Results of the Fall 2016 Data Curation Pilot

Release Date: March 15, 2017

Authors: Lisa R Johnston, Jake Carlson, Cynthia Hudson-Vitale, Heidi Imker, Wendy Kozlowski, Robert Olendorf, and Claire Stewart.

Introduction

The Data Curation Network began the planning phase of our project with a one-year grant from the Alfred P. Sloan foundation in May 2016. The project will develop a shared staffing model for curating research data that draws from the expertise across multiple institutions in order to broaden the depth and breadth of curation services beyond what a single institution might offer alone. In the fall of 2016, we conducted two rounds of pilots involving data curation workflows. Our primary goals were to 1) identify what our (actual) individual curation practices were in order to compare curation steps taken, 2) begin to establish what training network curators would need, and 3) identify any issues, misaligned expectations, and/or conflicts with the goals of the project. Our project reports and outcomes are posted to the project website https://sites.google.com/site/datacurationnetwork/.

Methodology

For a successful Data Curation Network model to emerge, the team first needed a good understanding of the current curation practices across our institutions: Cornell University, Penn State University, University of Illinois at Urbana-Champaign, University of Michigan, University of Minnesota, and Washington University in St. Louis. Team members are in leadership positions for data curation services and in the best position to accurately articulate their local practices. To better understand the differences in how each institution curates data, we created a pilot scenario, run in two rounds, to objectively answer this question.

Round One Methodology

In the first round of the pilot, six DCN team members (Table 1) were asked to curate the same dataset. The pilot dataset #1 was selected from a repository external to our six institutions so we might have less bias, treat it objectively, and be open with our curatorial actions and comments. To create additional objectivity, the data files (one Microsoft Excel file) and the metadata (exported as XML and saved as a text file) from pilot dataset #1 were downloaded and saved to shared google drive folder that all members could access.

Table 1: Data Curation Network Team Members Assigned to Curate Pilot Dataset #1

DCN Team Member	Institution	Job TItle	Disciplinary Subject Expertise
Wendy Kozlowski	Cornell University	Data Curation Specialist for Cornell eCommons	earth and environmental sciences, engineering
Heidi Imker	University of Illinois	Director of the Research Data Service and the Illinois Data Bank	biochemistry, organic chemistry, bioinformatics, microbiology
Lisa R. Johnston	University of Minnesota	Research Data Management Curation Lead for the Data Repository for the University of Minnesota (DRUM)	physics and astronomy, astronomical image files (e.g., FITS), raw/instrument data
Jake Carlson	University of Michigan	Research Data Services Manager for Deep Blue Data	political science, east asian studies
Rob Olendorf	Penn State University	Research Data Librarian for Science	bioinformatics, organismal biology, ecology and evolutionary biology, simulations, software code, R, Matlab, complex data sets
Cynthia Hudson-Vitale	Washington University in St. Louis	Digital Data Outreach Librarian and Lead of the Digital Research Materials Repository	qualitative health data

The assignment of pilot dataset #1 was made via email (Figure 1) to 6 DCN team members on September 15, 2016 with a four week deadline of completion set to October 15, 2016. Note that the team members were asked to imagine that the assigned dataset matched their particular skills and expertise. Since the dataset was an ecological dataset this was not the case for most team members. Team members were instructed to curate the dataset and record all actions taken in detail, possibly by using a screencast tool such as Jing, and to complete a report with the following four sections:

1. Curation actions taken to curate the data according to your local practice
2. A list of issues or problems that you would ask the data author
3. A list of issues or problems that you would ask the originating repository
4. A list of red flags or issues with the dataset (e.g. didn't have software to open files).

Finally, the DCN team member was asked to reflect on the assignment and answer a 15 question survey (Appendix A) capturing their experiences.

Figure 1: Sample email delivered to 6 members of the DCN team on September 15, 2016

assigning them the pilot dataset #1 for curation.

New DCN Curation Assignment

Curator: Jake Carlson (University of Michigan)
Pilot Dataset ID: hdl:1957/28175
Pilot Dataset Location: Shared Google Drive Folder (link removed)
Deadline for Curation Assignment Completion: **October 15, 2016**

Dear Jake,

Based on your expertise and availability as a curator in the Data Curation Network a new dataset has been assigned to you and the corresponding files and metadata may be accessed at the above shared directory. Since this is a controlled "pilot" of the data curation network please do the following steps in this order to complete the assignment by the deadline:

1. Using a screencast capture tool (e.g., Jing), record and verbally describe any and all actions that you would take to curate the dataset according to your local practice. In particular, note any local curation terminology, the chronological order of curation actions, and the time it takes to curate the dataset.
2. Draft a written report with the following sections:
    1. Curation actions: An annotated transcription of your curation actions taken with the dataset.
    2. Email to the Author: An email (using language you would send to the author) with any questions or comments that would be necessary to complete the curation process and finalize the submission.
    3. Email to the originating repository: An email (using language you would send to the curator) with any questions or comments that would be necessary to complete the curation process and finalize the submission.
    4. Red Flags and Issues: A list of issues that either prevented you from doing certain actions that you would have liked to or any potential “game stoppers” in the curation process.
3. Upload your report into this folder (link removed) on or before the assignment deadline of October 15, 2016.
4. Take this post-curation survey (link removed) to capture your experiences with the assignment.

If you have any questions, please notify me as soon as possible,
Lisa Johnston
Coordinator, Data Curation Network

Round Two Methodology

In order to expand the pilot and engage with data curator staff across our institutions, the DCN Team generated a list of 1-2 potential data curators and their subject expertise (Table 2). We identified 11 individual data curators with a good variety of skills and disciplinary expertise and, using this information, we matched (as well as possible) each curator to a dataset that originated from a different DCN institutional repository than their own. Like Round One, these pilot data sets (numbered #2-12) may or may not have been previous curated (e.g., self submission from a researcher with little to no curator review or intervention).

Table 2: Institutional affiliation and expertise of the eleven data curator staff that were each

assigned DCN pilot datasets #2-12 in Round 2. Curation staff names not included.

Dataset Assigned	Institution	Curator Job Title	Curator Expertise
#2	Cornell University	Research Data and Environmental Science Librarian	Environmental sciences
#3	Penn State University	Social Science Librarian	Census data, government data, PPI, sensitive data
#4	Penn State University	Geospatial Data Services Librarian	GIS, ESRI products, R-Statistical, Python for geospatial applications, GDAL
#5	University of Illinois	Data Curator	Digital humanities, video/audio, TEI, linked data, Health sciences, sensitive data, PII
#6	University of Illinois	Data Curator	Statistical survey packages (SPSS, SAS, R)
#7	University of Michigan	Biological Sciences Librarian; Interim Chemistry Librarian	Biological sciences, genomics
#8	University of Michigan	Informationist, Data Standards and Terminologies	clinical health, health informatics
#9	University of Minnesota	Digital Repository Specialist	Textual formats, XML, code
#10	Washington University in St. Louis	Data Specialist	Matlab, Mathematica, R
#11	Washington University in St. Louis	GIS & Data Project Manager; Anthropology Librarian	GIS, ISO 19115 metadata format, RDF/XML
#12	University of Minnesota	Scientific Data Curator	Genomics data, databases, Excel, software code

On October 24, 2016, two emails were sent by the project lead to kick off Round Two. The first email was sent to DCN team members and the 11 selected data curators introducing the Round 2 pilot and establishing expectations for curators to receive their assignments via separate email. Curator staff were asked to take any action needed to review the dataset just as they would if it had been submitted to their repository. Curators were also warned that each of the repositories use different repository software, so the layout, design, and/or metadata schema may not be familiar to them. The second email was sent to the individual data curators and the DCN team member from that institution. Here the same assignment instructions given in the first round were repeated for this Round Two cohort. Assignments were given with a deadline of submitting their curation reports back to the project team by November 18, 2016, or three weeks (see Figure 2). The same survey instrument was used to capture the experiences of curators in

the Round Two pilot.

Figure 2: Sample email message delivered to 11 data curators across the DCN institutions on October 24, 2016.

New DCN Curation Assignment

Curator: ## name removed## (Cornell University)
Pilot Dataset ID: 0002
Assignment URL: (link removed)
Type: Tabular
Subject: Ecosystem Science & Management
Discipline: Agricultural and Natural Sciences
Pilot Dataset Location: Penn State University
Deadline for Curation Assignment Completion: November 18, 2016

Dear Curator,

Based on your expertise and availability as a curator in the Data Curation Network a new dataset has been assigned to you and the corresponding files and metadata may be accessed at the above shared directory. Since this is a controlled "pilot" of the data curation network please do the following steps in this order to complete the assignment by the deadline:

1. Using a screencast capture tool (e.g., Jing), record and verbally describe any and all actions that you would take to curate the dataset according to your local practice. In particular, note any local curation terminology, the chronological order of curation actions, and the time it takes to curate the dataset.
2. Draft a written report with the following sections:
    1. Curation actions: An annotated transcription of your curation actions taken with the dataset.
    2. Email to the Author: An email (using language you would send to the author) with any questions or comments that would be necessary to complete the curation process and finalize the submission.
    3. Email to the originating repository: An email (using language you would send to the curator) with any questions or comments that would be necessary to complete the curation process and finalize the submission.
    4. Red Flags and Issues: A list of issues that either prevented you from doing certain actions that you would have liked to or any potential “game stoppers” in the curation process.
3. Upload your report with this filename "DCNPilot_0002" into this folder (link removed) on or before the assignment deadline of Friday November 18, 2016.
4. Take this post-curation survey (link removed) to capture your experiences with the assignment.

If you have any questions, please notify me as soon as possible,
Lisa Johnston
Coordinator, Data Curation Network

Coding the Data Curation Reports

After receiving all the reports from Round One and Round Two, each report was analyzed and the individual curation steps described in the reports were coded. To do this, the project lead labeled the steps described in each report with one of the “Data Curation Activities” defined by the DCN project (Appendix B lists the data curation activities and their definitions). If the step

could not be coded with an existing activity definition, a new code was created. For example, here is the text of one step described in a curation report: “Received the pop up message – “Workbook is protected and cannot be changed” multiple times during this review. Need to ask about removal of this limitation of use.” To code this step, a spreadsheet entry for this statement was made and included the following:

1. the statement text (raw, unedited)
2. the curators’ institution name,
3. the order in the report or local workflow for which this step appeared, and
4. the Data Curation Activity code for the statement.

Assigning the activity code was a highly subjective process. In the above example, the text was coded as “Rights Management,” defined as *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).* This coding process was repeated for each data curation report in round 1. Round 2 was coded by several DCN team members and the project lead reviewed each coding to aid with consistency. As a result, there were several new Data Curation Activities that emerged from both rounds. These were:

- Five activities added to the DCN list after running Round One of the pilot: Deidentification, File renaming, Data Cleaning, Interoperability, and Restructure.
- An additional 9 activities were generated but have not yet been added to the list and still need definitions: Communicate with Data Author, Disk Image, Inspect files, Inspect metadata, Transfer the Assignment, Virus Scan, Working Copy, Expertise Match (confirm assignment), and Link Checking.

Once each statement in the reports were coded, the information was analyzed based on the occurrence (e.g., was this step included or missing from each report) and frequency (e.g., the number of occurrences in each report) of each data curation activity used for a dataset or used by an institution. A special note on how the frequency of activities were counted. The number of occurrences for each step were normally counted in all cases except for two, Inspect Metadata and Inspect Data files. Since the reports frequently indicate these activities and in a variety of ways (e.g., a statement such as “Next I began to review the data2 tab of the data.xls file…”), these two Data Curation Activities were only counted once per report and assumed to occur, even if not explicitly stated.

Results

In Round One, all six of the assignments made to the six members of the DCN team were completed and all within the three week deadline. Additionally four team members (66.6%) completed the post-curation survey form. In Round Two, ten out of eleven assignments made to data curation staff were returned (91%) for pilot data sets #2-12 and only three out of the eleven (27%) were returned within the three week deadline. Additionally eight curation staff (73%) completed the survey form. These numbers are presented in Table 3.

Table 3: Results of the two rounds of data curation pilots.

Round	**Dataset Assignment**	**Curator Group**	**Assignment Completion Rate**	**Met Deadline?**	**Responded to Post-curation Survey?**
Round 1	All were assigned to curate dataset #1	6 project team members (see table 1)	100%	100%	66.6%
Round 2	Individuals matched with data sets #2-12	11 curation staff across partner institutions (see table 2)	91%	27%	73%

The detailed results of the Pilot Round 1 and Round 2 are presented in three parts.

1. Analysis of the Data Curation Reports: First, our analysis of the data curation reports revealed the frequency, order, and recurrence of the data curation actions taken for pilot data set #1 and for pilot data sets #2-12.
2. Correspondence with Network Curators: Second, the mock correspondence between the network curators and data authors or local curators highlighted realistic issues inherent to the pilot datasets.
3. Post-Curation Survey Results: Third, our post-curation survey revealed several questions and concerns about the process as well as helped to gauge overall confidence and satisfaction in curator involvement with the DCN activity as piloted.

Analysis of the Data Curation Reports

PIlot Dataset #1: The quantitative aspects of these reports, written by six of the DCN team members that were assigned pilot dataset #1, are more comparable since they represent actions taken for the *same* dataset (see Table 4). The reports were analyzed for the following:

- word counts, which ranged from 654 words to 2004 (more likely due to the style of communication for the team member than of anything representing completeness or quality),
- the number of steps described in the reports, which ranged from 21 to 38 steps taken (average of 31 steps for curating dataset #1),
- the number of questions in the mock emails, which ranges from 13 questions to no email response (shown as n/a), and,
- the time spent on the curation assignment, which ranged from 1 hour to 2-5 hours.

Table 4: Qualitative comparison of data curation reports for pilot dataset #1

Pilot Dataset	Curator Institution	Report Word Count	Count of Curation Steps	Count of Questions		Time Spent (Est.)
				Author	Curator	
#1	Cornell	945	38	n/a	n/a	1 hour
#1	Illinois	844	21	3	8	1-2 hours
#1	Minnesota	1592	29	13	11	2.25 hour
#1	Michigan	2004	37	2	3	2-5 hours
#1	Penn State	654	32	7	n/a	1 hour
#1	Wash U.	666	29	0	2	1-2 hours

The coded data curation activities revealed in the data curation reports for pilot dataset #1 show four groupings of steps. (Results are presented in Table 5).

- **Everyone did this:** There were 4 activities mentioned by all six DCN team members for this dataset: Documentation, File Format Transformations, Quality Assurance, and Communicate with Author. (Note: As mentioned above, the Data Curation Activities “Inspect Files” and “Inspect Metadata” were assumed taken by default for all data sets and only counted once).
- **Most did this**: An additional 8 activities were mentioned by 3-4 DCN team members: Restructure, Metadata, Risk Management, Terms of Use, File renaming, Curation Log, Working Copy, and Contextualization.
- **Some did this:** Finally 15 data curation activities were only mentioned by 1-2 members: Data Cleaning, Rights Management, Persistent Identifier, Versioning, File Inventory or Manifest, Data Visualization, Arrangement and Description, Contact Information, Disk Image, File validation, Secure Storage, Virus Scan, Transfer, Expertise Match, and Link Checking.
- **No one did this**: The remaining 28 activities were not mentioned for pilot dataset #1: Authentication, Cease Data Curation, Chain of custody, Code review, Conversion (Analog), Data Citation, Deidentification, Deposit agreement, Discovery Services, Embargo, Emulation, File Auditing, File download, Full-Text Indexing, Indexing or Cataloging, Interoperability, Metadata Brokerage, Migration, Peer-review, Repository Certification, Restricted Access, Selection, Software Registry, Succession Planning, Technology Monitoring and Refresh, Transcoding, Use Analytics

Table 5: The data curation activities mentioned in the curation reports for pilot data set #1

Institutional Occurrence	Data Curation Activity	Freq. of Occurrences						Total Count	Average per Inst.
		C	WU	IL	P	MN	MI		
Activities Mentioned by All 6 DCN Team Members									
6	Documentation	5	2	6	6	3	11	33	5.5
6	Quality Assurance	1	4	1	3	9	14	32	5.3
6	File Format Transformations	3	1	2	1	3	1	11	1.8
6	Communicate with Author	2	2	2	2	2	1	11	1.8
6	Inspect Files	1	1	1	1	1	1	6	1.0
6	Inspect Metadata	1	1	1	1	1	1	6	1.0
Activities Mentioned by 5 DCN Team Members									
Activities Mentioned by 4 DCN Team Members									
4	Restructure	7	0	4	7	0	3	21	5.3
4	Metadata	5	1	0	2	5	0	13	3.3
4	Risk Management	0	1	1	2	1	0	5	1.3
4	Terms of Use	0	1	1	1	1	0	4	1.0
Activities Mentioned by 3 DCN Team Members									
3	File Renaming	2	2	0	1	0	0	5	1.7
3	Curation Log	2	1	0	0	1	0	4	1.3
3	Working Copy	2	1	0	0	1	0	4	1.3
3	Contextualization	1	2	1	0	0	0	4	1.3
Activities Mentioned by 2 DCN Team Members									
2	File Inventory or Manifest	0	2	0	0	0	2	4	2.0
2	Data Cleaning	2	0	0	0	0	1	3	1.5
2	Rights Management	2	0	0	0	0	1	3	1.5
2	Persistent Identifier	1	0	0	1	0	0	2	1.0
2	Versioning	1	0	0	0	0	1	2	1.0
Activities Mentioned by 1 DCN Team Member									
1	Data Visualization	0	0	0	3	0	0	3	3.0
1	Transfer	0	2	0	0	0	0	2	2.0
1	Arrangement and Description	0	0	0	1	0	0	1	1.0
1	Contact Information	0	1	0	0	0	0	1	1.0
1	Disk Image	0	1	0	0	0	0	1	1.0
1	File Validation	0	1	0	0	0	0	1	1.0
1	Secure Storage	0	1	0	0	0	0	1	1.0
1	Virus Scan	0	1	0	0	0	0	1	1.0
1	Expertise Match	0	0	0	0	1	0	1	1.0
1	Link Check	0	0	1	0	0	0	1	1.0

Pilot Data Sets #2-12: Data sets #2-12 represent different data sets, so differences in quantitative aspects may be due to this, as well as differences in curator writing style. For this round 10 out of the 11 reports were returned and the results are compared in Table 6. Analyzing the activities taken to curate these data sets, the following trends were observed:

- Everyone did this (100%, all 10 out of 10): Inspect Files/Metadata
- Most did this (~90%, 9 out of 10): Quality Control
- Some did this (~50%, 4-6 out of 10): Documentation, Communicate with Author, Contextualization, Metadata, and File Inventory or Manifest
- Few people did this (~25%, 2-3 out of 10: Code review, File validation, Working Copy, File Format Transformations
- Only one did this (~10%, 1 out of 10): Chain of custody, Rights Management, Risk Management, Curation Log, File renaming, Interoperability, Restructure, Contact Information, Data Citation, Embargo, Full-Text Indexing, File Auditing, Link Checking, Researched File Format

Table 6: Qualitative comparison of data curation reports for pilot datasets #2-12

Pilot #	Curator Institution	Word Count	Activity Count	Data Curation Activities and their Frequency	Time Spent (Est.)
#2	Cornell	610	13	3 - Documentation, File Format, Transformations 2- Quality Control 1 - Communicate with Author, File renaming, Inspect Files/Metadata, Working Copy	n/a
#3	Penn State	n/a	n/a	*Report not completed*	n/a
#4	Penn State	876	7	2- Documentation 1 - File Inventory or Manifest, Inspect Files/Metadata, Metadata, Quality Assurance	2 hours
#5	Illinois	2372	10	2 - Documentation, Code Review 1 - Contextualization, Inspect Files/Metadata, Interoperability, Quality Assurance, Restructure	n/a
#6	Illinois	2142	32	15 - Researched File Format 4 - Quality Assurance 2 - Metadata 1 - Code Review, Communicate with Author, Contact Information, Contextualization, Data Citation, Documentation, Embargo, Inspect Files/Inspect Metadata, Right, Management, Risk Management	1.5 hours
#7	Michigan	680	4	1- Contextualization, File Auditing, Inspect Metadata/Inspect Files, Metadata	45 mins.
#8	Michigan	458	9	3 - Quality Assurance	5 hours

				2 - File Inventory or Manifest 1 - Communicate with Author, Inspect Metadata/Inspect Files, File validation	
#9	Minnesota	1439	7	2 - Quality Assurance 1 - Communicate with Author, Contextualization, File Inventory or Manifest, Inspect Metadata/Inspect Files, Link Check	4 hours
#10	Wash U	287	28	10 - Code review 7 - Quality Assurance 3 - Documentation 1 - File Inventory or Manifest, File Format Transformations, File validation, Full-Text Indexing, Inspect Metadata/Inspect Files, Metadata, Working Copy	n/a
#11	Wash U	251	15	8 - Documentation 2- Contextualization, Quality Assurance 1 - Metadata, Inspect Metadata, Inspect Files	n/a
#12	Minnesota	588	6	1 - Chain of Custody, Communicate with Author, Curation Log, Inspect Files/Metadata, Quality Assurance	2.5 hours

Correspondence with Network Curators

Second, the reports included mock emails to the data author and the curator at the pilot dataset's home institution. The correspondence provides a realistic glimpse into the way that the DCN might function. Some themes emerged and are presented below.

Curator Tone Varied in Emails to Data Authors

- Informal Tone
    - #4 "Hello Sir/Madam – this email is not spam."
    - #4 "I'm sure that you probably hate me and my associates for pointing out inconsistencies in your data after you've already been rewarded for publishing them. "
    - #4 "Sincerely, Your Data Curation Troll"
- Overly Formal Tone: "It is an honor and privilege to have your data accepted to the Data Curation Network and this will ensure that they are preserved, accessible and useful beyond the life of your project."
- Really Nice Tone: "Thank you for your submission to Deep Blue Data! I've reviewed the two Python scripts and the JSON file that you provided with your deposit. Thank you for including so many comments within your code and for the substantial usage documentation for the runner tool. That makes evaluation lot easier on our end."

Feelings Varied about the Curation process

- Defeatist? Feelings

    - "There were no game-stoppers with these data – they were actually in better shape than most geospatial datasets from my experience. That said, if the data producers are not responsive to the questions and requests posed to them, I think it should be withdrawn from consideration – not b/c it would then lack the information required for informed reuse, but rather b/c it's not worth the effort to clean up other people's messes for them."
- Ownership feelings
    - "I would not want my name associated with this dataset as a curator unless the researcher did a pretty substantial effort to revise the dataset, and I would guess would take a lot of back and forth with me as curator. If no work was done, we probably wouldn't want the DCN branding associated with it either. It's pretty rough."
- Pragmatic Feelings
    - "To be maximally useful, this data requires quite a bit of work, I think it will be easy to overwhelm them to the point where they would do nothing with the data at all. I hope to see some improvements in to dataset though, however minor, and I don't want to make scare them away from your resources either."

Acceptance w/o curation was an Overwhelming Yes

- "Incidentally, as did [originating repository], we would accept this dataset even without additional clean-up in the Illinois Data Bank, but we would try to lightly caution the researcher that it looks like a bit of mess and it won't be the best representation of them as a researcher."

Connection with Data Authors varied

- "Honestly, I'd offer to meet with the submitter to talk about / work with them on cleanup before presenting them with the list of suggested improvements. Face to face is less overwhelming when there's as much needed as is in this submission."
- "Would you have a few minutes to chat on the phone in the next week or so? I would be happy to discuss in more detail and we can go over some data formatting and documentation best practices that would help make the data more understandable in the future."

Post-Curation Survey Results

The post-curation survey was used to capture curator experiences for both rounds. The survey was included in the assignment as a Google Form and asked curators to rate the quality of the dataset they were curating, indicate their level of comfort with the data's discipline and primary file format types, and rate their confidence and enjoyment in the experience. (The full survey questions are presented in Appendix A.) In Round One, 4 out of 6 DCN team members completed the survey (66.6%) noting their experience in curating pilot dataset #1. However one DCN team members' survey results were invalid because they referred to a different dataset (the report from this team member was for the correct dataset). Therefore only three team member results for dataset #1 are presented here in Table 7a and 7b.

Table 7a: DCN Team ratings of "the quality of the dataset on receiving" for Pilot Dataset #1

Scale	**1**	**2**	**3**	**4**	**5**
Quality	Very High Quality	High Quality	Average	Low Quality	Very Low Quality

Pilot Dataset*	**Overall package**	**The structured data file(s)**	**The documentation file (if any)**	**The author supplied metadata**
#1	Low Quality	Low Quality	*blank*	Average
#1	Low Quality	Average	Very Low Quality	Average
#1	Low Quality	Low Quality	Very Low Quality	Very Low Quality

* One of the four DCN team member's score was inadvertantly for a different dataset and not valid for comparison.

Table 7b: DCN Team ratings of their expertise vs. level of comfort, confidence, and enjoyment with curating data sets #1

Scale	**1**	**2**	**3**	**4**	**5**
Comfort	Very Comfortable (this is my area of expertise)	Confortable	Average	Less Comfortable	Not Comfortable (this area is very intimidating)
Confidence	Very Confident (I feel this dataset was curated very well)	Confident	Average	Less Confident	Not Confident (I feel this dataset was not curated well)
Enjoyment	Very Enjoyable (I found this curation assignment energizing)	Enjoyable	Average	Less Enjoyable	Not Enjoyable (I found this curation assignment frustrating)

Pilot*	**Curator's Expertise**	**Dataset discipline**	**Primary file type**	**What was your level of <insert> when curating the dataset?**	
#1	biochemistry, organic chemistry, bioinformatics, microbiology	ecology	excel	Comfort (Discipline)	Average
				Comfort (File type)	Comfortable
				Confidence	Average
				Enjoyment	Average
#1	astronomical image files (e.g., FITS), raw/instrument data	Biological sciences, marine biology	Excel	Comfort (Discipline)	Not Comfortable
				Comfort (File type)	Very Comfortable
				Confidence	Very Confident
				Enjoyment	Very Enjoyable

#1	qualitative health data	Biological species	spreadsheet	Comfort (Discipline)	Average
				Comfort (File type)	Very Comfortable
				Confidence	Very Confident
				Enjoyment	Very Enjoyable

* One of the four DCN team members' score was for a different dataset and not valid.

In Round Two, 8 out of the 10 curators that turned in their assignment also completed the survey (80%). In the same survey used by the DCN Team members, curators for datasets #2-12 were asked to rate the quality of the dataset they were curating, indicate their level of comfort with the data's discipline and primary file format types, and rate their confidence and enjoyment in the experience. (The full survey questions are presented in Appendix A.) Responses are presented in Tables 8a and 8b and discussed in more detail in the Discussion section of this report.

Table 8a: Curator ratings of "the quality of the dataset on receiving" for Pilot Dataset #2-12

Scale	**1**	**2**	**3**	**4**	**5**
Quality	Very High Quality	High Quality	Average	Low Quality	Very Low Quality

Pilot Dataset*	**Overall package**	**The structured data file(s)**	**The documentation file (if any)**	**The author supplied metadata**
#2	Average	High Quality	Very Low Quality	Low Quality
#4	Very High Quality	High Quality	*blank*	High Quality
#5	Average	Low Quality	Average	Low Quality
#6	Low Quality	Average	*blank*	Low Quality
#7	High Quality	Average	Average	Average
#8	Very High Quality	Very High Quality	Very High Quality	Very High Quality
#9	Average	Average	Very Low Quality	Low Quality
#12	Very High Quality	Very High Quality	Very High Quality	Very High Quality

*Survey data for data sets #3, #10, and #11 not available.

Table 8b: Curator expertise vs. level of comfort, confidence, and enjoyment with curation assignment for data sets #2-12

Scale	**1**	**2**	**3**	**4**	**5**
Comfort	Very Comfortable (this is my area of expertise)	Comfortable	Average	Less Comfortable	Not Comfortable (this area is very intimidating)
Confidence	Very Confident (I feel this dataset was curated very well)	Confident	Average	Less Confident	Not Confident (I feel this dataset was not curated well)
Enjoyment	Very Enjoyable (I found this curation assignment energizing)	Enjoyable	Average	Less Enjoyable	Not Enjoyable (I found this curation assignment frustrating)

Pilot*	**Curator's Expertise**	**Dataset discipline**	**Primary file type**	**What was your level of <insert> when curating the dataset?**	
#2	Environmental sciences	Environmental Science or Dendrology	Excel data file	Comfort (Discipline)	Very Comfortable
				Comfort (File type)	Very Comfortable
				Confidence	Very Confident
				Enjoyment	Very Enjoyable
#4	GIS, ESRI products, R-Statistical, Python for geospatial applications, GDAL	Geospatial	*blank*	Comfort (Discipline)	Very Comfortable
				Comfort (File type)	Very Comfortable
				Confidence	Very Confident
				Enjoyment	Average
#5	Digital humanities, video/audio, TEI, linked data, Health sciences, sensitive data, PII	computational sociology and game theory	JSON data file and python scripts	Comfort (Discipline)	Average
				Comfort (File type)	Very Comfortable
				Confidence	Confident
				Enjoyment	Enjoyable
#6	Statistical survey packages (SPSS, SAS, R)	Political science	.sql	Comfort (Discipline)	Comfortable
				Comfort (File type)	Not Comfortable
				Confidence	Less Confident
				Enjoyment	Not Enjoyable
#7	Biological sciences, genomics	Biology: genomics	ASN.1	Comfort (Discipline)	Comfortable
				Comfort (File type)	Average
				Confidence	Confident
				Enjoyment	Enjoyable

#8	clinical health, health informatics	Public Health	*blank*	Comfort (Discipline)	Comfortable
				Comfort (File type)	Comfortable
				Confidence	Average
				Enjoyment	Average
#9	Textual formats, XML, code	Library and Information Science	XML/RDF (resource description framework file)	Comfort (Discipline)	Comfortable
				Comfort (File type)	Very Comfortable
				Confidence	Confident
				Enjoyment	Enjoyable
#12	Genomics data, databases, Excel, software code	Crop and Soil Sciences	Tabular data in csv format	Comfort (Discipline)	Average
				Comfort (File type)	Very Comfortable
				Confidence	Very Confident
				Enjoyment	Average

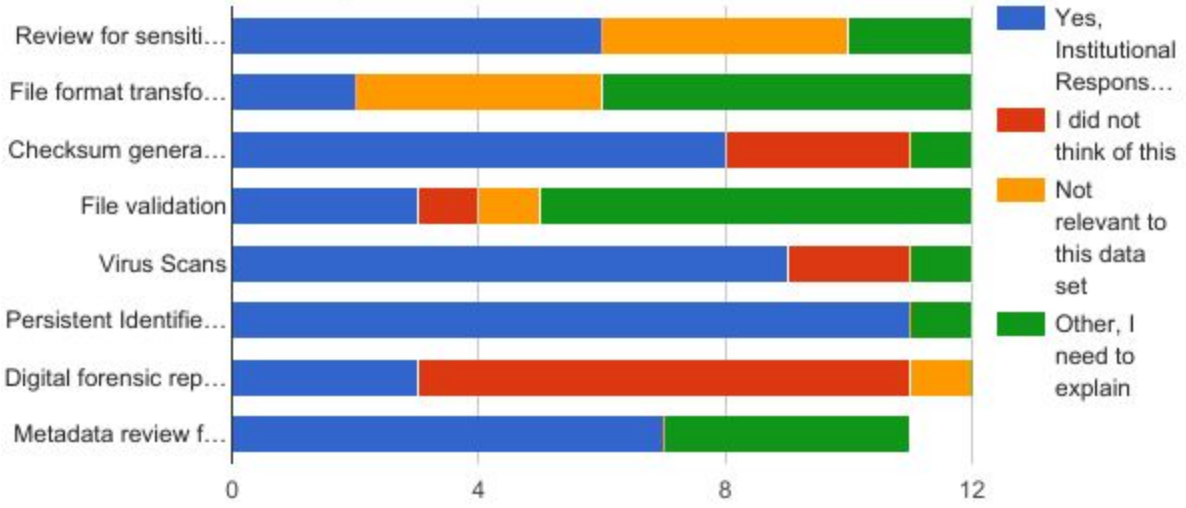
*Survey data for data sets #3, #10, and #11 not available.

Figure 3 shows the responses for all curators (n=12) curators regarding their assumptions about "What curation activities (if any) did you assume were done by the home institution?" for the following activities:

- Review for sensitive/legally protected data (eg. disclosive human subjects data, personally Identifiable information, credit card information)
- File format transformations to preservation-friendly formats
- Checksum generation
- File validation (e.g., do the formats match the extensions, do they render in appropriate software, are the corrupt, etc.)
- Virus Scans
- Persistent Identifier generation (e.g., Datacite DOI)
- Digital forensic reports
- Metadata review (e.g. is the author supplied metadata accurate and meets repository expectations)
- Other

Minting persistent identifiers was assumed done by the local institution by nearly all curators. The rest were not agreed on. One respondent sums this question up nicely in the comments for this question: "Most of the ones I assumed happened at the home institution weren't active assumptions but inherent ones based on my expectations of what a repository does (e.g., checksums, virus scan, [persistent identifier]). Otherwise, I assumed the "human stuff", like metadata review and sensitivity, was my responsibility. For me file validation is one of those things that sits in the middle...and I didn't actively do it like "I'm going to do file validation now!" but I opened up the file to review it and in doing so found out unintentionally that it was, in fact, an sql file."

Figure 3: Chart of responses to survey question “What curation activities (if any) did you assume were done by the home institution?” (n=12)



Finally, curators were asked to share any additional thoughts on the entire process. Here are their raw comments:

> ***The mix of human expertise (which is what I think the network is about) and the repo technology is really throwing me. As a network, will we be assigning DOIs if it doesn't have one? Will the DCN be collecting deposit agreements? *Maintaining* documentation? Re: Indexing... Verify all metadata provided by the author and crosswalk to descriptive and administrative metadata compliant with a standard format for repository interoperability... wouldn't we have to ask the intended repo to change their schema perhaps???***
>
> ***Though I'm well read on the subject, I'm relatively new at hands-on curation (most of my work has been been with data management issues), hence my middle of the road response on confidence. I probably missed some things that I would not have if I had more experience. I did enjoy the experience though and I'm looking forward to debriefing with other members of the DCN Team.***
>
> ***For the DCN to be successful, the level of curation needs to be deep and of substance. I would expect that subject expertise would add additional value to the dataset.***
>
> ***The data curation activity was fun, but it almost invariably leads to questions/things that are needed from the data producer and getting these things is incredibly difficult -- even when the producer is ""on the hook"" having received, or still receiving money from the person asking.***
>
> ***If data curation does not lead to better quality, documented, and preserved data then the curation activity will become less fun the more that you do it.***

Although I know quite a bit about this sort of data, I have never before actually curated a dataset for a repository, so I was uncertain of all the steps I was to perform. I just performed the steps that I thought were needed to ensure consistency, completeness, and usability of the data files, documentation, and metadata. Because I don't know much about DCN, I don't really understand the relationship between the original repository and the new repository, and this raised many questions for me, e.g., what the value is of having a duplicate data set in two different repositories? Is the intention actually for DCN to just be a catalog of datasets that are actually only housed at the originating repository, or to also have a copy at DCN? Wouldn't duplicating a dataset already in another repository lead to version control issues? What is the purpose of the letters to the prior repository and the author, when these have apparently already been curated and are already publically available? Is the purpose to have them make any additions or corrections to the current dataset on file in the original repository, prior to cataloging them in the DCN? Or to have an updated dataset actually located at the DCN? Does the curator of the original repository need a more detailed explanation of this project, and why the dataset was selected, in order to have more context in which to understand an email from the DCN? Or are they already participants in this project and expecting communication from the DCN?"

Very interesting dataset! And I found some issues/errors.

I have a lot of notes in my report!

Discussion

The results of the Data Curation Network Pilot Round 1 and Round 2 demonstrated several things: First, it generated a list of data curation activities that were already performed in local curation practice across our six institutions. Second, it highlighted where there are steps taken by only some of the DCN institutions, that perhaps should be taken by all (with the appropriate training). Third, the pilots provided a real-world context for identifying the pitfalls or challenges that the DCN may face once implemented. Based on these results, we report several key recommendations that the resulting DCN model should take into account.

Recommendation 1: Assignments to curators prioritize file format and software expertise over discipline when necessary.

When a curator worked with a data file type that they were not comfortable with they were not as confident with the result and found the experience less enjoyable. For example, Curator for dataset #6 expressed frustration over not being able to open the data file type without the necessary software, and after multiple attempts to download open source tools, the necessary expertise to use the tools. Therefore curator experts need to bring a general knowledge for a discipline, but more important is their deep expertise with software and domain file formats. Professional development or training in these areas should also be considered.

Recommendation 2: Train all curator staff to understand the purpose of the DCN and develop a baseline understanding of the roles and expectations for their work.

Based on the level of detail and number of steps taken by the DCN Team members that curated dataset #1 (average number of steps = 31) vs the the staff that curated data sets #2-12 (average number of steps = 13), there was a demonstratively stronger level of engagement by the DCN Team than the curation staff. The DCN model include in-person training to create buy-in and build shared expectations for how the DCN should operate. For example, the curator for pilot dataset #4 thought they was reviewing for acceptance into the DCN, and emailed the local institution curator "Dataset [#4] is worthy of consideration in the Data Curation Network. It is documented well enough to support informed reuse (level II Data Curation Network metadata standard), but not re-creation (level III standard). The data are geospatial and of likely high-value to current and future researchers." The curator goes on to describe to the local curator "#4 "[If] the data producers are not responsive to the questions and requests posed to them, I think it should be withdrawn from consideration – not b/c it would then lack the information required for informed reuse, but rather b/c it's not worth the effort to clean up other people's messes for them."

Recommendation 3: Centralize all DCN correspondence and perform routine checks on all submissions before assigning to DCN curator.

A coordinator role, separate from the DCN Curators, should perform routine checks (risk, rights, file inventory/manifest, and file audit) and open all files and check for integrity issues before sending the assignment to the appropriate curator. This step would have alleviated frustration felt for example by curator of dataset #7 who stated: "My inability to unzip the deposited dataset was an immediate red flag. This could either be the result of a file integrity issue, which would need to be corrected by a re-deposit, or lack of access to the correct decompression utility, which suggests that the deposit metadata needs to specify the necessary utility."

Recommendation 4: Place the control over how to work with the researcher up to the discretion of the local curator, not with the DCN.

Communication and back-and-forth between the researcher was uneven based on the results of our pilot. Some curators emailed descriptive questions others replied to the data author requesting an in-person meeting or a phone call follow-up. The timeframe sought by the curator for a response by the researcher was also an issue. For example the deadline for a response to one curators' comments was by "tomorrow...for further consideration by the Data Curation Network." This expectation may be unreasonable for many institutional cultures. The DCN should place the burden of communicating needed changes and facilitating this response on the

local curator. This not only would allow allow for variations in local culture and level of local curator support, but it would also alleviate any missed opportunity cost of researchers working with and collaborating with the local data curation staff and facilitate new relationships developed out of close collaboration through the data curation process.

Recommendation 5: Create levels of curator criteria for curators to aim for rather than allowing curators to fall into the "never ending" quest for high standards.

Curation activities could tend to be "never ending" and therefore certain minimum levels of curation must be set and activities prioritized. The time that curators spend on the assignments in our pilot ranged up to 5 hours. The curator of dataset #4 had an interesting idea of the data that was "documented well enough to support informed reuse (level II Data Curation Network metadata standard), but not re-creation (level III standard)." Additionally, the curator's email to the author of pilot dataset #2 suggested "We strive to have the datasets in eCommons meet the FAIR data principles (Findable, Accessible, Interoperable, Reusable). In order to do that, I'd love to have a conversation with you about your submission to be sure it is as complete and well-described as possible. In the meantime, I have included notes in the attached document regarding your dataset. Please contact me when the suggested changes have been made and we can work to get your data ingested into the XXX collection."

Recommendation 6: A workflow for how data should flow through the Data Curation Network should take into account activities that are common for DCN most curators and train DCN curators for activities that are not common.

Some curation activities were more common across our sample than others. This difference can be accounted for by the type of data being curated (e.g., code review is not a necessary step in the absence of code, etc.). However many activities probably could have been undertaken by all curators with the appropriate training and workflows in place. The frequency of activities taken in both rounds of our pilot are illustrated in Table 9.

Table 9: Activities taken in the pilots Round 1 (dataset #1) and Round 2 (datasets #2-12).

Activities	Round 1 (n=6)	Round 2 (n=10)	Total (n=16)	All 16, 100%	Most 11-16, > 66%	Some 4-10, 66-25%	Few, 2-3, <25%	One, 1%
Inspect files	6	10	16	1	0	0	0	0
Inspect metadata	6	10	16	1	0	0	0	0
Quality Assurance	6	9	15	0	1	0	0	0
Documentation	6	6	12	0	1	0	0	0

Communicate with Author	6	5	11	0	1	0	0	0
Metadata	4	5	9	0	0	1	0	0
Contextualization	3	5	8	0	0	1	0	0
File Format Transformations	6	2	8	0	0	1	0	0
File Inventory or Manifest	2	4	6	0	0	1	0	0
Working Copy	3	2	5	0	0	1	0	0
Risk Management	4	1	5	0	0	1	0	0
Restructure	4	1	5	0	0	1	0	0
Curation Log	3	1	4	0	0	1	0	0
File renaming	3	1	4	0	0	1	0	0
Terms of Use	4	0	4	0	0	1	0	0
File validation	1	2	3	0	0	0	1	0
Code review	0	3	3	0	0	0	1	0
Rights Management	2	1	3	0	0	0	1	0
Link Checking	1	1	2	0	0	0	1	0
Data Cleaning	2	0	2	0	0	0	1	0
Contact Information	1	1	2	0	0	0	1	0
Persistent Identifier	2	0	2	0	0	0	1	0
Versioning	2	0	2	0	0	0	1	0
Expertise Match	1	0	1	0	0	0	0	1
Virus Scan	1	0	1	0	0	0	0	1
Arrangement and Description	1	0	1	0	0	0	0	1
Interoperability	0	1	1	0	0	0	0	1
Chain of custody	0	1	1	0	0	0	0	1
Data Citation	0	1	1	0	0	0	0	1
Disk Image	1	0	1	0	0	0	0	1

Embargo	0	1	1	0	0	0	0	1
File Auditing	0	1	1	0	0	0	0	1
Full-Text Indexing	0	1	1	0	0	0	0	1
Secure Storage	1	0	1	0	0	0	0	1
Specialized Format Step	0	1	1	0	0	0	0	1
Transfer	1	0	1	0	0	0	0	1
Data Visualization	1	0	1	0	0	0	0	1

Recommendation 7: Data curation activities taken should differentiate between the role of the local repository curators versus the role of the Data Curation Network curator.

There were several data curation steps that were assumed to be the responsibility of the local data curation services, and in fact many of these rely on technical or policy decisions that may differ across institutions. For example, the appraisal of the files for local deposit. The pilots unveiled a separation of responsibilities between the Local Repository Curator and the Data Curation Network curation. How the data curation activities might possibly fall into these two roles as local vs network curation responsibilities are illustrated in table 10.

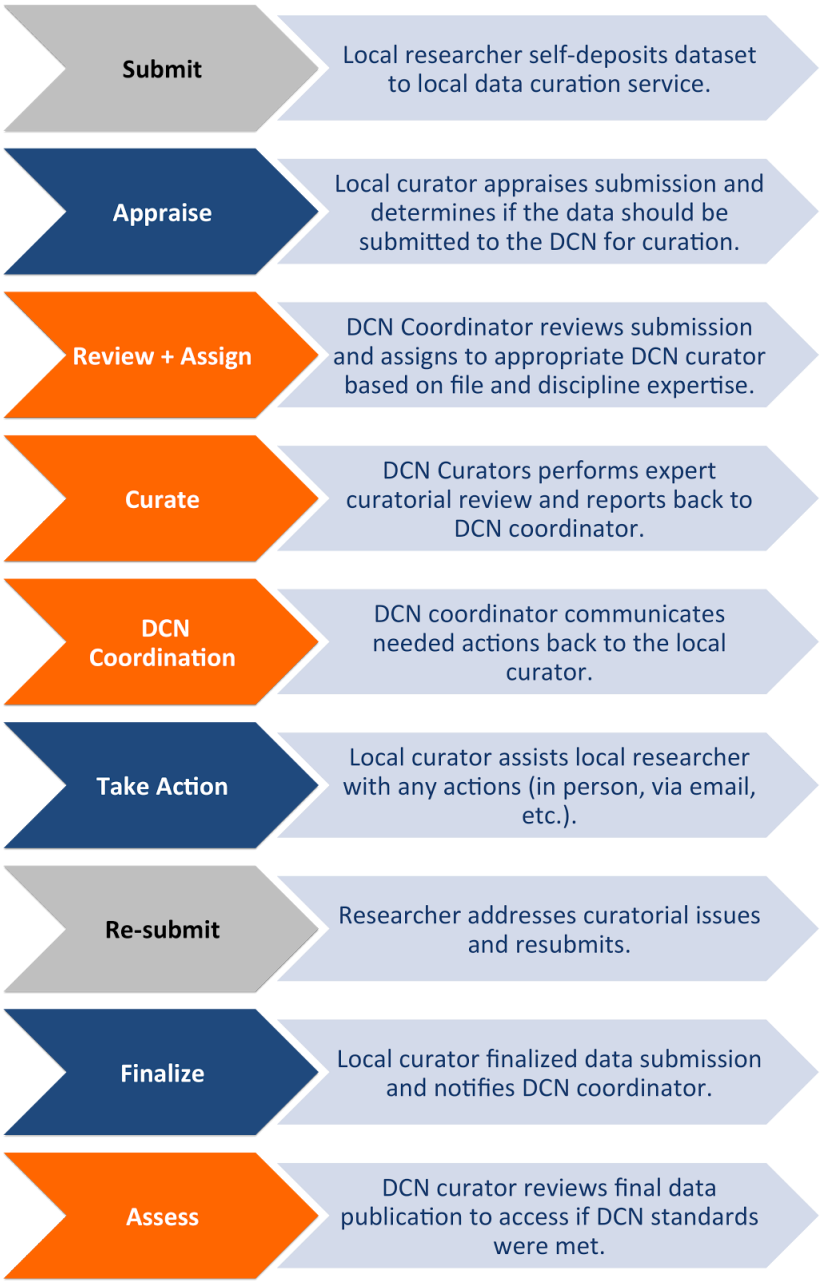
Table 10: Responsibilities for data curation activities separated by role in the Data Curation Network

Local Repository Curator Role		**Data Curation Network Curator Role**	
Deidentification Authentication Deposit agreement Appraisal / Selection Contact Information Risk Management Chain of Custody	Transfer to DCN Link Checking Virus Scan Persistent Identifier Secure Storage Terms of Use Embargo (if needed) Discovery Services	File Inventory File Validation Expertise Match Curation Log Working Copy Inspect files Inspects Metadata Documentation Metadata Rights Management	Quality Assurance Interoperability Data Cleaning Restructure Code Review Arrangement and description Contextualization File Format Transformations

Conclusion

Based on the findings of this pilot data curation exercise, the Data Curation Network now has a better understanding of the roles of the local and DCN curators and how the model for implementing the DCN might function. One possible outcome, taking into account the lessons learned here, is presented below.

Figure 4: Possible workflow for how data might flow through the Data Curation Network



Appendixes

Appendix A: Survey Questions Used in Round 1 and Round 2 of the DCN Curation Pilot

Google Forms Survey Title: DCN Curation Pilot: Curator Response
This survey will capture the responses from curators who participate in the Year 1 activity "Pilot the curation of a dataset". Please be as descriptive as possible.

1. What is your institution? (select one)
    - Cornell
    - Illinois
    - Michigan
    - Minnesota
    - Penn State
    - Wash U

2. What is your role in the Data Curation Network? (select one)
    - DCN Team Member
    - Staff member at partner institution

3. Dataset Assignment Identifier (text box)

4. How would you rate the “quality” of the dataset on receiving the assignment? (1-5 scale, 1 "Very High Quality” and 5 "Very Low Quality")
    - Overall package
    - The structured data file(s)
    - The documentation file (if any)
    - The author supplied metadata

5. How much time did you spend curating the dataset? (select one)
    - Less than one hour
    - 1-2 hours
    - 2-5 hours
    - 5-8 hours
    - 8-16 hours
    - More than 16 hours
    - Other:

6. Explain (Text box. *Help Text: How did you measure the time involved? Ex. Was is over several days or all in one sitting?*)

7. What was the subject area or discipline of the dataset? (Text box)

8. What was your level of comfort with the dataset's subject area or discipline? (Scale 1-5, 1 “Very Comfortable (this is my area of expertise)” to 5 “Not Comfortable (this area is very intimidating)”)

9. What was the primary file type found in the dataset? (Text box)

10. What was your level of comfort with the dataset's primary file type? (Scale 1-5, 1 “Very Comfortable (these files are my area of expertise)” to 5 “Not Comfortable (these file formats are very intimidating)”)

11. What curation activities (if any) did you assume were done by the home institution? (matrix, 1 “I assumed this was done already”, 2 “I assumed this will be done later”, 3 “I did not think of this”, 4 “ I thought of this, but it is not relevant to this dataset”, 5 “I need to explain…”)

- Review for sensitive/legally protected data (eg. disclosive human subjects data, personally Identifiable information, credit card information)
- File format transformations to preservation-friendly formats
- Checksum generation
- File validation (e.g., do the formats match the extensions, do they render in appropriate software, are the corrupt, etc.)
- Virus Scans
- Persistent Identifier generation (e.g., Datacite DOI)
- Digital forensic reports
- Metadata review (e.g. is the author supplied metadata accurate and meets repository expectations)
- Other

12. Explain your assumptions (Textbox)

13. What was your level of confidence when curating the dataset? (1-5 scale, 1 “Very Confident (I feel this dataset was curated very well)” to 5 “Not Confident (I feel this dataset was not curated well)”)

14. What was your level of enjoyment when curating the dataset? (1-5 scale, 1 “Very Enjoyable (I found this curation assignment energizing) to 5 “Not Enjoyable (I found this curation assignment frustrating)”)

15. Any other thoughts about this pilot activity? (textbox)

Appendix B: Definitions of Data Curation Activities

Definitions were written by the Data Curation Network team in the fall of 2016 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 2017 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: The 47 data curation activity definitions used by the Data Curation Network project

Data Curation Activity	**Definition**
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).
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.
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.
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.
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.
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.
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.
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.
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.
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.
Data Cleaning	A process used to improve data quality by detecting and correcting (or removing) defects & errors in data.
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.
Deidentification	Redacting or removing personally identifiable or protected 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.
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).
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).
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.
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)
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.
File download	Allow access to the data materials by authorized third parties.
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.
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.
File renaming	To rename files in a dataset, often to standardize and/or reflect important metadata.
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.

Full-Text Indexing	Enhance the data for discovery purposes by generating search-engine-optimized formats of the text inherent to the data.
Indexing	Verify all metadata provided by the author and crosswalk to descriptive and administrative metadata compliant with a standard format for repository interoperability.
Interoperability	Formatting the data using a disciplinary standard for better integration with other datasets and/or systems.
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).
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.
Migration	Monitor and anticipate file format obsolescence and, as needed, transform obsolete file formats to new formats as standards and use dictate.
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.
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.
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.
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).
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.
Restructure	Organize and/or reformate poorly structured data files to clarify their meaning and importance.
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).
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.
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.
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.
Software Registry	Maintain copies of modern and obsolete versions of software (and any relevant code libraries) so that data may be opened/used overtime.
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.
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.
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).
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).
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.
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.