IJDC | *Conference Pre-print*

Data Communities: Empowering Researcher-Driven Data Sharing in the Sciences

Rebecca Springer
Ithaka S+R

Danielle Cooper
Ithaka S+R

Abstract

There is a growing perception that science can progress more quickly, more innovatively, and more rigorously when researchers share data with each other. However many scientists are not engaging in data sharing and remain skeptical of its relevance to their work. As organizations and initiatives designed to promote STEM data sharing multiply – within, across, and outside academic institutions – there is a pressing need to decide strategically on the best ways to move forward. In this paper, we propose a new mechanism for conceptualizing and supporting STEM research data sharing. 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 the findings of four large-scale qualitative studies of research practices conducted by Ithaka S+R, as well as 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 empower *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 offering recommendations for ways forward.

Submitted 16 December 2019 ~ Accepted 19 February 2020

Correspondence should be addressed to Rebecca Springer, Ithaka S+R, 6 East 32nd Street 10th Floor, New York, NY 10016. Email: rebecca.springer@ithaka.org

This paper was presented at International Digital Curation Conference IDCC20, Dublin, 17-19 February 2020

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. The IJDC is published by the University of Edinburgh on behalf of the Digital Curation Centre. ISSN: 1746-8256. URL: http://www.ijdc.net/

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 as the standard benchmark for data sharing success (Wilkinson et al, 2016). There is also an emerging scholarly literature evaluating the efficacies of some of these policies.[2] By contrast, many scientists are not engaging in data sharing and remain skeptical of its relevance to their work (Long, M.P. and Schonfeld, R.C. 2013, Cooper, D. et al. 2017, Cooper,D., Springer, R. et al., 2018)

As organizations and initiatives designed to promote STEM data sharing multiply – within, across, and outside academic institutions – there is a pressing need to decide 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 paper, we propose a new mechanism for conceptualizing and supporting STEM research data sharing.[3] 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 the findings of four large-scale qualitative studies of research practices conducted by Ithaka S+R, (Long and Schonfeld 2013, Cooper 2017, Cooper, Daniel et al., 2017, Cooper and Springer, 2018) as well as 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 empower *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 offering recommendations for ways forward.

[1] This paper is based on Cooper and Springer (2019)

[2] A limited selection: Stodden (2013), Akers and Dotty (2013), Roche (2015), Herold (2015), Shen (2017), Vasilevsky (2017), Blassime et al (2018), Wallach et al Naudet et al (2018), Wiley (2018), Federer et al (2018); Couture et al (2018); Scholler et al (2019).

[3] 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. Christine L. Borgman (2012) 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: Alison Callahan et al. (2017) 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. 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 Hartley et al (2017). We also drew inspiration from the history of arXiv, a scholarly community based on sharing preprints as opposed to datasets (Ginsparg, 2011) 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 (Lave and Wenger 1991, Leonelli and Ankeny 2015, Ankeny and Leonelli 2016).

Defining the Data Community: Case Studies

A number of initiatives can already be considered data sharing "success stories." We begin by asking: 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 explore three examples of successful data sharing initiatives:

- Cambridge Structural Database[4] (crystal structures) (Groom et al 2016, Long and Schonfeld 2013)

- FlyBase[5] (drosophiliae gene and genome sequences) (Crosby et al 2007, Ankeny and Leonelli, 2016)

- DesignSafe-CI[6] (natural hazards engineering data) (Faniel and Jacobsen 2010, Rathje et al 2017)

We conclude that all three of these examples involve the creation, or growth, of what we call a "data community."

User data from FlyBase makes it clear that 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. 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. This is important because studies of data sharing tend to either lump all science researchers together or speak of "disciplinary" cultures and standards. Thinking about data sharing in terms of data communities rather than disciplines can allow us to represent scholarly activities more accurately, as scientific research is becoming increasingly interdisciplinary and grant-funded projects bring together scholars from diverse backgrounds to tackle complex issues.

Characteristics of Successful Data Communities

Having determined that the data community is the most useful unit of analysis for understanding data sharing, we turn to describing some of the features of successful data communities. 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.

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.

[4] https://www.ccdc.cam.ac.uk/solutions/csd-system/components/csd/
[5] https://flybase.org/
[6] https://www.designsafe-ci.org

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. It is also important to note that creating sustainable organizational models may require a balancing act between preserving the integrity of data communities and avoiding infrastructure replication.

Absence or Mitigation of Technical Barriers

The second feature of established data communities is that they tend to 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. 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.

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."[7] 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. However, the established data 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. Publisher and funder requirements, too, are likely to be most effective when they are built on a foundation of community norms. 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.

Emergent Data Communities: Definition and Example

Building on our study of established data communities, we argue that those who want to support data sharing in the sciences need to look for opportunities to empower data communities built around scholars' existing practices and interests. We call these opportunities *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

[7] 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

"intellectual property." They are already engaged in haphazard or ad hoc types of data sharing, such as putting data on their laboratory 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.

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. A review of eleven interviews conducted with air pollution specialists at a variety of institutions reveals evidence that the desire for a better way to share air quality data is widespread in this subfield – and that there is real potential, given the right tools, for an air quality data sharing community to emerge. We present this evidence in order to show how relatively straightforward technical solutions – principally the establishment of a repository that meets the community's particular needs – could enable the organic growth of a full-fledged data sharing community.

Conclusion: 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. 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.

Further 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.

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. We make suggestions specific to funders, professional societies, publishers, and information technologists.

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. 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.

Acknowledgements

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.

References

Akers, K. and Jennifer Doty, J. (2013) "Disciplinary Differences in Faculty Research Data Management Practices and Perspectives," *The International Journal of Digital Curation* 8.2: 5-26, DOI: 10.2218/ijdc.v8i2.263

Ankeny, R.A. and Leonelli, S. "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.

Blasimme, A., et al., (2018) "Data Sharing for Precision Medicine: Policy Lessons and Future Directions," *Health Affairs* 37:5, 702-9, DOI: 10.1377/hlthaff.2017.1558

Borgman, C. (2012) "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.

Callahan, A. et al. (2017) "Developing a Data Sharing Community for Spinal Cord Injury Research," *Experimental Neurology* 295 (2017): 135-43, DOI: 10.1016/j.expneurol.2017.05.012.

Callaghan, S. 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

Cooper, D. et al. (2017) "Supporting the Changing Research Practices of Agriculture Scholars," Ithaka S+R, DOI: 10.18665/sr.303663, 24-26

Cooper, D. Daniel, K. et al., (2017) "Supporting the Changing Research Practices of Public Health Scholars," Ithaka S+R, Dec. 14, 2017, 10.18665/sr.305867

Cooper, D., Springer, R. et al.,(2018) "Supporting the Changing Research Practices of Civil and Environmental Engineering Scholars," Ithaka S+R, DOI: 10.18665/sr.310885, 22-28.

Cooper, D. and Springer, R. (2019) "Data Communities: A New Model for Supporting STEM Data Sharing," *Ithaka S+R*, DOI: 10.18665/sr.311396.

Couture, J.L., et al. (2018), "A Funder-Imposed Data Publication Requirement Seldom Inspired Data Sharing," *PLOS ONE* 13:7, e0199780, DOI: 10.1371/journal.pone.0199789

Crosby, M.A. et al., "FlyBase: Genomes by the Dozen," Nucleic Acids Research 35:1 (Jan. 2007): D486-91, DOI: 10.1093/nar/gkl827

Faniel, I.M. and Jacobsen, T.E. "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

Federer, L. et al (2018) "Data Sharing in PLOS ONE: An Analysis of Data Availability Statements," *PLOS ONE* 13:5, e0194768, DOI: 10.1371/journal.pone.0194768

Figshare (2016) "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;

Figshare (2018) "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.

Freymann, J.B. et al., "Image Data Sharing for Biomedical Research–Meeting HIPAA Requirements for De-Identification," *Journal of Digital Imaging* 25 (2012): 14-24.

Ginsparg, P. (2011) "It Was Twenty Years Ago Today…" arXiv:1108.2700v2 [cs.DL], Sept. 13, 2011.

Groom, C.R et al., "The Cambridge Structural Database," Acta Crystallographica B72 (2016): 171-79, DOI: 10.1107/S2052520616003954

Hartley, J. et al., (2019) "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.

Herold, P., (2015) "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

Lave, J. and Wenger, E. *Situated Learning: Legitimate Peripheral Participation* (Cambridge, 1991)

Leonelli, S. and Ankeny,R.A. "Repertoires: How to Transform a Project into a Research Community," *BioScience* 65:7 (July 2015): 701-08, DOI: 10.1093/biosci/biv061

Long, M.P. and Schonfeld, R.C, (2013) "Supporting the Changing Research Practices of Chemists," Ithaka S+R, Feb. 26, 2013, DOI: 10.18665/sr.22561, 29-31

Naudet, F. et al. (2018) "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

Neylon, C. "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.

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

Roche, D. (2015) "Public Data Archiving in Ecology and Evolution: How Well Are We Doing?" *PLOS Biology* 13:11, e1002295, DOI: 10.1371/journal.pbio.1002295

Shen, Y., (2017) "Data Sharing Practices, Information Exchange Behaviors, and Knowledge Discovery Dynamics: A Study of Natural Resources and Environmental Scientists," *Environmental Systems Research* 6:9, DOI: 10.1186/s40068-017-0086-5

Sholler, D et al., (2019) "Enforcing Public Data Archiving Policies in Academic Publishing: A Study of Ecology Journals," *Big Data & Society* DOI: 10.1177/2053951719836259.

Socha, Y.M. 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.

Stodden,V., Guo, P., and Ma, Z., (2013) "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

Tenopir, C. et al., Data Sharing by Scientists: Practices and Perceptions," *PLOS One* 6:6, e21101 (June 2011), DOI: 10.1371/journal.pone.0021101;

Vasilevsky, N.A. et al., (2017) "Reproducible and Reusable Research: Are Journal Data Sharing Policies Meeting the Mark?" *PeerJ* e3208, DOI: 10.7717/peerj.3208

Wallach, J.D., Boyack, K.W. and Ioannidis, J.P (2018) "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

Wiley,C. (2018) "Data Sharing and Engineering Faculty: An Analysis of Selected Publications," *Science & Technology Libraries* 37:4 (2018), DOI: 10.1080/0194262X.2018.1516596

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