Data Curation Network Special Report
"Results of the Fall 2016 Researcher Engagement Sessions"
Version 1.0
Release Date: March 9, 2017
Authors: Lisa R Johnston (PI), Jake Carlson, Cynthia Hudson-Vitale, Heidi Imker, Wendy Kozlowski, Robert Olendorf, and Claire Stewart.

Introduction

The Data Curation Network is a project to develop a shared staffing model for curating research data that draws from the expertise across multiple institutions. This network model will allow institutions to broaden the depth and breadth of curation services beyond what a single institution might offer alone. The project planning phase began with a one-year grant from the Alfred P. Sloan foundation in May 2016. The results presented here represent an activity conducted in the DCN's first year to seek input from researchers on how data curation services fit into their research workflow and data management needs. Our project reports and outcomes are posted to the project website https://sites.google.com/site/datacurationnetwork.

In the fall of 2016 the Data Curation Network team held six focus group sessions across six academic institutions to determine what data curation activities were important for researchers, what activities where they applying themselves, and how satisfied were they with the results of those efforts. In short, we aimed to identify the challenges faced by researchers with regard to data curation. As an outcome of these focus group sessions, the process uncovered several "gaps" in highly valued data curation activities that researchers do not engage in for their data (or do not engage in as satisfactory as they would like to). These potentially "high impact" or "value add" activities will be of particular importance for the Data Curation Network once implemented.

The results of this research will allow libraries to develop more focused and useful services for their researchers. It will also help them gain a better understanding about the importance of data curation activities from the perspective of researchers, to what extent researchers value data curation activities, how data curation activities are valued differently across disciplines, and where the greatest gaps of support for highly valued data curation activities may fall.

Literature Review

The role of data curation was still an emerging topic within the library science, archival, and information sciences disciplines just a few years ago and very few academic libraries were successfully offering data curation services at all according to a study on research data services in academic libraries in 2011.[1] More recently Kouper et. al. (2017) provide an empirical analysis of research data services by North American Research libraries' (ARL).[2] Their findings indicate that the concept of data curation is found in less than 15% of institutions surveyed and is typically viewed as an advanced library service.

While studies of researcher attitudes toward data curation and management is not new, many focus high level curation services and data management needs (see McLure et. al., 2014 and Parham et. al., 2012),

[1] Tenopir, C., Birch, B., and Allard, S., Academic Libraries and Research Data Services: Current Practices and Plans for the Future. ACRL White Paper, June 2012 http://www.ala.org/acrl/sites/ala.org.acrl/files/content/publications/whitepapers/Tenopir_Birch_Allard.pdf
[2] Kouper, I., Ishida, M., Williams, S., Fear, K., Kollen, C. (2017). "Research Data Services Maturity in Academic Libraries." in Curating Research Data (Lisa R. Johnston, Ed.), American Library Association, Association of College and Research Libraries. http://hdl.handle.net/10150/622168.

without going into great detail on specific treatments and activities for curating the digital asset.[3] Many of these surveys use existing tools and frameworks for assessing faculty needs, such as the Data Curation Profiles or the Data Asset Framework.[4] While useful tools for assessing needs for institutional RDS, they lack a mechanism to collect feedback on researchers current practices for these treatments and a self-assessment on their satisfaction for these treatments. One CLIR study however does approach researcher attitudes directly and provides a number of comparable insights to this study.[5]

Along a parallel path, much has been written regarding the “competencies” for data curators and librarians working with data. Research by Madrid surveyed multiple panels of experts, using the Delphi Method, to develop consensus around competencies for digital curators.[6] Results of this research identified twenty high-level competencies for digital curators, including: “plans, implements, and monitors digital curation projects”; “selects and appraises digital documents for long term preservation”; and “verifies the provenance of the data to be preserved and ensures that it is properly documented”; among others. Librarians who work specifically with data have been found to need similar skills by Schmidt & Shearer.[7]

To better define the activities involved with data curation, the DCN team reviewed work by the DigCcurr program which provides a useful matrix of curation themes and ideas, however without sufficiently detailed definitions.[8] Work by Bowden et. al. focused on curator views of the DigCurr activities in the Digital curation gap project.[9] Their focus groups provided a good template for our work with researchers. In order to incorporate activities important to the digital repository community, the TRAC assessment tool provided some insight, but the language was jargon-laden and does include a researcher assessment of needs.[10] And the Digital Curation Center data lifecycle model and several reference sources (listed in Appendix A) paved the way for defining the Data Curation Activities used in our research.[11]

[3] McLure, M. & Level, A. V. & Cranston, C. L. & Oehlerts, B. & Culbertson, M. "Data Curation: A Study of Researcher Practices and Needs." portal: Libraries and the Academy, vol. 14 no. 2, 2014, pp. 139-164. Project MUSE, doi:10.1353/pla.2014.0009; Susan Wells Parham, Jon Bodnar, and Sara Fuchs Supporting tomorrow’s research: Assessing faculty data curation needs at Georgia Tech Coll. res. libr. news January 2012 73:10-13, http://crln.acrl.org/content/73/1/10.short.

[4] Witt, M., Carlson, J., Brandt, D. S., & Cragin, M. H. (2009). Constructing data curation profiles. International Journal of Digital Curation, 4(3), 93-103, doi:10.2218/ijdc.v4i3.117; Jones, S., Ball, A., & Ekmekcioglu, Ç. (2008). The data audit framework: A first step in the data management challenge. International Journal of Digital Curation, 3(2), 112-120, doi:10.2218/ijdc.v3i2.62.

[5] Jahnke, L. M., Asher, A., & Keralis, S. (2012). The problem of data: data management and curation practices among university researchers. Council on Library and Information Resources, Washington, DC. Chicago, https://www.clir.org/pubs/reports/pub154/pub154.pdf.

[6] Madrid, M. M. (2013). A Study of Digital Curator Competences: A survey of experts. The International Information & Library Review, 45(3), 149-156. http://dx.doi.org/10.1016/j.iilr.2013.09.001.

[7] Schmidt, B & Shearer, K.. Librarians' Competencies Profile for Research Data Management. Joint Task Force on Librarians’ Competencies for E-Research and Scholarly Communication. June 2016. https://www.coar-repositories.org/files/Competencies-for-RDM_June-2016.pdf.

[8] Lee, C. 2009. Matrix of Digital Curation Knowledge and Competencies (Overview), June 17, 2009 (Version 13), DigCCurr Project, https://ils.unc.edu/digccurr/digccurr-matrix.html.

[9] Bowden, H., Lee, C., Tibbo, H. “Closing the Digital Curation Gap Focus Groups Report,” June 28, 2011, London, UK. http://digitalcurationexchange.org/cdcg/sites/default/files/CDCG_FocusGroupReport.pdf.

[10] Center for Research Libraries (CRL) and Online Computer Library Center (OCLC). Trustworthy Repositories Audit and Certification (TRAC): Criteria and Checklist. Chicago: CRL and Dublin, OH: OCLC, http://www.crl.edu/sites/default/files/d6/attachments/pages/trac_0.pdf.

[11] Digital Curation Center (DCC), “DCC Curation Lifecycle Model,”

Methodology

Between October 21, 2016 and November 18, 2016 the authors of this report engaged 91 researchers across six focus group sessions, termed as "Data Curation Roundtable" sessions, held at the following academic institutions: Cornell University, Penn State University, University of Illinois at Urbana-Champaign, University of Michigan, University of Minnesota, and Washington University in St. Louis. The participants represented a good mix of experience (faculty, graduate student, post-doc) and discipline (see table 1). Each session lasted 1 ½ hours over lunch, which was provided by the DCN project in exchange for their participation.

Table 1: Disciplinary representation at the six researcher engagement sessions

Institution	Cornell	Wash U	Illinois	Penn State	Minnesota	Michigan	Totals
Date of Session	2016-10-11	2016-10-25	2016-10-27	2016-11-04	2016-11-14	2016-11-18	**6**
Sciences & Engineering	9	6	10	5	11	8	**49**
Social Sciences	6	1	2	1	1	4	**15**
Humanities	0	1	1	1	0	2	**5**
Staff/Service Providers*	5	3	5	4	1	0	**16**
Medical	0	0	0	0	0	4	**4**
Total	20	11	18	11	13	16	**91**

*Service providers, such as campus-based IT staff and library staff, as well as library and information science faculty were grouped into this category.

These sessions sought to directly engage with the communities who produced data and those who are likely to make use of datasets (the designated community), to better understand the value of data curation. What the team learns from these sessions will be incorporated into the shared staffing curation model and will be used to ascertain the success of the DCN project. The goals of the focus group sessions were to answer these questions:

1. What data curation activities researchers see as important or having value to themselves or to their communities of practice?
2. How, to what extent, and why researchers engage in data curation activities themselves as a normative part of their research workflows?
3. What are the barriers preventing researchers from doing so (time, personnel, knowledge, money, equipment, other resources)?
4. What level of satisfaction do researchers have with their current data curation treatments?

By developing an understanding of what curation activities researchers value, the DCN will develop and deliver services that are in-line with real world needs and expectations.

Definitions of Data Curation Activities

In preparation for the session, the DCN team defined 47 data curation activities relevant to our curation services and best practices by consulting a number of sources and the full list of definitions are presented

http://www.dcc.ac.uk/resources/curation-lifecycle-model; for the history and development of this model see Sarah Higgins, "The DCC Curation Lifecycle Model," International Journal of Digital Curation 3, no. 1 (2008): 134–40, doi:10.2218/ijdc.v3i1.48, where data are defined on p137.

in Appendix A. In addition, several key definitions were presented that the beginning of each session; these terms were:

- Data Curation: The encompassing work and actions taken by curators of a data repository in order to provide meaningful and enduring access to data.
- Data Repository: A digital archive that provides services for the storage and retrieval of digital content.
- Data: Facts, measurements, recordings, records, or observations about the world collected by scientists and others, with a minimum of contextual interpretation. Data may be any format or medium (e.g., numbers, symbols, text, images, films, video, sound recordings, drawings, designs or other graphical representations, procedural manuals, forms, data processing algorithms, or statistical records.).

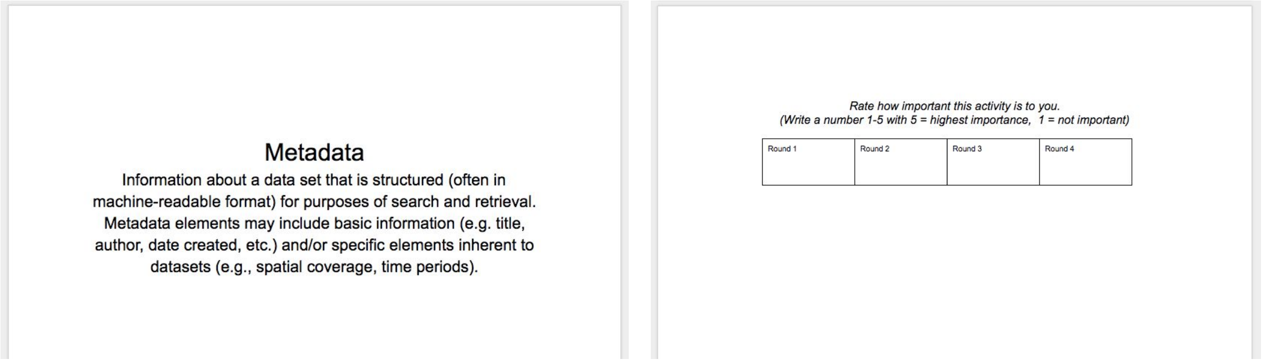
Each session was broken into three parts. First we used a card swapping and ranking exercise that asked researchers to rank the importance of data curation activities for their data. Second, we used a paper-based survey instrument to collection the researcher's' levels of engagement and satisfaction with those same data curation activities. Third, we engaged researchers in facilitated focus group discussion around the barriers and challenges of applying the top five most highly ranked data curation activities in their individual workflows. Each part is described in more detail below.

Part 1: Rating of Importance for Data Curation Activities

To address the first question, the DCN team first asked researchers to rank the importance of 35 data curation activities, and 35 out of the 47 were selected for this exercise.[12] Not all the activities were presented at each of the six session and it was up to the local DCN team member to select the subset of activities to focus on. To keep the exercise engaging, the activities were printed individually on a 5x8 card with the definition of the activity on the front and a score sheet on the back (see figure 1). The researcher was given 2-4 cards at a time and asked to read the definition and then rank that activity's importance from 1 (lowest) to 5 (highest). Once each card in their hand was ranked the researchers were asked to exchange their cards with another researcher in the room and repeat the ranking exercise for the new card. This was repeated for four rounds. Since there were 2 or 3 copies of the same card circulating around the room, researchers were advised to trade with those who had cards they had not ranked previously. After a quick total of all four rounds, this exercise provided our team with a priority list of data curation activities that were used as the focus of the group discussion through the session. The master card desk is provided as a supplemental file to this report.

[12] Twelve activities defined by the DCN were not ranked at any of the researcher engagement sessions; these were Arrangement and Description, Authentication, Cease Data Curation, Conversion (Analog), Deposit agreement, File download, File renaming, Indexing, Restructure, Selection, Succession Planning, and Transcoding

Figure 1: The front and back of an example card used in the card ranking activity.



Metadata

Information about a data set that is structured (often in machine-readable format) for purposes of search and retrieval. Metadata elements may include basic information (e.g. title, author, date created, etc.) and/or specific elements inherent to datasets (e.g., spatial coverage, time periods).

Rate how important this activity is to you.
(Write a number 1-5 with 5 = highest importance, 1 = not important)

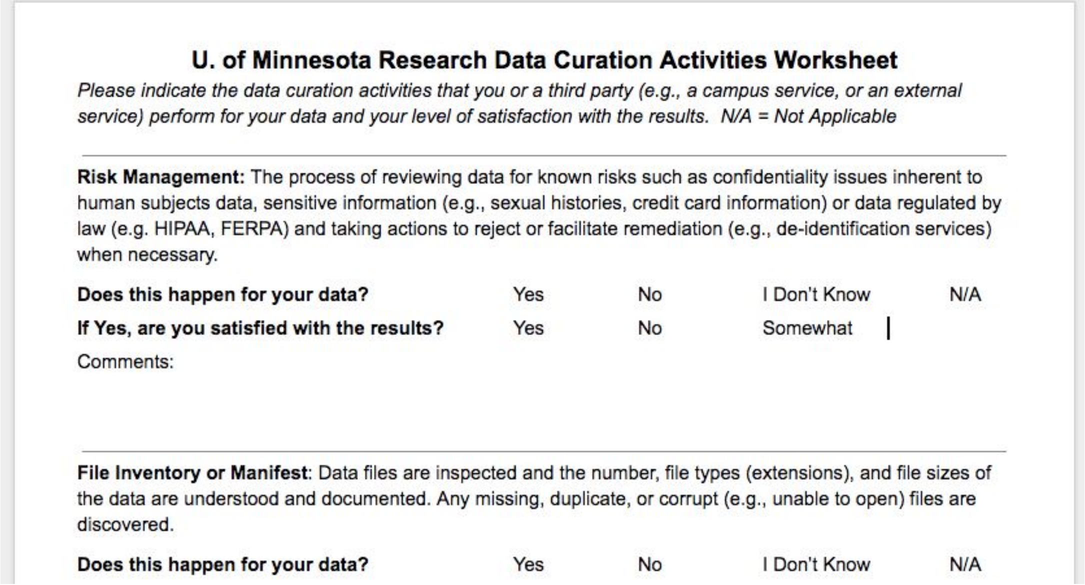
Round 1	Round 2	Round 3	Round 4

Caption: The DCN team reviewed several authoritative sources for definitions of data curation activities and formed a list of 47 activities and definitions. The full list of data curation activities and their definitions is presented as an appendix to this report.

Part 2: Engagement and Satisfaction with Data Curation Activities

To address the second question, a worksheet (example displayed in figure 2) with 18-20 of the selected data curation activities was handed out to the group of researchers and they were asked for each activity "Does this happen for your data?" and "If Yes, are you satisfied with the results?" For each activity, there was also a space for comments. The worksheet is available as a supplemental file to this report. Of the 47 data curation activities, 35 activities were chosen by team members to be further assessed on the worksheet exercise, with the selection and order varied at each institution.[13]

Figure 2: Worksheet instrument used to gauge researcher satisfaction with Data Curation Activities.



U. of Minnesota Research Data Curation Activities Worksheet

Please indicate the data curation activities that you or a third party (e.g., a campus service, or an external service) perform for your data and your level of satisfaction with the results. N/A = Not Applicable

Risk Management: The process of reviewing data for known risks such as confidentiality issues inherent to human subjects data, sensitive information (e.g., sexual histories, credit card information) or data regulated by law (e.g. HIPAA, FERPA) and taking actions to reject or facilitate remediation (e.g., de-identification services) when necessary.

Does this happen for your data? Yes No I Don't Know N/A

If Yes, are you satisfied with the results? Yes No Somewhat

Comments:

File Inventory or Manifest: Data files are inspected and the number, file types (extensions), and file sizes of the data are understood and documented. Any missing, duplicate, or corrupt (e.g., unable to open) files are discovered.

Does this happen for your data? Yes No I Don't Know N/A

Caption: To better understand how data curation activities happen for data, researchers were asked to provide comments describing how and by whom (themselves or a third party) a particular activity occurred

[13] In addition to the 12 activities not chosen for the card ranking activity listed in footnote 1, the additional three activities missing from the worksheets were: Curation Log, Emulation, and Interoperability.

or to explain why they were or were not satisfied with the results.

Part 3: Barriers and Challenges to Researcher Engagement in Data Curation Activities
Finally, to answer the third question the session allowed ample time to discuss the most highly ranked data curation activities in greater detail. Breaking out into groups, the tables described their current practices for engaging with the top ranked data curation activities, the challenges and barriers to this work, and how or by whom these services were obtained. The notes were captured by the DCN team members in attendance or by support from a library staff member from that institution.

The DCN had several assumptions going into these sessions that we wanted to test. During an early planning meeting the team identified seven barriers that may prevent the DCN from being successful. Three of these barriers may be better understood by the researcher focus group sessions were:

- Barrier 1: The value of data curation is not easy to measure and/or may be unknown.
- Barrier 2: Complex and evolving ecosystem of differing expectations-functional v. domain curation – researcher needs – funder needs.
- Barrier 3: It can be better to do it yourself. There may be a missed opportunity cost of library consultation with local researchers when using DCN.

Results

The six sessions generated results for each of our three questions:

1. What data curation activities researchers see as important or having value to themselves or to their communities of practice?
2. How, to what extent and why researchers engage in data curation activities themselves as a normative part of their research workflows? and
3. What are the barriers preventing researchers from doing so (time, personnel, knowledge, money, equipment, other resources)?

Part 1 Results: Importance of Data Curation Activities

First, the card ranking exercise revealed the most valued (highest level of importance) data curation activities overall, by institution, and by disciplinary groupings. All of the 35 activities ranked at least a 2.5 out of 5 level of importance. Table 2 shows how activities were ranked and the frequency of how many times the activity was ranked (higher number is proportional with our confidence in the ranking). Figure 3 shows the variety of rankings by discipline. No major disciplinary differences were found.

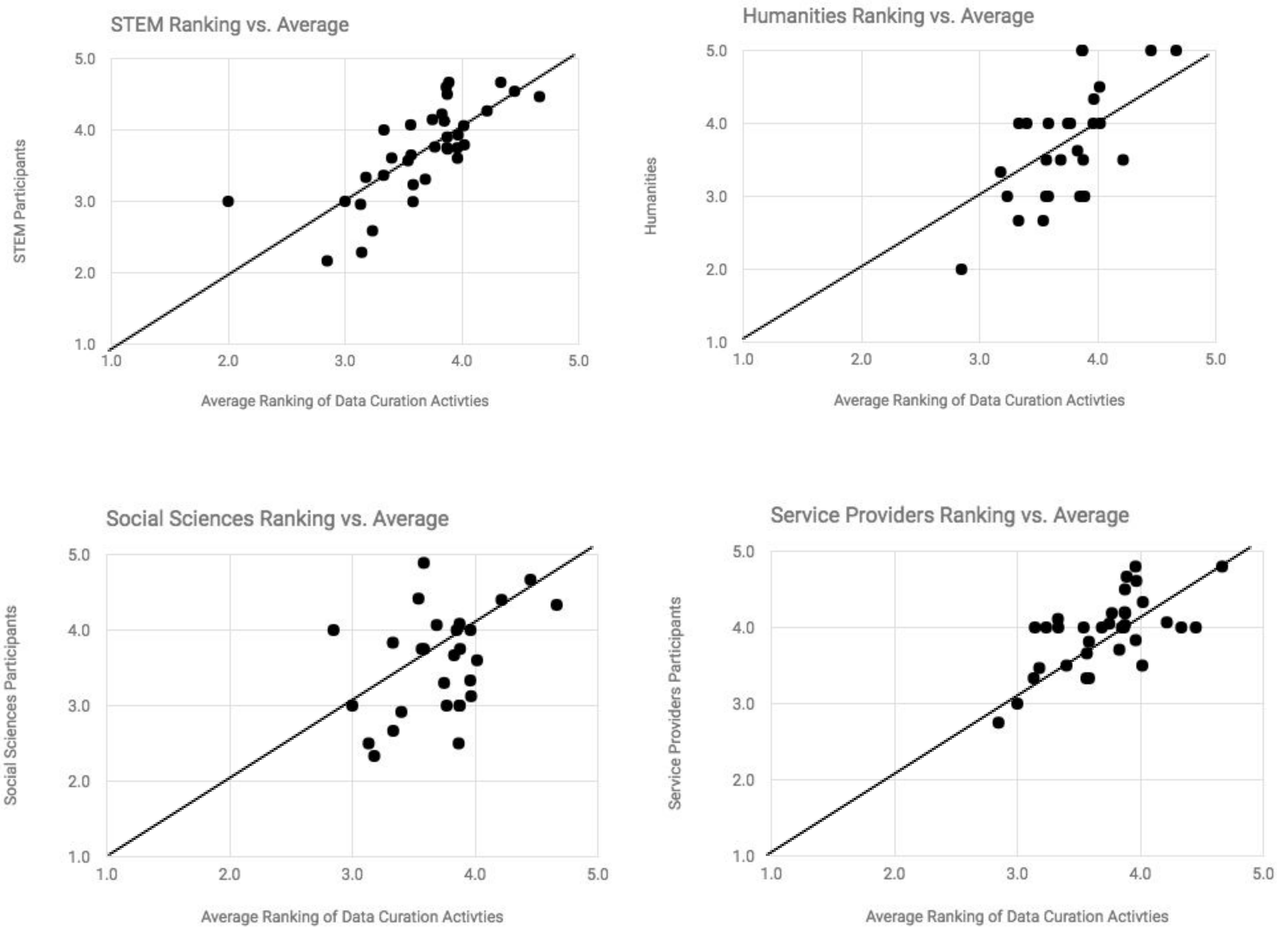
Table 2: The 35 data curation activities as ranked by researchers across six focus group sessions.

Rank	**Data Curation Activity**	C	WU	IL	P	MN	MI	Count of Sessions	Average Ranking	Range*
Ranking = 5 Highest Level of Importance "Most Important"										
1	Documentation	✓	✓	✓	✓	✓	✓	6	4.6	4.92 - 3.5
2	Chain of custody		✓					1	4.5	n/a
3	Secure Storage	✓	✓		✓		✓	4	4.4	5 - 3.88
4	Quality Assurance	✓	✓	✓	✓	✓		5	4.3	4.63 - 3.88
5	Persistent Identifier	✓	✓	✓	✓	✓	✓	6	4.3	4.75 - 4

6	Discovery Services				✓			1	4.3	n/a
7	Curation Log				✓			1	4.1	n/a
8	Technology Monitoring /Refresh			✓				1	4.1	n/a
9	Software Registry		✓				✓	2	4.1	4.25 - 3.88
10	Data Visualization		✓			✓		2	4.0	4 - 4
11	File Audit	✓		✓	✓			3	4.0	4.25 - 3.5
12	Metadata	✓		✓	✓	✓	✓	5	4.0	4.38 - 3.38
Ranking = 4 out of 5 Level of Importance “Very Important										
13	Versioning	✓	✓	✓	✓	✓	✓	6	3.9	4.75 - 3.38
14	Contextualize	✓	✓	✓	✓	✓	✓	6	3.9	4.56 - 3.25
15	Code review	✓	✓	✓	✓	✓	✓	6	3.9	4.5 - 2.88
16	File Format Transformations	✓	✓	✓	✓	✓		5	3.8	4.5 - 3.25
17	Interoperability				✓	✓		2	3.8	4.38 - 3.25
18	Data Cleaning		✓			✓		2	3.8	4 - 3.5
19	Embargo	✓	✓	✓	✓	✓	✓	6	3.7	4.13 - 3.25
20	Rights Management	✓			✓	✓	✓	4	3.7	4.25 - 3
21	Risk Management	✓		✓	✓	✓	✓	5	3.6	3.88 - 3
22	Use Analytics	✓	✓	✓	✓	✓	✓	6	3.6	4.13 - 3
23	Peer-review		✓	✓		✓		3	3.5	4.75 - 2.58
24	Terms of Use	✓		✓	✓	✓		4	3.5	3.63 - 3.38
25	Data Citation	✓		✓	✓		✓	4	3.5	4.08 - 2.75
26	File validation	✓		✓	✓		✓	4	3.4	4 - 3
27	Migration		✓				✓	2	3.4	3.88 - 2.83
28	File Inventory or Manifest	✓		✓	✓	✓		4	3.2	3.5 - 2.75
29	Metadata Brokerage	✓		✓	✓	✓	✓	5	3.2	4 - 2.63
30	Deidentification	✓		✓	✓	✓		4	3.1	4.25 - 2.13
31	Repository Certification			✓				1	3.0	n/a
Ranking = 3 out of 5 Level of Importance “Important”										
32	Emulation			✓	✓			2	2.9	3.13 - 2.63
33	Restricted Access		✓			✓		2	2.6	2.88 - 2.38
34	Correspondence						✓	1	2.5	n/a
35	Full-Text Indexing					✓		1	2.5	n/a
Ranking = 2 out of 5 Level of Importance “Less Important”										
Ranking = 1 out of 5 Level of Importance “Not Important”										

* Range represents the highest and lowest average rating given per institution.

Figure 3: Variation of Disciplinary Response from the Average Ranking of Data Curation Activities (STEM, Social Sciences, Humanities, and Service Providers).



Part 2 Results: Engagement and Satisfaction with Data Curation Activities

Second, the worksheet exercise revealed which activities researchers engaged in, what techniques were being used, and their levels of satisfaction in the results. Out of the 90 participants, 87 turned in their worksheet (Minnesota, Washington University and Michigan had 13, 11, and 18 participants respectively however only 12, 10, and 16 participants tuned in their handout due to leaving early, etc.) and therefore these individuals (and their disciplinary identity if known) were counted as “did not answer.” Additionally, note that the response “Sometimes” was a coded answer applied when a participant circled both yes and no. In total, 32 of the Data Curation Activities were analyzed by researchers in this exercise.

Figure 4: Overall Percentage of Researcher Responses to “Does this [Data Curation Activity] Happen for Your Data?” (Total =100%)

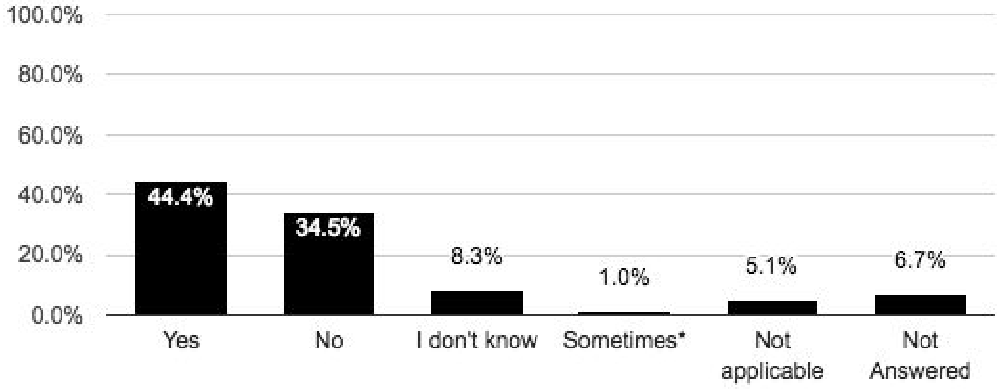


FIgure 5: Overall Percentage of Researcher Responses to “If Yes [this data curation activity happens for your data], are you satisfied with the results?” (Total, not including comments =100%)

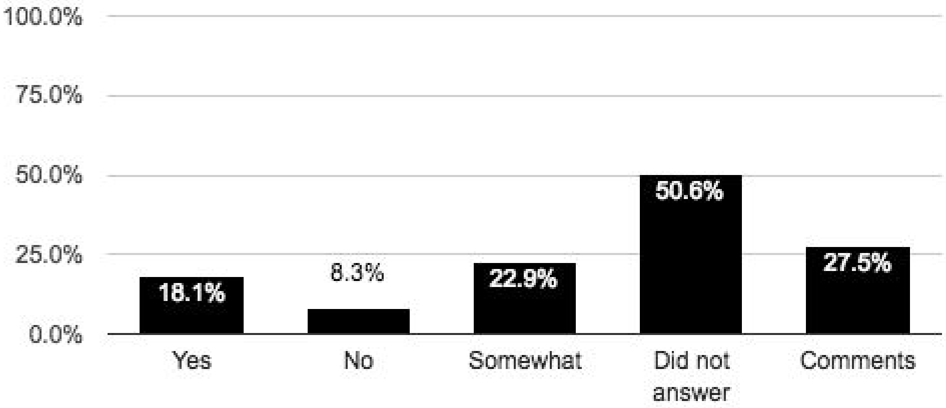


Table 3: Overall Responses to Worksheet Question: “Does this activity happen for your data?”

Data Curation Activity	Count of Responses	"Yes"	"No"	"I Don't Know"	Sometimes*	N/A	Not Answered
Documentation	91	80%	9%	3%	1%	2%	4%
Secure Storage	60	75%	17%	2%	0%	2%	5%
Chain of custody	11	64%	9%	9%	0%	9%	9%
Metadata	80	63%	24%	6%	1%	3%	4%
File Inventory or Manifest	62	58%	29%	8%	2%	0%	3%
Data Visualization	24	58%	25%	0%	0%	8%	8%
Versioning	91	56%	30%	7%	1%	2%	4%
File Format Transformations	91	55%	27%	8%	1%	3%	5%
Quality Assurance	91	52%	29%	7%	3%	3%	7%

Data Citation	67	49%	37%	7%	0%	1%	4%
Data Cleaning	11	45%	9%	9%	9%	18%	9%
Deidentification	62	44%	27%	3%	0%	16%	10%
Embargo	91	43%	38%	8%	0%	4%	7%
Risk Management	80	43%	33%	3%	1%	15%	6%
Use Analytics	91	42%	35%	10%	1%	2%	10%
Terms of Use	62	42%	34%	6%	2%	6%	10%
Software Registry	29	41%	38%	3%	0%	7%	10%
Code review	91	38%	34%	10%	1%	11%	5%
Contextualize	91	38%	45%	7%	1%	4%	4%
Restricted Access	24	38%	38%	0%	0%	17%	8%
Persistent Identifier	91	37%	44%	9%	2%	2%	5%
Peer-review	42	36%	38%	12%	0%	2%	12%
Rights Management	51	35%	31%	12%	2%	4%	16%
Technology Monitoring and Refresh	18	33%	39%	22%	0%	0%	6%
Contact Information	18	28%	33%	11%	0%	11%	17%
Full-Text Indexing	13	23%	69%	0%	0%	0%	8%
File validation	67	22%	49%	21%	0%	4%	3%
Metadata Brokerage	80	21%	51%	14%	0%	3%	11%
Discovery Services	11	18%	36%	18%	0%	27%	0%
File Audit	49	16%	57%	22%	2%	0%	2%
Repository Certification	18	11%	50%	17%	0%	11%	11%
Migration	29	7%	62%	10%	0%	10%	10%

Table 4: Overall Responses to Worksheet Question: “If Yes, Are You Satisfied with the Results?”

Data Curation Activity	**Count of Responses**	**Yes, Satisfied**	**No, not satisfied**	**Somewhat satisfied**	**Not Answered**	**Comments (% and Count)**	
Secure Storage	60	38%	3%	18%	40%	32%	19
Metadata	80	29%	8%	31%	33%	38%	30
File Format Transformations	91	29%	5%	21%	45%	27%	25
Chain of custody	11	27%	0%	36%	36%	18%	2
Documentation	91	26%	10%	46%	18%	41%	37
Embargo	91	24%	4%	16%	55%	20%	18
File Inventory or Manifest	62	23%	3%	37%	37%	35%	22

Data Citation	67	22%	12%	21%	45%	30%	20
Code review	91	22%	5%	14%	58%	23%	21
Deidentification	62	21%	5%	16%	58%	19%	12
Risk Management	80	21%	5%	23%	51%	40%	32
Restricted Access	24	21%	8%	4%	67%	17%	4
Persistent Identifier	91	19%	11%	33%	37%	30%	27
Peer-review	42	19%	5%	19%	57%	21%	9
Repository Certification	18	17%	0%	6%	78%	11%	2
Use Analytics	91	16%	12%	20%	52%	24%	22
Terms of Use	62	16%	15%	16%	53%	23%	14
Full-Text Indexing	13	15%	15%	8%	62%	0%	0
Quality Assurance	91	14%	4%	27%	54%	27%	25
Software Registry	29	14%	10%	21%	55%	45%	13
Data Visualization	24	13%	4%	33%	50%	21%	5
Versioning	91	13%	12%	37%	37%	32%	29
Rights Management	51	12%	8%	18%	63%	39%	20
Metadata Brokerage	80	11%	13%	18%	59%	29%	23
Migration	29	10%	10%	0%	79%	17%	5
Data Cleaning	11	9%	0%	45%	45%	18%	2
Contextualize	91	8%	14%	24%	54%	31%	28
Contact Information	18	6%	11%	17%	67%	39%	7
File validation	67	6%	7%	9%	78%	9%	6
File Audit	49	2%	14%	14%	69%	12%	6
Technology Monitoring and Refresh	18	0%	6%	33%	61%	39%	7
Discovery Services	11	0%	9%	18%	73%	0%	0

Comments in the worksheet provided rich detail as to how researchers were applying data curation activities and/or their difficulties in obtaining such services. These are presented as an appendix to this report.

Part 3 Results: Barriers and Challenges to Researcher Engagement in Data Curation Activities

Third, our focus group discussions gave us insights into the barriers and challenges faced by researchers engaged with data curation activities. In each session we focused on five of the top ranked data curation activities for that group.

Table 5: Discussion focus areas per institution

Top Ranked Activities**	**Cornell**	**Wash U**	**Illinois***	**Penn State***	**Minn.**	**Mich.**	**Total**
Contextualization		X					1
Documentation	X	X	X	X	X	X	6
File auditing	X						1
File Format Transformations			X				1
File validation						X	1
Interoperability				X			1
Metadata	X		X	X		X	4
Peer-Review		X			X		2
Persistent Identifier		X	X		X		3
Quality assurance	X		X		X		3
Secure storage	X			X		X	3
Software registry						X	1
Versioning		X		X	X		3

** The top ranked data curation activities that were the focus of our discussions may not align perfectly with the actual top ranked activities for that institution due to how the activity ranking were calculated in the session, on the fly and by hand. Yet, the five areas selected gave our group immediate feedback on their rankings and a focus for our following discussions.

* The identified areas are approximate based on notes taken during the session as the actual top 5 used in the session was written on a white board in the room and not explicitly documented.

Illinois Focus Group Discussion Summary

Conversation in the room was free-flowing. People did seem to somewhat self-assemble at tables where they knew people, so we had a table that was had the bulk of the health sciences attendees, another with a natural history background, and another with most of the engineering attendees. However, people from other areas were mixed in throughout.

At the health sciences table, the one thread of the discussion revolved around being surprised at the low ranking that others had given to “de-identification.” With human subjects being core to their research, one of the participants was mortified that someone at another table ranked it as “3,” and two others at the table also expressed bafflement. One attendee shared that they had to provide raw MRI data to collaborators at Harvard and they were concerned about the possibility of facial reconstruction and subsequent ability to identify the research subjects. The proposed solution was to make those accessing the data at Harvard sign an agreement saying they promised not to attempt identification. The researcher expressed dissatisfaction that solution relied on conscientious behavior and believed the resolution left much room for failure. This sharing concern led into another thread at the table about publication of data prior to completing all the analyses and publications. In the areas that these researchers represent, the fields are highly competitive and there was concern about being scoped and losing out on publications. One participant expressed feelings that reduced publications would decrease future grant competitiveness for the faculty and unit, but also impact their ability to recruit talented graduate students and postdocs who relied on publication output to demonstrate their productivity, skills, and creativity. Others concurred.

When the conversation was focused on what data curators could contribute, participants were happy to offload as much as possible, e.g. PIDs were seem as important to data that is published and not something that the researchers themselves were interested in figuring out themselves. Another table expressed a similar sentiment, further indicating that "trust" was currently not an issue with external services and believed that others could be counted on to do a good job. In regards to disclosure of sensitive data, one participant at the health sciences table was interested in there being an "authority" on campus to turn to for situations such as the MRI example.

Michigan Focus Group Discussion Summary

The discussion varied across the tables but several themes emerged. One theme was the balance between a desire to improve data management and curation practices with the amount of time and effort it would take to do so. For example, documentation was another important activity that nearly everyone engaged in, but fewer attendees indicated they were satisfied with the results. Good documentation was seen as a crucial element in the immediate use of the data and the potential reuse of the data by others. However, attendees noted a wide variation in the quality of documentation produced. Standardization would make it easier for others within and outside of the lab to read and understand, but attendees also recognized the need for flexibility with documentation to accommodate project and individual needs. The amount of consideration needed to develop standardized policy and practices for data with accommodations for deviations is daunting for researchers, especially if they do not feel confident in their knowledge of data management and curation issues.

Another theme that emerged from this event was an acknowledgement that more investment in curating data is needed. For instance, attendees who engage in or support developing software or scripts to use with the data mentioned that the process for maintaining software may be haphazard. A lack of protocols, formal processes or tools for data make quality assurance a challenge.

Finally, data curation is a new or emerging area for attendees and for their research community. Many of them have not had to address curation activities such as file validation, file format transformations yet, though they are seen as important for future consideration. Attendees indicated that they or their research team were at different stages of managing, sharing or curating their data which accounted for some variation in their assigning importance to activities. Use analytics, for example, had particularly wide variance with attendees who were actively sharing data giving it a high importance ranking and attendees who were not yet sharing data ranking it lower. Generally, curation activities that would directly benefit the researchers, such as a persistent identifier and contextualization to link the data and research outputs, were of particular interest even if they were not given a high ranking of importance currently.

Discussion

Our research on researcher attitudes toward data curation activities answered our three questions. We identified what data curation activities researchers see as important or having value to themselves or to their communities of practice. In this way, developing an understanding of what researchers value will help us to develop and deliver services that are more in line with real world needs and expectations. Next we determined how, to what extent and why researchers engage in data curation activities themselves as a normative part of their research workflows. And finally, we identified where are the gaps in highly valued data curation activities that researchers do not engage for their data (or engage as completely as they would like to) and what some of the barriers were preventing researchers from doing so (time, personnel, knowledge, money, equipment, other resources).

Finding 1: Most Data Curation Activities Were Rated as Important or Very Important

Only four activities out of 34 ranked below a 3 on a 5 point scale for importance (see figure 6). These were:

- "Emulation" of legacy system configurations in modern equipment in order to ensure long-term usability of data.
- Providing some data with "Restricted Access" in order to maintain the privacy of research subjects.
- Maintaining "Correspondence or contact information" for the data authors in order to facilitate connection with third-party users.
- "Full-Text Indexing" the data for discovery purposes by generating search-engine-optimized formats of the text inherent to the data.

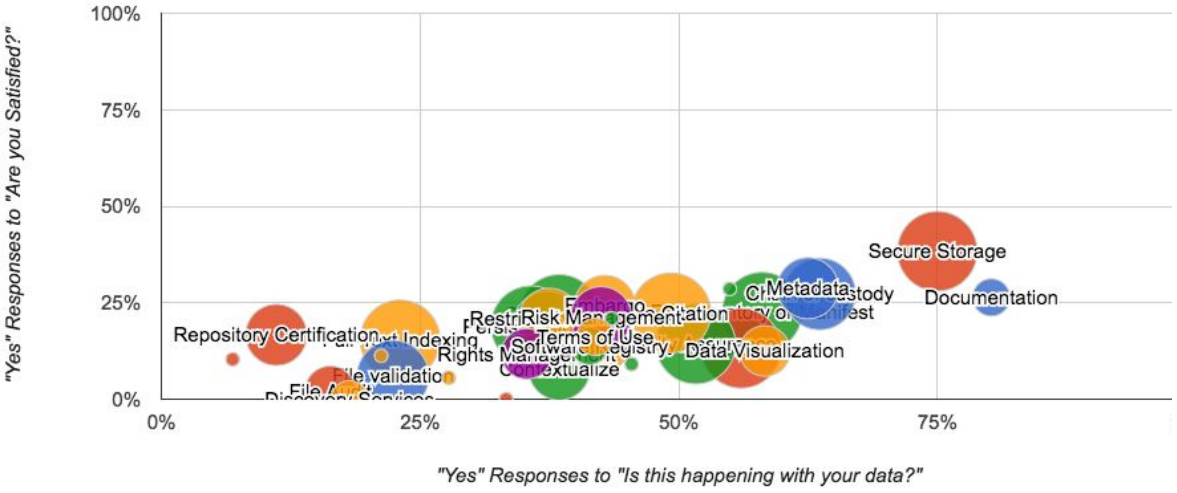
Figure 6: Results of the Average Ranking of Importance for Activities that were ranked by the Data Curation Network Focus Groups (5= highest importance, 1 = not important)

"Very Important" Average Ranking of 4.0 - 4.9	"Important" Average Ranking of 3.0 - 3.9	"Less Important" Average Ranking of 2.0 - 2.9	"Not Important" Average Ranking of 1.0 - 1.9
Documentation, Chain of custody, Secure Storage, Quality Assurance Persistent Identifier, Discovery Services, Curation Log, Technology Monitoring and Refresh, Software Registry, Data Visualization, File Audit, Metadata	Versioning, Contextualize, Code review, File Format Transformations, Interoperability, Data Cleaning, Embargo, Rights Management, Risk Management, Use Analytics, Peer-review, Terms of Use, Data Citation, File validation, Migration, File Inventory or Manifest, Metadata Brokerage, Deidentification, Repository Certification	Emulation, Restricted Access, Correspondence, Full-Text Indexing	

Finding 2: No Data Curation Activity was Satisfactorily Happening for a Majority

It is interesting to note that no data curation activity was happening in ways that satisfied the majority of our participants. The activity that came closest was Secure Storage that was happening for 75% of our researchers and satisfied 38% of our researchers (see figure 7).

Figure 7: Percent of Researchers that Use Data Curation Activities vs. Satisfaction with the Results (size of the circles indicate the number of groups weighing in, from 1 to 6).



Finding 3: Gaps or Areas of Opportunity for Data Curation Activities Exist

Another striking result is the gap in data curation activities that are very important (4 out 5 from the card ranking activity) but that are either not happening and not happening in a satisfactory way for a majority of our researchers. As noted in the previous section, the 12 activities that were ranked at least a 4 or higher on a 5-point scale are: Documentation, Chain of custody, Secure Storage, Quality Assurance Persistent Identifier, Discovery Services, Curation Log, Technology Monitoring and Refresh, Software Registry, Data Visualization, File Audit, and Metadata. Table 6 looks at the results of the Data Curation Network findings for these activities more closely.

The results of the data curation network research engagement sessions indicate several gaps in support and/or areas of opportunity for data curation service providers. For example, some data curation activities were ranked very important (rated 4 out of 5) but were *not happening for majority of researchers*. Service providers may consider investing and/or heavily promoting these important service areas that are not reaching the researchers that value them, including:

- minting and managing persistent identifiers (37% said happens),
- providing research data discovery services (18% said happens),
- monitoring and refreshing the technology housing data (33% said happens),
- maintaining a software registry (41% said happens), and
- providing tools and support for auditing file integrity (16% said happens).

For the highly ranked data curation activities that *were happening for a majority of our researchers*, better tools and or best practices might be welcome as no data curation activity was satisfying the majority of researchers who engaged in it; these were:

- creating adequate documentation (only 26% satisfied),
- tracking the provenance and chain of custody for data (only 27% satisfied),
- providing secure storage (only 38% satisfied),
- performing quality assurance for data (only 14% satisfied),
- visualizing research data (only 12.5% satisfied), and

- creating and or applying metadata (only 29% satisfied).

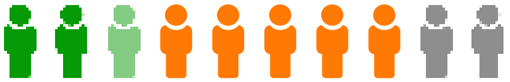
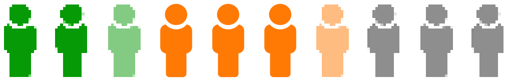
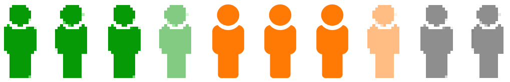
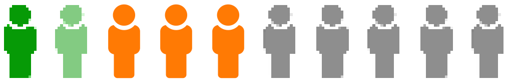
Table 6: Very Important Data Curation Activities vs. Level of Engagement and Satisfaction

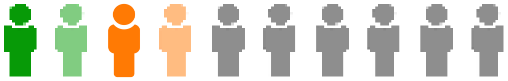
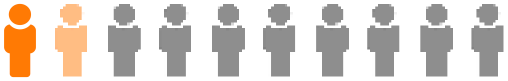
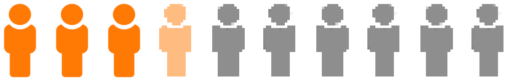
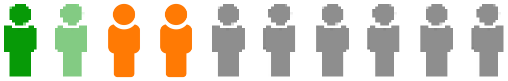
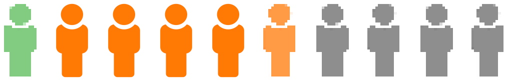
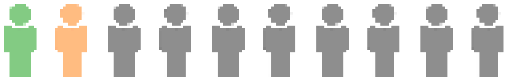
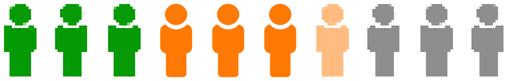
Responses >75%	**Responses 50-74%**	**Responses 25-49%**	**Responses <25%**

"Does this activity happen for your data?"			**If Yes, Are you Satisfied? (percent of total)**			
Data Curation Activity	**Rating**	**"Yes, this happens"**	**Yes**	**No**	**Some-what**	**N/A**
Documentation	4.6	80.2%	26.4%	9.9%	46.2%	17.6%
Secure Storage	4.4	75.0%	38.3%	3.3%	18.3%	40.0%
Chain of custody	4.5	63.6%	27.3%	0.0%	36.4%	36.4%
Metadata	4.0	62.5%	28.8%	7.5%	31.3%	32.5%
Data Visualization	4.0	58.3%	12.5%	4.2%	33.3%	50.0%
Quality Assurance	4.3	51.6%	14.3%	4.4%	27.5%	53.8%
Software Registry	4.1	41.4%	13.8%	10.3%	20.7%	55.2%
Persistent Identifier	4.3	37.4%	18.7%	11.0%	33.0%	37.4%
Technology Monitoring and Refresh	4.1	33.3%	0.0%	5.6%	33.3%	61.1%
Discovery Services	4.3	18.2%	0.0%	9.1%	18.2%	72.7%
File Audit	4.0	16.3%	2.0%	14.3%	14.3%	69.4%

* The data curation activity "Curation Log" was also highly ranked at 4.1 out of 5 but it was unintentionally missing on the worksheet and therefore engagement and level of satisfaction results are not available.

Figure 8: Percent of Positive Satisfaction ("Yes, I'm Satisfied" = green) versus "Yes this happens" = orange) on a 100% scale (grey) for the Very Important ranked Data Curation Activities

Documentation	"Yes this happens" = 80% "Yes, I'm Satisfied" = 26%	
Chain of custody	"Yes this happens" = 64% "Yes, I'm Satisfied" = 27%	
Secure Storage	"Yes this happens" = 75% "Yes, I'm Satisfied" = 38%	
Quality Assurance	"Yes this happens" = 52% "Yes, I'm Satisfied" = 14%	

Persistent Identifier	"Yes this happens" = 37% "Yes, I'm Satisfied" = 18.7%	
Discovery Services	"Yes this happens" = 18% "Yes, I'm Satisfied" = 0%	
Technology Monitoring and Refresh	"Yes this happens" = 33% "Yes, I'm Satisfied" = 0%	
Software Registry	"Yes this happens" = 41% "Yes, I'm Satisfied" = 14%	
Data Visualization	"Yes this happens" = 58% "Yes, I'm Satisfied" = 13%	
File Audit	"Yes this happens" = 16% "Yes, I'm Satisfied" = 2%	
Metadata	"Yes this happens" = 63% "Yes, I'm Satisfied" = 29%	

Finding 4: Partnership Opportunities Outside of the Library

The "Data Curation Roundtable" event attracted participation from researchers of a wide range of skills and abilities. Many researchers were struggling to manage their data using custom solutions and ad hoc methods - seeking new data curation techniques. However another group emerged: those who hold archival responsibilities for robust data services and archives. Participants from NASA, Roper Center, ICPSR, Biodiversity Heritage Library, and faculty at ISchools represent cohorts of skilled curation staff with valuable knowledge and perspectives that would enrich the library staff in a Data Curation Network. These responses from "managers" of data centers may have complicated the responses a bit. We need to address how to engage with this group who does data curation for researchers already (and may or may not be based in academic institutions). We need to understand what their incentives are to work with the DCN and have a role in our future work.

Finding 5: Disciplinary "Hubs" vs Local Support

In our focus groups we heard the theme that networked approaches to data curation services are ok as long as there is trust and high quality. Unfortunately, academic institutions have a "revolving door" of users, therefore it can be a challenge for the library to provide comprehensive and up to date outreach to everyone that needs our services. It may be possible to work with larger, national, and disciplinary groups to provide levels of "peer review" for data, rather than only curatorial review. Likewise, partnering with technology tools that provide mechanisms to support open, reproducible scholarship may be one key to success (e.g., GIT, R, OSF). Finally, we heard that institutional data repositories may not be meeting the needs of researchers to share their research, since data only tells part of the story. Where do the reports, grant proposals, and other forms of scholarship live on and help contextualize the larger project? Possibly consider working with other impact metric tools such as SHARE, VIVO, PURE, and SciVal Experts to complete the full story of data curation.

Conclusion

The results of our engagements with researchers have provided the Data Curation Network with a number of key recommendations. These findings will be used to build a model for how the Data Curation Network may enable academic institutions to broaden the depth and breadth of curation services beyond what a single institution might offer alone. The results presented here represent one activity of the DCN's first year to seek input from researchers to better understand how data curation services fit into their research workflow and data management needs.

Supplemental Files

Supplemental data files

- Data tables available as Excel Data file
- Cards used in the ranking exercise (word doc) available at Master Card Deck
- Worksheet (word doc) available at Worksheet Template

Appendix

A. Definitions of 47 Data Curation Activities and Rankings
B. Raw comments from the DCN Engagement Worksheet for highly ranked activities

Appendix A: Definitions of Data Curation Activities and Rankings by our Researchers

Definitions were written by the Data Curation Network team by consulting the following sources: The CASRAI Dictionary (http://dictionary.casrai.org/Main_Page), the Research Data Aliance (RDA) Terms Definition Tool (http://smw-rda.esc.rzg.mpg.de/index.php/Main_Page), the Digital Curation Center (DCC) Glossary (http://www.dcc.ac.uk/digital-curation/glossary), Data Curation Steps from the forthcoming book "Curating Research Data, Volume Two: A Handbook of Current Practice" (http://hdl.handle.net/11299/183502), the ICPSR Glossary of Social Science Terms (http://www.icpsr.umich.edu/icpsrweb/ICPSR/support/glossary), the Research Data Canada Glossary (https://www.rdc-drc.ca/glossary/), the Digital Preservation Coalition Glossary (http://handbook.dpconline.org/glossary), and the Society of American Archivists Terms Glossary (http://www2.archivists.org/glossary/terms).

Table A1: Definitions of 47 data curation activity from the Data Curation Network project (alphabetical order)

Data Curation	Definition	Rank of

Activity		**Importance**
Arrangement and Description	The re-organization of files (e.g., new folder directory structure) in a dataset that may also involve the creation of new file names, file descriptions, and the recording of technical metadata inherent to the files (e.g., date last modified).	Not Ranked
Authentication	The process of confirming the identity of a person, generally the depositor, who is contributing data to the data repository. (e.g., password authentication or authorization via digital signature). Used for tracking provenance of the data files.	Not Ranked
Cease Data Curation	Plan for any contingencies that will ultimately terminate access to the data. For example, providing tombstones or metadata records for data that have been deselected and removed from stewardship.	Not Ranked
Chain of custody	Intentional recording of provenance metadata of the files (e.g., metadata about who created the file, when it was last edited, etc.) in order to preserve file authenticity when data are transferred to third-parties.	2
Code review	Run and validate computer code (e.g., look for missing files and/or errors) in order to find mistakes overlooked in the initial development phase, improving the overall quality of software.	15
Contextualize	Use metadata to link the data set to related publications, dissertations, and/or projects that provide added context to how the data were generated and why.	34
Conversion (Analog)	In effort to increase the usability of a data set, the information is transferred into digital file formats (e.g., analog data keyed into a database). Note: digital conversion is also used to convert “fixed” data (e.g., PDF formats) into machine-readable formats.	14
Correspondence	Keep up-to-date contact information for the data authors and/or the contact persons in order to facilitate connection with third-party users. Often involves managing ephemeral information that will change over time.	Not Ranked
Curation Log	A written record of any changes made to the data during the curation process and by whom. File is often preserved as part of the overall record.	7
Data Citation	Display of a recommended bibliographic citation for a dataset to enable appropriate attribution by third-party users in order to formally incorporate data reuse as part of the scholarly ecosystem.	25
Data Cleaning	A process used to improve data quality by detecting and correcting (or removing) defects & errors in data.	18
Data Visualization	The presentation of pictorial and/or graphical representations of a data set used to identify patterns, detect errors, and/or demonstrate the extent of a data set to third party users.	10
Deidentification	Redacting or removing personally identifiable or protected	30

	information (e.g., sensitive geographic locations) from a dataset prior to sharing with third-parties.	
Deposit agreement	The certification by the data author (or depositor) that the data conform to all policies and conditions (e.g., do not violate any legal restrictions placed on the data) and are fit for deposit into the repository. A deposit agreement may also include rights transfer to the repository for ongoing stewardship.	Not Ranked
Discovery Services	Services that incorporate machine-based search and retrieval functionality that help users identify what data exist, where the data are located, and how can they be accessed (e.g., full-text indexing or web optimization).	6
Documentation	Information describing any necessary information to use and understand the data. Documentation may be structured (e.g., a code book) or unstructured (e.g., a plain text "Readme" file).	1
Embargo	To restrict or mediate access to a data set, usually for a set period of time. In some cases an embargo may be used to protect not only access, but any knowledge that the data exist.	19
Emulation	Provide legacy system configurations in modern equipment in order to ensure long-term usability of data. (E.g., arcade games emulated on modern web-browsers)	32
File Audit	Periodic review of the digital integrity of the data files and taking action when needed to protect data from digital erosion (e.g., bitrot) and/or hardware failure.	11
File download	Allow access to the data materials by authorized third parties.	Not Ranked
File Format Transformations	Transform files into open, non-proprietary file formats that broaden the potential for long-term reuse and ensure that additional preservation actions might be taken in the future. Note: Retention of the original file formats may be necessary if data transfer is not perfect.	16
File Inventory or Manifest	The data files are inspected periodically and the number, file types (extensions), and file sizes of the data are understood and documented. Any missing, duplicate, or corrupt (e.g., unable to open) files are discovered.	28
File renaming	To rename files in a dataset, often to standardize and/or reflect important metadata.	Not Ranked
File validation	A computational process to ensure that the intended data transfer to a repository was perfect and complete using means such as generating and validating file checksums (e.g., test if a digital file has changed at the bit level) and format validation to ensure that file types match their extensions.	26
Full-Text Indexing	Enhance the data for discovery purposes by generating search-engine-optimized formats of the text inherent to the data.	35

Indexing	Verify all metadata provided by the author and crosswalk to descriptive and administrative metadata compliant with a standard format for repository interoperability.	Not Ranked
Interoperability	Formatting the data using a disciplinary standard for better integration with other datasets and/or systems.	17
Metadata	Information about a data set that is structured (often in machine-readable format) for purposes of search and retrieval. Metadata elements may include basic information (e.g. title, author, date created, etc.) and/or specific elements inherent to datasets (e.g., spatial coverage, time periods).	12
Metadata Brokerage	Active dissemination of a data set's metadata to search and discovery services (e.g., article databases, catalogs, web-based indexes) for federated search and discovery.	29
Migration	Monitor and anticipate file format obsolescence and, as needed, transform obsolete file formats to new formats as standards and use dictate.	27
Peer-review	The review of a data set by an expert with similar credentials and subject knowledge as the data creator for the purposes of validating the soundness and trustworthiness of the file contents.	23
Persistent Identifier	A URL (or Uniform Resource Locator) that is monitored by an authority to ensure a stable web location for consistent citation and long-term discoverability. Provides redirection when necessary. E.g., a Digital Object Identifier or DOI.	5
Quality Assurance	Ensure that all documentation and metadata are comprehensive and complete. Example actions might include: open and run the data files; inspect the contents in order to validate, clean, and/or enhance data for future use; look for missing documentation about codes used, the significance of "null" and "blank" values, or unclear acronyms.	4
Repository Certification	The technical and administrative capacities of the repository undergo review through a transparent and well-documented process by a trusted third-party accreditation body (e.g., TRAC, or Data Seal of Approval).	31
Restricted Access	In order to maintain the privacy of research subjects without losing integral components of the data, some data access will be protected and/or mediated to individuals that meet predefined criteria.	33
Restructure	Organize and/or reformate poorly structured data files to clarify their meaning and importance.	Not Ranked
Rights Management	The process of tracking and managing ownership and copyright inherent to a data set as well as monitoring conditions and policies for access and reuse (e.g., licenses and data use agreements).	20

Risk Management	The process of reviewing data for known risks such as confidentiality issues inherent to human subjects data, sensitive information (e.g., sexual histories, credit card information) or data regulated by law (e.g. HIPAA, FERPA) and taking actions to reject or facilitate remediation (e.g., de-identification services) when necessary.	21
Secure Storage	Data files are properly stored in a well-configured (in terms of hardware and software) storage environment that is routinely backed-up and physically protected. Perform routine fixity checks (to detect degradation or loss) and provide recovery services as needed.	3
Selection	The result of a successful appraisal. The data are determined appropriate for acceptance and ingest into the repository according to local collection policy and practice.	Not Ranked
Software Registry	Maintain copies of modern and obsolete versions of software (and any relevant code libraries) so that data may be opened/used overtime.	9
Succession Planning	Planning for contingency, and/or escrow arrangements, in the case that the repository (or other entity responsible) ceases to operate or the institution substantially changes its scope.	Not Ranked
Technology Monitoring and Refresh	Formal, periodic review and assessment to ensure responsiveness to technological developments and evolving requirements of the digital infrastructure and hardware storing the data.	8
Terms of Use	Information provided to end users of a data set that outline the requirements or conditions for use (e.g., a Creative Commons License).	24
Transcoding	With audio and video files, detect technical metadata (min resolution, audio/video codec) and encode files in ways that optimize reuse and long-term preservation actions. (E.g, Convert QuickTime files to MPEG4).	Not Ranked
Use Analytics	Monitor and record how often data are viewed, requested, and/or downloaded. Track and report reuse metrics, such as data citations and impact measures for the data over time.	22
Versioning	Provide mechanisms to ingest new versions of the data overtime that includes metadata describing the version history and any changes made for each version.	13

Appendix B: Raw Comments from Researchers That Participated in the DCN Engagement Focus Groups for the Top Rated Activities

The comments of the top rated (most important) data curation activities are provided here in

alphabetical order. These are:
#1 Documentation (4.6 out of 5)
#2 Chain of custody (4.5 out of 5)
#3 Secure Storage (4.4 out of 5)
#4 Quality Assurance (4.3 out of 5)
#5 Persistent Identifier (4.3 out of 5)
#6 Discovery Services (4.3 out of 5) - No comments were provided
#7 Technology Monitoring and Refresh (4.1 out of 5)
#8 Software Registry (4.1 out of 5)
#9 Data Visualization (4 out of 5)
#10 File Audit (4 out of 5)
#11 Metadata (4 out of 5)

(C=Cornell researcher, P=Penn State researcher, I = Illinois researcher, WU = Washington University in St. Louis researcher, MI = Michigan research, MN = Minnesota research).

Chain of custody (n=2, 18.2% of those presented this worksheet option)

WU_STEM_3 - In development

WU_STEM_6 - I do this for my own research but other researchers are not very good at this and I have to track it down

Data Visualization (n=5, 20.8% of those presented this worksheet option)

WU_STEM_1 - New version of sequence data are [illegible] to NCBI, which has its own versioning policy

WU_STEM_6 - This needs to happen more often

MN_SS_1 - For publications

MN_STEM_7 - When this has occurred the visualization is in the accompanying journal article.

MN_STEM_9 - manual effort

Documentation (n=37, 40.7% of those presented this worksheet option)

C_STEM_1 - Inconsistent but trending in right direction

C_STEM_2 - metadata files accompany the raw data files

C_STEM_4 - Always seems like a chore to do this and effort (time) being spent to get students, collaborators, and myself to do this. Consistent format and guide to assemble this would help.

C_STEM_5 - Updating this is a challenge because we often fail to recognize that we made changes in procedure.

C_STEM_8 - We use README files, but I think there could be more push to include code used to develop figures, clacs, etc.

C-SS_1 - Same as documentation questionÉ curation is lacking for long-term utility and reuse for data

C-SS_4 - Don't know, too new in position.

C-SS_6 - I should

C_Staff_1 - Need standards for consistency

C_Staff_3 - Data info is understood by PI/faculty researcher.

WU_STEM_3 - not yet

WU_STEM_7 - We are working to incorporate DOIs into our standard process.

I_Staff_1 - My data needs to be better documented, but most (not all) of my paid-for data is reasonably documented.

I_SS_1 - Documentation is internal (for lab staff & PIs) so far. would be helpful to have template for data deposit purposes.

I_STEM_3 - COULD ALWAYS ["ALWAYS" is underscored] BE BETTER

I_STEM_4 - our group does this on our own, but there isn't a standardized method & we could do better.

I_Staff_4 - I COULD DEFINITELY BENEFIT FROM MORE DOCUMENTATIONS ON THE UTILITIES' END.

I_STEM_5 - [illegible] but I try to.

I_STEM_6 - all is to vague & probably unobtainable but I aim for the > 80% most common use cases

I_STEM_8 - different research groups vary in how well they do this themselves

P_5 - Our workflows in this area are a bit ad hoc, potentially leading to inconsistencies

P_6 - AGAIN ALL DONE AD HOC

MN_Staff_1 - This could be done better. Genebank is an exception

MN_SS_1 - Inconsistent-I have done for some projects but not others

MN_STEM_3 - we use readme files

MN_STEM_5 - Haven't undertaken this yet

MN_STEM_9 - No standard way to do this

MN_STEM_10 - DRUM has been very helpful, helping to create readme's when needed

MI_1 - ENG - Some individuals are meticulous. Most are only in the last 1-2 years becoming aware of the problem.

MI_2 - ENG - No standard practice for students or mechanism to ensure quality.

MI_3 - HUM - Much more to do with limited staff. Running into trade off of documentation vs. work.

MI_4 - HUM - This happens some of the time, but greater consistency of application would be desirable.

MI_5 - MED - Need to get researches to have better documentation

MI_6 - SCI - Standardizing workflow and data management procedures

MI_7 - SCI - Very time consuming, variable / changing information collected.

MI_8 - SS - Codebooks are made from scratch for each dataset, makes standardization difficult.

MI_9 - SS - Need a lot more standardized practices for this.

File Audit (n=6, 12.2% of those presented this worksheet option)

C_STEM_3 - This happens during a project, but not after.

C_STEM_5 - Haven't been concerned, though perhaps
should be.

C-SS_1 - Would be great.

C_Staff_1 - But I assume so.

I_Staff_2 - No sustainability!

P_5 - Probably Need to Address This

Metadata (n=30, 37.5% of those presented this worksheet option)

C_STEM_4 - Need to standardize metadata format and apply more regularly to projects (not created for all data at present).

C_STEM_6 - Often done after the fact, if ever. Not all projects include a req. or incentives for metadata creation.

C_STEM_7 - For some collections yes, for others no.

C-SS_1 - A little, but crappily so. This I VITAL, and a key reason why I'm here. I was considering developing our own database for this (a terrible idea).

C-SS_4 - Don't know, too new in position.

C-SS_6 - I don't use technical metadata, but instead use the file title to keep track of this.

C_Staff_1 - Not sure what all would be required across disciplines

C_Staff_2 - Currently expanding metadata collected

C_Staff_3 - PI gets this info as an artifact of how data is collected and stored (stored by researcher and date-stamped).

I_Staff_3 - DataONE

I_STEM_4 - We don't really have any plan for this type of thing to my knowledge.

I_Staff_5 - [from Do you] IN DEVELOPMENT [the researcher provided his own answer] HAVE FIELDS SELECTED, PLANNING 4 POPULATIONS & [the rest of the comments are not scanned]

I_STEM_6 - Somewhat because there are so many partially overlapping formats, none are sufficient & most difficult to use

I_STEM_8 - important!

P_5 - AGAIN, WORKFLOWS POTENTIALLY INCONSISTENT

P_6 - SORRY TO SOUND LIKE A BROKEN RECORD, BUT ALL DONE AD HOC

P_11 - Manually entered with some options

MN_SS_1 - Happens if I chose to do it, not machine readable. So inconsistent across projects

MN_STEM_2 - yes, if a readme counts, no if specific "metadata" file format [the researcher added an arrow pointing to the Documentation question]

MN_STEM_5 - I need to create, but have not done this type of work previously

MI_1 - ENG - We are starting this process through tool we developed called "Signac".

MI_2 - ENG - Varies by research cohort

MI_3 - ENG - Hit or miss

MI_4 - HUM - Could be better about this!

MI_5 - MED - We provide more general metadata for the Health System Data Warehouse, not machinable (?). Could do better, esp. with their enterprise data.

MI_6 - MED - This would hugely facilitate progress across student turnover.

MI_7 - MED - Have to be careful about HIPPA.

MI_8 - SCI - Currently building geospatial database

MI_9 - SS - Not machine readable.

MI_10 - SS - Not yet. Forthcoming

Persistent Identifier (n=27, 29.7% of those presented this worksheet option)

C_STEM_8 - Currently it’s up to the student

C-SS_1 - Might be hard to automate in our case, but I'd like some QA as part of the metadata process. But not so bossy as to mandate fake entries in irrelevant fields.

C_Staff_3 - Risk of detection of access is acceptable to PI, including potential loss of ALL data.

WU_STEM_1 - visualizations are part of PowerPoint presentations and [illegible] publications

WU_STEM_7 - We do this sometimes, even though it is beyond the scope of our requirements.

I_Staff_1 - I'd like to provide this in the future for my own data (that I provide to others after textual analysis).

I_STEM_5 - I hope it's long-term enough! + others don't always use (put data on lab web page etc.)

I_STEM_6 - 1. not cited in practice 2. difficult to version 3. no mechanism to relate upstream/downstream citation

I_STEM_8 - important!

P_1 - We are certainly working on this issue

P_3 - open source [illegible] but [illegible]

P_9 - only through publication DOI ...

P_11 - I think we have PURLs

MN_SS_1 - Each researcher on own

MN_STEM_3 - up to the researcher

MN_STEM_9 - manual effort

MI_1 - ENG - See "data citation" response.

MI_2 - HUM - DOI w/o version control defeats the goal of iteration, adding new data and refinement.

MI_3 - HUM - This tends to happen more recently but it is not the case that data put on the web in the past had this.

MI_4 - MED - Has not been a focus for UMHS Enterprise data sets, to date.

MI_5 - MED - This would be highly desirable if we were more actively sharing our data.

MI_6 - MED - Hoping Blue Data repository will take care of this

MI_7 - MED - 3rd party use is mandated by individual labs currently

MI_8 - SCI - I have not used these yet

MI_9 - SCI - Good idea, I use other sites - arcadis, etc.

MI_10 - SCI - Some 3rd party repositories will do this for us

MI_11 - SS - Provided by Open ICPSR but not currently supported after upgrade to system, will be challenging if distribute data in other ways.

Quality Assurance (n=25, 27.5% of those presented this worksheet option)

C_STEM_4 - This probably could be a bigger iss for 'proprietary' or 'embargoed' data (e.g.

State of NY sometimes hold data from NYS funded projects as private until they approve release -- after undefined, lengthy typically, period).

C-SS_4 - Don't know, too new in position.

C_Staff_1 - Not done across all.

WU_STEM_6 - I don't always have time to do all of this

I_Staff_1 - Paid service providers are generally good at this task.

I_STEM_3 - WOULD BENEFIT FROM INCREASED Q.A. [illegible] MANY [illegible].

I_Staff_3 - I think this is important but labor-intensive. Maybe tools could help...

I_STEM_4 - We do this sometimes but no clear procedure.

I_Staff_5 - CONSIDERING THIS IS ALL "ME" AT THIS POINT, BACKUP TO 3 DIFF. SOURCES + BOX

I_STEM_5 - It needs to! So hard to use others' materials.

I_STEM_8 - different research groups do this themselves

P_5 - Challenge of Resources and Tims. A lot of "As is" for the third party user

MN_STEM_3 - Our data is non-commercial, we get contracted a lot for permission for commercial use. Not sure what the alternative is.

MN_STEM_9 - not considered too deeply

MI_1 - ENG - We need this. Currently rely on individuals to do for their own datasets - inadequate.

MI_2 - ENG - This varies by research group.

MI_3 - ENG - There's a lot of data, couldn't do an exhaustive evaluation.

MI_4 - HUM - 1- We are working on building a colverent(?) end to end workflow. 2- ICPSR is also helping with this QA downstream.

MI_5 - HUM - This is done pretty well

MI_6 - MED - Data Quality is a service we're thinking about offering to the Enterprise, but have not done yet.

MI_7 - MED - Quick, one click validation would be very useful for ongoing collection.

MI_8 - MED - Need better documentation by researchers.

MI_9 - MED - QA could always be improved

MI_10 - SCI - Again strong lab protocols reduce problems

MI_11 - SS - Very cursory review by one other staff member, no formal process for quality check by anyone other than primary data processor. We do have one dataset archived by ICPSR general archive that has this process.

Secure Storage (n=19, 31.7% of those presented this worksheet option)

C-SS_4 - Don't know, too new in position.

C_Staff_1 - Consistency need across all data set.

C_Staff_3 - Use not happened yet.

WU_STEM_1 - Data files are backup periodically but no routine checks were performed other than checking the checksums of files

P_5 - we do this pretty well

P_7 - HAD TO LEARN THIS THE HARD WAY

P_8 - Could be better

P_9 - No fixity checks

MI_1 - ENG - Done by individuals (CrashPad) for current operations, but long term archives are not maintained in an accessible way, if at all.

MI_2 - ENG - Because some faculty are required to make their data available and visible for processing. Not all have this in a location that is backed up / redundant. No central service for SSH/SFTP archive that is equal or cheaper than do it yourself.

MI_3 - ENG - Redundant Back Up systems

MI_4 - ENG - I assume it works

MI_5 - HUM - Mbox and ICPSR / Fedora backed.

MI_6 - HUM - This is where I think a University wide group like Deep Blue Data can help

MI_7 - MED - Could always be improved

MI_8 - MED - This is a big focus at HITS / UMHS

MI_9 - SCI - This is an active work in progress and varies widely among procedures

MI_10 - SCI - routine checks not performed on integrity of data backed-up

MI_11 - SCI - Can be difficult to access quickly

Software Registry (n=13, 44.8% of those presented this worksheet option)

WU_STEM_1 - code are altered in a version control [illegible] (eg. git, subversion) with releave tags for the different versions.

MI_1 - ENG - Done for our major software distribution but self-maintained and tested only upon deposit of software.

MI_2 - ENG - Due to campus wide security concerns and centralized software licensing that we participate in. Older versions of software are not available to my knowledge.

MI_3 - ENG - Version control of software is generally important in our line of work

MI_4 - HUM - Made it a requirement that some architecture for the project was made open and available for iteration , working in public access.

MI_5 - HUM - I know that this is performed by some people I work with - but hardly everyone

MI_6 - MED - Clinical Systems that were retired for M. Chart only archived the data - application code not similarly archived.

MI_7 - MED - We use Git and GitHub, but it's rather ad hoc and poorly integrated with data containers

MI_8 - MED - Don't store versions of outside software, ex. Word, excel

MI_9 - MED - Main code version controlled, many ancillary scripts are also needed.

MI_10 - SCI - In my case, saving format determines readability with future software upgrades

MI_11 - SCI - It should happen

MI_12 - SS - Maintain in SPSS, stata and csv, would like to keep dictionary to have csv be read into additional software.

Technology Refresh and Monitoring (n=7 , 38.9% of those presented this worksheet option)

I_Staff_1 - I assume this happens.

I_Staff_3 - I don't know who does a formal periodic review..

I_STEM_4 - Our data so far hasn't been large enough to need additional data storage.

I_Staff_5 - TOO NEW OF A PROJECT

I_STEM_5 - Mostly test files so reasonably stable.

I_STEM_6 - Only done informally based on need funding for this is lacking.

I_STEM_8 - our unit does some & research groups do their own too

Versioning (n=29, 31.9% of those presented this worksheet option)

C_STEM_8 - I know we don't do versioning of metadata, but does Box have capabilities?

C-SS_1 - Yes! Very helpful for reconstructing in revising analysis. Now it's all ad hoc. :(Very Important.

C-SS_6 - I just change the file name.

C_Staff_1 - Not done consistently

WU_STEM_5 - We require documentation. But with a wide range of user abilities we still get questions.

WU_STEM_6 - same as above

I_Staff_2 - manual implementation causes inconsistency!

I_STEM_4 - We do this but with no clear cohesive procedure.

I_Staff_4 - THE BALANCE BETWEEN A NEW DATA SET AND A NEW DATA POINT IS SOMEWHAT NEBULOUS.

I_Staff_5 - THIS IS 1ST ITERATION of DATA SET, BUT I VERSION MY DOCS

I_STEM_5 - I know I should use git, but I have no idea how to install it on the cluster.

I_STEM_6 - had to do. Mostly, subsets are archived as snapshots

I_STEM_8 - important

P_6 - TO A CERTAIN DEGREE THIS IS DONE, BUT CERTAINLY NOT UNIVERSAL

P_9 - Done through an electronic notebook

P_11 - I Haven't looked at early versions

MN_SS_1 - Happens sporadically as I find the time

MN_STEM_3 - We only version code (github). We do not have an equally good method for versioning data

MN_STEM_7 - This was done for one project. I was involved with but not necessarily following a standardized system

MN_STEM_9 - Hasn't really come up yet.["yet" is underscored.]

MN_STEM_10 - I unfortunately have not used your service enough to answer fully, but versioning is very important for our Lab.

MI_1 - ENG - Done only through version control software and tar balls.

MI_2 - ENG - Varies by research cohort

MI_3 - HUM - We aren't tracking versions in depth but we do have ability to update.

MI_4 - MED - This can be a problem

MI_5 - MED - Don't version image data - probably not necessary

MI_6 - MED - We use version control software, could look at for data.

MI_7 - SCI - Building currently

MI_8 - SS - Open ICPSR does versioning and we track as well but description of changes isn't always present.