IJDC | *General Article*

The Role of Data in an Emerging Research Community: Environmental Health Research as an Exemplar

Danielle Pollock
Simmons University

An Yan
University of Washington

Michelle Parker
University of Tennessee, Knoxville

Suzie Allard
University of Tennessee, Knoxville

Abstract

Open science data benefit society by facilitating convergence across domains that are examining the same scientific problem. While cross-disciplinary data sharing and reuse is essential to the research done by convergent communities, so far little is known about the role data play in how these communities interact. An understanding of the role of data in these collaborations can help us identify and meet the needs of emerging research communities which may predict the next challenges faced by science. This paper represents an exploratory study of one emerging community, the environmental health community, examining how environmental health research groups form, collaborate, and share data. Five key insights about the role of data in emerging research communities are identified and suggestions are made for further research.

Submitted 4 Feb 2019 ~ *Revision received* 27 April 2022 ~ *Accepted* 5 May 2022

Danielle Pollock, Simmons University, 300 The Fenway, Boston, MA 02115. Email: danielle.pollock@simmons.edu

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

Introduction

Data play a crucial role in scientific research and discovery (Borgman, 2015). In the past researchers were often limited to using data collected either individually or within their research team; technical advances and the move to open sharing of scientific data have enabled researchers to move beyond this approach, not only by facilitating new forms of data gathering, but also by facilitating the discovery and reuse of data gathered by others--including others working outside one's own domain--in new scientific research (British Academy & The Royal Society, 2017; Parsons et al., 2011).

A key element of domains working together is the ability to access data from one another. The information challenges and scientific opportunities presented by expanding availability of scientific data at both large and small scale are numerous (Borgman, 2015; Hey, Tansley, and Tolle, 2009). Sharing and using datasets across domains, while often essential to research involving the convergence of scientific disciplines, can be especially challenging (Parsons et al., 2011; Sá & Grieco, 2016). This research studies the role of data in the "converging" environmental health community by studying these researchers' use of data stored in disciplinary-based repositories outside their own discipline. Following the trail left by the data provides a snapshot of how these interdisciplinary research groups formed, collaborated, and shared data. Understanding the role of data in the formation and work of these teams can help us identify and meet the needs of emerging research communities which may predict the next challenges faced by science. The use of datasets as a starting point for understanding data use in convergence research is a unique approach that adds to our understanding of researchers and provides insights into how the study of dataset usage could help identify emerging areas of science and the emerging communities that are forming to study these areas.

The multitude of data challenges researchers face include managing, sharing, discovering, and reusing data, and these challenges can be particularly notable when it comes to sharing or reusing data beyond one's own discipline (Borgman, 2015; Tenopir et al., 2011; Tenopir et al., 2015; Pasquetto, Randles, & Borgman, 2017; Sá & Grieco, 2016). Making data available supports the foundation of science—the reproducibility of scientific results. It can also lead to entirely new ways of producing scientific research including combining data from different fields to produce new discoveries, insights and conclusions which may lead to new streams of research (Hey et al., 2009; OECD, 2015; Sharp, Hockfield, & Jacks, 2016).

The National Science Foundation (2017) characterizes convergence as "the deep integration of knowledge, techniques, and expertise from multiple fields to form new and expanded frameworks for addressing scientific and societal challenges and opportunities." Convergence research is problem-driven and has much in common with concepts of interdisciplinary, multidisciplinary or transdisciplinary research in that it is highly collaborative, bringing researchers from multiple disciplines and their methods, tools, and data together to approach a single compelling scientific problem or challenge (National Research Council, 2014). Convergence research, however, goes beyond these in that knowledge and methods are not just shared across disciplines, but deeply integrated and unified in ways that may lead to the emergence of new research paradigms and new research communities (National Science Foundation, 2017; Sharp et al., 2016).

This convergence of domains and new streams of research result in a phenomenon that we term emerging research communities. Emerging research communities are those that have begun to converge around new areas of science and new scientific challenges. They are characterized by a high degree of collaboration and by the formation of professional connections and networks across traditional disciplinary boundaries. This paper reports on research conducted to identify the information behaviors exhibited by an emerging research community when interacting with open scientific data.

Statement of the Problem

An extensive literature review found that little is known about the role of data in emerging research communities or how these communities are interacting with data repositories during the research process. For example, environmental health is an emerging research community that is conducting a substantial amount of new and innovative research focusing on the impact of changes in the environment on human health and wellbeing (Bright et al., 2012). Environmental health research uses data and methods from environmental science, the health sciences, and other fields including the social sciences, data science, and engineering (Hoover, Renauld, Edelstein, & Brown, 2015; National Institute of Environmental Health Sciences, 2012). An understanding of the role data plays in these collaborations can help us identify and meet the needs of emerging research communities.

This paper reports findings about the role of data in the emerging environmental health research community from the perspective of health research that is using data collected by environmental researchers. We address the following questions:

RQ1: Are researchers using open data exposed by disciplinary-based repositories that are outside their own discipline?

RQ2: What role does data play in the formation of research groups that address environmental health challenges?

RQ3: How do environmental health research groups share data?

Background on Environmental Health

We chose environmental health as our exemplar for examining the role of data in a convergence research environment because the environment and human health are intertwined, making this an important area of research with potentially high impact on human life and well-being. Adverse environmental conditions account for a large fraction of global death and disability. Prüss-Ustün et al.(2016) estimate that 23% of 12.6 million global deaths in 2012 were caused by modifiable risks associated with air, ultraviolet, noise, occupational risks, the built environment, man-made climate change, behavior related to the availability of safe water and sanitation facilities, and other environmental factors; 22% of the disease burden[1] could be avoided if those risks were removed. Though these environmental health risks are a global concern, vulnerable subpopulations in developed countries and individuals in low-income countries are disproportionately affected by the health risks brought about by adverse environmental conditions, often due to factors such as increased exposure to environmental pollutants, lack of local health infrastructure, and lack of basic resources such as clean drinking water and sanitation (Elliott, 2011; McMichael et al., 2008).

Numerous definitions of environmental health have been developed (U.S. Department of Health and Human Services, 1998). This study uses the definition provided by the World Health Organization (WHO) (1993): "Environmental health comprises those aspects of human health, including quality of life, that are determined by physical, chemical, biological, social and psychosocial factors in the environment. It also refers to the theory and practice of assessing, correcting, controlling and preventing those factors in the environment that can potentially affect adversely the health of present and future generations."

Environmental health research varies in both scope and scale, and includes topics such as identifying environmental-attributable altering on genetics of pathogens, determining specific links between adverse health effects and exposure to pollutants, assessing the impact of development on human health and the environment, and tracking and predicting global scale disease patterns under climate change (Costello et al., 2009; Kearny et al., 2015; Liu et al.,

[1] Disease burden in DALYs, a combined measure of years of life lost due to mortality and years of life lost due to disability. See: http://www.who.int/healthinfo/global_burden_disease/metrics_daly/en/

2012; Prüss-Ustün et al., 2016). Environmental health issues and their solutions are complex, and require the expertise of researchers from multiple domains and disciplines (Costello et al., 2009; Manlove et al., 2016). Environmental health research is often organized by broad topic area or issue of concern, and conducted by groups of researchers from domains including ecology, biology, chemistry, public health, social science, earth science, and economics (Bright et al., 2012; Hoover et al., 2015; National Environmental Health Association, 2016; National Institute of Environmental Health Sciences, 2012). The one constant is the crucial nature of the data that describe the environment itself. Each of these groups may have different approaches to interacting with environmental data.

As with other types of convergence research, the success of environmental health collaboration depends on a number of factors, including institutional and organizational support for such research; funding, space, and resources for collaboration activities; appropriate venues for the publication of research results; tools for the collection, management, and sharing of data resulting from the research; and researchers' own skills and experience working as collaborating members of multidisciplinary teams (Hicks et al., 2010; Lungeanu et al., 2014; Mauz, et al., 2012; Porter et al., 2012; Reichman, 2004; Sharp et al., 2016). Challenges that exist for convergence research teamwork can include recruitment of team members with needed expertise; continued funding and institutional support; challenges involving the establishment of team roles and guidelines for sharing credit for research work; and the development of a shared vision, shared understanding of research goals, and common vocabularies, sets of practices, and standards. All of these can be difficult, as disciplinary cultural differences may not only impact research practice, but also how underlying research questions are formulated and understood (Liu et al., 2012; Mauz et al., 2012; Parsons et al., 2011). For collaborations that are international in scope, language barriers can also be a concern (Liu et al., 2012).

The Role of Data in Environmental Health Research

The environmental health community is characterized by an integrated and data-driven investigation approach (Bright et al., 2012). The increasing availability of environmental data—satellite observational data, in-situ measurements, model outputs, reanalysis data, biotic surveys, and social science data—has generated numerous new scientific discoveries in human health research by stimulating spatial thinking and providing resources to uncover biological and ecological interaction between human and environment (Overpeck et al.,2011).

Multiple challenges currently exist for the collection, management, and sharing of environmental health data. The establishment of shared practices and standards for data collection at the beginning of a project—as well as agreement on what data should be collected—is crucial, yet disciplinary differences and the number and variety of variables in play in environmental health research can make this consensus difficult (Mauz et al., 2012). Due to the nature of the work, data are often extremely heterogeneous and deciding how to appropriately manage and store all the data resulting from a project is also a challenge (Brooks, et al., 2016; Parsons et al., 2011).

Most importantly for this study is the fact that environmental health research makes secondary use of previously collected data. A current lack of shared standards and practices for metadata creation, data quality assurance, and quality control can make finding and reusing data difficult (Hendrickx, et al., 2014; Kearny et al., 2015; Lake et al., 2010; Mattes et al., 2004; Palmer et al, 2016).

Convergence research areas, including environmental health research, involve the reuse and integration of disparate and often extremely large datasets, each of which were collected and described based on the norms and disciplinary practices of scientists involved in the previous research (Palmer et al., 2016; Sharp et al., 2016). Convergence researchers often lack the tools, methods, and training to locate needed datasets, to properly interpret and integrate datasets, to determine the quality of datasets generated by others, and to provide appropriate metadata for the datasets they themselves create in the course of their research (Hendrickx, et al., 2014; Kearny et al., 2015; Mattes et al., 2004; Wendt, et al., 2015). The lack of data integration can

lead to inability to make connections between environmental factors and health outcomes and to determine appropriate questions for further research, and possible indirect impacts of adverse conditions and changes in the environment (Brooks et al., 2016; Kearney, 2015; Liu et al. 2012). New tools for cross-disciplinary data discovery and integration are being developed, and information professionals as well as experts trained in helping facilitate multidisciplinary team science, such as staff members at synthesis centers created to support such work, can provide assistance with these challenges (Bright et al., 2012; Palmer et al., 2016; Roco & Bainbridge, 2013), but there is still much to be understood about how these teams currently interact with data and existing data repositories and services.

Methods

We take the approach of "following the data" in order to identify researchers who are engaging in convergence science and working in an "open science" environment. We chose environmental health research as a focus for three reasons. First, it is a data-intensive research community that is heavily dependent on secondary use of data. Second, it is a convergence research community bringing together data from many disciplines. Third, the research is important to society and depends on data that is reliable and trusted.

The first step we took was to use the definition of environmental health provided by the WHO (1993) to identify the primary and proximate disciplines that supply the data used for analysis by environmental health researchers. Here, the primary discipline is human health and the proximate discipline is environmental science.

Next, we identified repositories in the proximate discipline that make reliable environmental data easily identifiable and accessible. We used Data Observation Network for Earth (DataONE) as a federated access tool to search 30 member repositories holding over 402,066 individual data files and 227,339 metadata files of environmental data as of June 2016 (DataONE, n.d.-a).

The third step was to review what proximate discipline datasets had been used to produce environmental health publications. The repositories identified health science publications that had used their datasets. We also searched Web of Science and Google Scholar for environmental health articles published in environmental health or human health-related publications, including journals such as J*ournal of Public Health*, *Population & Environment*, and *Environmental Health Perspectives*.

The fourth step was to review the list of publications and identify articles that matched the following three criteria: 1) the research fit the definition of an environmental health study, 2) the research was published within the last 10 years, and 3) the datasets used in the research were available from any of the DataONE repositories. After reviewing 14,000 results, there were twelve articles that fit these three criteria.

All twelve articles had multiple authors and we found that 32 of these authors had their email addresses available online. Each of these 32 authors was contacted via email with an invitation to participate in an interview. The first-round invitation email allowed a week for a response and the following week, a second invitation email was sent to those who had not responded. This resulted in five participants scheduling an interview. These five participants were co-authors of five different papers resulting in contact with 41.7% of the teams represented by the paper population. These five participants were based in the United States. To gain an international perspective, a sixth participant who had published environmental health research using an environmental dataset was invited based on engagement as a DataONE community member. The sample size is determined to be sufficient since it represents a third of teams responsible for the paper population and since analysis of the interviews determined that saturation was reached (Fusch & Ness, 2015). The sample of researchers was sufficient for the population specified and is not generalized to the full environmental health community.

Each co-author participant was a member of an environmental health research team. The participants represented a diversity of institutional types, domains and professional roles. Interview times were set to be convenient for the participant, and each participant chose their preferred mediated communication tool, either online via Skype and WebEx or via phone. After obtaining informed consent, the interview was conducted using a semi-structured protocol to elicit responses. The protocol included fifteen questions that addressed a range of team issues starting with how groups are established, how they share data within and outside of the group, and how the group deals with datasets large and small. Each interview was recorded and the duration of the interviews varied between 23 - 44 minutes.

Table 1. Participants

Participant	Institutional Type	Domain	Professional Role
1	University	Bioengineering, metabolism in health and disease [Primary]	Research Assistant Professor
2	University	Statistics, biomedical data science [Primary]	Associate Professor
3	Federal Agency	Human health impacts of pathogens in the environment [Primary]	Research scientist
4	Research institute	Environmental transmission of infectious disease [Primary]	Researcher
5	University	Air pollution, aerosols and occupational health [Primary]	Associate Professor
6	Federal Agency	Environmental geochemistry [Proximate]	Research scientist

The interview recordings were transcribed by members of the research team. After the transcription process was completed, an initial round of open coding was completed on half the interview transcripts to identify broad themes in the data and develop the codebook. Following codebook development, a second round of coding was done on all transcripts. At least two researchers coded each transcript to help ensure validity and consistency of findings (Merriam & Tisdell, 2016). When there was disagreement on what the respondent indicated in his or her response the researchers discussed the discrepancy and came to a consensus on how to classify the responses. This brought the intercoder reliability to 100%.

The research protocol was approved by the university Institutional Review Board prior to beginning data collection. Limitations of the study include that a lack of fully established practices and norms for data citation may mean that published articles in environmental health were not linked to the datasets that informed the research, as discussed below (Mooney & Newton, 2012). Articles not linked to an environmental health dataset would also not have been included in the sample. Additionally, there are limits created by identifying DataONE as a search tool that could be used by convergence researchers to attain easy access to a large number of repositories holding open environmental science data. This approach allowed us to identify data that was used in these repositories, but limited our ability to follow datasets held in other open repositories that may have been used in environmental health research during the time period under study. Finally, while our focus was on identifying whether environmental

health researchers had used datasets exposed by an environmental data repository, we do not know how the authors discovered the data they used in their work.

Results

RQ1: Are researchers using open data exposed by disciplinary-based repositories that are outside their own discipline?

Using data repositories to identify researchers for our study, we found that the repositories reviewed for this study are exposing data that are being used by environmental health researchers who are outside, but proximate, to the environmental science discipline. While we do not know if DataONE served as a source of that data for these particular research groups, our findings do indicate that these repositories do expose data of value to the environmental health community. While the number of articles found that met the research criteria sounds small, it reflects three issues: (1) datasets and the published articles they informed are not always linked; (2) datasets were not mandated to be made available during the full ten-year period of our search; and (3) data repositories may not be aware of all the articles that have used their data, especially if these articles are outside the repository's own discipline.

RQ2: What role does data play in the formation of research groups that address environmental health challenges?

Data plays a role in environmental health teams from the earliest stages of team formation. Researchers noted they used a problem-driven approach to forming research teams, including reaching out to the owners of existing data needed for a project. "I would reach out to people, whoever owns the data, whoever have large data with large sample size...I would reach out to the owners of those [data] and have them on my team" (Participant 2). They also mentioned the need to reach out to potential team members based on the requirements of the project grant and to fill identified gaps in knowledge or expertise.

While not all researchers specifically mentioned the role of data in team formation, two consistent themes emerged that were also found to be related to participants' ability to share data once a team was formed: relationship-building and reputation. Formation of research groups around environmental health challenges relied heavily on active principal investigators who guided group composition. All scientists in our sample mentioned the principal investigator (PI) as the person largely responsible for the formation of teams, though other team members may play a role as well by suggesting possible collaborators. Often members of convergence teams are invited to participate based on a previous relationship with the PI or with another team member with whom they have worked in the past. In the words of one respondent, "Sometimes, it is based on just people we already know…we heard a presentation by them, and we were impressed with what they did. But very often you build up a relationship working with people in another project. Hence you already know who you want to work with, because you've already sort of have a team ready to go, if the funding becomes available. Those alliances are built upon many years" (Participant 6).

When invitations were extended to researchers without a prior relationship to other team members, the importance of reputation within the scientific community was stressed, including reputation based on scholarly publications and presentations, which are direct products from collecting and analyzing data, as well as media coverage of one's work. A researcher said "you can kind of tell through someone's published literature what kind work they do and what quality work they do" (Participant 1). In addition to reputation established via formal scientific communication, team members were selected based on reputation spread via personal networks. "I reach to people who I worked with before," said one interviewee who had assembled teams,

"I ask them who they know and who they would recommend" (Participant 5). These strategies align with previous findings on data sharing, in which researchers were found to rely on personal networks, reputation, and publication records in deciding whether to share or reuse data (Faniel & Jacobson, 2010; Kaye et al., 2009).

These answers speak to data's role in the emerging community of environmental health researchers, even as there are many existing relationships between collaborating researchers, and to the importance of the formal and informal scholarly reputation built on previous scholarly works that relied on collecting and using high quality data.

RQ3: How do environmental health research groups share data?

Data is the coin of the realm and a key for collaboration, as evidenced by the fact that all participants shared data among the members of their teams. All the researchers named at least one technology or tool that supported their data-sharing. Sharing platforms such as Google Drive or Dropbox were most frequently mentioned. Also frequently mentioned was the more traditional approach of spreadsheets, as well as email and publicly accessible data repositories. The use of presentation software suggests data is also being shared in summary form after analysis. Other sharing tools included Access databases, FTP/LTP sites, and cloud servers (Table 2).

Table 2. Data Sharing Tools Used by Participants' Teams

Data Sharing Tools	Recognition Level
Sharing platform, e.g. Google Drive, Dropbox; Spreadsheets	High
Email; Presentation software, e. g. PowerPoint; Public data repository, e.g. Dryad	Moderate
Access databases; Cloud servers; Data center; External databases; FTP/LTP sites; Open source websites, e.g. Forgenet, Github; Physical hard drive; Conference call	Low

Similar to data sharing, team communication uses email and real-time meetings through conference calls. Other tools such as Skype or virtual meeting spaces were specified although they were used mostly as a proxy for in-person meetings. While such technologically-mediated communication methods were not always seen as ideal, they were seen as sometimes necessary, particularly for geographically distributed teams. "Face-to-face is always better but it is not practical to always go face-to-face" observed one interviewee (Participant 5), while another noted, "to get enough done, to keep things moving, the Internet becomes a necessity because otherwise, I would just be on the run all of the time" (Participant 1).

Half the participants noted data sharing requirements or restrictions from publishers, institutions, or funding agencies as determinants for if and how project data would be shared. Data typically were not shared outside the research team prior to publication of a research article at all, although exceptions could be made with permission. Open-source publication was noted as one method of sharing data. Data publication, sharing, and storage venues included open-source websites, public data repositories, external databases, and data centers that handle large datasets.

Environmental health research faces of the challenge of making sure data from different domains will be interoperable, especially when working with large datasets. Surprisingly, all the researchers, regardless of domain, said they had no data interoperability issues. This may be because the repositories housing the data we found are already addressing interoperability. It could also be that the datasets these researchers are using are more homogenous than

anticipated, or that these needs are being met by technical experts on the team or within team members' organizations. Only two of the six participants dealt with what they considered large datasets, and they used massively parallel computing networks, university computers, and statistical programs such as SAS and MATLAB as tools that helped them deal with these datasets. One also mentioned the university's IT department as a source of help for sharing large datasets.

When researchers spoke of sharing data, most focused on interpersonal challenges and processes rather than technological issues. Similar to team formation, team member expertise and the trust of other team members in that expertise played a substantial role. Most participants said that data were shared and interpreted via direct explanation, with several noting that other members of the team accepted the interpretation and explanation given by the team member(s) who were experts on the data. Not all teams share raw data and not all team members are necessarily involved in data analysis and processing. Multiple researchers spoke of sharing summary or conceptual data among team members rather than raw data, while only one discussed sharing the full information about data processing via a conference call to go through all steps.

To share data, researchers observed that time and patience are required, as is the need to establish a common language so that technical terms can be understood. For example, to share context across domains, respondents spoke of the importance of using plain language and less jargon, explaining concepts until they are understood, and using analogies. "Basically," said one participant, "you have to find a common language, so we're all using the same terms, you know. But once you establish that, we are speaking the same vocabulary, we understand the terms we're using, people can work from different fields together quite well" (Participant 6). This is not needed if team members share a common background already. One researcher said, "I think the communication part is the hardest part, because everyone…you know, every discipline has got its own way of communicating" (Participant 3). Researchers noted these issues can be addressed by having a skilled communicator on the team with knowledge of multiple domains.

Shared goals, particularly the need to use the data to publish a research paper, result in a natural bond between team members. Maintaining cohesion within the team requires openness and transparency on the part of team members. The importance of trust for idea sharing was specifically noted with one researcher stating "it depends on you know, how much you trust the members of the group. So at the beginning, people were…well this person can be trusted with what we do, and what we want to do, but some details were…not really discussed. But then, you know through time, meeting after meeting, we learned to know each other, and so now we are working, and are good collaborators" (Participant 4).

Discussion and Conclusion

Society benefits from domains converging to examine a scientific problem (National Research Council, 2014). The National Science Foundation (NSF) is one of many agencies to recognize the advantages of this integration of disciplinary knowledge and problem-solving approaches and in 2017 included growing convergence research as one of its 10 Big Ideas for future NSF investments (National Science Foundation, 2017). Understanding the role of data in convergence research collaborations can help information professionals support successful collaborations.

Environmental health was chosen as an exemplar of an emerging research community in this study because it is one in which networks of collaborating researchers from multiple disciplines have begun to converge around scientific challenges that substantially impact human health and quality of human life. In order to find successful environmental health research teams who used data made available through the proximate domain of environmental studies, we searched established environmental data repositories that provided open access for their data holdings. We focused on datasets that were identified as being used in published environmental

health journal articles. This approach provided evidence that data from a proximate domain was used by environmental health researchers. We could then contact individuals who could shed light on the role of data in their environmental health research teams. The value in our approach is that by "following the data" we are finding researchers who are engaging in convergence science and working in an "open science" environment. Based on the results, we offer evidence that supports these five key insights about the role of data in environmental health research and characteristics of successful convergence research teams. We also offer some thoughts for the future, as we see a need for continuing research in this area in order to better understand the role of data in convergence communities.

- **Sharing data presents interpersonal and technical challenges.** While some participants mentioned technical challenges related to data sharing, the majority did not experience interoperability issues and were able to share data via the tools available to them. Researchers were more focused on the interpersonal challenges of data sharing, including the challenges involved in making sure data are understood when team members come from multiple disciplines. Learning more about how researchers find their proximate discipline datasets is a focus for future research.

- **Data expert team members play a vital role in data sharing by being data mediators.** Team members identified as data experts are vital to the sharing of data within environmental health teams. Within teams, data experts may be the only members working directly with raw data and may be called upon to interpret that data for others, requiring them to be skilled communicators across disciplinary boundaries. For example, most researchers in this study said data were shared within the team through the data expert. We identify this role as being a *data mediator* since these may be the only team members interacting directly with other team members and the raw data itself. We saw that other team members often trust the data mediator to provide their main conduit to the data they will use for analysis and interpretation.

- **Reputation builds trust in the data and data mediators.** Reputation and trust, the same factors that shape team formation and communication, are vital in shaping how environmental health researchers interact with data, particularly in cases where data mediators bear the responsibility of interpreting data for non-expert team members. A reputation for producing quality research in one's area of expertise, for collaborating well with others, and for being a skilled communicator across disciplines can lead to a particular team member being invited to participate in a research project or trusted to fill the role of data mediator for other members of that team.

- **Trust within a team is needed for a successful collaboration as evidenced by data sharing.** Trust shaped how and by whom data was communicated. Not all team members were involved directly in data processing and analysis, with some team members choosing to trust data mediators or other subject matter experts on the team. As researchers continue to collaborate, researcher information behavior may move from a practice in which one or more trusted researchers have a direct relationship with the data and are able explain the results of data analysis, to a practice where teams develop a shared ownership of the data and a shared vocabulary and language around it, even when they are coming from different paradigms and disciplines. Further study is warranted to understand the role trust plays in this process.

- **Convergence research is problem-driven, facilitating an immediate affinity among the people who have been drawn together and providing a focus for creating a scholarly product from the data, such as a research article.** Researchers' relationships with data are a crucial factor because of the problem-driven nature of environmental health research. Team formation and composition are a result of the problem-centric focus, and for example, teams may form as a result of

investigators identifying existing datasets needed for addressing a problem and reaching out to the owners of those datasets. A repeating theme in the interviews was the focus on the goals for funded research, specifically the underlying research problems and shared goals. Overarching outcomes such as improving science or our knowledge base were important, however equally important were more specific outcomes such as reporting on the analysis of the data in a published paper.

Thoughts for the Future

Data play an essential role in convergence research, including a role in the relationship-building needed for this work, and could be thought of as a "team member" in terms of their importance to the research enterprise and influence on team dynamics. Our exploratory study provides insight into the role of data in the emerging community of environmental health research. Continuing research can lead to additional insights into the specifics and strength of the social role data play in convergence collaborations.

There is clearly an important role for information professionals and organizations involved in data management. For example, these professionals can help teams in emerging communities share the context of their data both prior to and following publication. Information professionals can also help teams locate data from proximate fields that are held in widely distributed data repositories and provide training for using tools to find data (e.g.DataONE), manage data (e.g. DMPTool, Dash), and analyze data (e.g. ArcGIS, JMP, MATLAB).

Observations from this data are potential themes for future research. Such research could explore the role of funding entities and organizations such as synthesis centers in fostering data-intensive convergence collaborations, and methods for using data as a starting point to identify environmentally-adjacent research communities as they emerge. There are also many opportunities for examining how other emerging research communities find and work with existing data. While this study focused on environmental health as an exemplar community and used DataONE as an access tool to locate data held in disciplinary-based repositories, we envision our approach of "following the data" being used with other open data repositories and other proximate disciplines, including those in the social sciences, as a starting point.

Acknowledgements

This study was funded as part of the National Science Foundation, Division of Cyberinfrastructure, Data Observation Network for Earth (DataONE) NSF award #1430508 under a Cooperative Agreement, William Michener, P. I.

References

Borgman, C. L. (2015). Big data, little data, no data: Scholarship in the networked world. Cambridge, MA: MIT Press.

Bright, P, R., Buxton, H. T., Balistrieri, L. S., Barber, L. B., Chapelle, F. H., Cross, P. C. … Winton, J. R. (2012). USGS environmental health science strategy—Providing environmental health science for a changing world. Retrieved from http://pubs.usgs.gov/of/2012/1069/of2012-1069.pdf

British Academy & The Royal Society. (2017). Data management and use: Governance in the 21st Century. Retrieved from https://royalsociety.org/topics-policy/projects/data-governance/

Brooks, C. F., Heidorn, P. B., Stahlman, G. R., & Chong, S. S. (2016). Working beyond the confines of academic discipline to resolve a real-world problem: A community of scientists discussing long-tail data in the cloud. First Monday, 21(2).

Costello, A., Abbas, M., Allen, A., Ball, S., Bell, S., Bellamy, R., … others. (2009). Managing the health effects of climate change. The Lancet, 373(9676), 1693–1733.

DataONE. (n.d.). Current Member Nodes. Retrieved from https://www.dataone.org/current-member-nodes#uploads

Elliott, L. (2011). Environmental change and health human dimensions, ethics and global governance. Human Health and Global Environmental Change, 7.

Faniel, I. M. & Jacobsen, T. E. (2010). Reusing scientific data: How earthquake engineering researchers assess the reusability of colleagues' data. Computer Supported Cooperative Work, 19(3-4). doi:10.1007/s10606-010-9117-8

Fecher, B., Friesike, S., & Hebing, M. (2015). What drives academic data sharing? PLoS ONE, 10(2). doi:10.1371/journal.pone.0118053

Fusch, P. I., & Ness, L. R. (2015). Are we there yet? Data saturation in qualitative research. The Qualitative Report, 20(9), 1408.

Hendrickx, D. M., Boyles, R. R., Kleinjans, J. C. S., & Dearry, A. (2014). Workshop report: Identifying opportunities for global integration of toxicogenomics databases, 26–27 June 2013, Research Triangle Park, NC, USA. Archives of Toxicology, 8, 2323–2332. doi:10.1007/s00204-014-1387-3

Hey, T., Tansley, S., & Tolle, K. M. (2009). The fourth paradigm: Data-intensive scientific discovery. Redmond, WA: Microsoft Research.

Hicks, C. C., Fitzsimmons, C., & Polunin, N. V. C. (2010). Interdisciplinarity in the environmental sciences: barriers and frontiers. Environmental Conservation, 37, 464–477. doi:10.1017/S0376892910000822

Hoover, E., Renauld, M., Edelstein, M. R., & Brown, P. (2015). Social science collaboration with environmental health. Environmental Health Perspectives, 123(11), 1100–1106. doi:10.1289/ehp.1409283

Kaye, J., Heeney, C., Hawkins, N., de Vries, J., & Boddington, P. (2009). Data sharing in genomics: Re-shaping scientific practice. Nature Reviews Genetics, 10(5). doi:10.1038/nrg2573

Kearney, G. D., Namulanda, G., Qualters, J. R., & Talbott, E. O. (2015). A decade of environmental public health tracking (2002-2012): progress and challenges. Journal of Public Health Management and Practice, 21, S23-S35.

Kim, Y. (2017). Fostering scientists' data sharing behaviors via data repositories, journal supplements, and personal communication methods. Information Processing & Management, 53, 871-885. doi:10.1016/j.ipm.2017.03.003

Lake, A. A., Burgoine, T., Greenhalgh, F., Stamp, E., & Tyrrell, R. (2010). The foodscape: classification and field validation of secondary data sources. Health & Place, 16(4), 666-673.

Liu, H. Y., Bartonova, A., Pascal, M., Smolders, R., Skjetne, E., & Dusinska, M. (2012). Approaches to integrated monitoring for environmental health impact assessment. Environmental Health, 11(1), 1.

Lungeanu, A., Huang, Y., & Contractor, N. S. (2014). Understanding the assembly of interdisciplinary teams and its impact on performance. Journal of Informetrics, 8, 59–70. doi:10.1016/j.joi.2013.10.006

Manlove, K. R., Walker, J. G., Craft, M. E., Huyvaert, K. P., Joseph, M. B., Miller, R. S., … Cross, P. C. (2016) "One Health" or three? Publication silos among the One Health disciplines. PLoS Biology, 14(4): e1002448. doi:10.1371/journal.pbio.1002448

Mattes, W. B., Pettit, S. D., Sansone, S. A., Bushel, P. R., & Waters, M. D. (2004). Database development in toxicogenomics: Issues and efforts. Environmental Health Perspectives, 112(4), 495.

Mauz, I., Peltola, T., Granjou, C., Van Bommel, S., & Buijs, A. (2012). How scientific visions matter: Insights from three long-term socio-ecological research (LTSER) platforms under construction in Europe. Environmental Science & Policy, 19, 90-99.

McMichael, A. J., Friel, S., Nyong, A., & Corvalan, C. (2008). Global environmental change and health: Impacts, inequalities, and the health sector. BMJ, 336(7637), 191-194. doi:10.1136/bmj.39392.473727.AD

Merriam, S., & Tisdell, Elizabeth J. (2016). Qualitative research: A guide to design and implementation (4th. ed.). San Francisco, CA : Jossey-Bass.

Mooney, H., & Newton, M. (2012). The anatomy of a data citation: Discovery, reuse, and credit. Journal of Librarianship and Scholarly Communication, 1(1). doi:10.7710/2162-3309.1035

National Environmental Health Association. Definitions of Environmental Health |: NEHA. (2016). Retrieved June 20, 2016, from http://www.neha.org/about-neha/definitions-environmental-health

National Institute of Environmental Health Sciences. (2012). Advancing Science, Improving Health: A Plan for Environmental Health Research National Institute of Environmental Health Sciences, Strategic Plan 2012-2017. Retrieved from https://www.niehs.nih.gov/about/strategicplan/strategicplan2012_508.pdf

National Research Council. (2014). Facilitating transdisciplinary integration of life sciences, physical sciences, engineering, and beyond. Washington, D.C.: National Academies Press. doi:10.17226/18722

National Science Foundation. (2017). Dear colleague letter: Growing convergence research at NSF (NSF 17-065). Retrieved from https://www.nsf.gov/pubs/2018/nsf18058/nsf18058.jsp

OECD. (2015). Making Open Science a Reality (OECD Science, Technology and Industry Policy Papers, No. 25). Paris, France: OECD Publishing. doi:90/10.1787/5jrs2f963zs1-en

Oushy, M. H., Palacios, R., Holden, A. E. C., Ramirez, A. G., Gallion, K. J., & O'Connell, M. A. (2015). To share or not to share? A survey of biomedical researchers in the U.S. Southwest, an ethnically diverse region. PLoS ONE, 10(9). doi:10.1371/journal.pone.0138239

Overpeck, J. T., Meehl, G. A., Bony, S., & Easterling, D. R. (2011). Climate data challenges in the 21st century. Science, 331(6018), 700-702. doi:10.1126/science.1197869

Palmer, M. A., Kramer, J. G., Boyd, J., & Hawthorne, D. (2016). Practices for facilitating interdisciplinary synthetic research: The National Socio-Environmental Synthesis Center (SESYNC). Current Opinion in Environmental Sustainability, 19, 111-122. doi:10.1016/j.cosust.2016.01.002

Parsons, M. A., Godøy, Ø., LeDrew, E., De Bruin, T. F., Danis, B., Tomlinson, S., & Carlson, D. (2011). A conceptual framework for managing very diverse data for complex, interdisciplinary science. Journal of Information Science, 37(6), 555-569.

Pasquetto, I., Randles, B., & Borgman, C. (2017). On the reuse of scientific data. Data Science Journal, 16. doi:10.5334/dsj-2017-008

Porter, A. L., Garner, J., & Crowl, T. (2012). Research coordination networks: Evidence of the relationship between funded interdisciplinary networking and scholarly impact. BioScience, 62(3), 282–288. doi:10.1525/bio.2012.62.3.9

Prüss-Ustün, A., Wolf, J., Corvalán, C., Bos, R., & Neira, M. (2016). Preventing disease through healthy environments: A global assessment of the burden of disease from environmental risks. Geneva, Switzerland: World Health Organization. Retrieved from http://apps.who.int/iris/bitstream/10665/204585/1/9789241565196_eng.pdf

Reichman, O. J. (2004) NCEAS: Promoting creative collaborations. PLoS Biology, 2(3): e72. doi:10.1371/journal.pbio.0020072

Roco, M. C., & Bainbridge, W. S. (2013). The new world of discovery, invention, and innovation: convergence of knowledge, technology, and society. Journal of Nanoparticle Research, 15(9), 1946. doi:10.1007/s11051-013-1946-1

Sá, C., & Grieco, J. (2016). Open data for science, policy, and the public good. Review of Policy Research, 33, 526-543. doi:10.1111/ropr.12188

Sharp, P., Hockfield, S., & Jacks, T. (Eds.) (2016). Convergence: The future of health. Cambridge, MA: MIT. Retrieved from http://www.convergencerevolution.net/2016-report/

Specht, A., Guru, S., Houghton, L., Keniger, L., Driver, P., Ritchie, E. G., . . . Treloar, A. (2015). Data management challenges in analysis and synthesis in the ecosystem sciences. Science of the Total Environment, 534, 144-158. doi:10.1016/j.scitotenv.2015.03.092

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A. U., Wu, L., Read, E., ... & Frame, M. (2011). Data sharing by scientists: practices and perceptions. PLoS One, 6(6), e21101. doi:10.1371/journal.pone.0021101

Tenopir, C., Dalton, E. D., Allard, S., Frame, M., Pjesivac, I., Birch, B., ... & Dorsett, K. (2015). Changes in data sharing and data reuse practices and perceptions among scientists worldwide. PLoS One, 10(8), e0134826. doi:10.1371/journal.pone.0134826

U.S. Department of Health and Human Services. (1998, 20 November). An ensemble of definitions of environmental health. Retrieved from https://health.gov/environment/DefinitionsofEnvHealth/ehdef2.htm

Volk, C. J., Lucero, Y., & Barnas, K. (2014). Why is data sharing in collaborative natural resource efforts so hard and what can we do to improve it? Environmental Management, 53, 883-893. doi:10.1007/s00267-014-0258-2

Wendt, A., Kreienbrock, L., & Campe, A. (2015). Zoonotic disease surveillance–inventory of systems integrating human and animal disease information. Zoonoses and Public Health, 62(1), 61-74.

World Health Organization. (1993). WHO | Public health, environmental and social determinants of health (PHE). Retrieved June 29, 2016, from http://www.who.int/phe/en/ .