Scholars,” 27; Cooper, “Agriculture Scholars,” 25.

[31] Cooper, “Agriculture Scholars,” 25.

[32] Grace Agnew, “Can I Trust this Data? Selecting Data for Reuse and Other Dilemmas of the Research Scientist,” presentation at the Coalition for Networked Information Fall 2018 Membership meeting, December 10-11, 2018, https://www.youtube.com/watch?v=9gwLyimdBzg&feature=youtu.be, at 11:16.

[33] In fact, there are several data communities focused on aspects of cancer research, including those which utilize the Cancer Genome Atlas, https://cancergenome.nih.gov/; the Cancer Human Biobank, https://biospecimens.cancer.gov/about/cahub/default.asp; and the Cancer Imaging Archive, http://www.cancerimagingarchive.net/.

At this point, the reader may be wondering whether the concept of a “data community” is simply a rebranding of the domain-specific digital repository. This is not the case. While most of today’s best-developed data communities congregate through online platforms and repositories, it is also possible to identify pre-digital data communities, some of which are still thriving. One example is the data community of botanists, biologists, ecologists, and taxonomists who make use of herbaria. A herbarium is a facility which stores and catalogs physical plant specimens. Herbaria began as quasi-scientific collections in the seventeenth century. Today, scientists who publish discoveries of previously unknown plants must deposit a specimen, known as a holotype, in a herbarium according to a standardized procedure.[34] In this way – and similarly to other successful data communities, such as the scientists who use the CSD – depositing accomplishes the dual functions of sharing information and laying claim to innovations. Scientists also reference the collections in herbaria in order to identify specimens of known species. Although some herbaria make images or indices of their collections available online, including through JSTOR’s Global Plants database and the Smithsonian’s online Botany Collections,[35] there is no single, centralized way to search across the thousands of physical plant collections stored in different herbaria across the globe, and depositing physical specimens remains the only way to share data with the community. Nevertheless, there is clearly a vast community of scholars who collectively supply and use “data” (in the form of physical specimens) according to community norms.

[34] “What is a Herbarium?” *Purdue Herbaria*, 2015, accessed Mar. 15, 2019, https://ag.purdue.edu/btny/Herbaria/Pages/What-is-an-Herbarium-and-what-does-it-do.aspx; Nicholas J. Turland et al., eds., *International Code of Nomenclature for Algae, Fungi, and Plants (Shenzhen Code) Adopted by the Nineteenth International Botanical Congress Shenzhen, China, July 2017*, Regnum Vegetabile 159 (Glashütten, 2018), DOI: 10.12705/Code.2018.

[35] See https://plants.jstor.org/ and https://collections.nmnh.si.edu/search/botany/.

Characteristics of Successful Data Communities

Having determined that the data community is the most useful unit of analysis for understanding data sharing, we can now turn to describing some of the features of the three successful or "established" data communities described above.[36] These features fall into three categories: bottom-up development, absence or mitigation of technical barriers to sharing, and community norms.

Bottom-Up Development

All three of the featured data communities have relatively long histories which begin with small-scale collaborations and communications among researchers. Long-term funding and organizational support allowed those involved in small-scale data sharing efforts to gradually take advantage of new technologies, both of data production and of data storage and sharing. The communities expanded as researchers noticed the benefits their colleagues derived from sharing their data – sometimes serendipitously, sometimes through direct advocacy – and began to do the same. And, as discussed below, publisher and funder mandates reinforced developing community norms.

It is more effective to identify low-tech and small-scale ways in which scholars are already sharing information – and then concentrate efforts on facilitating and improving those existing activities.

Today, because the technology required to create a data sharing platform is widely available, it may be tempting to think that simply creating a platform will stimulate data sharing. This seems not to be the case. It is more effective to identify low-tech and small-scale ways in which scholars are already sharing information – and then concentrate efforts on facilitating and improving those existing activities. For instance, our research

[36] The characteristics of successful data communities identified in this section can also be described using institutional theory, an established sociological theory which posits coercive, normative, and mimetic influences on behavior (in this case, journal and funder requirements, community norms, and organic growth, respectively). See Youngseek Kim and Jeffrey M. Stanton, "Institutional and Individual Influences on Scientists' Data Sharing Practices," *Journal of Computational Science Education* 3 (Jun. 2012): 47-56.

on civil and environmental engineering found that researchers are manually extracting data from the figures and graphs published with research articles using ImageJ analysis software and, in the case of one lab group, by printing out graphs and measuring their coordinates with a ruler.[37] Researchers use these data to benchmark their own experimental results or include them in meta-analyses. There is clearly a demand in at least some areas of civil and environmental engineering for the research data associated with publications to be made available in downloadable format.

Photo courtesy of Rebecca Springer

The example of FlyBase highlights an important tension between the ways in which data communities develop and their long-term sustainability. We have seen that data communities grow around relatively narrow research interests, mirroring researchers' professional networks. FlyBase and the handful of other communities dedicated to model organisms exist successfully alongside GenBank, an enormous, cross-species database of genetic sequences. This makes little sense from the standpoint of technical and financial efficiency, yet these model organism platforms continue to exist because they meet the needs of specific data communities. But it is also possible to find examples of data communities merging in order to pool their resources. VertNet, a project to facilitate the sharing of biodiversity data for vertebrate species, was created from four discrete biodiversity data communities: FishNet (fish), MaNIS (mammals), HerpNET (reptiles) and ORNIS (birds).[38] Creating sustainable organizational models may require a balancing act between preserving the integrity of data communities and avoiding infrastructure replication. In the conclusion of this brief, we discuss how data communities cultivated at the ground level must eventually be supported by large, well-resourced initiatives in order to ensure long-term sustainability.

[37] Cooper and Springer, "Engineering Scholars," 27.

[38] See http://www.vertnet.org/about/about.html.

Absence or Mitigation of Technical Barriers

The second feature of established data communities is that they share data that is technically easy to upload, transfer, and reuse. Specifically, the data files are not extremely large; they do not contain sensitive or personal information; they are shared in standardized file formats that are intelligible to the community; and they can be sufficiently contextualized to enable reuse. In some cases, the emergence of a data community may be closely tied to technological developments which capture essential metadata and make standardization easier: for instance, the success of the CSD is due in part to the development and widespread adoption of the .CIF file format.[39] This is not to say that larger, sensitive, or more complex datasets should not be shared more widely. But those looking to make the greatest impact on data sharing should start by focusing their energy on supporting the growth of communities where the technical and ethical barriers to sharing are lowest – or on developing technical solutions that lower those barriers and promote standardization.[40]

Community Norms

Finally, it is important to observe how data sharing is motivated or rewarded in established data communities. Much of the discussion around how to motivate data sharing has focused on making shared datasets “citable,” either as they exist in repositories or through presentation in “data papers.”[41] The logic runs that data citations might develop a perceived value comparable to that of article citations, and researchers will therefore share their data in order to directly bolster their career prospects. There is some evidence to suggest that the prospect of having their data cited by others would motivate STEM researchers to share their data, although how this reward would balance against perceived costs is more difficult to predict.[42] However, the established data

[39] Colin R. Groom et al., “The Cambridge Structural Database,” *Acta Crystallographica* B72 (2016): 171-79, DOI: 10.1107/S2052520616003954, 173.

[40] For an example of progress toward such a technical solution to facilitate the sharing of biomedical image data, see John B. Freymann et al., “Image Data Sharing for Biomedical Research--Meeting HIPAA Requirements for De-Identification,” *Journal of Digital Imaging* 25 (2012): 14-24.

[41] Callaghan, “Making Data”; Socha, “Out of Cite.” See also the ongoing work of the CODATA-ICSTI Data Citation Standards and Practices task group: http://www.codata.org/task-groups/data-citation-standards-and-practices.

[42] Carol Tenopir et al., Data Sharing by Scientists: Practices and Perceptions,” *PLOS One* 6:6, e21101 (June 2011), DOI: 10.1371/journal.pone.0021101; “The State of Open Data: A Selection of Analyses and Articles about Open Data, Curated by Figshare,” Oct. 2016, DOI: 10.6084/m9.figshare.4036398.v1, 13; “The State of Open Data 2018: A Selection of Analyses and Articles about Open Data, Curated by Figshare,” *Digital Science*, Oct. 2018, DOI: 10.1371/journal.pone.0021101, 9.

communities described above grew even absent the widespread uptake of standardized data citation. Rather, data communities thrive when they cultivate formal or informal norms through which data sharing comes to be expected within the community.

In fact, research has shown that scholars respond better to institutional motivators to share their work, such as tenure and promotion incentives, when community sharing norms have already been established.[43] To be sure, there are other reasons why standardized data citation might be beneficial to the STEM community, including the possibility that the integration of data citations into search tools like Google Scholar, Google Dataset Search, and the Data Citation Index could significantly increase discoverability.[44] And it is certainly important that shared datasets be associated with stable identifiers such as DOIs, and that data "authors" be credited when their work is reused. But data citation must be coupled with efforts that focus on cultivating cultural norms – supporting scholars in sharing the data that they can most easily recognize as useful to themselves and their colleagues.

Publisher and funder requirements, too, are likely to be most effective when they are built on a foundation of community norms.[45] For example, submitting a crystal structure to the CSD is an important way to record it as one's own discovery, but our interviews of chemistry researchers did not indicate that CSD entries carry anything close to the value of peer-reviewed research articles for tenure and promotion.[46] Rather, most of the relevant journals either require researchers to submit, or submit themselves, structures relating to any published articles. These journal requirements were not just created out of thin air: they formalize a culture of sharing that "has always been the norm for the crystallographic community."[47] By contrast, journal-imposed data sharing requirements

[43] Atreyi Kankanhalli, Bernard C. Y. Tan, and Kwok-Kee Wei, "Contributing Knowledge to Electronic Knowledge Repositories: An Empirical Investigation," *MIS Quarterly* 29 (Mar. 2005): 113-43, at 131. Note that this research encompassed the archiving of academic "knowledge," including gray literature, as opposed to data exclusively. See also the survey of 23 participants at a workshop on FAIR spinal cord injury data which rated, on average, establishing discovery and journal requirement compliance as the most important incentives for sharing data, and citation of data and associated papers as the least important: Callahan, "Developing a Data Sharing Community," 137. We are unaware of more formal studies which ask researchers to rank motivations for, as opposed to barriers to, sharing data.

[44] See https://scholar.google.com/; https://toolbox.google.com/datasetsearch; https://clarivate.com/products/web-of-science/web-science-form/data-citation-index/.

[45] Cameron Neylon, "Compliance Culture or Culture Change? The Role of Funders in Improving Data Management and Sharing Practice amongst Researchers," *Research Ideas and Outcomes* 3:e21705 (19 Oct. 2017), DOI: 10.3897/rio.3.e21705.

[46] Long and Schonfeld, "Chemists," 32.

[47] Groom, "Cambridge Structural Database," 172.

in fields which did not have similar traditions of sharing have garnered low compliance rates.[48] For instance, fMRIDC, a now-defunct neuroimaging repository, experienced significant pushback from researchers when several journals instituted a data sharing requirement on its behalf.[49] This is yet another reason why those interested in facilitating data sharing should seek to identify emergent data communities where an ethos of – and rationale for – sharing information is already developing.

Those who want to support data sharing in the sciences need to look for opportunities to grow data communities around scholars' existing practices and interests.

To summarize, successful data communities grow over relatively long timescales, as scholars who were already working cooperatively discover new production and sharing technologies to make data sharing easier. They focus on data that is relatively easy to transmit and reuse. And they don't rely on citations to motivate sharing, but rather operate within formal and informal community norms. It follows that those who want to support data sharing in the sciences need to look for opportunities to grow data communities around scholars' existing practices and interests. We call these opportunities "emergent data communities."

Emergent Data Communities

An emergent data community may not be much of a community at all – yet. Instead, it is a loosely connected group of scholars who all work with a particular type of data, often linked by professional relationships through multiple degrees of separation. These scholars generally have an interest in sharing data with each other and using each other's data. They recognize the benefits of data sharing to their own research agendas, to their colleagues, and/or to their field or even society more broadly, and are not be overly concerned with guarding their own "intellectual property." They are already engaged in haphazard or ad hoc types of data sharing, such as putting data on their laboratory

[48] Federer, "Data Sharing in PLOS ONE"; Couture, "Funder-Imposed Data Publication Requirement"; Naudet, "Data Sharing and Reanalysis"; Sholler, "Enforcing Public Data Archiving Policies."

[49] Russel A. Poldrack and Krzysztof J. Gorgolewski, "Making Big Data Open: Data Sharing in Neuroimaging," *Nature Neuroscience* 17:11 (Nov. 2014): 1510-1517, DOI: 10.1038/nn.3818, 1510.

websites, providing supplemental data files for articles they publish, or sending data to their colleagues when asked personally. And the types of data that these scholars generally work with are relatively easy to transmit and reuse. Specifically, they adhere to standard file types; have manageable file sizes; and are not of a sensitive or proprietary nature. (Examples of the challenges to data sharing posed by nonstandard and large file types and sensitive data abound, particularly in clinical medicine.[50]) In what follows, we present one example of how these conditions should allow for relatively straightforward technical solutions – principally the establishment of a repository that meets the community's particular needs – to enable the organic growth of a full-fledged data sharing community.

Case Study: Air Pollution Researchers

One example of an emergent data community can be identified from the interviews conducted with civil and environmental engineering scholars during an Ithaka S+R study.[51] The resulting report quotes a scholar who researches air quality, commenting on biologists' use of gene sequence databases. "They put it somewhere, and then it's readily available to all researchers," the scholar mused. "And I keep thinking, 'Well, we need to be doing that in air pollution.'"[52] A review of ten other interviews conducted for the same project with air pollution specialists across institutions reveals evidence that the desire for a better way to share air quality data is widespread – and that there is real potential, given the right tools, for an air quality data sharing community to emerge.

None of the air pollution researchers interviewed for the project were hostile to the idea of making their data widely available, and one reported that they already make their data available to other researchers via "external websites," apparently purpose-built by the laboratory group. Moreover, most of the air pollution researchers interviewed articulated a desire to either share their own data or more easily obtain other researchers' data. One likely reason for this is that air pollution researchers are already accustomed to making extensive use of data they did not collect themselves. Government agencies such as the EPA and other organizations provide large quantities of air quality data online, and many air pollution researchers depend on this information. Crucially, the vast majority

[50] Chokhi, Parker and Kwiatkowski, "Data Sharing"; Freymann, "Image Data Sharing"; Cooper and Springer, "Engineering Scholars," 25-26. See especially Poldrack and Gorgolewski, "Making Big Data Open," 1511, for a discussion of the trade-offs between the research potential of large, unprocessed neuroimaging datasets and the costs of making them available relative to smaller, processed datasets.

[51] Cooper and Springer, "Engineering Scholars." In what follows, examples and quotations without citations are from interviews collected for the project which were not included in the final report. See ibid., 4-7 for the project scope and methodology.

[52] Ibid., 27.

of air quality data is neither sensitive nor proprietary, and it is often stored in basic .CSV file formats. It is possible that this dependence on public online sources has helped researchers recognize the benefits of having access to air quality data produced by others.

Indeed, the air pollution researchers interviewed expressed several reasons why they believed that improved data sharing would be a positive innovation. The same researcher who enviously described a colleague's participation in a gene sequence data community also articulated altruistic motives for wanting to make their data more easily available to colleagues: "It would be really nice, instead of [data] just sitting now on my hard drive until I delete it, to put it somewhere that somebody might eventually like to look at it in a different way." Another scholar told a more pointed story about how the lack of a coherent data community caused them to accidentally duplicate a colleague's work. "Just recently, I paid twelve thousand dollars for [a] ... company to generate, for me, meteorological data. ... Then, a couple of weeks ago, [I learned] there's actually a professor in the physics department who is running that same model to generate the same data. And, it's very, very nice and [they] probably would have given the data to me. But, how would I know this?" One can imagine that if there were a single repository where scientists shared air quality data, this researcher would have been able to easily find and access her colleague's data. In addition, this example highlights how an air quality data community could usefully include researchers from fields other than environmental engineering, such as physics.

Many interviewees mentioned that an increasing number of journals in the air pollution subfield are requiring that the data underlying published articles be made freely available. Several also mentioned similar requirements of funders such as the National Science Foundation and Alfred P. Sloan Foundation.[53] However, it appears that in the field of air pollution – as in other fields[54] – journal and funder data sharing mandates have resulted only in partial and haphazard compliance. For instance, interviewees mentioned providing links to Google Drive files containing datasets as a method of fulfilling journal requirements. Several expressed a desire for more "stable" platforms in which to deposit their data. And one researcher expressed bewilderment at the proliferation of institutional and journal repositories, only some of which meet the requirements of funders.[55] Nevertheless, the general attitude of interviewees toward data

[53] See https://nsf.gov/bfa/dias/policy/dmp.jsp and https://sloan.org/storage/app/media/files/application_documents/proposal_guidelines_research_trustee_grants.pdf, 7.

[54] Federer, "Data Sharing in PLOS ONE"; Couture, "Funder-Imposed Data Publication Requirement"; Naudet, "Data Sharing and Reanalysis"; Sholler, "Enforcing Public Data Archiving Policies."

[55] Cooper and Springer, "Engineering Scholars," 28.

sharing requirements – despite confusion around compliance – was not hostile. Rather, these mandates may be helping to acclimate air pollution researchers to the need to integrate data management and processing into their workflows. One scholar described how they now prepare "archives" of all the data underlying each of their publications in processed form as a matter of course. This suggests that, were a more robust air quality data sharing community to emerge, journals and funders could play a role in setting and enforcing data sharing norms, as is the case with crystallography, for instance.

It seems clear that air pollution researchers would benefit from technical and resource support for sharing air quality data with one another, and that such support would have a strong potential to facilitate the development of a full-fledged data community. At the most basic level, a repository is needed. In order to fully meet researchers' needs, this repository would need to provide an assurance of long-term, stable archiving as well as tools for version control. A more ambitious project might be to work with government agencies that publish air quality data to ensure interoperability with the repository, or even to make government data, including historical data, directly available through the repository. Pittsburgh's Breathe Project represents an interesting step in this direction.[56]

Instead of pouring resources into "build it and they will come" strategies, we should take a ground-up approach to data sharing support.

Identifying emergent data communities requires careful attention to the practices and attitudes of scholars at a granular level, spotting patterns within and across disciplines as well as across institutions. Large-scale, collaborative, qualitative studies, such as those conducted by Ithaka S+R, are one way to achieve this.[57] Despite the effort required, we believe that teasing out these areas of potential development and structuring support accordingly will pay significant dividends. Instead of pouring resources into "build it and they will come" strategies, information professionals, publishers and policy makers should take a ground-up approach to data sharing support. Over time, data communities will multiply as scientists recognize the benefits their colleagues enjoy from sharing information with one another – as the air pollution researcher did when reflecting on genetics data communities and musing that "we need to be doing that." Crucially, such

[56] See https://breatheproject.org/.

[57] See https://sr.ithaka.org/our-work/research-support/.

approaches will require significant work across institutional boundaries and therefore outside the traditional scope of the academic library, IT center, or research office. We outline possibilities for implementing cross-institutional solutions in the conclusion.

Counter-Arguments

We recognize that advocating for a bottom-up, community-scoped approach to supporting research data sharing runs against the grain of much of the work that is currently being done around this issue. Below, we offer responses to two likely counter arguments.

Will an incremental approach limit our ability to truly disrupt the scientific research system?

A number of large-scale initiatives are currently underway to promote and enable STEM data sharing, including the National Institute of Health's Data Commons and Canada's Federated Research Data Repository.[58] Perhaps the most ambitious project is the European Union's European Open Science Cloud (EOSC).[59] This long-term project aims to create a cross-disciplinary research data commons utilized by all scientists working across EU member states. Unsurprisingly, there are diverse and significant technical and policy challenges – not to mention funding requirements – to be met.[60]

These large-scale initiatives are important for articulating shared visions, informing policy and fueling research and innovation. However, we believe that the most effective research support interventions are those that map closely onto current research practices and expressed support needs, because scholars will more readily adopt them. By contrast, rushing ahead to create large-scale infrastructures is likely to backfire. Without sufficient buy-in among scholars, expensive repositories will remain under-utilized, and mandated data sharing will be perceived as simply another burden imposed on already-strained researchers. A bottom-up approach – characterized by close attention to scholars' practices and organic community growth – is needed in order to achieve truly systemic change.

[58] See https://nihdatacommons.us/ and https://www.frdr.ca/repo/.

[59] See https://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud.

[60] For an in-depth discussion, see highlights from presentations at the 2017 EOSC summit: https://www.youtube.com/watch?v=qIn0homyLm4.

Notably, some supporters of the EOSC have already recognized this. GO FAIR is an international organization that supports bottom-up initiatives toward achieving the expansive vision of the EOSC by implementing FAIR data principles.[61] The organization primarily operates through implementation networks, small stakeholder groups each dedicated to addressing a specific domain or technical challenge. These implementation networks are one promising vehicle for identifying and supporting emergent data communities.

Could institutional repositories – which libraries have already poured money and staff hours into – be adapted for data archiving and sharing?

Over the past two decades, many academic libraries have invested considerable staff and financial resources into developing and utilizing institutional repositories. Most of these repositories were initially built to store and make available academic gray literature, such as unpublished reports, dissertations, and preprints of articles destined for publication. However, in recent years, librarians have begun exploring ways to repurpose institutional repositories as platforms to enable data sharing; an extensive literature details the successes and challenges of these efforts.[62]

While institutional repositories serve other useful purposes, we believe that they are a vehicle ill-suited to the support of data communities that cross institutional boundaries. By its nature, the institutional repository segregates information according to the college or university at which it was created, while at the same time bringing together the vast array of different types of data created within and across the institution's departments. While this may fulfill an institution's preservation mandate, such repositories are not

[61] See https://www.go-fair.org/.

[62] Select examples: Philip M. Davis and Matthew J. L. Connolly, "Institutional Repositories: Evaluating the Reasons for Non-use of Cornell University's Installation of DSpace," *D-Lib Magazine* 13:3-4 (March-Apr. 2007), DOI: 10.1045/march2007-davisIJDC; Dorothea Salo, "Innkeeper at the roach motel," *Library Trends* 57 (2008): 98-123, http://digital.library.wisc.edu/1793/22088; "The Research Library's Role in Digital Repository Services," *Association of Research Libraries*, 2009, http://www.arl.org/storage/documents/publications/repository-services-report-jan09.pdf; Cragin, "Data Sharing"; Beth Plale et al., "SEAD Virtual Archive: Building a Federation of Institutional Repositories for Long-Term Data Preservation in Sustainability Science," *International Journal of Digital Curation* 8:2 (2014): 172-180, DOI: 10.2218/ijdc.v8i2.281; Dong Joon Lee, "Practices of Research Data Curation in Institutional Repositories: A Qualitative View from Repository Staff," *PLOS ONE* 12:3, e0173987 (2017), DOI: 10.1371/journal.pone.0173987; "Rethinking Institutional Repository Strategies: Report of a CNI Executive Roundtable Held April 2 & 3, 2017," *Coalition for Networked Information*, May 2017, https://www.cni.org/wp-content/uploads/2017/05/CNI-rethinking-irs-exec-rndtbl.report.S17.v1.pdf; Lisa R. Johnston et al., "How Important Are Data Curation Activities to Researchers? Gaps and Opportunities for Academic Libraries," *Journal of Librarianship and Scholarly Communication* 6, eP2198 (2018), DOI: 10.7710/2162-3309.2198; Sayeed Choudhury, Gregory E. Monaco, and Thomas Hauser, "Research Data Curation: A Framework for an Institution-Wide Services Approach," *EDUCASE*, May 2018, https://library.educause.edu/resources/2018/5/research-data-curation-a-framework-for-an-institution-wide-services-approach.

conducive to supporting the way scientific research is conducted, with researchers from different institutions working together on focused projects. As described below, academic librarians are uniquely positioned to spearhead data sharing initiatives; they have the opportunity to redirect their efforts in this area away from institutional repositories and toward strategic initiatives that support data communities across institutional boundaries.

Ways Forward

Thinking about data sharing in terms of data communities can help librarians, information technologists, scholarly communications professionals, and research funders create more dynamic and strategic support services that reflect the way scientists work. Data communities, today usually facilitated through online repositories, are fluid, networked groups that focus on specific types of data and cut across disciplinary and institutional boundaries. They grow socially, enabled by policies and technologies but ultimately fueled by a recognition among researchers of the benefits of sharing data. The concept of a data community points toward several avenues for action: we must concentrate existing data sharing efforts on building data communities from the ground up. Three principle action steps are needed in order to accomplish this.

Growing and Multiplying Data Communities: Three Steps

First, more research into the current practices, attitudes, and expressed desires of researchers is needed in order to identify emergent data communities and tailor supports to their unique needs. Crucially, this research must eschew the institution and discipline as categories of analysis, instead recognizing that STEM scholars work in interdisciplinary and multi-institutional clusters around specific datasets – data communities. Qualitative research methods are well suited to uncovering the subtleties of scholarly predispositions. Network analysis has the potential to illuminate the informal professional relationships through which data is shared, particularly in emergent data communities.[63] And quantitative research into the rates at which available datasets are actually used – and the factors which best predict their reuse – holds great promise.[64]

[63] Santo Fortunato et al., "Science of Science," *Science* 359:6379, eaao0185 (Mar. 2, 2018), DOI: 10.1126/science.aao0185.

[64] Lisa Federer, "FAIR Data at the National Library of Medicine and National Institutes of Health," keynote presentation at the Drexel-CODATA FAIR-RRDM Workshop, Philadelphia, PA, Mar. 31, 2019.

Second, there is a need for a variety of stakeholders to work toward developing technical solutions that make cumbersome or heterogeneous data types more easily shareable, since data communities tend to grow most successfully around data that can be easily reused. The following stakeholders may have particularly important roles.

- Rather than imposing sweeping data sharing requirements, **funders** should focus on funding projects that facilitate data reuse in specific fields. In some cases, this will involve improvements to the ways in which information is collected and transmitted by scientific instruments, or the creation of new file formats, content guidelines and metadata standards. In other cases, it will mean the development of entirely novel technologies, such as the Personal Health Train, to enable remote research on sensitive datasets.[65]

- **Professional societies** may have a role to play in advocating for standardization of both file formats and metadata properties according to the norms that develop within emerging data communities. But in so doing, they must be mindful of the interdisciplinarity of these communities, encouraging members to adopt data curation practices that adequately contextualize data for reuse in multiple research contexts.

- **Publishers** of disciplinary and interdisciplinary journals should be on the lookout for opportunities to reinforce existing and emerging community norms around sharing data through mandated data publication and citation.

- In addition to standardizing formats to facilitate human-to-human sharing, **information technologists** must continue to work toward data and metadata standards that maximize machine readability. And **digital curation experts** should weigh in on issues of long-term preservation and stability, especially since some researchers worry about the longevity of repositories that rely on government funding.[66]

Third – and building on the foundations of research and technological development – established and well-resourced organizations must cooperate with, and indeed rely on, small-scale, on-the-ground initiatives in order to grow data communities. Seen from the opposite perspective, new data communities are likely to be susceptible to the adverse effects of organizational and staffing changes and funding reductions. Support from larger organizations is crucial to ensuring that data communities, once established, can to continue to flourish. We have already mentioned the GO FAIR implementation networks as a promising start in this direction. Another possibility is for data communities to enlist the support of organizations like the Educopia Institute or the American Association for the Advancement of Science's Center for Scientific Collaboration and Community Engagement, which provide training and resources to

[65] Johan von Soest et al., "Using the Personal Health Train for Automated and Privacy-Preserving Analytics on Vertically Partitioned Data," *Studies in Health Technology and Informatics* 247 (Jan. 2018): 581-85.

[66] Cooper and Springer, "Engineering Scholars," 28.

help build and sustain communities from the ground up.[67] It also is possible to envision a variety of other organizations playing key roles, including government-funded laboratories; multi-institution research networks; publishers; and platforms capable of hosting discrete communities, akin to Figshare or the Social Science Research Network.[68] As data communities emerge and mature, these larger organizations can look for ways to provide long-term infrastructure and encourage greater standardization and interoperability, while preserving the shared identities and norms that allowed each community to develop in the first place.

Academic Libraries

It is also worth reflecting on the need to rethink the role of academic libraries specifically in supporting data sharing. The academic institution – and, by extension, its library – is not always the appropriate scale on which to address the challenges that scientists face. Librarians who want to effectively support scientists must find creative ways to contribute their expertise within the broader, cross-institutional, interdisciplinary ecosystem of scientific research. Awareness is the first step: librarians should understand to which data communities scientists at their institutions belong. This awareness may lead to opportunities to support scientists who are seeking to grow data communities in their fields. The web and organizational infrastructure that supports many data communities is "hosted" at particular institutions; for instance, DesignSafe-CI was developed by a multi-institution team of investigators led by researchers at the University of Texas at Austin, and its cloud storage and analysis capabilities are thanks to the same university's Texas Advanced Computing Center.[69] Librarians who are aware of nascent initiatives to foster data communities at their own institutions could contribute their expertise by, for instance, advising on issues of intellectual property and copyright.

A more radical approach might see librarians getting involved in supporting data sharing outside their own institutions. Successful data communities rely on dedicated staff who help researchers ensure that the data they share with others is properly formatted and contextualized. As Katherine Akers and Jennifer Green have suggested, academic

[67] See https://educopia.org/ and, in particular, Katherine Skinner, "Community Cultivation: A Field Guide," *Educopia Institute*, Nov. 2018, https://educopia.org/wp-content/uploads/2018/11/CommunityCultivationFieldGuide.pdf; see https://www.aaas.org/programs/center-scientific-collaboration-and-community-engagement.

[68] See https://figshare.com/ and https://www.ssrn.com/index.cfm/en/.

[69] See https://www.tacc.utexas.edu/research-development/tacc-projects/designsafe.

librarians could work for domain repositories as “local curators.”[70] Libraries could recognize such work as a form of professional development for promotion and tenure, akin to conference presentations and research. The Data Curation Network, already piloting a model to link data curation experts across partner institutions, is an excellent candidate to organize these efforts.[71] The best opportunities for librarians to leverage their unique expertise in designing information systems to support science research may lie outside the university altogether.

Looking Beyond the Academy

A variety of stakeholders have a role to play in helping scientists to think outside the box – or more aptly, outside the academy – when it comes to sharing and acquiring data. There is certainly a need for better discovery mechanisms to enable researchers to locate data generated by governments, nonprofit organizations, and corporations online, an issue which lies outside the scope of this paper.[72] More pertinently, it is possible to point to successful “citizen science” initiatives, in which non-academic enthusiasts contribute data that is useful to them and to scientists, in addition to benefiting from online community interactions.[73] For example, eBird, based at the Cornell University Laboratory of Ornithology, has used artificial-intelligence-enabled crowdsourcing to collect and make freely available over 100 million bird sightings per year, incentivizing community engagement by allowing birders to curate their own bird sighting lists, naming “eBirders of the Month,” and giving away prizes to top contributors – a user-centered approach originally designed through direct consultation with the birdwatching community.[74]

Finally, we have focused this issue brief on enabling data produced by researchers – sometimes referred to as the “long tail” of research data – to be used by other

[70] Katherine G. Akers and Jennifer A. Green, “Towards a Symbiotic Relationship Between Academic Libraries and Disciplinary Data Repositories: A Dryad and University of Michigan Case Study,” *International Journal of Digital Curation*, 9:1 (2014): 120-31, DOI: 10.2218/ijdc.v9i1.306, 124-25.

[71] See https://datacurationnetwork.org/.

[72] For a theoretical perspective on the role libraries might take, see Lorcan Dempsey, “Library Collections in the Life of the User: Two Directions,” *Liber Quarterly* 26:4 (2016): 338-59, DOI: 10.18352/lq.10170.

[73] National Academies of Sciences, Engineering, and Medicine, *Learning through Citizen Science: Enhancing Opportunities by Design* (Washington, DC, 2018). DOI: 10.17226/25183.

[74] See https://ebird.org/home. Matthew Loy, “eBird 2009: A Two-sided Market for Academic Researchers and Enthusiasts,” *Ithaka S+R*, July 14, 2009, http://sr.ithaka.org/?p=22390; idem, “eBird 2011: Driving Impact through Crowdsourcing, Case Study Update 2011,” *Ithaka S+R*, Oct. 6, 2011, DOI: 10.18665/sr.22373.

researchers.[75] But it is also true that data pertaining to many fields, from medicine to transportation to human behavior, is increasingly produced and controlled by private corporations. The problem of researcher access to commercial data is significant, and is only destined to intensify. For example, researchers in agriculture and civil and environmental engineering wish that academic libraries could purchase access to proprietary datasets[76] – a desire that will likely go unfulfilled given the existing strains on library budgets. And concerns around privacy, ethics, data security, and preservation in the context of the private sector abound, with regulators struggling to keep pace. In order to maximize innovation and promote research ethics in STEM fields, efforts to support the sharing of academic research data must be coupled with advocacy around the ethical use of proprietary data.

Final Thoughts

For all today's technological affordances, data sharing remains fundamentally a social activity. As a variety of stakeholders consider strategies to meet the evolving needs of STEM scholars – and to guide scientific research toward greater rigor and innovation – they must focus on understanding and supporting research practices from the ground up. An important first step is the development of evidence-based conceptual models, such as the data communities model presented here, to describe scholarly activity – models which can then be tested, refined, and implemented as new information comes to light. We look forward to further exploring these models with others invested in the future of STEM research support.

[75] Adam R. Ferguson et al., "Big Data from Small Data: Data-Sharing in the 'Long Tail' of Neuroscience," *Nature Neuroscience* 7:11 (Nov. 2014): 1442-48.

[76] Cooper, "Agriculture Scholars," 16; Cooper and Springer, "Engineering Scholars," 18.