ITHAKA S+R

Data Communities: Data Sharing from the Ground Up

By: Danielle Cooper

COVID-19, the Limits of & Possibilities for Remote Research

What (Many) Researchers Can Do

- Conduct literature reviews
- Analyze pre-existing datasets
- Review/organize materials (e.g. lab notebooks)
- Connect remotely with collaborators and peers
- Share data and findings

What (Some) Researchers Cannot Do

- Run experiments with site-dependent equipment and/or subjects (e.g. animal specimens)
- Maintain specimens
- Visit field sites
- Engage with human subjects in-person

COVID-19 data community

Influenza Virus Genetics

GISAID (Global Initiative on Sharing All Influenza Data) is an interdisciplinary organization supporting a repository of genetic data and related projects

Understanding data communities can help us support sharing across institutional boundaries, mirroring how scientists actually work.

About Ithaka S+R

We **provide research and strategic guidance** to help the academic and cultural communities serve the public good and navigate economic, technological, and demographic change.

.

We are a U.S. based not-for profit that conducts a triennial survey of faculty complimented by qualitative deep dives into teaching and research. **ithaka.sr.org**

We do our **work in partnership with libraries, scholarly societies, and publishers** to understand the specific needs of researchers.

Published and upcoming issue briefs and blog posts synthesize research on topics like **data communities** and the **organization of research data services.**

.

Research Support Services Program

At Ithaka S+R we **study the practices of scholars by discipline** or thematic area using a unique collaborative, qualitative methodology.

Completed Studies:

- Agriculture (2017)
- Art History (2013)
- Asian Studies (2018)
- Chemistry (2014)
- Civil & Environmental Engineering (2018)
- History (2012)
- Indigenous Studies (2019)
- Language and Literature (2020)
- Public Health (2017)
- Religious Studies (2017)

Agenda

Identifying Data Communities
Exploring Emergent Data Communities
Supporting Data Communities

Identifying Data Communities

The Data Sharing Landscape

Institution Driven
Researchers across disciplines are encouraged to deposit their datasets in institutional repositories.

Collaborative groups are working to **share curation expertise** and make data sets discoverable across institutions.

Compliance Driven
Funders and publishers require researchers to deposit datasets when their articles are published.

Generalist repositories are targeting this type of data sharing.

Community Driven
Researchers form communities around the sharing and reuse of certain types of data.

Community-centric sharing usually takes place via **domain repositories**.

What makes data sharing work?

Our qualitative research shows that **data sharing is fundamentally social**. We have examined data sharing **success stories**, such as:

A **data community** is a formal or informal network of researchers who share and reuse a certain type of data.

Data sharing success stories

Genetics

- GenBank (NCBI)
- Species-specific databases such as FlyBase
- Viral genomic sequencing via GISAID

Neuroimaging

- Neuroimaging Tools and Resources Clearinghouse
- OpenNeuro
- Donders Repository

How can we find and map data sharing communities?

As the infrastructure for sharing and tracking data outputs grows so do the opportunities to identify data communities.

The FREYA Project explored mapping scholarly networks in novel ways using dataset citations through its **PID Graph** by leveraging **DataCite** event data.

A not-quite data community?

Economics

Reproducibility is important to the field and data sharing does occur on a small scale. Access to code is just as important as the underlying data.

Sharing data from private entities and regulatory bodies is often heavily restricted.

Characteristics of successful data sharing communities

1. Bottom-up development
2. Community norms
3. Absence of mitigation of technical barriers

Exploring Emergent Data Communities

An **emergent data community** is a group of scholars who are enthusiastic about sharing and reusing a certain type of data, but haven't yet fully established the necessary processes and infrastructure.

Emergent data community

Air pollution research

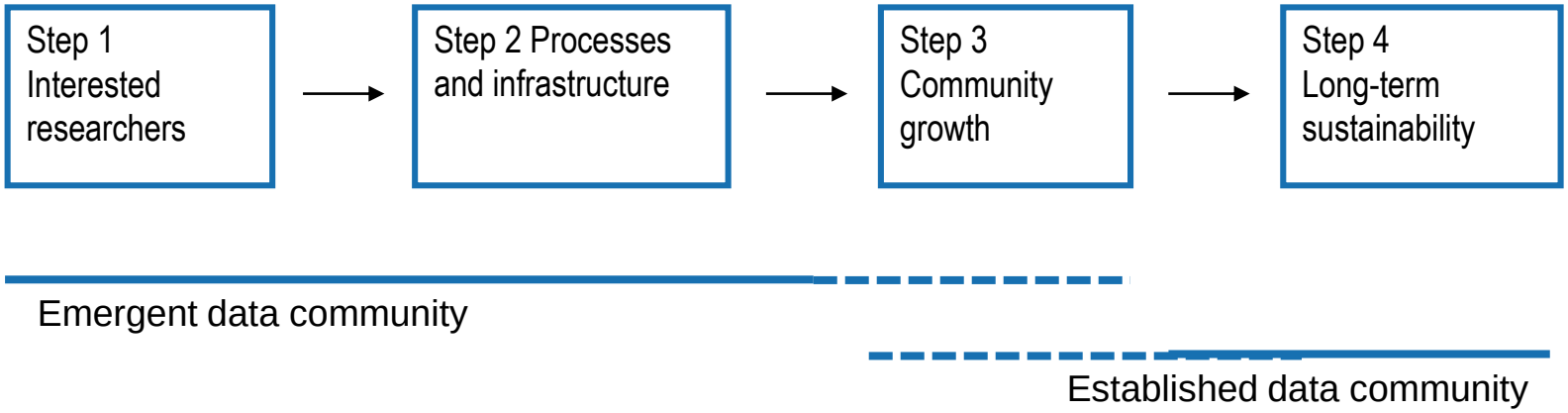
Ithaka S+R research suggests that some environmental engineers who work on air pollution are eager to **share and reuse air quality data,** but don’t have good ways to do so

Emergent data community

Spinal cord injury research data

Researchers are working to facilitate data sharing among scientists working in the multidisciplinary field of spinal cord injury (SCI) research.

The data community growth process



Supporting Data Communities

Ways Forward

- Data sharing can help overcome barriers to data collection in some research communities.
- The strategies of successful data communities provide examples of how to support data sharing effectively.
- Identifying and supporting emergent data communities is a strategy for leveraging technology to improve research workflows through data sharing.

What do data communities need?

- Help building or identifying existing repository infrastructure
- Technical and policy advice on metadata, vocabularies, preservation, privacy, etc.
- Guidance and advocacy for achieving organizational and financial sustainability
- Help getting the word out to researchers who might be interested in getting involved

RDA COVID-19 working group

- International working group of librarians and other research data management experts, formed in response to COVID-19
- Has released a comprehensive set of recommendations for sharing COVID-19 research data
- Has created a Zotero bibliography of COVID-19-related resources

Implications for data sharing support

- "Build it and they will come" approaches generally don't result in data communities
- Top-down mandates generally don't result in data communities
- Institutional and generalist repositories can provide infrastructure and curation support
- Librarians with institutional support remits have the challenge of supporting cross-institutional data communities

Upcoming Ithaka S+R Projects

Study: Supporting Big Data Research (2020-2021)

Ithaka S+R is working with with a cohort of 21 U.S. academic libraries to investigate the evolving research activities and support needs of scholars who work with big data across a variety of fields and methods.

Assessment: Data Service Organization (forthcoming)

How can scholars in data communities be best supported? We created a method to track how universities organize their data services and evaluated the U.S. landscape. We will be expanding this to evaluate the broader landscape in Canada, Europe, the UK, Australia and beyond. We welcome expressions of interest from institutions, organizations and individuals.

ITHAKA S+R

Further Reading

Issue Briefs

Data Communities: A New Model for Supporting STEM Data Sharing

Research Data Services in US Higher Education

Blog Posts

Emergent Data Community Spotlight Series

Contact Me

Danielle M. Cooper

Manager, Collaborations and Research

@dm_cooper

Danielle.Cooper@ithaka.org

ITHAKA S+R

Thank you