ISSN 2161-3974

JeSLIB 2020; 9(1): e1186
https://doi.org/10.7191/jeslib.2020.1186

Journal of eScience Librarianship
putting the pieces together: theory and practice

EScience in Action

Testing Our Assumptions: Preliminary Results from the Data Curation Network

Elizabeth Coburn and Lisa Johnston

University of Minnesota - Twin Cities, Minneapolis, MN, USA

Abstract

Objective: Data curation is becoming widely accepted as a necessary component of data sharing. Yet, as there are so many different types of data with various curation needs, the Data Curation Network (DCN) project anticipated that a collaborative approach to data curation across a network of repositories would expand what any single institution might offer alone. Now, halfway through a three-year implementation phase, we're testing our assumptions using one year of data from the DCN.

Methods: Ten institutions participated in the implementation phase of a shared staffing model for curating research data. Starting on January 1, 2019, for 12 months we tracked the number, file types, and disciplines represented in data sets submitted to the DCN. Participating curators were matched to data sets based on their self-reported curation expertise. Aspects such as curation time, level of satisfaction with the assignment, and lack of appropriate expertise in the network were tracked and analyzed.

Correspondence: Elizabeth Coburn: ecoburn@umn.edu

Received: April 2, 2020 **Accepted**: May 28, 2020 **Published**: September 9, 2020

Data Availability: Data associated with this article are available from the Data Repository for the University of Minnesota at: https://doi.org/10.13020/ak4d-ge34.

Disclosures: The authors report no conflict of interest.

JeSLIB 2020; 9(1): e1186
https://doi.org/10.7191/jeslib.2020.1186

Abstract Continued

Results: Seventy-four data sets were submitted to the DCN in year one. Seventy-one of them were successfully curated by DCN curators. Each curation assignment takes 2.4 hours on average, and data sets take a median of three days to pass through the network. By analyzing the domain and file types of first- year submissions, we find that our coverage is well represented across domains and that our capacity is higher than the demand, but we also observed that the higher volume of data containing software code relied on certain curator expertise more often than others, creating potential unbalance.

Conclusions: The data from year one of the DCN pilot have verified key assumptions about our collaborative approach to data curation, and these results have raised additional questions about capacity, equitable use of network resources, and sustained growth that we hope to answer by the end of this implementation phase.

JeSLIB 2020; 9(1): e1186
https://doi.org/10.7191/jeslib.2020.1186

Introduction

To address growing requirements for research data sharing, academic institutions have ramped up efforts to deliver a variety of data repository services. Regardless of whether the repository services are offered via locally -developed systems or third-party platforms, data curation plays an important role in data sharing. A data curator applies a special combination of professional ethics, technological skills, subject expertise, and an overarching concern for enabling reuse to the data publication pipeline. Given the complexity and variety of research data, it seems logical that repositories would benefit from an approach that "harnesses the expertise of well-aligned institutions that collectively provide data curation services to researchers in a multitude of disciplines, ensuring that valuable scholarly data sets are findable, accessible, interoperable and reusable, or FAIR" (Johnston et al. 2018, 130).

The Data Curation Network team, with support from the Alfred P. Sloan Foundation, set out to test this theory and published A Cross-Institutional Staffing Model for Curating Research Data by Johnston et al. in 2018 (herein referred to as the "DCN Model"). The Network has now been up and running for more than a year. In this paper we reflect on the initial assumptions made in the DCN Model and test them against the preliminary first- year results of collaboratively curating research data.

Background

The DCN initially launched with eight academic and general data repository partners from Cornell University, Dryad Digital Repository, Duke University, the University of Illinois, Johns Hopkins University, the University of Michigan, the University of Minnesota (lead), and Pennsylvania State University. After months of curator training, workflow testing, and infrastructure development, the DCN's three-year implementation phase hit a major milestone on January 1, 2019 when we put the Network into action (Figure 1). Two additional institutional partners, New York University and Washington University in St. Louis, joined the project in June 2019, mid-way through our first year of testing. Therefore, the first-year results of piloting our collaborative data curation service reflects the efforts of 10institutions from January-December 2019.

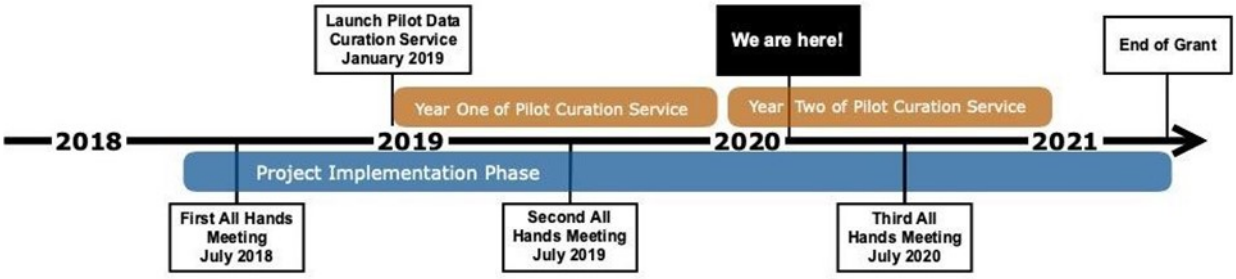

Figure 1: Timeline of DCN Project Implementation Phase

Sharing data curators across 10 distributed sites is made possible by a multi-tiered workflow (Figure 2). First, we designed the DCN service to fit seamlessly within a partner's existing workflow where they decide what data to send to the Network

JeSLIB 2020; 9(1): e1186
https://doi.org/10.7191/jeslib.2020.1186

for curation and are entirely responsible for the long-term storage, access, and preservation of that data. Second, our DCN Coordinator reviews data sets from partner repositories and matches them to an appropriate DCN curator based on the domain and file format. Third, all curators in the DCN are trained in standardized CURATE steps to help ensure that every data set—regardless of format, size, discipline, or complexity—receives standardized review. In every case, the DCN and local curators work closely with data providers to curate their data (Figure 3). DCN curators apply their specialized subject and format expertise to perform curation actions such as checking the data and documentation, running code, assessing and offering advice to mitigate risk (such as privacy disclosure or copyright concerns), and transforming files when appropriate. Lastly, all activity is tracked via our shared project management tool, Jira[1].

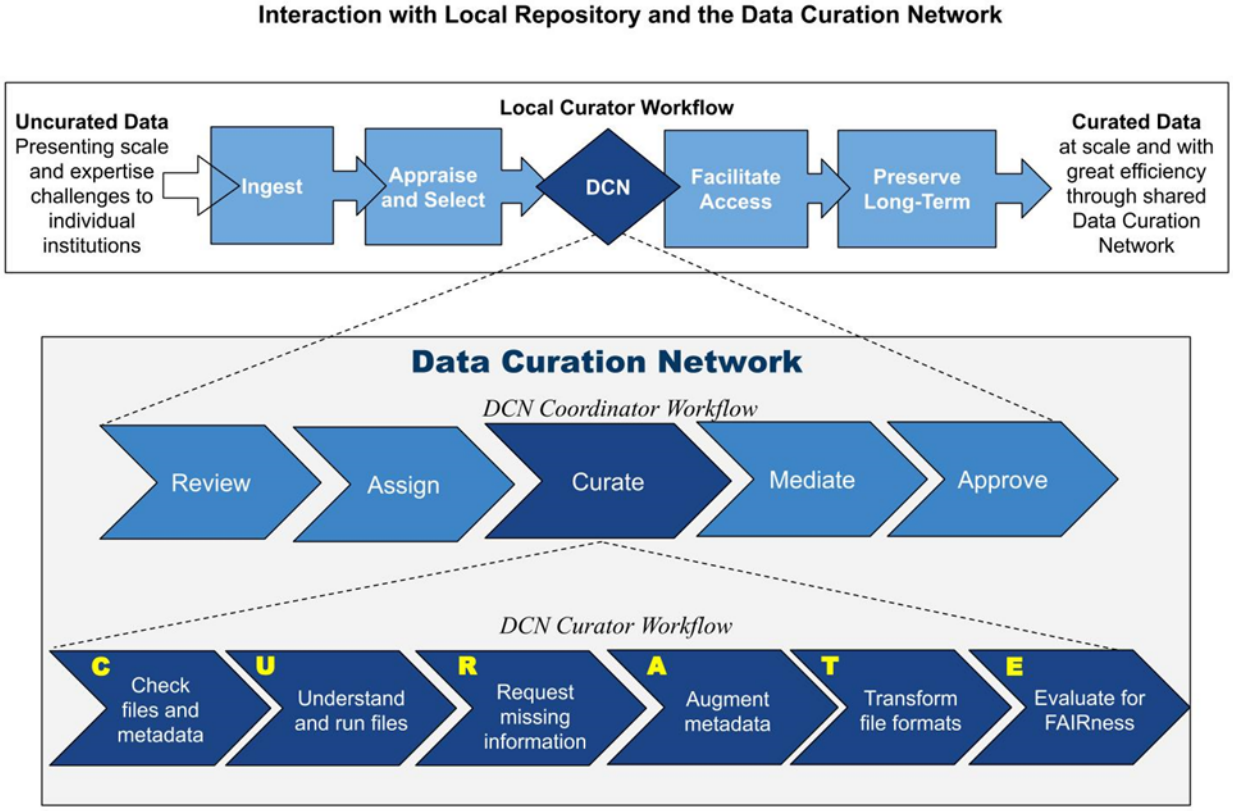

Figure 2: Diagram illustrating the DCN workflow and CURATE steps[2]

In our first year, 74 data sets passed through the DCN workflow. Curation requests from partner institutions remained steady throughout the year, averaging around six submissions per month (Figure 4). With these results, we test four assumptions originally made prior to launching the DCN, plus two unwritten assumptions that we've only recently realized.

1 https://www.atlassian.com/software/jira
2 Figure 2 reproduced from DCN Model 2018, 132.

JeSLIB 2020; 9(1): e1186
https://doi.org/10.7191/jeslib.2020.1186

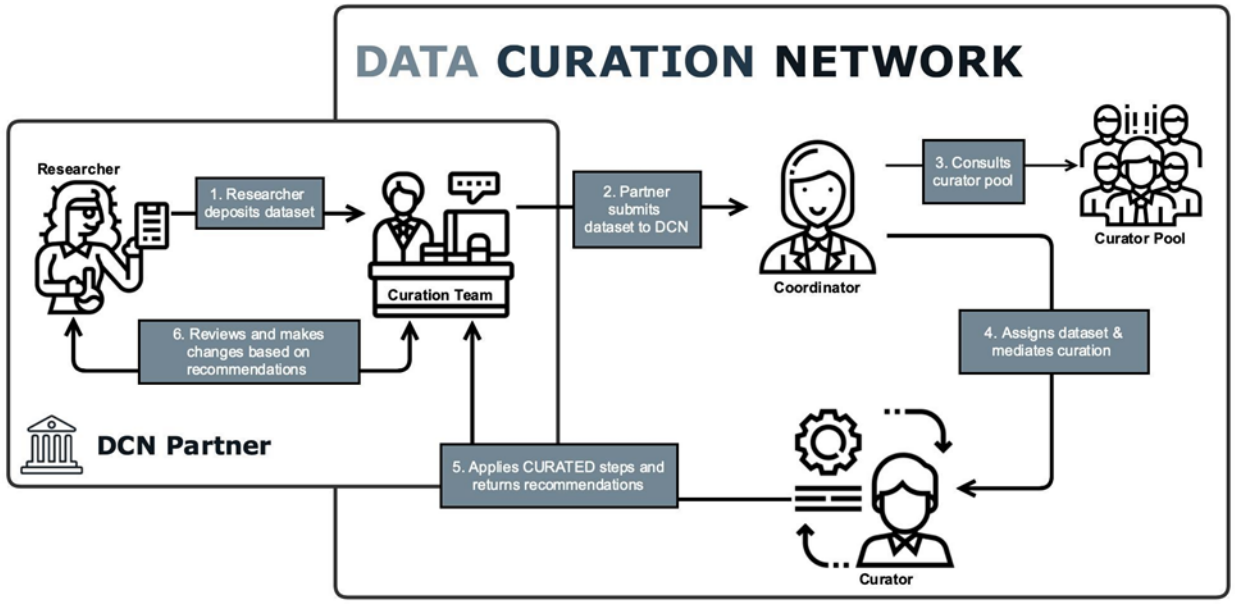

Figure 3: Diagram illustrating the collaboration between the DCN, DCN partners and researchers

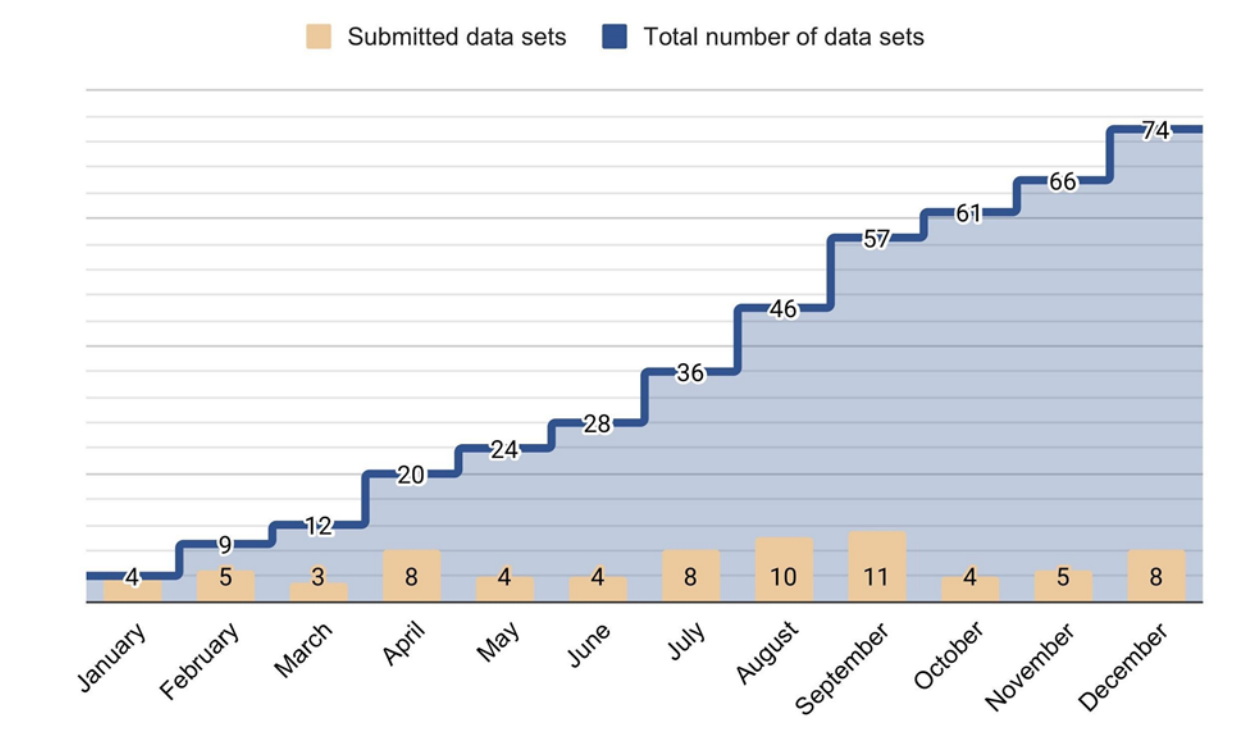

Figure 4: First 74 data sets curated by the DCN by month

JeSLIB 2020; 9(1): e1186
https://doi.org/10.7191/jeslib.2020.1186

Assumption #1: Many hands make light work…

> "Multiple data curation experts are needed to effectively curate the diverse data types a repository typically receives…" (DCN Model 2018, 127)

In a lot of ways, this was the main premise of the project. Looking at the results of year one, we saw a wide variety of data in terms of domain and data type. The largest proportion of data sets were from the life sciences and physical sciences/ mathematics (Figure 5). The most heavily represented disciplines in those subject areas were ecology and evolutionary biology (n=8), earth sciences (n=8), chemistry (n=5), and plant sciences (n=4).

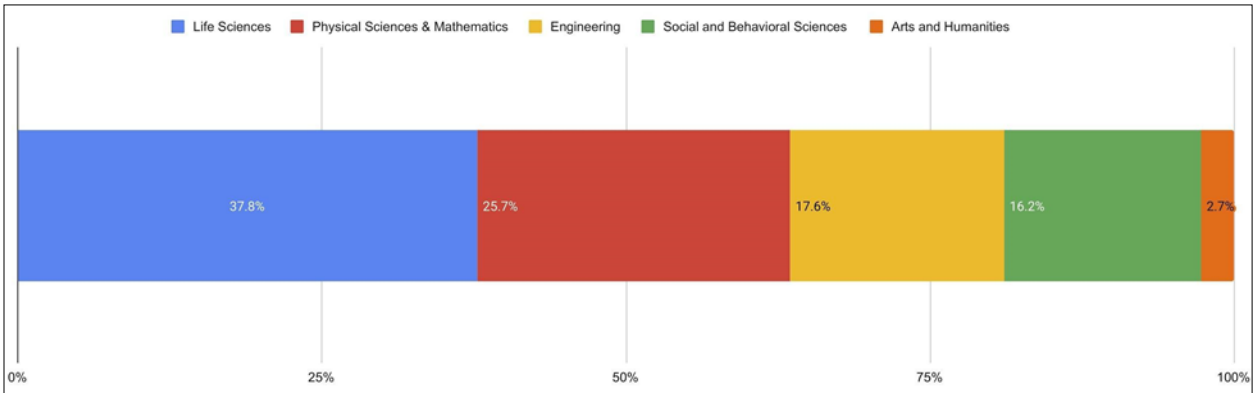

Figure 5: Breakdown of subject areas for the first 74 data sets curated by the DCN

Illustrated in figures 6-10 are comparisons of the total number of datasets submitted to the DCN for a particular discipline (blue) relative to the total number of DCN curators possessing the necessary expertise to curate data for each discipline (orange). Although the number of data sets versus the number of curators ("expert count") is not a direct one-to-one relationship, these illustrations provide an at-a-glance perspective of how well our curation strengths and weaknesses match demand.

The file types found in a data set play a major factor in the DCN's capacity, as anticipated by the DCN Model. Data sets often contain multiple data types, so we focused on tracking the primary data file type. Code was found most frequently, followed by tabular data in data sets submitted to the DCN in year one (Figure 11). MATLAB was the most common programming language for code (Figure 12).

It is clear, in reviewing the data from year one, that in order to successfully curate the wide variety (in terms of discipline, data type and file format) of data sets submitted, the Network requires a diversity of expertise. It will be interesting to see if the trends in these commonly-seen disciplines, data types, and file formats continue, or shift over time.

JeSLIB 2020; 9(1): e1186
https://doi.org/10.7191/jeslib.2020.1186

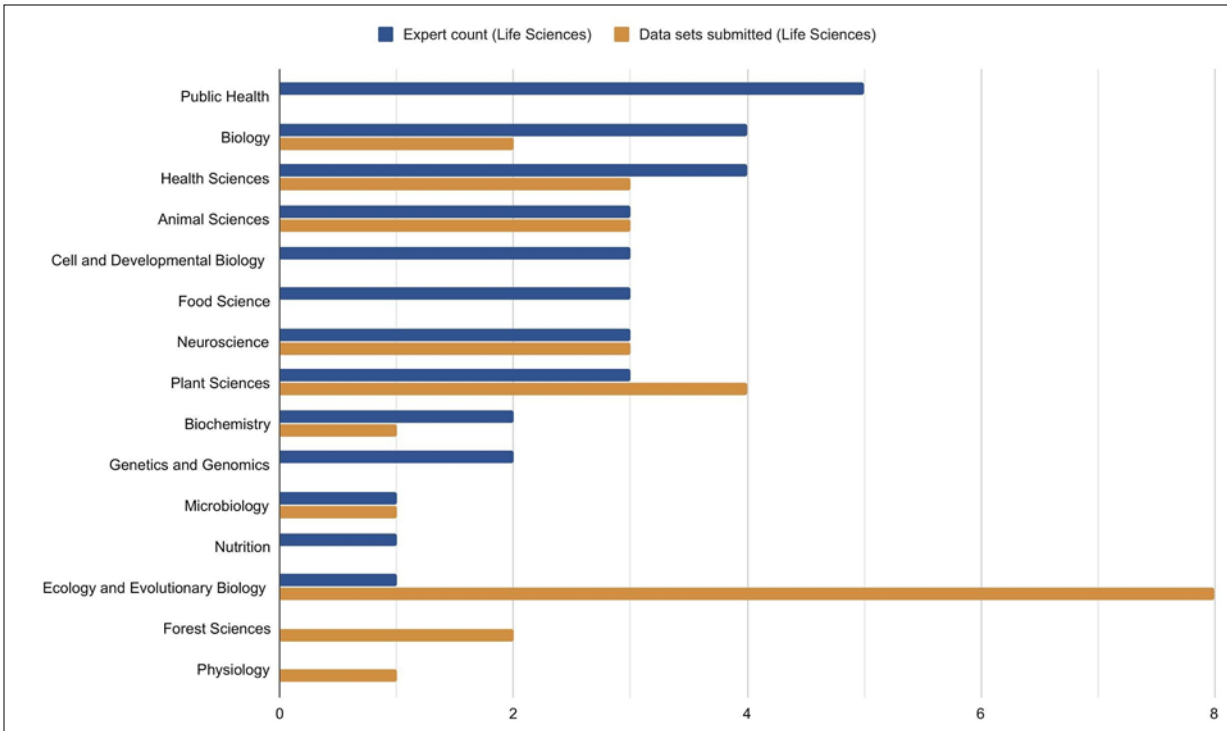

Figure 6: Life Sciences data sets submitted versus DCN curators with disciplinary expertise

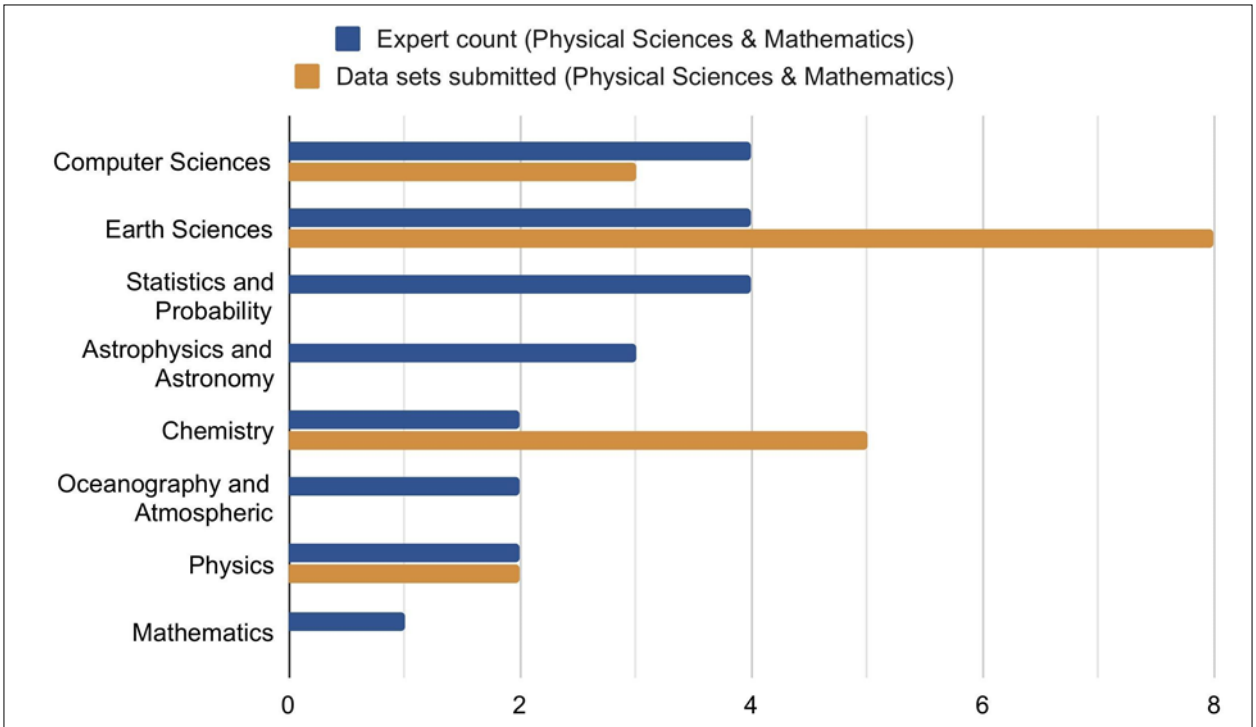

Figure 7: Physical Sciences & Mathematics data sets submitted versus DCN curators with disciplinary expertise

JeSLIB 2020; 9(1): e1186
https://doi.org/10.7191/jeslib.2020.1186

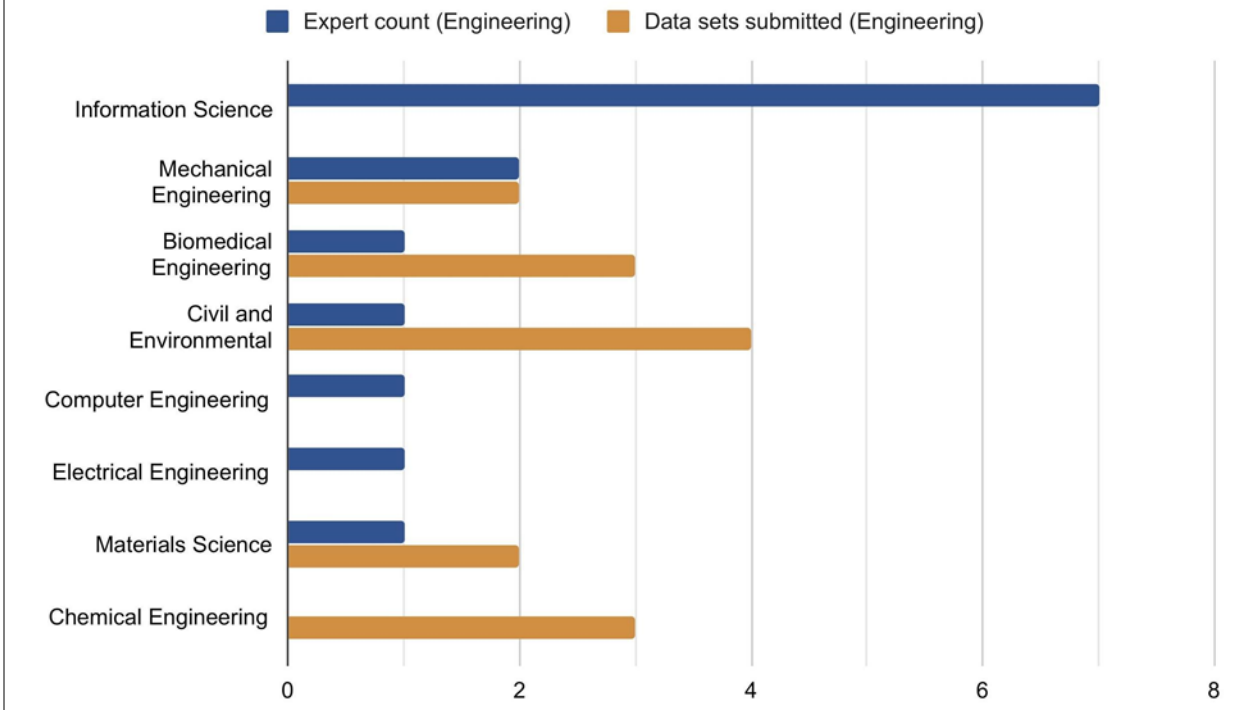

Figure 8: Engineering data sets submitted versus DCN curators with disciplinary expertise

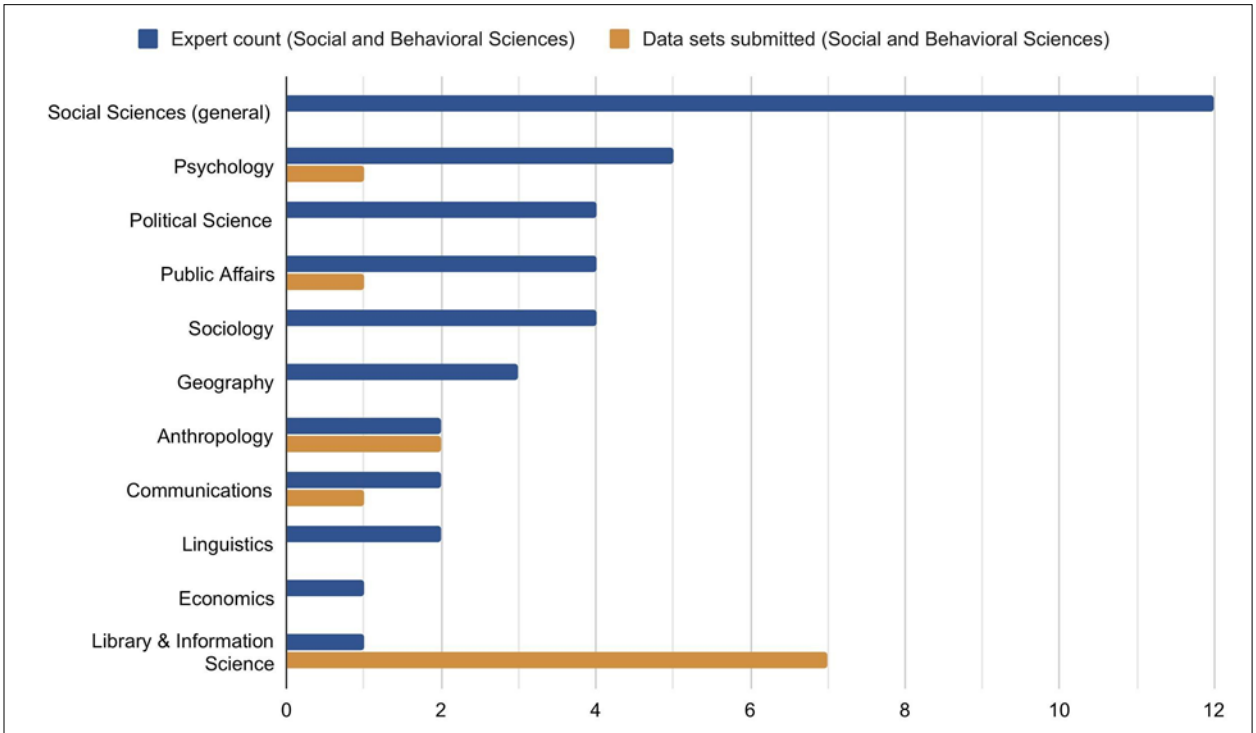

Figure 9: Social and Behavioral Sciences data sets submitted versus DCN curators with disciplinary expertise

JeSLIB 2020; 9(1): e1186
https://doi.org/10.7191/jeslib.2020.1186

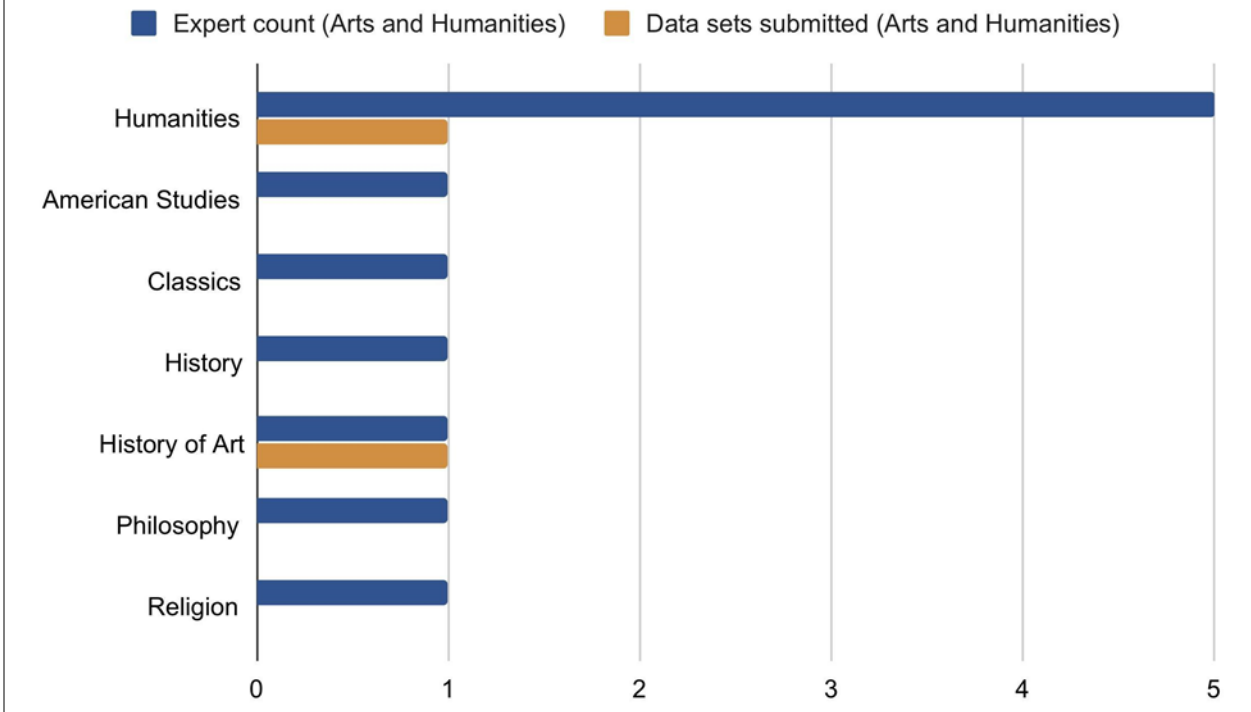

Figure 10: Arts and Humanities data sets submitted versus DCN curators with disciplinary expertise

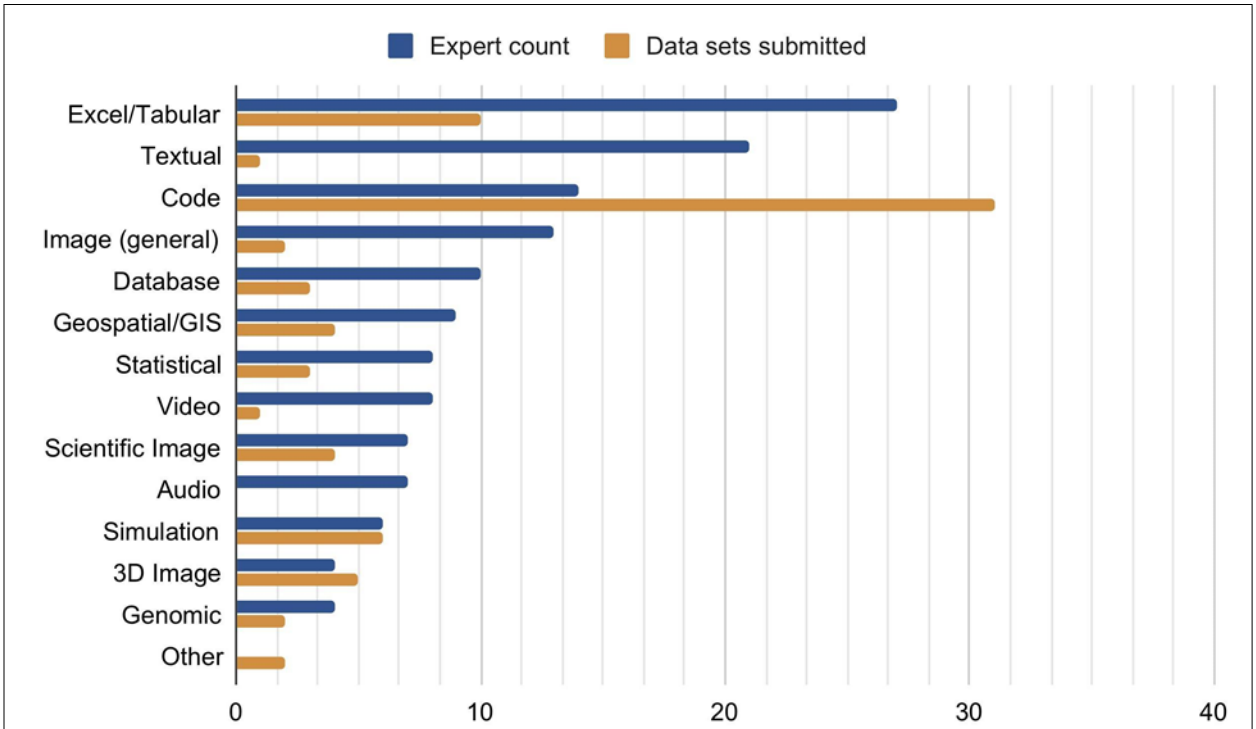

Figure 11: Data sets submitted compared to DCN expertise (count of DCN curators) by data type

JeSLIB 2020; 9(1): e1186
https://doi.org/10.7191/jeslib.2020.1186

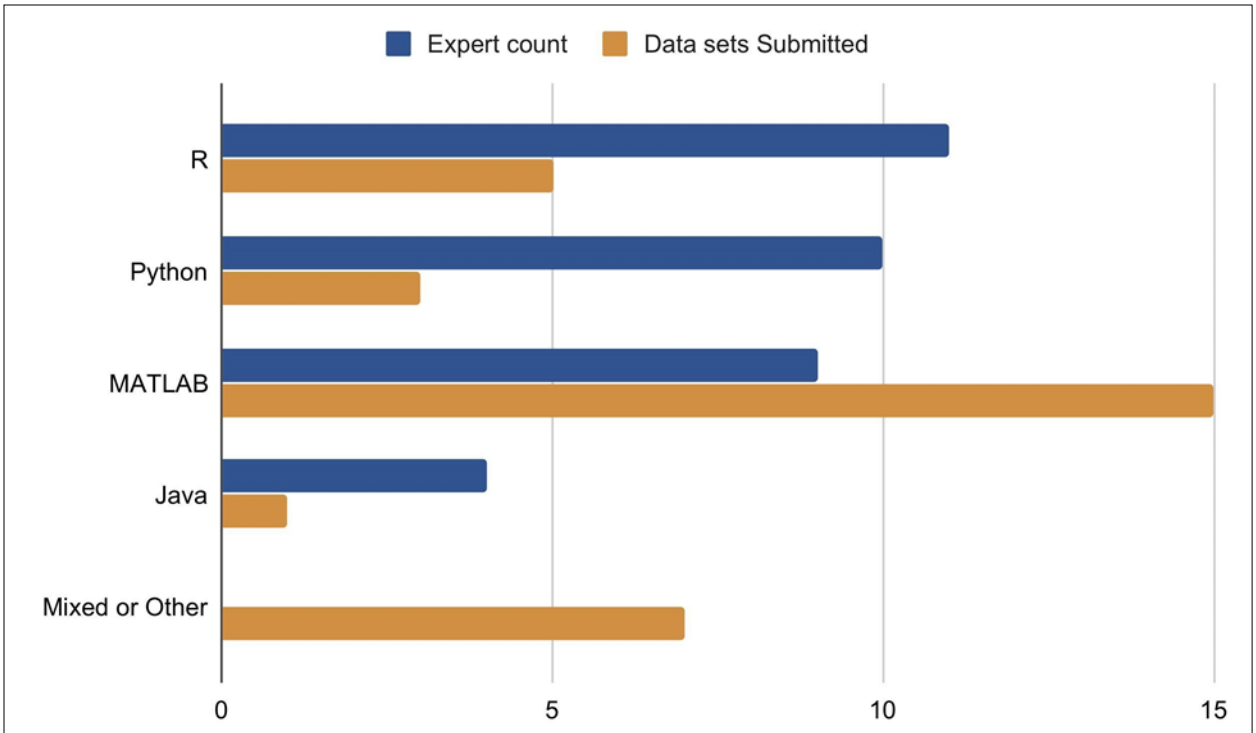

Figure 12: Code data sets submitted compared to DCN expertise (count of DCN curators) by programming language

It is clear, in reviewing the data from year one, that in order to successfully curate the wide variety (in terms of discipline, data type and file format) of data sets submitted, the Network requires a diversity of expertise. It will be interesting to see if the trends in these commonly-seen disciplines, data types, and file formats continue, or shift over time.

Assumption #2: Different Problems, One Shared Solution...

> "[DCN partners] will benefit from a collective approach that will allow them to supplement at peak times, access specialized capacity when infrequently-curated types arise, and stabilize service levels to account for local staff transition, such as during turn-over periods." (DCN Model 2018, 125-126)

Our partners decide which data sets to submit to the DCN, and when. Often data sets were submitted for more than one reason, the most common being a lack of technical expertise, a lack of local capacity, or a lack of disciplinary expertise (Figure 13).

JeSLIB 2020; 9(1): e1186
https://doi.org/10.7191/jeslib.2020.1186

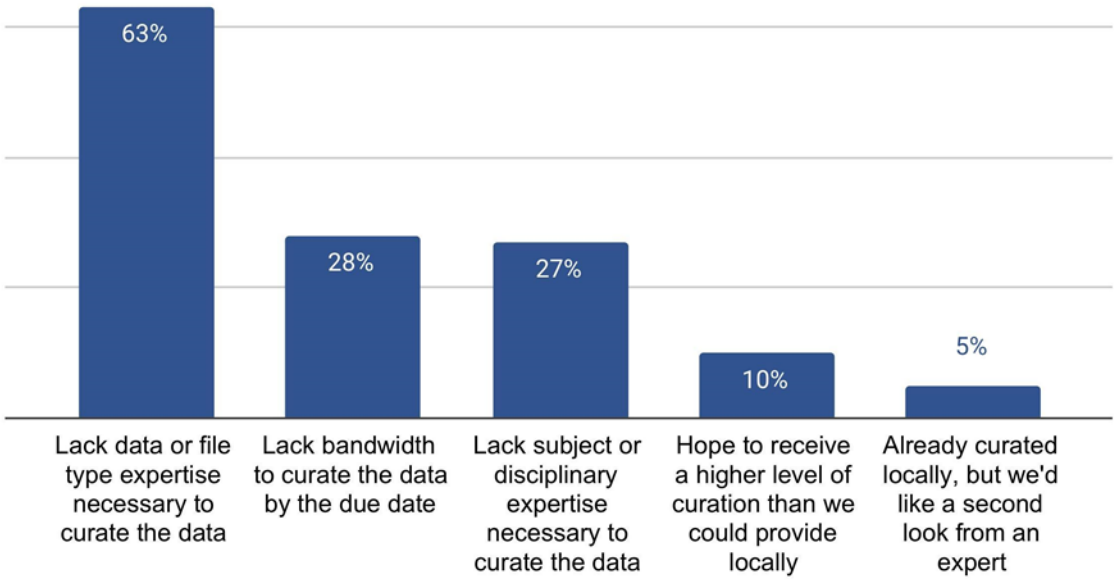

Figure 13: Reasons DCN partners submitted the first 74 data sets to the DCN (n=100 as multiple reasons were possible)

Assumption #3: It takes a village…

> "The implementation phase of the DCN will track trends in the types of domains or file types that come to the Network and work to recruit new institutions that might fill any gaps in expertise support. Capacity for curating data in the Network will grow as new partners join." (DCN Model 2018, 134)

We launched our pilot service with 21 data curators, each contributing between 1% and 5% of their FTE[3] staff time to the project (1,555 curation hours per year). By the end of year one we'd grown this number to 27 data curators as a result staffing transitions, new hires, and the addition of two new partner institutions midway through the year (Figure 14). This translates to about 2,000 data curation hours.

As the number of curators increases, our curation capacity increases; however, the collective expertise of the DCN fluctuates with each staff change. As a result, the DCN cannot always rely on retaining specific expertise and the DCN may at times become oversaturated with certain types of expertise (as illustrated in Figures 6-12). While demand also tends to be unpredictable, and prone to fluctuations, we're closely tracking trends in submissions and we plan to target our recruitment efforts for next year, accordingly.

3 FTE stands for Full-time equivalent. See: https://en.wikipedia.org/wiki/Full-time_equivalent

JeSLIB 2020; 9(1): e1186
https://doi.org/10.7191/jeslib.2020.1186

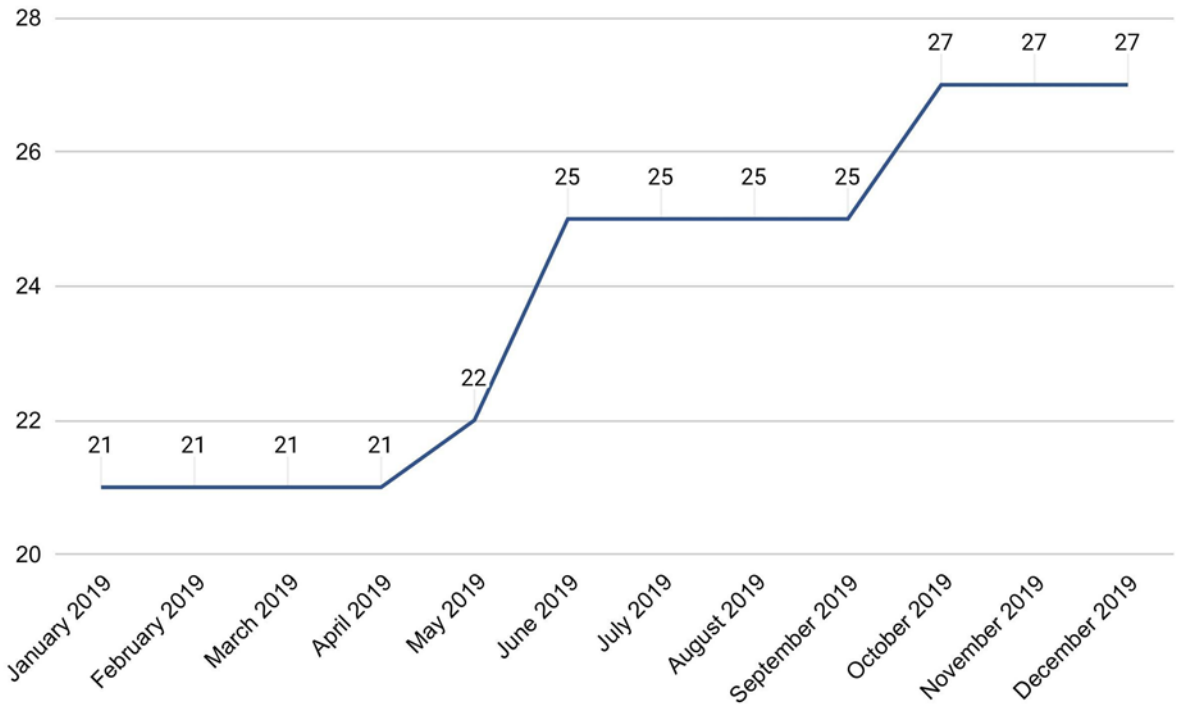

Figure 14: Year one growth in the number of curators in the DCN since launch of pilot service

Assumption #4: Good things take time...

> "...we found from our 2016-2017 metric tracking that curators spend an average of two hours to curate a data set (ranging from less than one hour to more than eight hours)." (DCN Model 2018, 134)

Consistent with our model report, our first 74 data sets showed DCN curators spending an average of 2.4 curation hours per data set (ranging from 0.5 hours min to 6 hours max). The total curation time for all 74 datasets was 167.5 hours. Based on our annual capacity (approximately 2,000 curation hours) we could curate roughly 833 datasets per year. However, this is unlikely as this figure does not take into account the variability of data sets submitted, the fluctuations in Network expertise, or curator availability.

Typically, data sets submitted to the DCN are given a due date of five business days. Curators almost always met their curation due date (only 1 in 74 assignments went past due), and often well in advance of the deadline (see an example month in Figure 16). The median turnaround time for data sets in year one was three days.

JeSLIB 2020; 9(1): e1186
https://doi.org/10.7191/jeslib.2020.1186

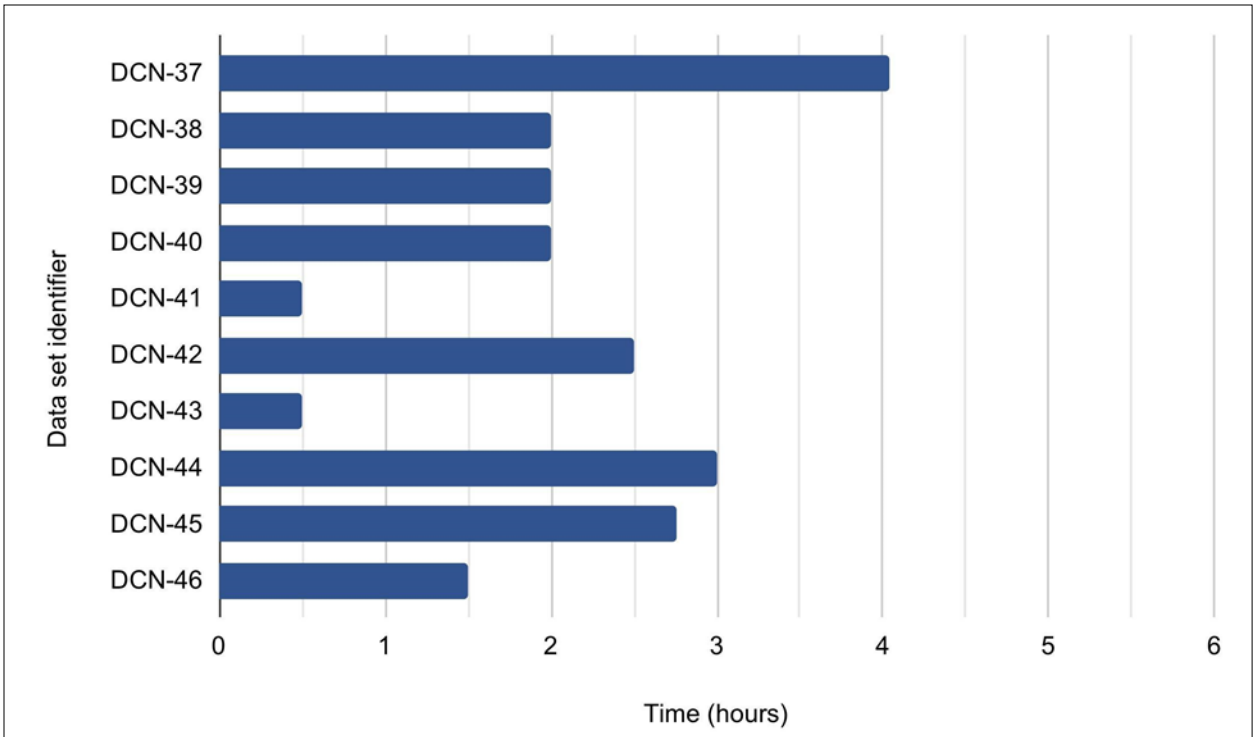

Figure 15: Amount of time (in hours) DCN curators spent curating August 2019 data sets

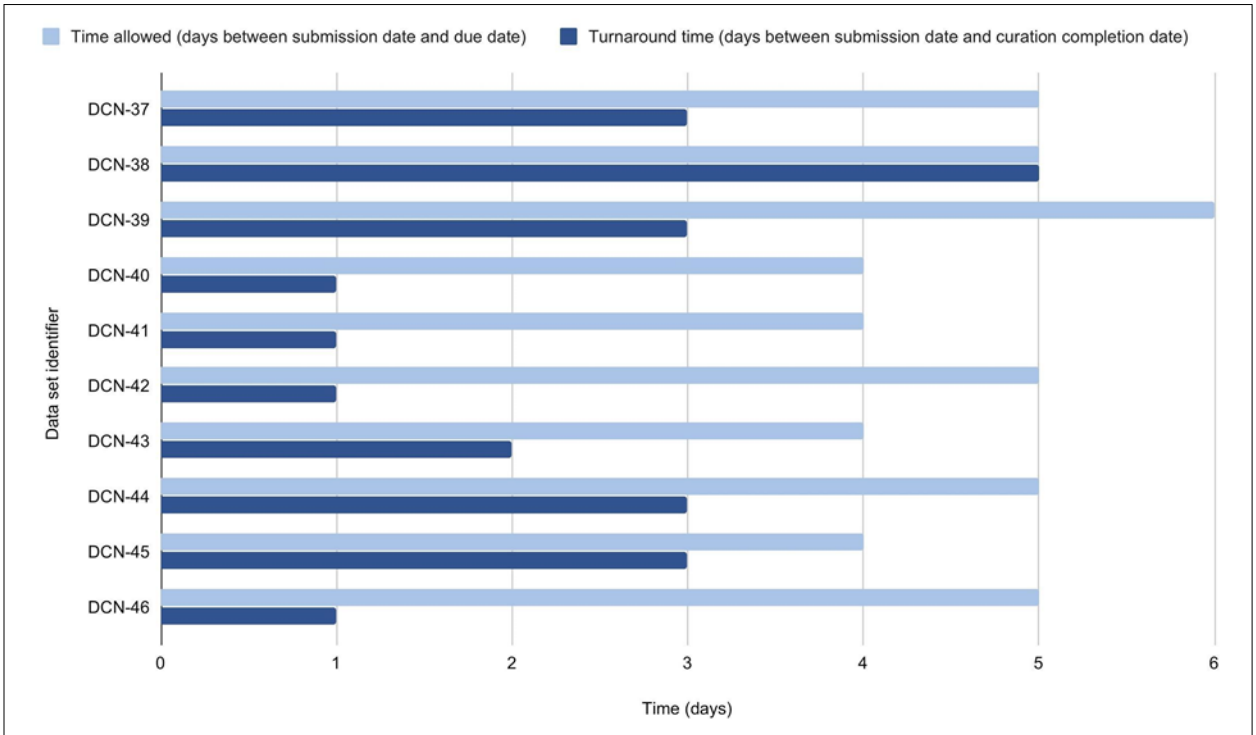

Figure 16: DCN turnaround time for the month of August, 2019

JeSLIB 2020; 9(1): e1186
https://doi.org/10.7191/jeslib.2020.1186

Unwritten Assumption #1: Matchmaking is easy…

The coordination of this process can be complicated by unpredictable outages, quick turnaround times, curator availability, and other factors. If a curator chooses to accept an assignment, they have until an expected due date (usually five days from submission) to curate the data set using our standardized CURATE steps[4] while tracking their work in Jira.

When a data set is submitted to the DCN, it is assigned to a curator as quickly as possible. Assignments are typically made through Jira and curators receive an email notification of each assignment. Curators are asked to respond to assignments (accepting or declining) as soon as possible or within 24 hours. After 24 hours data sets are reassigned, if possible, or in rare cases they are returned to the submitting institution. In year one, the median curator response time was about one hour. And on four occasions, a data set could not be assigned and was returned to the submitting institution for local curation.

The DCN coordinator's primary responsibility is to make sure the Network's curation activities run smoothly. The coordinator matches each dataset with an appropriate curator, and then mediates the curation process from start to finish. Among other things, this requires constant communication in order to maintain good rapport and fine-tune workflows. The DCN maintains multiple communication channels. In addition to Jira, we use Slack[5] (instant messaging), Google Groups (email), utilize surveys, and hold virtual meetings (weekly stand-ups and periodic meetings with each partner team). So far, these coordination methods have been successful based on survey feedback from curators.

Unwritten Assumption #2: Take a penny, leave a penny…

Although it doesn't appear in our model report, we had assumed a balance in how partners use and contribute to the network. For example, institutions that dedicate more curator time to the Network would receive more curator time from the Network. However, this balance was not always realized (Figure 17). The reasons for inequity are complex given the variability of curation time and expertise needed to curate data sets, the differences amongst partners in the volume of data received (Figures 18 and 19), the timing of submissions, local curation staff and local curation capacity, as well as local workflows and the actual or perceived barriers to submitting data sets to the DCN.

4 Table 2 Appendix of DCN Model report 2018, 139.
5 https://slack.com

JeSLIB 2020; 9(1): e1186
https://doi.org/10.7191/jeslib.2020.1186

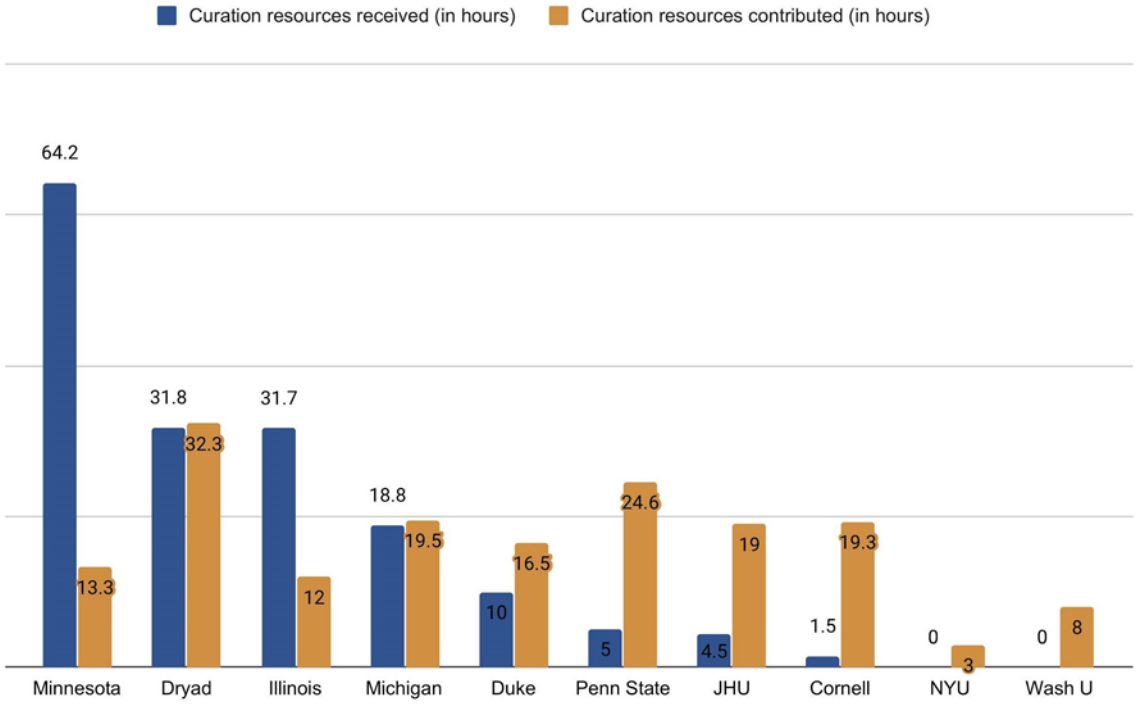

Figure 17: 2019 resources received by each partner compared to resources contributed by each partner (ex. Hours spent by DCN partners curating Minnesota's data sets compared to hours spent by Minnesota curating DCN partner data sets)

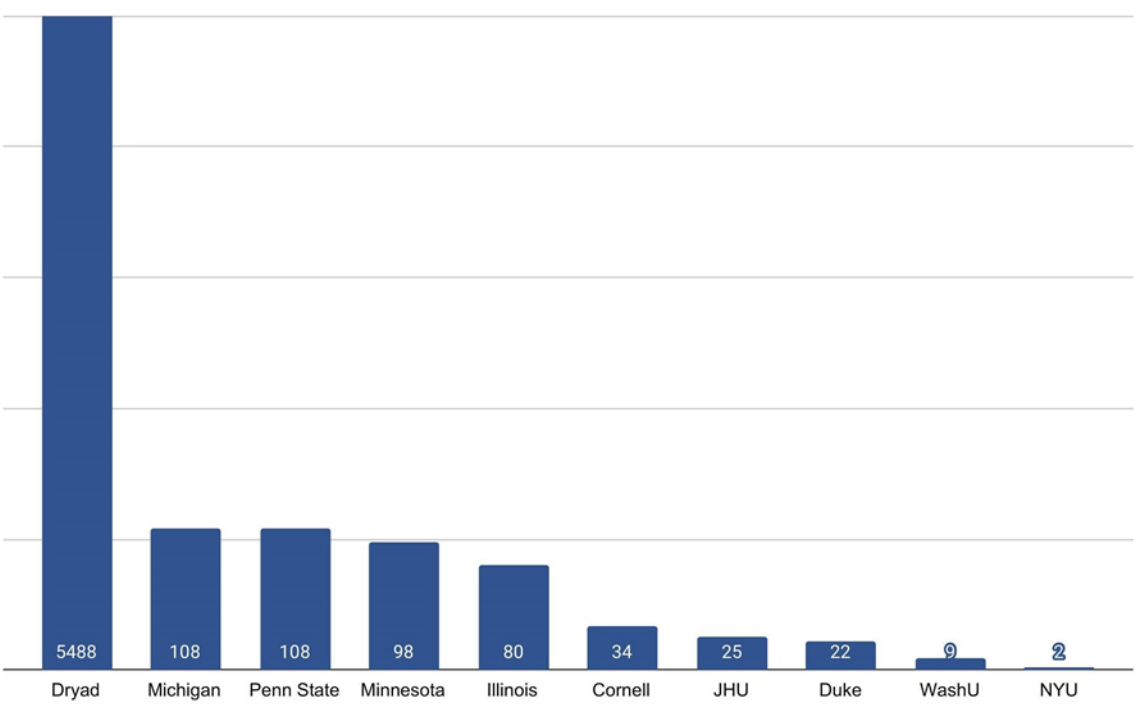

Figure 18: Total data sets received in 2019 by DCN partners *(Dryad not shown to scale)*

JeSLIB 2020; 9(1): e1186
https://doi.org/10.7191/jeslib.2020.1186

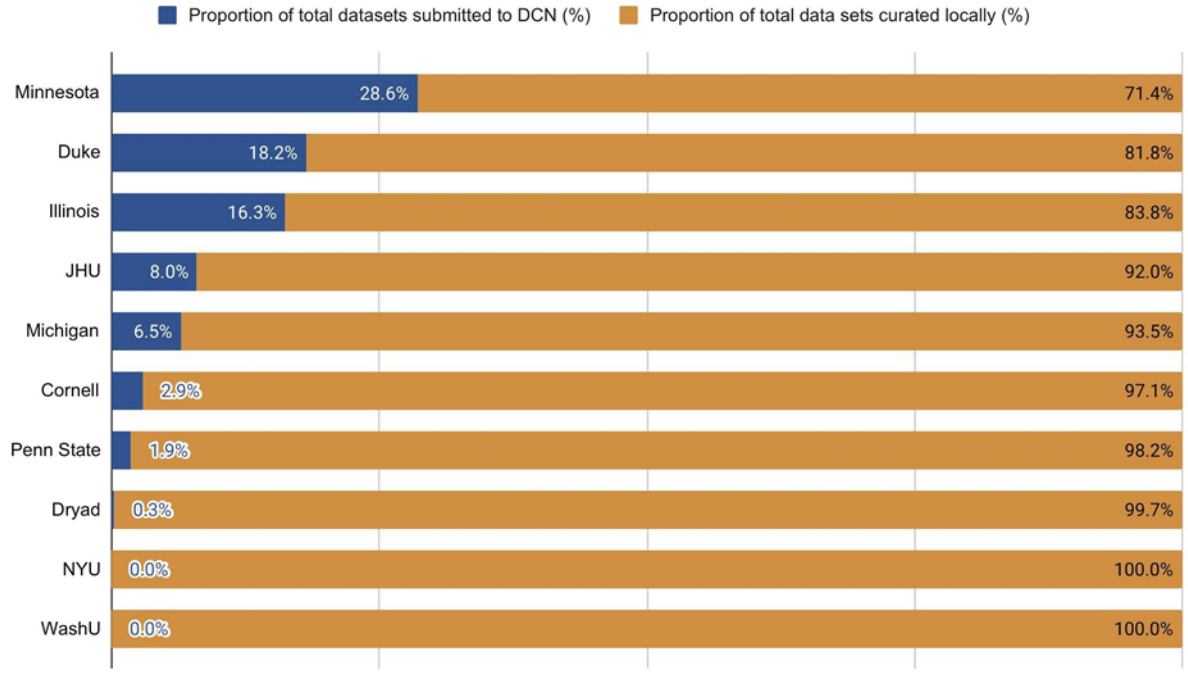

Figure 19: Proportion of total data sets received by partners that were submitted to the DCN for curation in 2019

Conclusion

We've made a lot of progress in our first year of testing the Network. With just over a year remaining in the project's implementation phase, we will continue to monitor the Network and to adapt our approach in hopes of becoming more efficient. Here are some of the changes happening in year two based on what we've learned so far:

1. Our ability to successfully assign and curate data sets in the DCN depends on several factors including the availability and expertise of curators. In the first 74 data sets submitted, we already see some gaps in expertise. For example, 10% of the data sets we received were from ecology and environmental sciences. While the DCN launched with three data curators with ecology expertise, including two with advanced degrees, we lost two of these curators due to normal staff turnover, and their replacements did not have the same backgrounds. Therefore, it will be difficult for our distributed project to rely on static expertise. One idea to explore is whether the DCN can help recruit and fund specific (needed) data curator roles that would collectively benefit the community, though this would require more in-depth exploration and buy-in from the partner members.

2. The ability to match a data set with an available curator is a key factor in our project's success. The DCN Coordinator plays a vital role in getting to know the curators, understanding their interests and expertise, and creating a vibrant virtual community for the DCN to flourish. For example, since the DCN positions the curator to accept or reject a particular assignment, and

JeSLIB 2020; 9(1): e1186
https://doi.org/10.7191/jeslib.2020.1186

allows them 24 hours to respond to an assignment, the time it takes to make a successful match can range from as little as an hour to as much as many days. We find that the coordinator must be in close contact with the curators to better gauge not only their skill levels, but also if the timing is right for a new assignment.

3. Curator satisfaction is a key concern of the project. For example, some curators may appreciate the exchange of knowledge by curating data from other institutions thereby freeing up time to focus on their own specialty area, while others may feel that this work is "in addition" to their regular load. Ensuring that curators' work is transparent and recognized is a core value of the DCN. One way we are trying to highlight the people behind the curation is through our website (Figure 20), where we tie curator expertise to the data sets they've curated (with a list of generalized actions taken). Exploring, understanding and addressing curator satisfaction and engagement will be a key activity in year two.

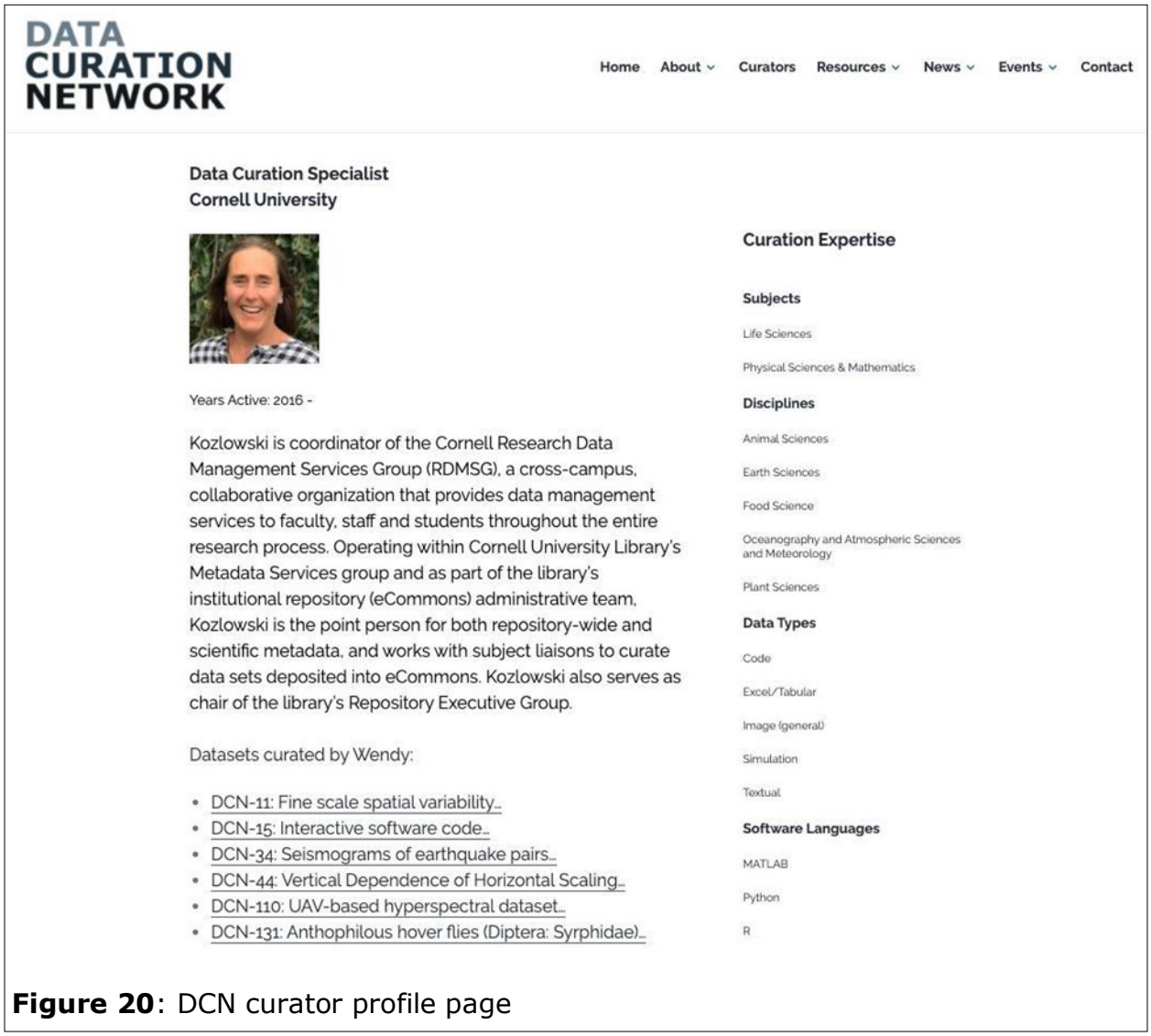

Figure 20: DCN curator profile page

JeSLIB 2020; 9(1): e1186
https://doi.org/10.7191/jeslib.2020.1186

4. Equitable distribution of Network resources amongst partners was a foundational assumption of the project. Perhaps the most important and common factors contributing to any imbalance are the domain and data type of the data set. Most of the data sets submitted to the DCN in year one contained software code. Code expertise, like most types of disciplinary and data type expertise, isn't equitably distributed across the DCN. This is precisely why the DCN exists, but we would like to explore any perceptions of imbalance more in year 2.

Finally, we were pleased that our experience in our first year revealed a less tangible benefit—that of being a part of a larger community for data curators. The individuals participating in the DCN project work together on common challenges, such as curating human subjects data, curating big data, and advocating for the value of curation to researchers and the larger data sharing community. We also work to expand capacity for data curation by offering workshops on specialized data curation and providing a platform for our peers to share their expertise through data curation primers—further expanding and solidifying this community. The DCN has enabled data repositories to collectively curate a wider variety of data than we could each curate alone, but along the way we have also formed a community of professional data curators who share our passion for enabling the reuse of research data.

Acknowledgments

We would like to thank the Alfred P. Sloan Foundation for its generous support of our project (Primary Award: G-2018-10072; Planning Award: G-2016-7044).

We would like to thank all past and current contributors to the Data Curation Network and acknowledge the time and ongoing commitment of this amazing community: Aditya Ranganath, Alexis Logsdon, Alicia Hofelich Mohr, Andrew Battista, Ashley Hetrick, Chen Chiu, Claire Stewart, Cynthia Hudson-Vitale, Dave Fearon, Debra Fagan, Dorris Scott, Elizabeth Hull, Erica Johns, Erin Clary, Henrik Spoon, Hannah Hadley, Heidi Imker, Hoa Luong, Jake Carlson, Janice Jaguszewski, Jennifer Darragh, Jennifer Moore, Joel Herndon, John Russell, Katie Wilson, Katie Wissel, Mara Blake, Marley Kalt, Melinda Kernik, Nathan Piekielek, Rachel Woodbrook, Robert Olendorf, Rich Yaxley, Sarah Wright, Seth Erickson, Shanda Hunt, Sophia Lafferty-Hess, Susan Borda, Tim McGeary, Tracy Teal, Valerie Collins, Wanda Marsolek, Wendy Kozlowski and Xuying Xin.

Data Availability

Data associated with this article are available from the Data Repository for the University of Minnesota at: https://doi.org/10.13020/ak4d-ge34.

References

Coburn, Elizabeth and Lisa R. Johnston. 2020. "Data supporting: 'Testing Our Assumptions: Preliminary Results from the Data Curation Network.'" Retrieved from the Data Repository for the University of Minnesota. https://doi.org/10.13020/ak4d-ge34

Johnston, Lisa R., Wendy Kozlowski, Cynthia Hudson-Vitale, Elizabeth Hull, Mara Blake, Claire Stewart, Jake Carlson, et al. 2018. "Data Curation Network: A Cross-Institutional Staffing Model for Curating Research Data." *International Journal of Digital Curation* 13(1): 125-140. http://dx.doi.org/10.2218/ijdc.v13i1.616