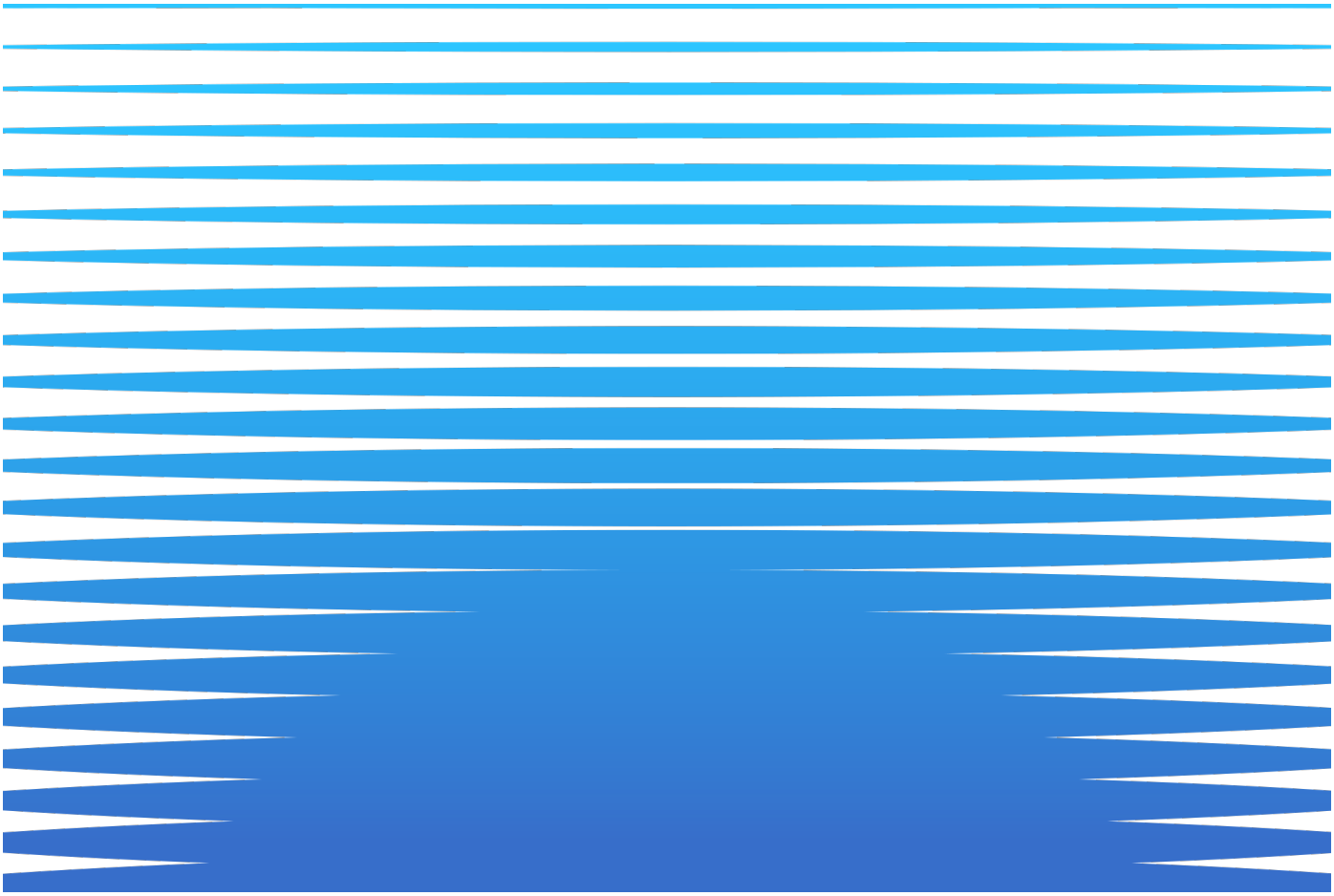
Leveraging Data Communities to Advance Open Science

Findings from an Incubation Workshop Series

Dylan Ruediger
Ruby MacDougall
Danielle M. Cooper
Jake Carlson
Joel Herndon
Lisa Johnston



ITHAKA S+R

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 with a mission to improve access to knowledge and education for people around the world. We believe education is key to the wellbeing of individuals and society, and we work to make it more effective and affordable.

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

This material is based upon work supported by the National Science Foundation under Grant No.2103433.

Table of Contents

Introduction

Several recent studies have indicated that large numbers of researchers in many STEM fields now accept the value of openly sharing research data. Yet, the actual practice of sharing data—especially in forms that comply with FAIR principles—remains a challenge for many researchers to integrate into their workflows and prioritize among the demands on their time.[1] In many disciplines and subfields, data sharing is still mostly an ideal, honored more in the breach than in practice.[2]

The barriers to open data sharing are numerous.[3] However, sustained funding from federal agencies in the United States including the NSF and NIH and important initiatives in other countries such as Canada's Tri-Agency Research Data Management Policy and the European Union's OpenAire, is creating a growing infrastructure for open sharing of research data, albeit one that highlights the tension between scientific research practices that are now regularly multi-national in scope yet exist within funding and regulatory structures determined largely by national entities.[4] In the US context, the most visible fruits of these efforts are the decentralized network of repositories that have become available to researchers in many fields and are now a

[1] The FAIR data principles hold that data must be findable, accessible, interoperable, and reusable, by both humans and machines. For an overview of FAIR principles, see Mark D. Wilkinson et al., "Comment: The FAIR Guiding Principles for Scientific Data Management and Stewardship," *Scientific Data* 3:160018 (15 March2016), https://doi.org/10.1038/sdata.2016.18.

[2] Dylan Ruediger, Danielle Cooper, et al., "Big Data Infrastructure at the Crossroads: Support Needs and Challenges for Universities," *Ithaka S+R,* 1 December 2021, https://doi.org/10.18665/sr.316121. Digital Science et al., "The State of Open Data 2021," *Digital Science*, 30 November 2021, https://doi.org/10.6084/m9.figshare.17061347.v1; Natasha Susan Mauthner and Odette Parry, "Open Access Digital Sharing: Principles, Policies and Practices" *Social Epistemology* 27, no. 1 (2013): 47–67, http://dx.doi.org/10.1080/02691728.2012.760663. For disciplinary differences, see also Michal Tal-Socher and Adrian Ziderman, "Data Sharing Policies in Scholarly Publications: Interdisciplinary Comparisons," *Prometheus* 36, no. 2 (2020): 116–34; Katherine G. Akers and Jennifer Doty, "Disciplinary Differences in Faculty Research Data Management Practices and Perspectives," *International Journal of Digital Curation* 8, no. 2 (21 November 2013): 5–26, https://doi.org/10.2218/ijdc.v8i2.263; Christine L. Borgman, "The Conundrum of Sharing Research Data," *Journal of the American Society for Information Science and Technology* 63, no. 6 (2012): 1059–78, https://doi.org/10.1002/asi.22634; C. Tenopir et al., "Research Data Sharing: Practices and Attitudes of Geophysicists," *Earth and Space Science* 5, no. 12 (2018): 891–902, https://doi.org/10.1029/2018EA000461; Bobby Lee Houtkoop et al., "Data Sharing in Psychology: A Survey on Barriers and Preconditions," *Advances in Methods and Practices in Psychological Science* 1, no. 1 (1 March 2018): 70–85, https://doi.org/10.1177/2515245917751886; Willem G. van Panhuis et al., "A Systematic Review of Barriers to Data Sharing in Public Health," *BMC Public Health* 14, no. 1 (5 November 2014): 1144, https://doi.org/10.1186/1471-2458-14-1144.

[3] Carol Tenopir et al., "Data Sharing by Scientists: Practices and Perceptions," *PLOS ONE* 6, no. 6 (29 June 2011): e21101, https://doi.org/10.1371/journal.pone.0021101; Laia Pujol Priego, Jonathan Wareham, and Angelo Kenneth S. Romasanta, "The Puzzle of Sharing Scientific Data," *Industry and Innovation* 29, no. 2 (7 February 2022): 219–50, https://doi.org/10.1080/13662716.2022.2033178; Natasha J. Gownaris et al., "Barriers to Full Participation in the Open Science Life Cycle among Early Career Researchers," *Data Science Journal* 21, no. 1 (19 January 2022): 2, https://doi.org/10.5334/dsj-2022-002; Greg Tananbaum and Michael M. Crow, "We Must Tear Down the Barriers That Impede Scientific Progress," *Scientific American*, 18 December 2020, https://www.scientificamerican.com/article/we-must-tear-down-the-barriers-that-impede-scientific-progress/.

[4] Innovation Government of Canada, "Tri-Agency Statement of Principles on Digital Data Management - Science.Gc.Ca" (Innovation, Science and Economic Development Canada), 21 January 2021, https://www.ic.gc.ca/eic/site/063.nsf/eng/h_83F7624E.html; David Moher and Kelly D. Cobey, "Ensuring the Success of Data Sharing in Canada," *FACETS* 6 (January 2021): 1534–38, https://doi.org/10.1139/facets-2021-0031; "OpenAIRE," accessed 29 June 2022, https://www.openaire.eu/; Najla Rettberg and Birgit Schmidt, "OpenAIRE: Supporting a European Open Access Mandate" *College & Research Libraries News*, 10 April 2017, https://doi.org/10.5860/crln.76.6.9326.

vital infrastructure for data sharing across many fields. As incentive structures have slowly shifted, the number of researchers taking advantage of these resources has also grown.

The existence of these repositories are necessary enabling conditions for data sharing, but their ability to transform researcher's practices around data depositing and sharing absent changes to incentive structures and the culture of research communities will remain uneven. Furthering the goals of open science requires convincing more researchers of the value of data sharing to themselves and to the community of researchers with whom they most tangibly identify. Creating and encouraging community norms that reward sharing is necessary because data sharing, especially FAIR (findable, accessible, interoperable, and reusable) compliant sharing, is hard work. Absent strong incentive and reward structures, researchers are often reluctant to take on this "extra" labor. Successful data sharing ultimately depends on cultural and social infrastructures as much as on technical infrastructures.

> Successful data sharing ultimately depends on cultural and social infrastructures as much as on technical infrastructures.

Despite the complexity of encouraging scholars to share data, many researcher communities are deeply invested in this work—some have been voluntarily building infrastructures for sharing for decades. For the past several years, Ithaka S+R has been exploring these groups, which we have defined as "data communities." Data communities are fluid networks of scientists who voluntarily exchange and reuse data across disciplinary boundaries to advance shared or complementary research goals, often organized around an online data repository.[5] Successful data communities leverage overlapping research agendas, and in some cases interpersonal relationships, to provide a purpose for making research data available to others that resonates with the goals and motivations of researchers in practical and concrete ways. As Martin Boeckhout, Gerhard A. Zielhuis, and Annelien L. Bredenoord observe, successful data sharing initiatives in the life sciences are "usually carried forward by closely-knit communities of practice which collaborate on more than just standards."[6] Learning from the success of these networks can help ensure that broader data sharing initiatives are responsive to the needs of scholars and provide insights into the importance of community building and shared research

[5] Danielle Cooper and Rebecca Springer, "Data Communities: A New Model for Supporting STEM Data Sharing," *Ithaka S+R*, 13 May 2019, https://sr.ithaka.org/publications/data-communities/. For case studies, see: Rebecca Springer, "Emergent Data Community Spotlight: An Interview with Dr. Vance Lemmon on Spinal Cord Injury Research," *Ithaka S+R*, 22 July 2019, https://sr.ithaka.org/blog/emergent-data-community-spotlight/; Danielle Cooper, "Emergent Data Community Spotlight: An Interview About Energy Modeling with the Open Energy Modelling Initiative," *Ithaka S+R*, 20 April 20 2021, https://sr.ithaka.org/blog/emergent-data-community-spotlight-openmod/; Rebecca Springer, "Emergent Data Community Spotlight III: An Interview with Kitty Emery and Rob Guralnick on ZooArchNet," *Ithaka S+R*, 19 September 2019, https://sr.ithaka.org/blog/emergent-data-community-spotlight-iii/; Rebecca Springer, "Emergent Data Community Spotlight II: An Interview with Felicity Tayler and Marjorie Mitchell on the SpokenWeb Project," *Ithaka S+R*, 10 September 2019, https://sr.ithaka.org/blog/emergent-data-community-spotlight-ii/.

[6] Martin Boeckhout, Gerhard A. Zielhuis, and Annelien L. Bredenoord, "The FAIR Guiding Principles for Data Stewardship: Fair Enough?" *European Journal of Human Genetics* 26, no. 7 (July 2018): 931–36, https://doi.org/10.1038/s41431-018-0160-0; Kathleen Gregory et al., "Lost or Found? Discovering Data Needed for Research," *Harvard Data Science Review* 2, no. 2 (30 April 2020), https://doi.org/10.1162/99608f92.e38165eb.

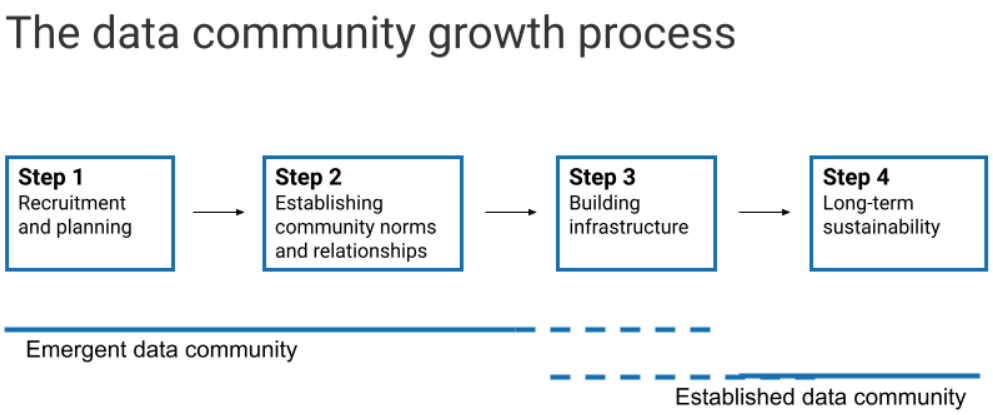
goals to create the organizational and cultural conditions necessary for widespread adoption of FAIR data sharing practices.

"Leveraging Data Communities to Advance Open Science," a multi-session incubation workshop made possible with generous funding from a National Science Foundation (grant no. 2103433), was designed to accomplish three core goals in support of building data infrastructures and community networks to advance open science, using the data communities framework as a guide.

1. To test the hypothesis that the data communities framework, originally designed to describe a specific type of data sharing initiative, could be used as a tool for identifying and supporting them. Could the concept be leveraged into a philosophy of action?
2. To foster cross-disciplinary and cross-institutional conversations about data management, curation, and preservation infrastructures and build ties between researchers and information professionals.
3. To assess how voluntary data sharing efforts might contribute to development of the NSF's efforts to develop metadata fields to maximize data discoverability and, where possible, machine readability, of metadata deposited to the NSF's Public Access Repository (PAR).

"Leveraging Data Communities to Advance Open Science" participants met during the period of January-March 2022. The following report provides a detailed account of the workshop activities and key findings about how the data communities framework can contribute to open data sharing among scientific communities. Three key findings are of particular interest to advocates for open data sharing:

- Data communities have distinct support needs at different points in their life cycle, but all data communities will benefit from treating their social infrastructure as a vital ingredient in successful data sharing efforts.
- Bringing together researchers from radically different disciplines and dedicated information professionals for conversations about data sharing challenges can provide opportunities to discern strategies for data sharing that cut across disciplines and fields.
- Domain repositories offer unique advantages for promoting social and cultural aspects of data sharing. However, both domain and generalist repositories will benefit from greater connections across platforms that will advance discoverability.



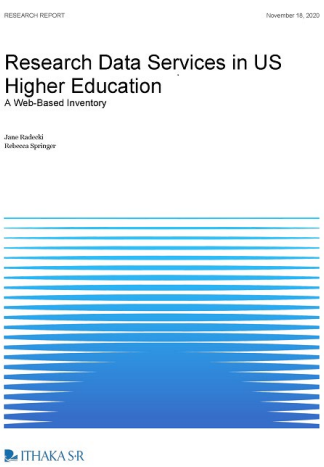
About the Workshop Organizers

Ithaka S+R is a nonprofit research organization that helps academic and cultural communities serve the public good and navigate economic, technological, and demographic change. Since 2000, we have closely monitored changing research practices, norms, and communities across disciplines. Much of this work has focused on how data-intensive researchers collect, organize, and share research data. Working in collaboration with university libraries, we have conducted over 1,500 semi-structured interviews with researchers at more than 107 institutions to explore evolving information and scholarly communication practices. Throughout these studies, we have focused attention on data management and sharing practices, particularly in relation to the growing acceptance of the goals of open science.

Highlights of Ithaka S+R's work on data-intensive research practices

Recent and in-progress reports

- Research Data Services in US Higher Education (2020)
- Big Data Infrastructure as the Crossroads (2021)
- What is a Research Core? (2021)
- Leveraging Data Communities to Advance Open Science workshops (2022)
- Teaching with Data in the Social Sciences (forthcoming)
- Data Sharing Practices in the Humanities (forthcoming)

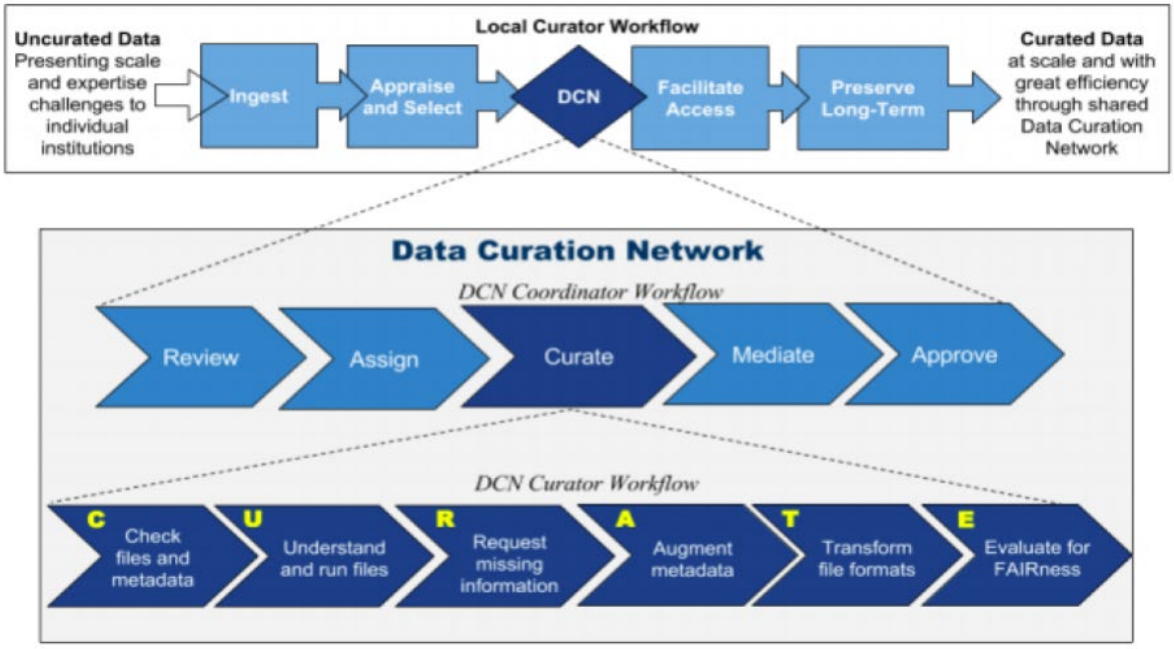


Foundational work

- US Faculty Survey (2000-present)
- Research and Teaching Support Services projects (2012-present)
- Data communities model and case studies (2019-present)
- Research enterprise (2020-present)

"Leveraging Data Communities to Advance Open Science" builds on these previous research projects, most notably our investigations of data communities. Our research on trends in data sharing have consistently highlighted the important role that data communities play in creating cultural conditions conducive for establishing the goals of open science. "Leveraging Data Communities to Advance Open Science" represents our first attempt to explore how this concept, originally designed as a descriptive category of analysis, can be used as an organizational framework for fostering multidisciplinary data sharing.

The Data Curation Network (DCN) is an organization comprised of 15 academic institutions and data repositories designed to empower researchers to publish high quality data in accordance with FAIR standards, advance the art and science of data curation by creating, adopting, and sharing best practices, and support the professional development of information professionals. Launched with support from the Alfred P. Sloan Foundation and the Institute for Museum and Library Services (IMLS), the DCN is now a self-sustaining member-based organization. Its core service offerings include providing hands-on, detailed data curation support by matching information professionals with research teams. The DCN pools expertise from across institutions to collectively and efficiently curate a wider variety of data types than any single member institution can offer.



DCN leaders and Ithaka S+R staff comprised the core organizing committee for the workshop. The organizing committee was responsible for designing the workshop, creating the call for applications, evaluating applications and selecting participants, meeting with participants, and organizing the workshop presentations and activities. Dr. Danielle Cooper, Dr. Dylan Ruediger, and Ruby MacDougall represented Ithaka S+R. Jake Carlson (University of Michigan), Dr. Joel Herndon (Duke University), and Lisa Johnston (formerly of the University of Minnesota, now at the University of Wisconsin) represented the DCN.

Grant Activities

"Leveraging Data Communities to Advance Open Science" was originally conceptualized as an in-person event that would have been held at the University of Michigan on February 28 and March 1, 2022. Due to surging COVID-19 cases, it became clear that it was neither safe nor viable to hold a workshop that required participants to travel and to meet in-person and, in early

January 2022, the organizing committee made the difficult decision to convert the workshop into a virtual event. This last-minute switch in modalities required significant reimagining of a program designed to take full advantages of the intimacy and collegiality that in-person meetings provide.

The in-person event was re-designed as a series of interactive workshops and meetings in the winter of 2022. The primary components of the workshop series were two initial consulting sessions devoted to information exchange and relationship building, an intensive two-day incubation workshop, a specialized working group session devoted specifically to the NSF's PAR, and a final reflection session to synthesize learnings.

Recruiting participants

To recruit scientists, the organizing committee solicited applications from representatives of existing or potential data communities (See Appendix 1 for the CFP). We sought applicants in two ways: by sharing the call for applications widely and through targeted encouragement to apply, with a goal of ensuring that our cohort would include scientists from a wide range of disciplines (focused on researchers in fields funded by the NSF) and include representation from established and emerging data communities. The combination of an openly advertised CFP and direct recruitment was designed to ensure a diverse pool of applicants in terms of gender, race, ethnicity, professional status, and other important measures of diversity, equity, and inclusivity.

Interested parties were invited to submit an application, in which they provided a brief narrative outlining how they met the criteria for participation. The organizing committee evaluated the applications using a scoring rubric based on five criteria to ensure a systematic process for evaluation:

- The extent to which the applicant represented an existing or potential data community
- The extent to which the applicants identified goals would be addressed by participating in the workshop
- The likelihood that the applicants participating would advance readiness for voluntary metadata sharing via the NSF's PAR
- The extent to which the applicant's data community included interdisciplinary and multi-institutional participation
- The extent to which the applicant demonstrated a commitment to diversity

Using these criteria, the organizing committee invited 14 research teams to participate in the workshop: all 14 accepted.

A core goal of the workshop series was to foster collaboration between academic researchers and information professionals. Information professionals—broadly defined as data librarians, data curators, computing staff, and research staff affiliated with departments or labs—are an important source of expertise on metadata, controlled vocabularies, preservation standards and

other issues relating to management and curation of data.[7] To build ties between researchers and information professionals, the organizing committee invited 14 individuals with data management expertise in domains relevant to the selected data communities to serve as advisors to the research teams. These individuals were selected by the organizing committee rather than through a call for participation in order to ensure the best possible match between the needs of the research teams and the expertise of the information professional advisor.

Together, these parallel selection processes resulted in a cohort of 54 individuals who participated in the workshop series: 40 team members representing 14 research teams and 14 information professionals. A complete list of the participating data communities is below. A brief description of their research and participating team members can be found in Appendix 2.

Participating Data Communities

- American Society of Agronomy
- Association of Religion Data Archives (ARDA)
- Center for Applied Internet Data Analysis (CAIDA)
- Center for Health Equity Research/Duke Clinical Research Institute
- IsoBank
- Maple River Dam Removal Project
- Materials Commons
- Montana CREWS Project
- Natural History Data
- ASU NCSU Phosphorus Sustainability
- Play and Learning Across a Year Project
- Polar Geospatial Center (PGC)
- Radiopharmaceutical Therapy Data (RTD)
- Artificial Intelligence Task Force of the Society of Nuclear Medicine and Molecular Imaging (SNMMI AI Task Force)

Consulting sessions

To prepare for the incubation workshop, each research team met for a pre-workshop data interview with their paired information professional and one organizing committee member. In these meetings, participants discussed the current state of the data community the group represents, their current data sharing practices or interests, and the issues that the group would like to address through attending the workshop. These meetings helped the groups clarify and

[7] 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 (1): 1-24, https://doi.org/10.7710/2162-3309.2198; Lisa R. Johnston et al., "Data Curation Network: A Cross-Institutional Staffing Model for Curating Research Data," *International Journal of Digital Curation* 13, no. 1 (31 December 2018): 125–40, https://doi.org/10.2218/ijdc.v13i1.616.

focus their data sharing goals and specify how collaboration with an information professional might advance those goals. The meetings also gave the information professionals and organizing committee members a chance to identify overlapping themes among groups that would be valuable to explore more deeply within the workshop. Following the pre-workshop meetings, the entire organizing committee and all participating information professionals met for a collective debriefing session to discuss high-level takeaways from the consulting sessions and to use knowledge gained from them to ensure that the incubation workshop program was aligned with researchers' interests and needs.

Incubation workshop

The incubation workshop, the centerpiece of our program, was held online over two days. The first day of programming focused on defining success. A plenary panel which brought together leaders from the Center for Open Science, Inter-university Consortium for Political and Social Research (ICPSR), and ZooArchNet to discuss characteristics of successful data communities across a variety of disciplines provided an initial framework for the event. The research teams then broke out into seven pre-assigned working sessions (each consisting of two research teams and two information professionals) to discuss the strengths and successes of their data communities. After discussion, the working groups reconvened in plenary for final reflections about the nature of success and ways to measure it.

Day two of the workshop was organized as a series of unconference sessions on the following topics:

1. Data size, formats, and interoperability
2. Metadata, discoverability, and platforms
3. Storage issues, infrastructure, and policy groups
4. Tools, automation, and workflows
5. Sensitive data, privacy, and ethics
6. Community engagement

Participants were free to join those conversations that were most germane to their data communities' challenges. The conversations provided opportunities for researchers to conceptualize problem solving strategies, with the hope that putting researchers in contact with individuals from dramatically different disciplinary backgrounds with similar problems would lead to innovative thinking that sometimes is difficult to accomplish when participants get too bogged down in the minutiae of their individual challenges.

As this brief description suggests, the workshop program focused heavily on creating an environment conducive to peer-learning and on staging increasingly focused conversations oriented towards collaborative problem solving.

PAR workshop and reflection session

Following the incubation workshop, the organizing committee held two additional group meetings, the first with Martin Halbert, NSF Science Advisor for Public Access, to discuss the NSF's data sharing goals and policies, particularly those related to the NSF's Public Access Repository (PAR). A final debriefing session for the information professional advisors provided an opportunity to reflect on what the information professionals learned during the conference and on how to build closer collaborative relationships with research communities in the future.

Finding Area One: Testing the Concept of Data Communities

The data communities framework can serve as an identity label

Few members of participating research teams were accustomed to thinking of themselves as members of a data community when they applied to participate in the workshop. However, they rapidly adopted the term to describe what they were trying to create. The facility with which researchers began to identify themselves and their colleagues as a community, and the idea that their repository could serve as a focal point for that community clearly resonated with participants, who were using the term fluidly by the end of the workshop. The term proved valuable in part for its flexibility: a data community can consist of a small cluster of highly specialized researchers such as those involved in the Radiopharmaceutical Therapy Data Community or a large and dispersed network of scientists and community members like those involved in the Maple River Dam Removal data community. Its emphasis on the social, cultural, and interpersonal relationships between researchers encouraged participants to consider managing human relationships and organizing people as well as data as a central component of successful data sharing efforts.

Data communities have different needs across their lifecycle

When we designed our workshop, we understood that it would be important to include representation from data communities at different levels of maturity, but our thinking on this point was essentially pedagogical: it seemed clear that the co-learning environment we hoped to facilitate would best be served if emerging data communities had the opportunity to learn from established ones. During the course of the workshop series, it became clear that data communities, like research data, go through a lifecycle or a sequence of stages from initial formation to the end of active life, with each stage requiring different levels and kinds of support. The lifecycle of a data community can be conceptualized as a spectrum:

Potential data communities are small groups of scholars working on overlapping problems who are highly motivated to share research data but lack the expert technical, curatorial, or infrastructure support necessary to share data openly with larger research communities. Research from Ithaka S+R suggests that sharing in these groups often happens through

informal networks and interpersonal relationships.[8] When these groups begin attempting to formalize data sharing practices and infrastructures, they transform into emergent data communities.

Emergent data communities have begun creating more formalized structures for widespread sharing of research data, often with a goal of creating a specialized repository. They may be developing protocols and policies and seeking institutional or grant funding. Emergent data communities may also have already established nascent repositories (or protocols for using generalist repositories), but at this stage are likely to be concerned with creating momentum and establishing an initial pool of depositors and reusers. Emergent data communities often require support focused on designing basic infrastructure and marketing and communication efforts designed to attract a user base and additional deposits of data sets relevant to the community.

ASU NCSU Phosphorus, an **emergent data community**, is exploring building a data repository, and they need help understanding how to accommodate many heterogeneous data types, including sensitive industrial data, from an interdisciplinary community of users.

Established data communities have created a platform for sharing, developed initial standards for data management and curation, and established a core group of users. Established data communities are likely to face challenges associated with sustainability and growth beyond that initial group of users. They may face financial challenges associated with sustaining repositories after start-up grant funds have been exhausted or of adequate staffing as user-bases grow and are likely to be piloting mechanisms to create diversified revenue streams. Growing numbers of users may require significant revisions to metadata standards and other adjustments to policies and protocols as new types of users seek to deposit data types and formats as they grow, while others reported needing help in teaching and incentivizing new users to follow their formatting standards. These are all problems associated with managing growth. However, established data communities are perhaps equally likely to struggle with stagnation associated with the failure to attract new community members. Endless growth is not a goal for most data communities, which depend after all, on their ability to have a core identity capable of mobilizing a community. However, without attracting new members and interest, the vitality of a data repository will decline over time.

[8] Dylan Ruediger and Danielle Cooper, "Big Data Infrastructure at the Crossroads: Support Needs and Challenges for Universities," *Ithaka S+R,* 1 December 2021, https://doi.org/10.18665/sr.316121; Danielle Cooper and Rebecca Springer, "Data Communities: A New Model for Supporting STEM Data Sharing," *Ithaka S+R,* 13 May 2019, https://doi.org/10.18665/sr.311396.

Michigan Natural History, a community working with a variety of natural history data, is an **established community** that can benefit from assistance in understanding the curation and storage a complex data type: 3D data. In addition to helping the team establish metadata on provenance of 3D images, an information professional can help them align research projects with the institutional goals of accessible, long-term preservation by evaluating storage options for 3D data.

Mature data communities have their own unique sets of challenges. Maturity can be associated with decline, as community members disengage and new deposits of data and/or the funding necessary to sustain a vibrant repository dry up. In such cases, the primary issue community members face may be locating a place for long term preservation of deposited data, perhaps in generalist repositories. However, decline is not an evitable result of maturation. Some mature communities, for example FlyBase, or WormBase will continue to expand and grow while retaining tight topical focuses.[9] Others, such as OpenNeuro will grow by expanding the scope of their deposits and data formats they support. In rare instances, domain repositories may reach the point where their scope and the numbers of researchers they are serving raise the prospect that they might become too large and diverse to effectively function as a data community. Funders, universities, and other entities involved in supporting data sharing will benefit from understanding where particular data communities are located on this spectrum and fine-tuning support offerings to challenges associated with the dynamics of growth and stagnation that characterize data communities across their lifecycle.

Metadata and ontologies are social infrastructure

In data sharing contexts, metadata and ontologies are often treated primarily as descriptive tools allowing for standardizing the structuring of repositories and facilitating discoverability and identification. Certainly, this is an important function of both terms. However, throughout the course of this workshop it became apparent that metadata and ontologies also embody community norms and shared—if often contestable—frameworks of understanding. They represent a moment of consensus in a conversation between members of a community about core questions of how knowledge is, or ought to be, organized and classified. One central lesson from the workshop was that the opportunities afforded by emphasizing metadata as a social *process* rather than a descriptive output, a call for engagement rather than a technical standard, provided important ways to engage new researchers and foster their investment in creating and maintaining a data sharing infrastructure.

[9] Madeline A. Crosby et al. "FlyBase: Genomes by the Dozen," *Nucleic Acids Research* 35, no. 1 (1 January 2007): D486–D491, https://doi.org/10.1093/nar/gkl827.

> Throughout the course of this workshop it became apparent that metadata and ontologies also embody community norms and shared—if often contestable—frameworks of understanding.

For IsoBank, a multi-organization data community focused on building a common repository for stable isotope data, metadata controls provide an opportunity to shape the **community's language** going forward. Using workshops, the community works with users to define the vocabulary and optimize the form/template. By setting a controlled vocabulary based on user input, the community grows stronger. The more the vocabulary is used, the more powerful it becomes.

Finding Area Two: Creating Collaborations Between Information Professionals/Data Communities

Information professionals are an important resource

Most of the data communities who participated in our workshop were struggling with data management, curation, and organization. These are all areas where information professionals are well equipped to provide guidance, yet only a few research teams had significant histories of collaboration with informational professionals. This was not surprising. Though university libraries and other campus units are making significant investments in providing data services to researchers and now commonly include dedicated information professionals on staff, structured opportunities for information professionals and research teams to interact are still relatively rare.[10] The lack of opportunities for deep engagement contributes to the shallow and uneven ties between many research teams and information professionals: many teams in our group reported little previous awareness of what information professionals could offer them. One of the core goals of our workshop was to provide a forum for these groups to interact and share knowledge, and to test the hypothesis that collaboration across these professional lines can pay dividends in the form of increased efficiency of data sharing.[11]

[10] A.M. Cox, et al., "Developments in Research Data Management in Academic Libraries: Towards an Understanding of Research Data Service Maturity," *Journal of the Association for Information Science and Technology* 68: 2182-2200, https://doi.org/10.1002/asi.23781; Rebecca Springer, "Counting Data Librarians," *Ithaka S+R*, 29 July 29 2019, https://sr.ithaka.org/blog/counting-data-librarians/.

[11] There is some precedent for one-off workshops that bring together scientists and information and technology professionals to discuss data sharing in a particular discipline, and these have been successful in creating domain-specific recommendations and action items. See, Alison Callahan et al., "Developing a Data Sharing Community for Spinal Cord Injury Research," *Experimental*

ARDA, established in 1998, is one of the longest running data communities that participated in the workshop series. They have a well-developed repository with good discoverability and a large number of users across disciplines and sectors. Nonetheless, they invited the two information professionals who worked with them to join their board to ensure the future direction of the community is guided by information professional expertise.

Information professionals, who are trained in file formats and data enrichment and have professional obligations to keep abreast of data sharing trends and developments, are well positioned to assess the particular needs of a data community and offer tailored advice on specific actions the data community can take to adopt best practices for data sharing.[12] In short, input from information professionals can help expand the data community's user base, create, improve or identify the required infrastructure, such as online repositories and metadata schemas, establish a plan for long-term preservation, and help advocate or guide the group to organizational and financial sustainability.

Ideally, collaboration between researchers and information professionals should begin during the initial stages of research, but, as we found in the workshop series, information professionals can also provide valuable assistance to data communities who are already well-established. For example, a common theme that emerged from the workshop was that data communities are often hyper aware of their own data challenges, but unlikely to fully understand how those challenges relate to or are the same as the challenges of other data communities. This can result in data communities working to create solutions from scratch, unaware of examples, models, and templates that might ease their workload and promote interoperability. One immediate value information professionals provided to our workshop was their ability to see the larger picture and capacity to help researchers to identify and contextualize their data challenges.

Neurology 295 (2017):135-43, DOI: 10.1016/j.expneurol.2017.05.012; E. Demir et al., "The BioPAX Community Standard for Pathway Data Sharing," *Nature Biotechnology* 28 (2010): 935-42, DOI: 10.1038/nbt.1666; Stacey Harper et al., "Nanoinformatics Workshop Report: Current Resources, Community Needs and the Proposal of a Collaborative Framework for Data Sharing and Information Integration," *Computational Science & Discovery* 6 (2013): 014008, DOI: 10.1088/1749-4699/6/1/014008; Jean-Baptiste Poline, "Data Sharing in Neuroimaging Research," *Frontiers in Neuroinformatics* 6 (2012), DOI: 10.3389/fninf.2012.00009.

[12] Jake Carlson and Mikala R Narlock, "Life Rafts in a Sea of Data: The Role of Librarians in Supporting Data Sharing," Retrieved from the University of Minnesota Digital Conservancy, 6 April 2022, https://hdl.handle.net/11299/226887.

CREWS, an interdisciplinary community that works across six higher-ed institutions as well as with government and business partners, was one of the few participating data communities with a dedicated information professional as a core part of their team. During one breakout session, CREWS talked about a pipeline structure they created that helps to curate researchers' data for final use. The CREWS pipeline model intrigued another community, also working across multiple institutions and struggling with coordinating protocols. This brief, targeted conversation between data communities from different disciplines introduced a roadmap for actionable steps to address a shared problem.

Cross-disciplinary conversations can help solve data sharing challenges

Contemporary research, particularly in STEM fields, is commonly—perhaps even usually—conducted across disciplinary lines, a trend that reflects the complexity of the urgent questions researchers are working to resolve. However, the ecosystem that supports research is still often siloed into narrower domains that hinder sustained, deep, communication between researchers from radically different disciplinary traditions.[13] This is especially reflected in how difficult it can be for researchers to use support and services they receive from their home institutions to conduct cross-institutional research projects. The inability to fluidly use funds or support across institutions creates a bureaucratic obstacle that often reinforces research siloes, and these siloes can limit data sharing in several ways. They can be mitigated, however, by building data sharing communities around common problems and shared interests rather than around disciplinary or institutional borders.[14] Even many data communities, however, are likely to be used by researchers from a relatively small group of aligned disciplines. This is an asset insofar as it provides a common purpose and identity necessary for the community to succeed. However, our workshop clearly demonstrated the value of bringing together researchers from widely disparate backgrounds for discussions about data sharing because it encouraged researchers to consider technical and social aspects of data sharing without getting bogged down in minutiae, looking for patterns and common ground. One key insight shared by several participants was that they discovered that data challenges that had seemed highly specific to their domain—challenges of metadata or ontology, for example—were in fact common to data sharing efforts regardless of discipline. Recognizing common data management, curation, and sharing challenges encouraged them to consider a wider range of examples and models for making progress on

[13] Andrew Freiband et al., "Undisciplining the University Through Shared Purpose, Practice, and Place, *Humanities and Social Sciences Communications* 9, 172 (2022), https://doi.org/10.1057/s41599-022-01195-4; Julie Thompson Klein, *Beyond Interdisciplinarity: Boundary Work, Communication, and Collaboration* (Oxford University Press: 2021) https://doi.org/10.1093/oso/9780197571149.001.0001.

[14] Yi Shen, "Data Sharing Practices, Information Exchange Behaviors, and Knowledge Discovery Dynamics: A Study of Natural Resources and Environmental Scientists," *Environmental Systems Research* 6, no. 9 (2017), DOI: 10.1186/s40068-017-0086-5.

their issues, and to consider how information professionals might contribute substantial value to their data sharing goals.

Finding Area Three: Informing the NSF's Public Access Repository

Domain specific repositories are better equipped to serve as community hubs than generalist repositories

The current system of data repositories is highly decentralized and diverse in terms of size and scope.[15] Large generalist repositories like Dryad, Figshare, Zenodo, and others have become important parts of the data sharing ecosystem. Institutional repositories, now hosted by most research universities, represent another form of generalist repositories. Generalist repositories have certain advantages—for example they can accept highly heterogeneous data sets, host data in emerging or narrow fields that can not support a domain repository. However, federal agencies such as the NSF and NIH prefer that researchers deposit to established domain repositories rather than general repositories when possible. Both agencies have provided numerous grants to establish new repositories when existing repositories do not serve a particular research communities' needs, which has contributed to the growth of a large number of specialized repositories.

A data communities framework reinforces the importance of domain repositories. Participants in our workshop generally preferred to deposit research data in repositories specific to their community and their needs. Researchers judged these repositories better able to support individualized metadata, ontologies, and file formats germaine to their specific research communities needs than generalist repositories. Moreover, specialized repositories appear better suited for serving as a community hub. Data sharing is a social as well as a technical phenomenon—and one clear advantage of specialized repositories is that they are better equipped than generalist repositories to serve as hubs for research communities. Researchers identify with disciplines and research problems first and are most motivated to share data with researchers who share their interests and with those who they can identify as members of their community. Data communities looking to grow their user base will benefit from leveraging the identification scholars form within their field as a way of encouraging engagement in data sharing.

[15] R.R. Downs, "Improving Opportunities for New Value of Open Data: Assessing and Certifying Research Data Repositories," *Data Science Journal* 20, no. 1: 1, http://doi.org/10.5334/dsj-2021-001.

Maple River Dam, a community comprised of diverse stakeholders (government, state biologists, Michigan DNR, USGS, nonprofit agencies, private landowners) is working storage solutions for their data. They have a unique data set, with a long-time frame and a wide taxonomic range, that they want to share but are stalled by number of challenges, not least among them trying to organize a backlog of data. They share some through a local repository at the bio-station which then gets pushed to larger repositories like DataOne. This can only work because it is **domain specific.** When they move to site-based rather than domain-based data, it is more difficult to conceptualize how the data will get shared.

The NSF's Public Access Repository (PAR) is a central repository where NSF-funded investigators are required to deposit peer-reviewed, published journal articles and juried conference papers. Data deposit to the PAR is not required, but the NSF is interested in encouraging the site's use as a directory of datasets that can provide centralized discovery for datasets held across the repository ecosystem. To meet this goal, the NSF is invested in understanding how to encourage data communities to prepare to share such metadata in the PAR. However, the factors that lead researchers to prefer decentralized, domain repositories, also limit the interest of data communities in contributing metadata to the PAR. Researchers understand the value of a centralized metadata repository as a means for enhancing discoverability and accessibility across fields. However, they generally showed little interest in investing time in contributing to it: their primary motivation for sharing was to ensure that researchers within their perceived community had access to data with which they had a common interest. Planning for the future of a nation-wide scientific infrastructure, such as the NSF's PAR and the NIH's Generalist Repository Ecosystem Initiative (GREI) will lead to advances in cross platform discovery but can exist in tension with these more immediate needs. For many researchers, the PAR was simply too far removed from the communities that they identified with to incentivize engagement.

Decentralized repositories and discoverability challenges

Decentralized repositories can tailor their offerings to specific data communities and acquire reputations in their field that promote their visibility. For example, in a recent Ithaka S+R study of data-intensive research practices across dozens of fields, many researchers reported coming to know—through word of mouth, citations, and informal exposure—which repositories were most relevant to their field.[16] However, the same forces that can help repositories become well known to their immediate audiences can also contribute to ongoing problems with the siloing of

[16] Dylan Ruediger and Danielle Cooper, "Big Data Infrastructure at the Crossroads: Support Needs and Challenges for Universities," *Ithaka S+R*, 1 December 2021, https://doi.org/10.18665/sr.316121.

scientific knowledge, a serious problem in an era when the most urgent scientific questions cut across fields and require convergent frameworks.[17]

Creating mechanisms for cross-platform discovery is essential to integrating scholarly research data from across fields. Progress on this front will require creating comprehensive methods of linking together the domain repositories that are most favorable to intra-community use with generalist repositories that are better equipped to promote inter-community exchange. Information professionals, who are more accustomed than researchers to thinking about the organization of data at a high level of abstraction for connecting with broader audiences, may also prove a fruitful source of advice to PAR and GREI and other efforts to advance these goals.

Challenge One: Testing the Concept of Data Communities

Aligning funding structures with sustainability

Research funding structures, largely organized around grants, present a significant challenge to the long-term sustainability of data communities. Grants typically cover costs associated with activities only during the grant's award period, a model that does not account for long term costs associated with continued stewardship of data. Perhaps the most important of these are the staff necessary to ensure the sustainability and vitality of a data community, but the long-term costs associated with archiving and maintaining data are also major challenges.

Incompatible Funding: A participating data community received funding to run workshops that teach people how to use their repository, but they did not have funding for full time staff who could be responsible for managing the repository. While waiting to hear about funding for staffing the repository, they were in a standstill for planning their next steps.

While the value of data sharing has been accepted by many scholars, funding structures often treat data sharing efforts as start-up ventures and restrict or prohibit funding necessary to sustain data communities as they develop into service and support organizations. Building sustainable funding based on diverse revenue streams remains a substantial challenge, though some mature data communities are experimenting with promising models such as membership fees, donations, and offering consulting and training services.

[17] R. D. Kush et al., "FAIR Data Sharing: The Roles of Common Data Elements and Harmonization," *Journal of Biomedical Informatics* 107 (1 July 2020): 103421, https://doi.org/10.1016/j.jbi.2020.103421.

International collaborations characterize much contemporary scientific research, yet the funding structures that support researchers are often national agencies that prohibit or complicate multinational funding. This challenge seems likely to grow as the scope and scale of research becomes globalized more quickly than do the funding structures necessary to support them.

Expanding users beyond the initial core

Some successful data communities rely on small, but highly active user bases. However, a common challenge facing maturing communities is how to expand beyond their initial user base, a step that is often necessary in order to avoid stalling out after initial periods of success. Often, the immediate challenge is understanding who their audience is and where to target their recruitment. Some participants struggled with generational differences, specifically a lack of interest in data sharing among senior scholars whose authority in their fields can limit momentum toward a cultural consensus on the importance of data sharing. Here again, information professionals are well-positioned to assist data communities in finding solutions. While they may not be able to create an entire cultural shift, information professionals do understand the broader landscape of data sharing outside of each specific data community, and they can help to connect data communities and researchers who may not otherwise know how to find each other. These challenges warrant future research on the life cycle of data communities that focuses on how the size and composition of data communities change over time and at different stages of maturity.

Incorporating new users

Successfully attracting new members can pose a different set of challenges for data communities. It is difficult for all communities, and especially those who are not well-resourced, to provide training, and to some degree socialization, for new members in the norms and workflows of the data community. As a result, new users often do not understand the importance of following the community's data management structure and tend to dump data, submit data incorrectly, or not follow the metadata schema. An influx of improperly structured data can overwhelm a data community, especially given the chronic lack of staff who can help new users learn and follow the community guidelines.

PGC developed a uniquely strong community of people to do the work necessary to support sharing their data. They curate and add value to the data before it gets to researchers, which is **labor intensive work** that facilitates re-use, but satisfying requests for data is an ongoing struggle.

Challenge Two: Creating Collaborations Between Information Professionals and Data Communities

Projects are multi-institutional, but staff are employed by one institution

Contemporary scientific research is often conducted by research teams that are distributed across several institutions and, in many cases, countries. However, information professionals are usually employed by, and expected to support and work on a single campus and are obligated to provide services to many different researchers and students simultaneously. This makes it challenging for them to invest in the kinds of long-term collaborations with researchers that, from a data management and curation perspective, would best serve the interests and needs of data communities.

Data support is labor intensive and difficult to scale

Robust data sharing communities often require extensive support offered to researchers, particularly as those communities grow beyond their initial users to include those who may be less familiar with that often complex and heterogeneous datasets that data communities typically share. This can lead to a paradox: a lack of staffing can hinder growth, and a lack of growth can make it difficult to hire the staff necessary to scale up the service offering necessary to attract new users.

Resisting the urge to overbuild

As data communities develop their data sharing infrastructure, they need to balance their existing capacity with future needs. One lesson from our workshop was that anticipating future needs—for example, building in extreme flexibility about which file formats a repository will support, or the quest to build comprehensive metadata, can inhibit the ability of a community to accomplish more reasonable and immediate goals. When building repositories or other infrastructure, communities would be advised to focus attention on more modest goals of providing a workable repository that meets the most important needs of the current data community. Data communities of all stages, but especially emergent communities, have real limits on the resources at their disposal and should build accordingly. However, this humility of purpose will necessarily exist in tension with the need to anticipate future needs that will fuel and sustain future community growth.

Center for Health Equity Research/Duke Clinical Research Institute developed comprehensive common data elements, but new participants do not readily adopt them. In an ongoing **process of conversion** rather than an immediate uptake, existing community members need to make appeals to new participants as to why the common data elements matter by emphasizing what can be done with them.

Challenge Three: Informing the NSF's Public Access Repository

Connecting domain and generalist repositories

Our workshop suggests that the PAR faces significant marketing challenges. Researchers recognize the value of a cross-cutting discovery tool but show little interest in doing extra work to support its development.[18] They often believe that they, and other researchers in their immediate field, already know where to locate relevant data. To become capable of linking repositories through metadata deposits, the NSF will need to invest significant resources into demonstrating that the utility of the PAR to researchers warrants significant investments of their time and will serve the needs of the communities with which they identify. The NSF will also need to continue efforts to encourage data communities to think beyond their members and issues and act in the interests of broader infrastructure goals. These efforts may need to be supplemented by mandates that require NSF-funded researchers to deposit metadata in the PAR, and investment in tools to monitor compliance.[19] Partnerships with generalist repositories, perhaps modeled on the NIH's GREI initiative, may be worth considering as options for outreach to researchers and models for promoting cross-platform discovery.

Aligning research cultures and incentive structures

While massive investments by federal agencies and other funders have created significant improvements in the infrastructure required for sharing, fewer gains have been made in the necessary step of aligning the culture of researchers in many disciplinary communities, and the incentive structures within which they operate, to the goals of open science.

[18] R. D. Kush et al., "FAIR Data Sharing: The Roles of Common Data Elements and Harmonization," *Journal of Biomedical Informatics* 107 (1 July 2020): 103421, https://doi.org/10.1016/j.jbi.2020.103421.

[19] On the importance of compliance mechanisms, see Jessica L. Couture et al., "A Funder-Imposed Data Publication Requirement Seldom Inspired Data Sharing," *PLOS ONE* 13, no. 7 (6 July 2018): e0199789, https://doi.org/10.1371/journal.pone.0199789.

The Materials Commons, a data community and domain-specific repository with a big data capacity, pioneered a streamlined approach to documenting and ingesting data into their repository that provides a good view of connections, links, and meaning across the contents of the deposit. The Materials Commons strongly suggests metadata that allows for easier exploration and discovery, but getting users to change their workflow to capture metadata is an ongoing struggle. Also, despite being integrated with Google Dataset Search, the Materials Commons faces challenges with **visibility and discoverability** because researchers face a range of repository options. Ensuring people are aware of the benefits of depositing, despite the initial work to properly curate the data, is an important and challenging step.

Data communities can play a powerful role in fomenting cultural change, since they focus attention on voluntary sharing of data, and thus on the ways that individuals and communities can create their own incentives for sharing with colleagues. However, most of the data communities in our workshop struggled with the uphill battle of academic incentive structures that continue to undermine data sharing efforts. Until promotion and tenure standards, financial models, and cultural prestige reflect the value of data sharing to the entire scientific community, efforts to promote data sharing will remain limited.

Recommendations

Data Community Organizers

Emergent

- Build ties with information professionals early in the process of establishing a data community, seeking expert advice about basic design elements including metadata, interoperability, and controlled vocabularies.
- Consider ways to maximize interoperability when defining metadata.
- Seek opportunities to fund longer-term staffing of information professionals into grant budgets and consider including information and data professionals as collaborators and advisory committees.
- Integrate marketing, communication and outreach strategies into the design process.
- Focus repository and other community infrastructure on immediate community needs that will encourage creation of an active user base.
- Implement improvements to repositories that comply with desirable characteristics of data repositories for federally-funded research, as identified by the National Science and Technology Council.

Established

- Provide training or tutorials to new members about workflows and community norms. When possible, conduct these trainings in group settings, to encourage community building.
- Invite community members to contribute to the process of refining metadata and ontologies.
- Invest time in planning for diverse revenue streams, cost modeling, and business models capable of providing sustainable long-term funding. Seek relevant expertise to maximize the effectiveness of these efforts.
- Build governance frameworks and organizational structures that maximize community participation in decision making around standards, oversight, and preservation.
- Weigh the impacts of expanding the scope and format of data deposits on the mission, values, and needs of your community.

Mature

- Develop long-term preservation strategies that will ensure continued access to research data should the original community host cease operations.
- Consider expanding scope to incorporate new community members and users.

Information Professionals and Data Service Providers

- Customize data support services to researchers needs and motives: in particular, consider that voluntary data sharing efforts associated with the data communities framework can point to different service models than those associated with mandatory data sharing requirements.
- Cultivate the competencies to support data communities across their life cycle, including the capacity to advise about business models and community building as vital components of successful, sustainable data sharing
- Participate in and continue building robust professionalization and networking organizations.
- Seek opportunities to engage with researchers on-campus or at other institutions working in your area of expertise.

Universities

- Develop staffing models at libraries and other key units that recognize and support information professionals' ability to make long-term contributions to cross-institutional research projects.
- Host forums for information professionals and researchers to meet and share knowledge.
- Invest in supporting professional organizations such as the Data Curation Network, Research Data Alliance, Research Data Access and Preservation Association, and the Campus Research Computing Consortium.
- Create user-friendly guides to campus data support services available on campus and increase coordinated outreach to research communities.

- Align promotion and tenure and other internal incentive structures with the now central role data curation, management, and sharing play in the production of knowledge.
- Expand existing data storage capacities available to researchers to support the long-term need to preserve and share research data.

Funders

- Provide grant funding for activities designed to identify/build a diversified, sustainable revenue or cost recovery model for data communities and associated repositories.
- Prioritize initiatives that connect data services providers with research communities and/or provide infrastructure solutions applicable to multiple data communities for funding.
- Encourage incorporation of community-building exercises and plans as an optional or required component of grant applications for researchers interested in building domain repositories.
- Invest in further research on social and cultural dynamics that support voluntary data sharing among researchers.
- Increase investments in multi-disciplinary and cross platform discovery tools to connect decentralized repository networks.
- Fund research into data communities that have created sustainable funding models and make resources available to grantees that assist them in planning for cost recovery, maintenance, and expansion after grant funding ends.

NSF PAR

- Articulate a more cohesive set of goals for the PAR and develop outreach and communication strategies to ensure research communities are aware of the value and role of the PAR.
- Consider requiring funded researchers to submit metadata to PAR as a condition of grant funding and develop procedures to monitor compliance.
- Explore automated mechanisms to render domain repositories in fields funded by the NSF discoverable in the PAR.

Generalist Repositories

- Coordinate with relevant agencies to develop infrastructure to share metadata across platforms.
- Develop programming and tools to support community-building activities by researchers seeking to deposit research data and organize data communities within generalist repositories.

Appendix 1

Call for Proposals: Leveraging Data Communities to Advance Open Science

Ithaka S+R welcomes applications from interested researchers to participate in *Leveraging Data Communities to Advance Open Science*, an NSF-funded workshop, developed in partnership with the Data Curation Network. Participants will gain expertise in best practices for facilitating data sharing across disciplinary and institutional boundaries, including developing shared metadata conventions and standardization of existing file formats or the creation of new formats for emerging data types (e.g., 3D data), building or improving repositories and other sharing infrastructures, and developing schemas that promote interoperability and machine readability. This workshop will support the development of infrastructures for data sharing within interdisciplinary data communities, informal or formal groups of scholars from STEM fields who voluntarily share data to foster scientific progress on a topic of mutual interest.

Building sustainable data communities requires a significant technical infrastructure that is best created through collaboration between scientists and information technology professionals. Unfortunately, these groups are often bifurcated by differing professional identities, and have relatively few opportunities for sustained dialog across disciplinary and professional perspectives, and institutional or geopolitical borders. The *Leveraging Data Communities to Advance Open Science* workshop will provide a rare forum for collaboration between information professionals and scientific researchers.

Successful applicants will receive one-on-one access to an information professional with specialized expertise who will help participants identify individualized solutions to the data sharing challenges they face during a remote meeting in fall 2021. They will also receive funding to participate in an NSF-sponsored two-day workshop in the spring of 2022. Together, the remote meeting and in person workshop will serve as incubators by providing a forum for focused dialogue between scientists and information professionals about how to combine their expertise to foster robust cultures of data sharing and data reuse within scientific communities. These sessions will also help generate recommendations for key metadata fields in the National Science Foundation's Public Access Repository (PAR).

Who is eligible to apply?

We welcome applications from teams of researchers who are involved in existing data communities, or who are interested in establishing one. Data communities are formal or informal groups of scholars who voluntarily share data across disciplinary and institutional boundaries for collective benefit and the advancement of science.

Project teams should include from two to five members. The composition of the team should reflect the diversity of roles required to successfully share data and thus include a combination of tenured- or tenure-track faculty, non-tenure-researchers, and postdocs, as well as information professionals, lab managers, or data management staff. We are particularly interested in teams that include individuals from multiple institutions and researchers from different disciplinary backgrounds.

What will participants do?

All members of the project team will be paired with an information professional for a mandatory virtual meeting, to be held in the fall of 2021. In addition, two members of the project team will attend the two-day workshop on February 28 and March 1, 2022, at the University of Michigan, Ann Arbor.

What support will Ithaka S+R and the Data Curation Network provide?

Ithaka S+R and the Data Curation Network will facilitate learning opportunities within the project cohort and pair each project team with an expert information professional who will provide participants with individualized advice about data management and sharing challenges they face. Project teams will receive full funding for two team members to attend the Spring 2022 workshop.

When are applications due?

Applications will be due on Sept 1, 2021. To apply, submit the following information as a single PDF or word document to Dylan.Ruediger@ithaka.org.

What do I need to do to apply?

Interested parties should submit the following information:

1. List the names, email addresses, institutional affiliation, and job title of all members of your project, and indicate who will serve as the primary contact. *Project teams must be composed of 2-5 individuals.*
2. What kinds of data does your data community wish to/currently share? (Approximately 100 words)
3. Provide a brief narrative describing the data community your team represents or would like to create. Please describe your data community's goals and methods for data sharing. If you have a web page or existing data repository, please include a link. (Approximately 250 words)

> We welcome applicants from established and emerging data communities, as well as from teams looking to create a new data community. Please be open about the status of your community, as our goal is to include communities at different stages of organization.

4. Provide a brief description of the main challenges your data community faces and what you hope to learn from participating in *Leveraging Data Communities*. (Approximately 250 words)
5. A brief description of the efforts your data community has taken or anticipates taking to demonstrate a commitment to encouraging racial, ethnic, gender, social, and other forms of diversity within your team and/or data community. (Approximately 200 words)

For questions about this workshop or further conversation about whether you are part of a data community, please contact Dylan Ruediger (Dylan.Ruediger@ithaka.org).

"Leveraging Data Communities to Advance Open Science" is supported by the National Science Foundation under Grant No. 2013433.

Appendix 2: Participant List

Participating Data Communities

American Society of Agronomy

Members of the American Society of Agronomy (ASA) conduct research at the interface of agriculture and the environment. Our data are diverse including measures of agronomic and environmental performance of crop production systems. Treatments often explore crop or soil management strategies including variation in fertilizer applications, tillage, crop varieties, irrigation, plant populations, etc. Most data sets include temporal and spatial variation. As a team representing academic publishers, the ASA team wants to develop guidance for how to prepare authors for data publication in FAIR formats.

Team Members: Sylvie Brouder, Purdue University; Jeff Volenec, Purdue University; Matt Wascavage, American Society of Agronomy; Kathy Yeater, American Society of Agronomy

Information Professional: Leslie Delserone, University of Nebraska

Association of Religion Data Archives (ARDA)

ARDA archives and disseminates more than 1,000 data collections that each include at least a few measures on religion. The vast majority are surveys of individuals, but also included are collections using nations, organizations, and other groups as the unit of analysis and collections relying on many different research designs. All of the collections are supported with extensive metadata. ARDA also supports databases used for historical timelines, family trees of religions, national constitutional clauses on religion, local GIS reports, and other online tools. The ARDA team is looking to build user and depositor communities in new disciplines.

Team Members: Roger Finke, Penn State University; Gail Ulmer, Penn State University

Information Professional: Renata Curty, University of California, Santa Barbara

Center for Applied Internet Data Analysis (CAIDA)

CAIDA collects, processes, curates and shares Internet measurements data. This includes (1) Internet topology data, (2) Internet traffic data, (3) Internet security-related measurements. In pursuit of the FAIR (findable, accessible, interoperable, reusable) principles of their scientific data infrastructure mission, CAIDA has invested significant effort to make their data sets and associated resources more accessible to other researchers. Their goals include safeguarding privacy of sensitive data and visibility challenges and measurement barriers associated with internet network data.

Team Members: KC Claffy, CAIDA; Bradley Huffaker, CAIDA; Alex Ma, CAIDA; Elena Yulaeva, CAIDA

Information Professional: Seth Erickson, Penn State University

Center for Health Equity Research/Duke Clinical Research Institute

The National Institutes of Health is fostering a multi-armed initiative known as the Rapid Acceleration of Diagnostics (RADx) for COVID-19. RADx-UP has funded over 80 research projects across the US, awardees of which report common data elements (CDEs) to the Coordination and Data Collection Center (CDCC), the working group from which these team members stem from. Data proposed are complex as they range in nature (i.e., qualitative and quantitative), source, and level of aggregation. Their goal is to expand the impact of RADx-UP data through an evolution from "centralized hub" into a contributing data community capable of supporting science and community members well into the future.

Team Members: Bhargav Srinivas Adagarla, Duke University; Karla Garcia-Rascon, University of North Carolina; Jayalalitha Krishnamurthy, Duke University; Michelle Song, University of North Carolina

Information Professional: Jen Darragh, Duke University

IsoBank

Stable isotope data have made pivotal contributions to nearly every discipline of the physical and natural sciences. While the pace of growth in the generation of isotopic data rivals that of genetics, the latter field is now driven by "omic"-based approaches that have produced ground-break discoveries based on publicly accessible centralized databases; the former has yet to reach this stage of development. Therefore, the IsoBank data community wishes to improve the manner in which stable isotope data are centralized and shared across academic disciplines

Team Members: Anna Dabrowski, University of Texas; Oliver Shipley, University of New Mexico

Information Professional: Brian Westra, University of Iowa

Maple River Dam Removal Project

Maple River Dam Removal Project collects abiotic and biotic measurements made pre and post dam removal on the Maple River in Pellston, MI. These data include abiotic parameters of the river (temperature, ph, flow rate, chemical composition, sediment structure, etc.) and biotic communities of organisms including fish, macro invertebrates/insects, algae, and plants. These data, while collected by multiple PI's and researchers across multiple years and techniques, need to be harmonized as a comprehensive and cohesive data resource for inferring change in the Maple River.

Team Members: Pat Kocelick, University of Colorado; Paul A. Moore, Bowling Green State; Jason Tallant, University of Michigan; Amy Schrank, University of Minnesota; Karie Slavik, University of Michigan

Information Professional: Amanda Rinehart, Ohio State University

Materials Commons

The Materials Commons project (materialscommons.org) serves the materials science community. Data shared by the materials science community includes output from experimental instruments such as microscope images, input and output files from computational software, and data analysis files. The Materials Commons supports all types of files and is also moving into serving very large datasets, such as synchrotron high energy diffraction methods data. Their focus is on strategies for encouraging data sharing by minimizing roadblocks and maximizing the payoffs that come from making data findable, shareable, and easier to analyze.

Team Members: John Allison, University of Michigan; Tracy Berman, University of Michigan; Brian Puchala, University of Michigan; Glenn Tarcea, University of Michigan

Information Professional: Fernando Rios, University of Arizona

Montana CREWS Project

Montana CREWS Project shares water quality data and the code for processing it. Shared data are primarily non-proprietary, non-sensitive tabular format representing laboratory analyses of aquatic environmental samples, field measurements corresponding to those samples, and high-frequency measurements of environmental signals from deployments of state-of-the-art water quality sensors. Their goal is to explore ways to integrate standardized metadata into the data collection process and improve data collection workflows.

Team Members: Venice Bayrd, Montana State University; Toby Koffman, Montana State University; Robert A. Payn, Montana State University; Qipei Shangguan, University of Montana; Claire Utzman, University of Montana

Information Professional: Jordan Wrigley, University of Colorado

Natural History Data

Natural history research and museum communities collect, archive, and disseminate digital 3D data derived from physical museum specimens. Specimens may be fossils or preserved modern animals and plants. Digital 3D data may be collected by surface scanning, photogrammetry, magnetic resonance imaging, or X-ray computed tomography. Natural History Data hopes to improve the value of these resources by developing solutions to challenges related to provenance, authenticity, and discoverability, sustainability, and preservation and intellectual property management.

Team Members: Doug Boyer, Duke University; Nelson Rios, Yale Peabody Museum; Adam Rountrey, University of Michigan; Cody Thompson, University of Michigan; University of Michigan; Kate Webbink, Field Museum

Information Professional: Xuying Xin, Penn State University

Phosphorus Sustainability

Phosphorus Sustainability is focused on advancing phosphorus sustainability through materials informatics, drawing on a broad range of contributing disciplines. It is a diverse and inclusive community, built around open-science and open-data models, that includes over thirty research groups and nine institutions and will grow over time. It plans to establish an internal data repository at NC State through which community participants will share and access data for all research projects.

Team Members: Rada Chirkova, North Carolina State University; Stevan Earl, Arizona State University; Alexey Gulyuk, North Carolina State University; Becca Muenich, Arizona State University; Kara Schatz, North Carolina State University

Information Professional: Sarah Wright, Cornell University

Play and Learning Across a Year Project

The PLAY consortium publicly shares video documentation of every aspect of the protocol and annotation scheme and of preliminary planning sessions. Raw data are shared with the PLAY consortium on Databrary: video data of one hour of natural infant-mother activity in the home; video home tours with room measurements; decibel recordings; family demographics; geocodes of homes; parent report data (digital questionnaires and video); and video annotation files (scored according to standardized protocols using the open-source video-coding software, Datavyu). They are looking for guidance on metadata and data schemes user for diverse data formats and types, and better discoverability.

Team Members: Karen Adolph, New York University; Rick Gilmore, Penn State University; Kasey Soska, Play & Learning Across a Year (PLAY) Project

Information Professional: Wind Cowles, Princeton University

Polar Geospatial Center (PGC)

PGC provides access to polar data through its website (https://www.pgc.umn.edu/data/) where it shares air photos, satellite imagery, maps, and elevation models. It currently distributes the publicly available data to our users via HTTP, web-based applications, and partnerships with ESRI, Google, and NASA. PGC's resources include large datasets (ArcticDEM and REMA) in Google Earth Engine. The PGC team is looking to build its user base to new communities and improve workflows as they build out a new web application to expand search capability.

Team Members: Mike Cloutier, University of Minnesota; Cole Kelleher, University of Minnesota; Charles Nguyen, University of Minnesota; Claire Porter, University of Minnesota

Information Professional: Reina Chano Murray, Johns Hopkins University

Radiopharmaceutical Therapy Data (RTD)

RTD works with medical imaging data relevant to performing patient specific radiation dosimetry calculation in Radiopharmaceutical Therapy (RPT). These include both clinically acquired patient images as well as realistic simulations (virtual patients) typically generated by Monte Carlo simulation of the imaging system. The imaging modalities relevant to our work are Nuclear Medicine functional imaging such as SPECT and PET and anatomical imaging such as CT and MRI. Both anatomical and functional images are required to perform highly patient specific dosimetry calculations. RTD's major challenges include privacy issues and providing data in formats useful to users.

Team Members: Yuni Dewarajara, University of Michigan; Carlos Uribe, BC Cancer; Benjamin Van, University of Michigan

Information Professional: Amy Nurnberger, MIT

The Artificial Intelligence Task Force of the Society of Nuclear Medicine and Molecular Imaging (SNMMI AI Task Force)

The Artificial Intelligence Task Force of the Society of Nuclear Medicine and Molecular Imaging (SNMMI) is composed of physicists, data scientists, and physicians interested in enhancing research on how artificial intelligence, machine-learning, and deep learning can be applied to nuclear medicine and molecular imaging. Currently nuclear medicine imaging is very much unrepresented in image datasets available for research, collaboration, and public access. The goal of their data community is to establish a collection of well annotated scans with pertinent clinical and outcome data.

Team Members: Tyler Bradshaw, University of Wisconsin; Joyita Dutta, Harvard Medical School; Arman Rhamim, University of British Columbia; Babak Saboury, National Institute of Health Clinical Center; Eliot Siegel, University of Maryland

Information Professional: Helenmary Sheridan, University of Pittsburgh

Participating Information Professionals

- Wind Cowles, Director, Research Data and Open Scholarship, Princeton University
- Renata Curty, Research Data Specialist, University of California Santa Barbara
- Jen Darragh, Senior Research Data Management Consultant, Duke University
- Leslie Delserone, Science and Research Data Services Librarian, University of Nebraska
- Seth Erickson, Research Data Librarian- Social Sciences, Penn State University
- Reina Chano Murray, Geospatial Data Curator and Applications Administrator, Johns Hopkins University
- Amy Nurnberger, Program Head, Data Management Services, and Interim Department Head, Data and Specialized Services, MIT
- Amanda Rinehart, Life Sciences Librarian, Ohio State University
- Fernando Rios, Research Data Management Specialist, University of Arizona
- Helenmary Sheridan, Data Services Librarian, University of Pittsburgh
- Brian Westra, Data Services, University of Iowa
- Sarah Wright, Life Sciences Librarian for Research, Cornell University
- Jordan Wrigley, Teaching Assistant Professor, Data Librarian, University of Colorado
- Xuying Xin, Data Analyst, Penn State University

Appendix 3: Main Workshop Schedule

Leveraging Data Communities to Advance Open Science: An Incubation Workshop

Feb 28: Building Successful Data Sharing Communities (11 AM to 2:30 PM Eastern)

Zoom

11:00-11:15 - Welcome and overview of data communities
Danielle Cooper, Associate Director, Libraries, Museums and Scholarly Communication, Ithaka S+R

11:15-11:30 - The value of collaboration between information professionals and data communities
Jacob Carlson, Director of Deep Blue Repositories, University of Michigan

11:30-11:45 - The NSF and data sharing communities
Martin Halbert, Science Advisor for Public Access, National Science Foundation

11:45-12:15 - "Lunch"

- We will provide an Uber Eats coupon to participating team members

12:15-1:15 - Plenary Panel: Success factors for sustainable data communities
Danielle Cooper (moderator)
Nici Pfieffer, Chief Product Officer, Center for Open Science
Amy Pienta, Associate Research Scientist, ICPSR
Robert Guralnick, Curator of Biodiversity Informatics, Florida Museum of Natural History

Break

1:20-2:00 - Working Groups: Identifying success factors for your community

- What are the strengths of your existing data community?
- What would "success" look like for your data community?

2:00-2:30 - Synthesis (in plenary): What is success and how do you get there?

March 1: Unconference: Tackling Organizational Challenges (1 PM - 3:30 PM Eastern)
Zoom

1:00-1:15: Welcome and run of show

1:15-2:00: Small group session A: data and documentation groups

Breakout rooms 1 and 4: data size, formats, interoperability

Breakout rooms 2 and 5: metadata, discoverability

Breakout room 3 and 6: platforms, storage issues

2:00-2:30: Break

2:30-3:15: Small group Session B: infrastructure and policy groups

Breakout rooms 1 and 4: tools, automation, workflow

Breakout rooms 2 and 5: sensitive data, privacy, ethics

Breakout rooms 3 and 6: community engagement

3:15-3:30: Concluding remarks

This material is based upon work supported by the National Science Foundation under Grant No.2103433.