When Enough Isn’t Enough: Rethinking Research Support in the Sciences through the Issue of Data Sharing

Danielle Cooper and Rebecca Springer

A major challenge facing academic libraries is offering services that align with scientists’ research support needs. This paper shares findings from Ithaka S+R on science scholars’ evolving research activities in order to consider how libraries can effectively structure scientific research support. The paper begins with a brief review of the landscape of academic library services for scientists alongside recent evidence on science scholars’ evolving research needs. From this review the authors identify data sharing as an area of particular need for further support and explore these needs in further detail. The paper concludes by considering challenges and opportunities for developing data sharing research supports in the sciences, including academic library policies and services.

Introduction

What would it take for university libraries to develop the same kind of buy-in with scientists as they have with humanists? A major impediment to this is the misalignment between institution-specific library services and scientific research support needs. Deeply understanding how scientists’ research activities alongside the constraints facing academic libraries and institutions more broadly is essential to designing effective research support services in the sciences.

This paper examines how science research is evolving in relation to library service models to identify the greatest challenges and opportunities for scientific research support. Following a review of the current library support service landscape in the sciences, as well as broader trends in scientists’ research activities and support needs, we take a deep dive into data sharing as an emerging area warranting further attention by information professionals. We contend that there are three interrelated issues that academic libraries with science research support missions must especially attend to. First, scientists’ data support needs exceed data management services. Second, the concept of “communities” as opposed to “disciplines” best captures how scientists work. And third, the locus of science activity and its associated research support needs often exceeds institution-specific scoping.

We conclude by considering how academic libraries may respond to these issues—including the possibility of conceeding that some of these issues must be met collaboratively or delegated to other non-library entities. We hope that this research will prompt further consideration of what role academic libraries are best positioned to take in supporting science research.

Library Services and Science Research Support Needs: The Current Landscape

Research support for the scientific disciplines, similar to other disciplines in the humanities, social sciences, and professions, continues to be carried out in many academic libraries through liaison service models. Reflecting
differences in scale between various scientific fields at different schools, institutions may have some or all of their science library services offered through a central library, a separate library with a mandate to serve some or all of the sciences, and/or more discipline- or department-specific libraries. Some of the major critiques to the liaison model include that it is challenging to scale while maintaining sufficiently specialized services at the disciplinary level, and that discipline-level scoping does not adequately reflect how scholars work with information.

In recognition that the liaison model in and of itself may not be sufficient to ensure appropriately personalized touchpoints for scholars, some academic libraries have developed embedded approaches to service. Embedded approaches are characterized by co-locating librarians in the physical spaces of faculty, and, by extension, tasking them with become involved in faculty activities in a more direct and ongoing nature. Similar to liaison roles, embedded librarianship is a model that is being enacted in academic librarianship more broadly, as opposed to science librarianship specifically.

However, there is also recognition that subject liaison, and more recently, embedded models, cannot merely be leveraged more effectively, as many have been trying to do for over a decade. Instead, they must be re-thought or discarded entirely. Of particular interest has been the possibility of expanding or developing more functional roles; in the sciences, this is especially represented in the discourse around e-science librarianship. “E-science” has a variety of definitions, but of particular relevance to academic libraries is that it represents a data-driven approach to research made possible through digital technologies—which in turn necessitates new research support services for working with data. This movement in how science research is conducted, coupled with new requirements from funders pertaining to data management, has led to the creation of new library roles and services specifically oriented towards data support, particularly data management support. Therefore, while these data support services may be offered under the banner of “e-science librarianship,” they more closely follow trends towards offering functionally-defined, as opposed to discipline-defined, services in libraries.

The scientific research process involves information-oriented activities necessitating research support at all points, including content discovery, access, and manipulation; information management; and scholarly communications. Since many academic libraries are invested in these areas, it is important to consider how common library interventions align (or misalign) with scientists’ research practices. For example, through a tri-annual survey of U.S. faculty and a series of discipline-specific studies on scholars’ research activities in a variety of scientific fields, Ithaka S+R research demonstrates that science scholars have experienced diminishing challenges with discovery of and access to published literature over time, whereas working with grey literature and data continues to represent a relatively greater challenge. In particular, data management is an ongoing challenge, with 90% of faculty reporting in the most recent U.S. faculty survey cycle that they organize their data on their computers. In discipline-specific studies of research practices—including studies of chemistry, agriculture, public health, and civil and environmental engineering—scholars further detailed ad-hoc information management practices that lead to data loss and other inefficiencies. By contrast, scientists’ scholarly communications practices remain relatively stable, with the need to publish in journals that are high ranking, relevant, and prestigious being a much higher priority than making that content available through mainstream open access channels.

While it is possible to speak generally across the sciences about needs to the discovery and communication of published content, or pertaining to some aspects of data management, these generalities notably do not extend to other areas of data practice, particularly data sharing. As findings from Ithaka S+R’s most recent project on research practices in civil and environmental engineering reflects, effective data sharing requires a trust between researchers, adequate contextual information, and usable file formats—all of which are difficult to provide through discipline-wide or institution-wide policies or services. Yet scholars recognize that data sharing is an area of particular potential for innovating research practice. As the following section will further explore, data
sharing may represent the limit of what is possible to achieve through conventionally scoped academic library support services.

Traditionally, library interventions in data sharing are focused on adapting the institutional repository to house research data—that is, they are focused at the institutional level. Scientists’ data sharing habits extend across and beyond institutions, but not enough attention has been paid to studying and learning from the ways that scholars are already sharing their data. First of all, 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 more comfortable with sharing data in a personal and ad-hoc manner, either with other scholars they know and trust or, most commonly, with collaborators. In other words, many scholars implicitly conceive of data sharing as a social activity, and when they share data they rely on networks of professional relationships. 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 Structural Database, are well known, whereas others, such as FlyBase and DesignSafe-CI, are relatively unheard of. What do these success stories have in common, and what can they teach us about how libraries should work strategically to support 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 (CSD).** This initiative, run by the Cambridge Crystallographic Data Center, catalogues information about the crystal structures, or arrangements of atoms, of known crystalline materials. 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. 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. Some researchers also submit crystal structures which have not yet been peer reviewed or published.

**FlyBase.** This database records gene and genome sequences, as well as species information, for the insect family *drosophilidae*, or the fruit fly. 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. 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.

**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. Its predecessor was the NEEShub (Network for Earthquake Engineering Simulation), which itself had origins in a California-based research network established in 1988. 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. In our research, 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.

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 publish their data online.

Each of these successful data sharing initiatives involves the creation, or expansion, of 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. Although in the pre-Internet era some researchers could cooperatively compare and share data by, for instance, depositing plant specimens in herbaria, today data communities are generally hosted on a website.

A data community is not the same thing as a discipline. In fact, the researchers who make up a data community often belong to many different disciplines. Consider the case of John FlyGuy: a search of the name “John” in FlyBase’s directory of “Fly People,” or directory members, returns researchers affiliated with twelve different disciplines (see Figure 1). And not all the researchers working in any one discipline will belong to the same data community—for instance, not all biologists are interested in fly genetics. Additionally, a researcher can belong to several data communities or no data communities. In short, a data community is not the same thing as a discipline.

This is important because studies of data sharing tend to either lump all science researchers together or speak of “disciplinary” cultures and standards. 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 and psychology, are more dedicated to the notion of “reproducibility” than others. But our research shows that attitudes toward and practices of data sharing vary just as much or more widely within disciplines than between disciplines. For example, in our reports on scientists’ research practices, most of the researchers interviewed expressed either skepticism or ambivalence toward sharing data on a wide scale. However, in both fields, 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. “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.

<table>
<thead>
<tr>
<th>Field</th>
<th>“Fly People” Named John</th>
</tr>
</thead>
<tbody>
<tr>
<td>Biological Sciences</td>
<td>3</td>
</tr>
<tr>
<td>Biology</td>
<td>16</td>
</tr>
<tr>
<td>Cell Biology/Molecular Biology/Biochemistry</td>
<td>9</td>
</tr>
<tr>
<td>Ecology and Evolution</td>
<td>1</td>
</tr>
<tr>
<td>Entomology</td>
<td>1</td>
</tr>
<tr>
<td>Genetics</td>
<td>4</td>
</tr>
<tr>
<td>Life Sciences</td>
<td>1</td>
</tr>
<tr>
<td>Medicine</td>
<td>2</td>
</tr>
<tr>
<td>Natural Sciences</td>
<td>1</td>
</tr>
<tr>
<td>Neurobiology</td>
<td>3</td>
</tr>
<tr>
<td>Physics</td>
<td>1</td>
</tr>
<tr>
<td>Statistics</td>
<td>1</td>
</tr>
<tr>
<td>Total</td>
<td>43</td>
</tr>
</tbody>
</table>
Thinking in terms of data communities can also help us recognize that science 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. 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. Additionally, the example of FlyBase suggests that smaller data communities built around relatively narrow research interests can be valuable to researchers. From the standpoint of technical efficiency, one might want to simply fold FlyBase and the several other data communities dedicated to model organisms into GenBank, an enormous, cross-species database of genetic sequences. Yet these platforms exist successfully alongside GenBank because they meet the needs of specific research communities.

**Conclusion**

Thinking about data sharing in terms of data communities can help librarians create more dynamic and strategic support services that reflect the way scientists work today. As discussed briefly above, there are a number of areas—including information discovery and access, data management, and scholarly communications—where a greater recognition of scientists’ support needs has the potential to improve libraries’ offerings. However, in this paper, we chose to focus specifically on the area of data sharing as an example of how critically examining scientists’ research practices can suggest transformative changes to the scope of scientific research support. Specifically, we argue that scientists’ data sharing practices can be best described with reference to communities. These “data communities,” today usually facilitated through an online repository, are fluid, social groups that focus on specific types of data and cut across disciplinary and institutional boundaries.

The organization of scientific data sharing according to “data communities”—rather than departments or institutions—has important practical implications for how library services are delivered. For example, as academic libraries ramp up their investments in data-oriented services, they should explore ways to structure these services around the communities, as opposed to disciplines, that will benefit. In fact, the framing of library data services exceeds the dichotomy of functional versus subject-specific. While data services focus on tasks that cut across disciplinary areas, scholars’ associated needs cannot be separated from the research contexts in which data is used.

Furthermore, it seems unlikely that institutional repositories will ever become the platform of choice for a substantial share of scientists who want to share their data with others. Setting aside the enormous technical challenges of building a database that effectively captures metadata and data in the vast array of formats used by scientists within any single discipline—much less across the university—it is clear that when scientists think data, they think in terms of specific research communities, not universities. And it is vital that data sharing services—and other research support services—be built in ways that reflect how scientists work. The three successful data communities highlighted above all have relatively long histories which begin with small-scale collaborations and communications among researchers. As internet and database technology developed, they were able to expand and become more dynamic. In other words, simply making an institutional repository available is not enough to stimulate data sharing, especially if that platform is misaligned with scientists’ existing research habits and networks.

Our research highlights how the 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. They are particularly well placed to advocate for bottom-up, researcher-centric solutions as a way to balance top-down interventions such as the European Union's ambitious European Open Science Cloud.\textsuperscript{26} 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 the cloud storage and analysis capabilities of DesignSafe-CI are thanks to the same university's Texas Advanced Computing Center. 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 altogether. Successful data communities rely on dedicated staff who help researchers ensure that the data they share with others is properly formatted and contextualized. In this model academic librarians skilled in metadata creation, for instance, would increasingly be encouraged to arrange secondments to organizations hosting data repositories so that they could work part-time in these capacities. Libraries could recognize such work as a form of professional development for promotion and tenure, akin to conference presentations and research. The best opportunities for librarians to leverage their unique expertise in designing information systems to support science research may lie outside the university altogether.

For all today's technological affordances, data sharing remains a social activity. As academic libraries consider the best ways to reconfigure their support services to meet the evolving needs of scientists, it will be important for the extra-institutional nature of these needs to be taken into account. Shoring up the profession, and its services, to support the extra-institutional and cross-disciplinary nature of data work is one possibility. Another is simply to develop a clearer vision of what an individual academic library can and cannot do. We look forward to further exploring those trade-offs with others invested in the future of science research support.

Endnotes
1. Ithaka S+R's 2015 Faculty Survey makes clear the differences in perceived importance of library functions among humanities, social sciences, science, and medicine faculty, with humanities faculty consistently rating the importance of library functions more highly: Christine Wolff, Alisa Rod, and Roger C. Schonfeld, "Ithaka S+R US Faculty Survey 2015," Ithaka S+R, April 4, 2015, https://doi.org/10.18665/sr.277685, 69.
2. While issues pertaining to research overlap with those pertaining to teaching, for the purposes of this paper we are focusing particularly on research because these activities warrant distinct supports. By extension, this paper focuses on academic libraries at institutions with mandates to support research, as opposed to teaching-focused mandates.
3. For a recent consideration of the issues facing library liaison programs specifically in the humanities, see Danielle Cooper and Roger Schonfeld, "Re-thinking Liaison Programs for the Humanities," Ithaka S+R, July 26, 2017, https://doi.org/10.18665/sr.304124.
5. E-science and other terms for data support services in library contexts may be grouped together or used interchangeably. See for example the title of the foundational report on the topic, from Catherine Soehner, Catherine Steeves, and Jennifer Ward, "E-Science and Data Support Services: A Study of ARL Member Institutions," Association of Research Libraries (2010), https://files.eric.ed.gov/fulltext/ED528643.pdf.
Danielle Cooper and Rebecca Springer


12. See https://www.cccd.cam.ac.uk/solutions/csd-system/components/csd/.
15. See https://www.designsafe-ci.org/.
19. The directory is at https://flybase.org/community/find.