2018

How Important Are Data Curation Activities to Researchers? Gaps and Opportunities for Academic Libraries

Lisa R. Johnston
University of Minnesota, ljohnsto@umn.edu

Jacob Carlson
University of Michigan, jakecar@umich.edu

Cynthia Hudson-Vitale
Washington University, cuv185@psu.edu

Heidi Imker
University of Illinois, imker@illinois.edu

Wendy Kozlowski
Cornell University, wak57@cornell.edu

See next page for additional authors

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

ISSN 2162-3309 10.7710/2162-3309.2198

RESEARCH

How Important Are Data Curation Activities to Researchers? Gaps and Opportunities for Academic Libraries

Lisa R. Johnston
Research Data Management/Curation Lead, University of Minnesota
Jacob Carlson
Research Data Services Manager, University of Michigan
Cynthia Hudson-Vitale
Data Servics Coordinator, Washington University
Heidi Imker
Director of the Research Data Service, University of Illinois
Wendy Kozlowski
Data Curation Specialist, Cornell University
Robert Olendorf
Science Data Librarian, Pennsylvania State University
Claire Stewart
Associate University Librarian for Research and Learning, University of Minnesota

INTRODUCTION Data curation may be an emerging service for academic libraries, but researchers actively "curate" their data in a number of ways—even if terminology may not always align. Building on past user-needs assessments performed via survey and focus groups, the authors sought direct input from researchers on the importance and utilization of specific data curation activities. **METHODS** Between October 21, 2016, and November 18, 2016, the study team held focus groups with 91 participants at six different academic institutions to determine which data curation activities were most important to researchers, which activities were currently underway for their data, and how satisfied they were with the results. **RESULTS** Researchers are actively engaged in a variety of data curation activities, and while they considered most data curation activities to be highly important, a majority of the sample reported dissatisfaction with the current state of data curation at their institution. **DISCUSSION** Our findings demonstrate specific gaps and opportunities for academic libraries to focus their data curation services to more effectively meet researcher needs. **CONCLUSION** Research libraries stand to benefit their users by emphasizing, investing in, and/or heavily promoting the highly valued services that may not currently be in use by many researchers.

Received: 05/05/2017 Accepted: 02/04/2018

Correspondence: Lisa R. Johnston, Science/Engineering Library; 108 Walter Library, University of Minnesota, Minneapolis, MN 55105, ljohnsto@umn.edu

IMPLICATIONS FOR PRACTICE

1. These findings represent opportunities for academic libraries to focus their data curation services to more effectively meet researcher needs.
2. Readers may gain a better understanding about the extent to which researchers value data curation activities, as well as how data curation activities are valued differently across disciplines.
3. Results indicate where the greatest gaps of support for highly valued data curation activities may fall.

INTRODUCTION

In the fall of 2016 the authors held six focus group sessions across six academic institutions to determine what data curation activities were important for researchers, what activities they were currently applying themselves, and how satisfied they were with the results of those efforts. In short, our research 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 in which researchers do not currently engage for their data (or do not engage as satisfactorily as they would like to). These findings represent opportunities for academic libraries to focus their data curation services to more effectively meet researcher needs.

This research was performed as part of the Data Curation Network (DCN) project funded by the Alfred P. Sloan Foundation aimed at developing a shared staffing model for curating research data. A white paper reporting the full results of this research was first published on our project website with all DCN project outputs (Johnston et al., 2017).

LITERATURE REVIEW

The role of data curation is still an emerging topic within the library science, archival, and information sciences disciplines. Just a few years ago, very few academic libraries were successfully engaging in any kind of data curation services, according to a study by Tenopir, Birch, and Allard (2012) on research data services in academic libraries. More recently, Kouper, Fear, Ishida, Kollen, and Williams (2017) provided an empirical analysis of research data services at North American research libraries, finding that data curation services were available in less than 15% of institutions surveyed and were typically viewed as an "advanced" library service.

While studies of researcher attitudes toward data curation and management are not new, many focus on high-level curation services and data management needs (McLure, Level, Cranston, Oehlerts, & Culbertson 2014; Parham, Bodnar, & Fuchs, 2012) or data sharing (Tenopir et al., 2011), without going into great detail on specific treatments and activities for curating digital assets. Many of these surveys use existing tools and frameworks for assessing faculty needs, such as the Data Curation Profiles (Witt, Carlson, Brandt, & Cragin, 2009) or the Data Asset Framework (Jones, Ball, & Ekmekcioglu, 2008). While useful tools for assessing needs for institutional research data services, they lack a mechanism to collect feedback on researchers' current practices for these treatments and assessment of their satisfaction for these treatments. A scoping review of 310 articles by Perrier et al. (2017) found that most research data management studies performed by academic institutions do not include direct interaction with data producers but instead rely on indirect methods such as self-reporting surveys and case studies by a third-party observer. Jahnke, Asher, and Keralis's 2012 CLIR study, however, does approach researcher attitudes directly via their method of ethnographic interviews with social sciences researchers at five institutions. Bardyn, Resnick, and Camina (2012) also provide a useful methodology from their focus groups with translational sciences researchers. Although our methods differ, these studies provide a number of comparable insights to this study, such as researchers' low satisfaction level with their data curation know-how and the lack of ability to perform curation actions on their data due to lack of time and a burdensome workload.

The lack of shared definitions for data curation terms has been an important area of discussion, recently prompting an Research Data Alliance (RDA) Working Group to task itself with establishing "a reference data terminology that can be used across communities and stakeholders to better synchronize conceptualization" (RDA, 2016). To pursue our question on which data curation activities are most important to researchers, the authors consulted several sources to obtain term definitions and rework them for our study participants, including the CASRAI Dictionary, the Research Data Alliance (RDA) Terms Definition Tool, the Digital Curation Center (DCC) Glossary, the ICPSR Glossary of Social Science Terms, the Research Data Canada Glossary, the Digital Preservation Coalition Glossary, and the Society of American Archivists Terms Glossary. Along a parallel path, much can be learned from reviewing "competences" for both data curators and researchers working with data. For example, research by Madrid (2013) surveyed multiple panels of experts, using the Delphi Method, to develop consensus around competencies for digital curators. The 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." Librarians who work specifically with data have been found to need similar skills by Schmidt and Shearer (2016). And twelve researcher-focused competencies

were explored in detail in the Data Information Literacy project (Carlson, Fosmire, Miller, & Nelson, 2011; Carlson & Johnston, 2015), which focused on the educational needs of graduate students across a variety of science disciplines.

To better define the activities involved with data curation, work by the DigCCurr program (Lee, 2009) provides a useful matrix of curation themes and ideas but does not supply them with sufficiently detailed definitions. Follow-up work by Bowden, Lee, and Tibbo (2011) focused on the curator views of DigCCurr activities in the Closing the Digital Curation Gap project (http://digitalcurationexchange.org/cdcg). Their focus groups provided a good template for the present study. To ensure the inclusion of activities important to the digital repository community, the TRAC assessment tool by the Center for Research Libraries (CRL) and the Online Computer Library Center (OCLC) (2007) was consulted, but the language is jargon laden and lacks a researcher assessment of needs. Finally, the Digital Curation Center's data lifecycle model (Higgins, 2008) and the Data Curation Handbook Steps (Johnston, 2017) paved the way for defining the Data Curation Activities used in our study.

METHODS

Between October 21, 2016, and November 18, 2016, the authors engaged with researchers, librarians, and research support staff across six focus group sessions, termed "Data Curation Roundtables," held at the following academic institutions: Cornell University, Penn State University, the University of Illinois at Urbana-Champaign, the University of Michigan, the University of Minnesota, and Washington University in St. Louis. The 91 participants represented a diverse mix of experience levels (e.g., faculty, graduate student, postdoc) and a variety of disciplines (see Table 1, hereafter "participants"). Each session lasted one and a half hours, with lunch provided for free in exchange for attendees' participation. Notably, participants were either recruited through direct invitation or attended the open session due to self-interest; in both cases, selection bias impacted the representation of the sample.

Institution	Cornell	Wash U	Illinois	Penn State	Minnesota	Michigan	Total
Date of Session	2016-10-11	2016-10-25	2016-10-27	2016-11-04	2016-11-14	2016-11-18	All 6 Sessions
Sciences & Engineering	9	6	10	5	11	12	53
Social Sciences	6	1	2	1	1	4	15
Humanities	0	1	1	1	0	2	5
Library and Information Science Faculty	0	0	5	0	0	0	5
Service Providers*	5	3	0	4	1	0	11
Total	20	11	18	11	13	18	91

Table 1. Disciplinary/professional distribution of participants at the six focus group sessions
*Service providers, such as IT staff and library staff, were grouped into this category.

These sessions sought to engage directly with both the communities that produced data and those that are likely to make use of data sets authored by others, to better understand the value of data curation. The goals of the focus group sessions were to answer these questions:

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

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

Definitions of Data Curation Activities

In preparation for the sessions, the authors identified and defined 47 data curation activities relevant to data curation services and best practices (see Appendix). In addition, at the start of each focus group session, several key definitions were presented to all participants to set the foundation for the event:

- 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 focus group session was broken into three parts corresponding to each of our questions. First, a card-swapping and rating exercise captured the participants' opinion of the importance of data curation activities for their data. Second, a paper-based survey instru-

ment collected their levels of engagement and satisfaction with those same data curation activities. Third, the authors engaged participants in a facilitated focus group discussion about the challenges of applying the top-five most highly rated data curation activities from the first exercise in their individual workflows. To aid consistency of our methods, one author [name removed] was present for all six sessions. The methodology for each part is described in more detail below.

Part 1: Rating the Importance of Data Curation Activities

To address the first question, the authors first asked participants to rate the importance of a selection of 18–20 data curation activities. Not all the activities were selected for the rating exercise, as it was up to the local facilitator to select the subset of activities to focus on depending on their local service offerings and interest.1 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 and supplementary file). Each participant was given two to four cards at a time, and then was instructed to read each definition and rate that activity's importance from 1 (lowest) to 5 (highest). Once each card in their hand was rated, the participants were asked to exchange their cards with another participant in the room and repeat for a total of four rounds each. Since there were two or three copies of the same card circulating around the room, participants were advised to trade with those who had cards they had not rated previously. A quick total of all four rounds yielded a priority list of data curation activities that were used as the focus of the group discussion throughout the session.

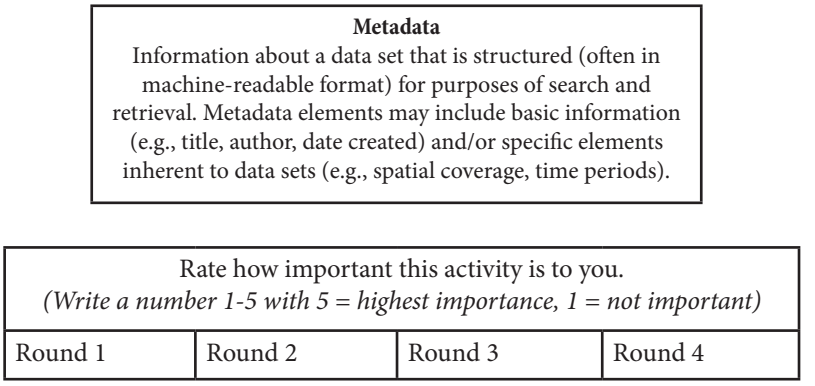
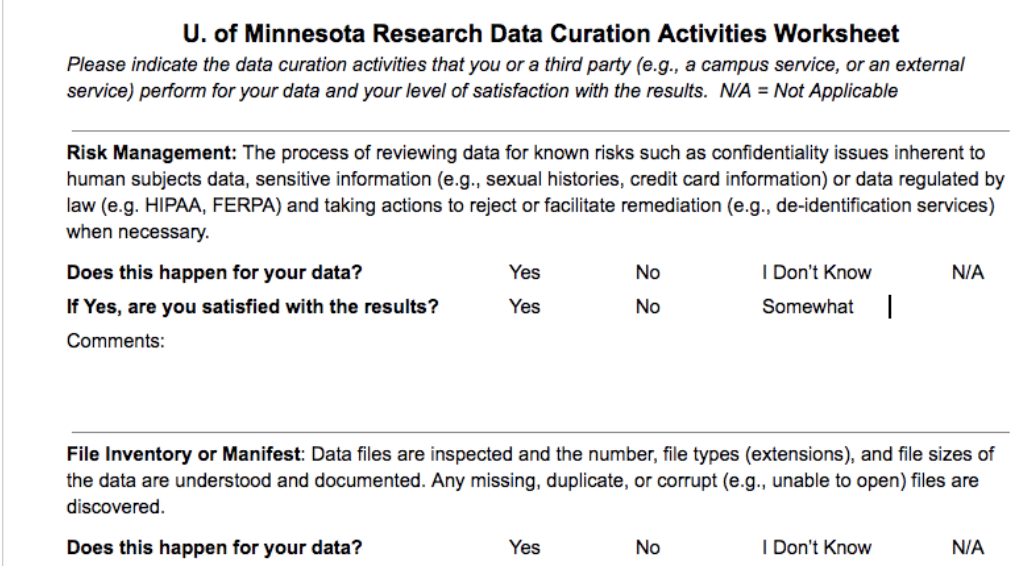


Figure 1. The front and back of an example card used in the importance-rating activity

[1] Twelve activities defined by the DCN were not rated at any of the researcher engagement sessions: arrangement and description, authentication, ceasing data curation, conversion (analog), deposit agreement, file download, file renaming, indexing, restructuring, selection, succession planning, and transcoding.

Part 2: Capturing Researcher Engagement and Satisfaction with Data Curation Activities

To address the second question, a worksheet (see Figure 2 and supplemental file) with 18–20 of the same data curation activities captured participant responses to the questions "Does this happen for your data?" and "If Yes, are you satisfied with the results?" along with space for comments. Of the 47 data curation activities, 32 were assessed using the worksheet exercise, with the selection and order varied at each institution according to the preference of the local author (e.g., service offerings already provided by that institution).[2]



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

Figure 2. Worksheet instrument used to gauge researcher satisfaction with data curation activities

To better understand how data curation activities were happening, researchers were asked to provide comments describing how and by whom (themselves or a third party) a particular activity occurred or to explain why they were or were not satisfied with the results.

[2] In addition to the 12 activities not chosen for the card-rating activity listed in footnote 1, the following three activities were not assessed with the worksheet exercise: curation log, emulation, and interoperability.

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

Finally, to answer our third question, the sessions allowed ample time to discuss the most highly rated data curation activities in greater detail. Breaking out into small groups of four to six, the researchers described their current practices for engaging with the top-rated data curation activities (resulting from Part 1), the challenges and barriers to this work, and the means by which these services were generally obtained. The notes were captured by the authors in attendance or by support from library staff members at that institution. The discussion sessions were designed to test several of our key assumptions as leads/directors of library-based data curation services:

- The value of data curation is not easy to measure and/or may be unknown,
- There exists a complex and evolving ecosystem of differing expectations with respect to research data such as functional vs. domain curation and researcher needs vs. funder needs, and
- It can be better or easier to just do it yourself when it comes to data curation.

RESULTS

The six sessions generated results for each of our three questions: First, what data curation activities do researchers see as important or having value, either to themselves or to their communities of practice? Second, how, to what extent, and why do researchers engage in data curation activities as a normative part of their research workflows? Third, what level of satisfaction do researchers have with their current data curation treatments, or what are the barriers preventing researchers from pursuing them (time, personnel, knowledge, money, equipment, other resources)?

Part 1 Results: Rating the Importance of Data Curation Activities

The card-rating exercise revealed the most important data curation activities for participants overall and by institution. Of the 35 activities, 31 rated received at least an average 3 out of 5 rating for importance. Table 2 displays how activities were rated in descending order of average importance and the frequency with which each activity was rated (NB: a higher count is proportional with our confidence in the rating with a minimal threshold of two groups for calculating the rating range).

Rank	Data Curation Activity	C	WU	IL	P	MN	MI	Count of Ratings	Average Rating	Rating Range*
Rating = 5 Highest Level of Importance "Most Important"										
1	Documentation	X	X	X	X	X	X	6	4.6	4.9 – 3.5
2	Chain of Custody		X					1	4.5	n/a
3	Secure Storage	X	X		X		X	4	4.4	5.0 – 3.9
4	Quality Assurance	X	X	X	X	X		5	4.3	4.6 - 3.9
5	Persistent Identifier	X	X	X	X	X	X	6	4.3	4.8 – 4.0
6	Discovery Services				X			1	4.3	n/a
7	Curation Log				X			1	4.1	n/a
8	Technology Monitoring Refresh			X				1	4.1	n/a
9	Software Registry		X				X	2	4.1	4.3–3.9
10	Data Visualization		X			X		2	4.0	4.0–4.0
11	File Audit	X		X	X	X	X	3	4.0	4.3–3.5
12	Metadata	X		X	X	X	X	5	4.0	4.9–3.9
Rating = 4 out of 5 Level of Importance "Very Important"										
13	Versioning	X	X	X	X	X	X	6	3.9	4.8–3.4
14	Contextualize	X	X	X	X	X	X	6	3.9	4.6–3.3
15	Code review	X	X	X	X	X	X	6	3.9	4.5–2.9
16	File Format Transformations	X	X	X	X	X		5	3.8	4.5–3.3
17	Interoperability				X	X		2	3.8	4.9–3.3
18	Data Cleaning		X		X			2	3.8	4.0–3.5
19	Embargo	X	X	X	X	X	X	6	3.7	4.1–3.3
20	Rights Management	X			X	X	X	4	3.7	4.3–3.0
21	Risk Management	X		X	X	X	X	5	3.6	3.9–3.0
22	Use Analytics	X	X	X	X	X	X	6	3.6	4.1–3.0
23	Peer Review		X	X		X		3	3.5	4.8–2.6
24	Terms of Use	X		X	X	X		4	3.5	3.6–3.4
25	Data Citation	X		X	X		X	4	3.5	4.1–2.8
26	File Validation	X		X	X		X	4	3.4	4.0–3.0
27	Migration		X				X	2	3.4	3.9–2.8
28	File Inventory or Manifest	X		X	X	X		4	3.2	3.5–2.8
29	Metadata Brokerage	X		X	X	X	X	5	3.2	4.0–2.6
30	Deidentification	X		X	X	X		4	3.1	4.3–2.1
31	Repository Certification			X				1	3.0	n/a
Rating = 3 out of 5 Level of Importance "Important"										
32	Emulation		X	X				2	2.9	3.1–2.6
33	Restricted Access	X			X			2	2.6	2.9–2.4
34	Correspondence						X	1	2.5	n/a
35	Full-Text Indexing					X		1	2.5	n/a
Rating = 2 out of 5 Level of Importance "Less Important"										
Rating = 1 out of 5 Level of Importance "Not Important"										

Table 2. The 35 data curation activities as rated by 91 participants across six focus group sessions (C=Cornell University, P=Penn State University, IL = University of Illinois, WU = Washington University in St. Louis, MI = University of Michigan, MN = University of Minnesota).

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

Part 2 Results: Engagement and Satisfaction with Data Curation Activities

The worksheet exercise revealed the activities in which researchers currently engaged, what techniques they used, and their levels of satisfaction with the results. Out of the 91 participants, 4 failed to turn in their worksheets (due to leaving early, etc.), and the missing worksheets were coded as "did not answer." Additionally, the response "Sometimes" was introduced as a coded answer applied only when a participant circled both yes and no. In total, 32 of the data curation activities were analyzed by participants in this exercise and 44% "Yes" responses to "Does this [data curation activity] happen for your data?" indicated that many data curation activities were happening for participants in a variety of ways (see Figure 3: documentation (80%), secure storage (75%), chain of custody (64%), metadata (63%), file inventory or manifest (58%), data visualization (58%), versioning (56%), file format transformations (55%), and quality assurance (52%) marked as "Yes, happening").

However, overall satisfaction with data curation activities was low, with only 18% responding positively to our question regarding satisfaction with the results of those activities (see Figure 4). More often participants who received data curation activities for their data were either not satisfied or only somewhat satisfied. No activity was occurring in a satisfactory way for a majority of participants. Secure storage came the closest at 39% satisfied, while efforts to create metadata and perform file format transformations satisfied 29% of our sample.

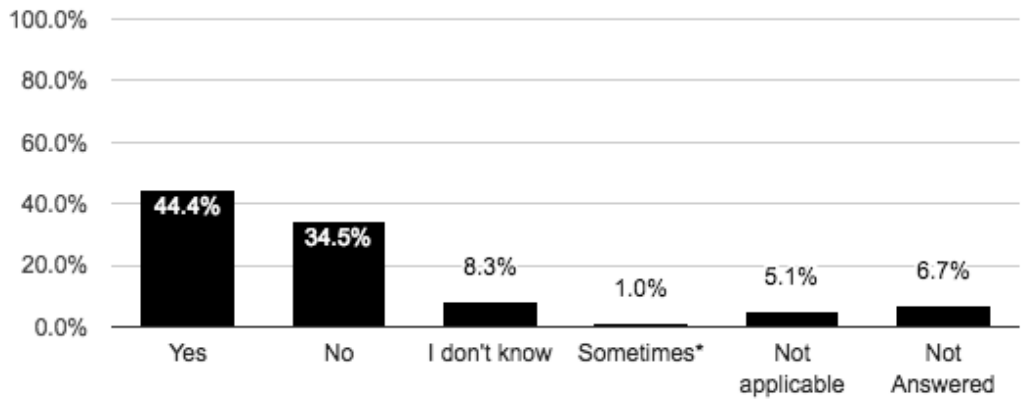


Figure 3. Overall breakdown of 91 participant responses to "Does this [data curation activity] happen for your data?" (Total =100%)

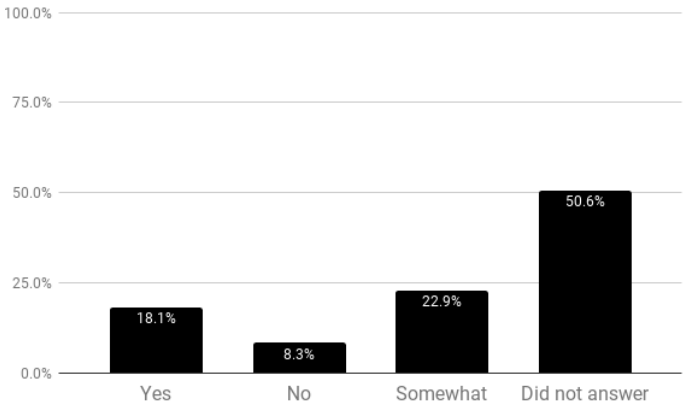


Figure 4. Overall breakdown of 91 researcher responses to "If yes [this data curation activity happens for your data], are you satisfied with the results?" (Total=100%)

Looking closer at the top-12 highly rated activities reveals key areas of opportunity for libraries, as many important data curation activities are not happening in a satisfactory way. Table 3 shows the worksheet responses for 11 of the 12 data curation activities that averaged a score of 4 or higher on a 5-point scale in Part 1, and these findings are explored in our discussion (the responses to all 35 activities are appended as a supplemental file).

"Does this activity happen for your data?"			If Yes, Are You Satisfied? (percent of total)			
Data Curation Activity	Rating	"Yes, this happens"	Yes	No	Somewhat	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 & 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%

Table 3. "Very important" data curation activities with recorded levels of engagement and satisfaction by 91 participants

* The data curation activity "Curation Log" was also highly rated 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.

Comments provided by participants in the worksheet provided additional detail as to how researchers were applying data curation activities and their difficulties in obtaining such services. Time was a factor for many researchers, with one citing, "Much more to do with limited staff. Running into trade off of documentation vs. work." For activities such as documentation and metadata, comments expressed a desire for more standards and templates: "[Documentation] 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." A few comments echoed the lack of standards and cited more ad hoc practice: "I don't use technical metadata, but instead use the file[name] title to keep track of this." Overall, comments expressed more instances of frustration than exemplars and demonstrated a desire for greater support in many data curation activities.

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 in data curation activities. In each session we focused on five of the top-rated data curation activities for that session. Two of the focus groups session discussions are profiled here and complement the results from Parts 1 and 2 by providing more granularity to the importance of data curation activities.

Case Study: University of Illinois Focus Group Discussion

Conversation in the room was free-flowing. Participants seemed to somewhat self-assemble at tables where they knew people, so we had a table with the bulk of the health sciences attendees, another with participants from 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, one thread of the discussion revolved around being surprised at the low rating that others had given to "de-identification." Given the importance of human subjects to health sciences research, one of the participants was mortified that someone at another table rated it as "3," and two others at the table also expressed bafflement. One attendee shared that they were asked to share raw MRI data with collaborators at [another institution], and they were concerned about the possibility of facial reconstruction and subsequent ability to identify the research subjects. A proposed solution was to make those accessing the data at [the other institution] sign an agreement saying they promised not to attempt identification, but the researcher expressed dissatisfaction that such a 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. The respective fields of the focus group participants are highly competitive, and there was concern

expressed about being scooped and losing out on publications. One participant expressed feelings that producing fewer publications would not only 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 seen 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 the disclosure of sensitive data, one participant at the health sciences table was interested in having an "authority" on campus to turn to for situations such as the MRI example.

Case Study: University of Michigan Focus Group Discussion 3

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 acknowledgment 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 software and scripting data make quality assurance a challenge.

[3] Excerpt from full case study report published online by Carlson (2017).

Finally, data curation is a new or emerging area for attendees and for their research communities. Many of them have not yet had to address curation activities such as file validation or file format transformations, though these 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 when assigning importance to activities. Use analytics, for example, had particularly wide variance: attendees who were actively sharing data gave it a high-importance rating, and attendees who were not yet sharing data rated it lower. Generally, curation activities that would directly benefit the researchers, such as persistent identifiers and contextualization to link the data and research outputs, were of particular interest in our group discussions, even if they were not given a high rating of importance.

DISCUSSION

Our focus groups on researcher attitudes toward data curation activities answered our three questions. We identified which data curation activities participants in our sample saw as important or having value to themselves or to their communities of practice. In this way, developing an understanding of which curation activities researchers value will help providers develop and deliver services that are more in line with real-world needs and expectations. Next we determined how, to what extent, and why our participants engaged in data curation activities themselves as a normative part of their research workflows. Finally, we identified gaps in highly valued data curation activities in which the sampled participants did not engage for their data (or engage as completely as they would like to) and some of the barriers preventing them from doing so.

Study limitations

Although well-suited to our purposes of examining the particular needs of researchers across the partner institutions designing a shared data curation service as part of the Data Curation Network project, our study presents some limitations for understanding researcher attitudes regarding data curation activities more generally. For example, the local facilitator chose which activities to include in the rating activity either in accordance with perceived local interest or in order to eliminate activities that might be difficult to offer across institutions. Therefore, as mentioned in footnote 1, twelve activities defined by the DCN were not rated at any of the researcher engagement sessions. Furthermore, only 4 activities out of 34 rated below a 3 on a 5-point scale for importance (see Figure 5). These were emulation, restricted access, correspondence or contact information, and full-text indexing. However, since our sample was composed of self-selected and invited attendees with interest or experience in data curation to our session titled "Data Curation Round-

table," the results may be more positive toward data curation topics in general, and we do not propose that these findings of importance are typical for all researchers.

"Very Important" Average Rating of 4.0–4.9	"Important" Average Rating of 3.0–3.9	"Less Important" Average Rating of 2.0–2.9	"Not Important" Average Rating 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	

Figure 5. Average rating of importance for 35 data curation activities

Levels of Importance and Satisfaction

Based on the results of the Part 2 worksheet exercise, our analysis found that no single data curation activity was happening in ways that satisfied the majority of our participants (see Figure 6). The activity that came closest was secure storage, which was occurring for 75% of our sample yet satisfied only 38%. Notably, two activities were found to satisfy a greater percentage of researchers than was reported for their data, repository certification and migration, possibly indicating that participants were satisfied with these activities not happening (see Figure 6).

Our study found gaps in support for data curation activities that are very important (average rating of at least 4 out of 5 in importance) but that are either not happening or not happening in a satisfactory way for a majority of our researchers (Figure 7). These may be areas of opportunity for libraries to invest in new services and/or heavily promote services that may already exist but are not reaching the researchers who value them:

- 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).

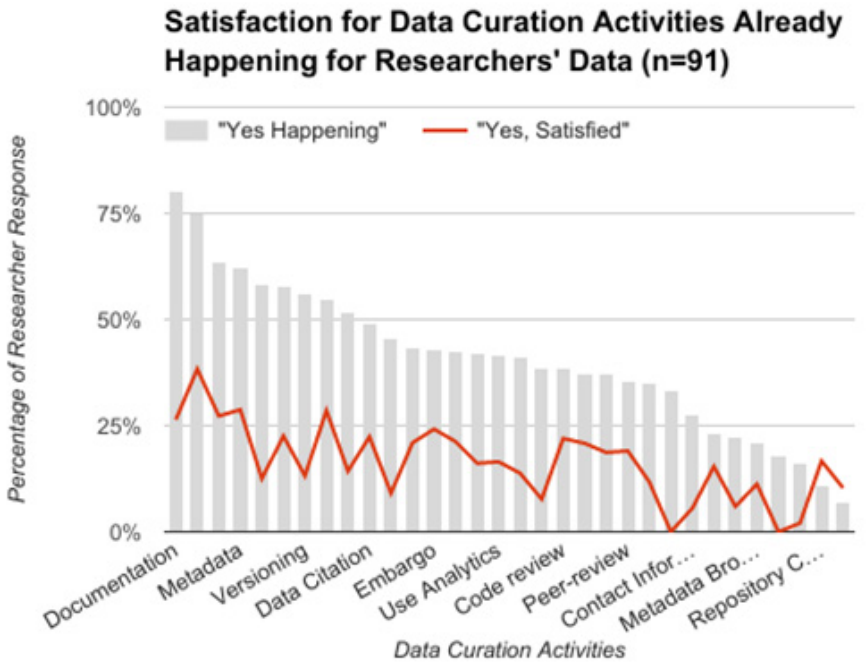
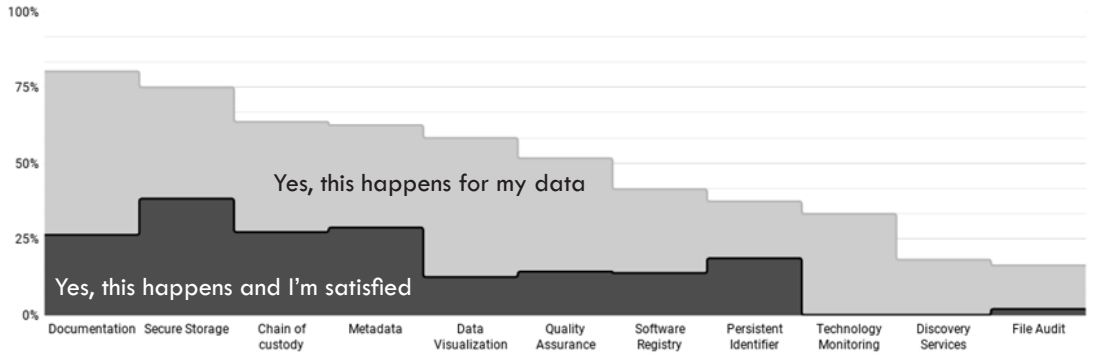


Figure 6. Visualization of worksheet responses indicating levels of satisfaction with data curation activities that were happening for participants' data

Similarly, several highly rated data curation activities were happening for a majority of our researchers, but researchers were not overwhelmingly satisfied with the results. Therefore, libraries might provide better tools and/or best practices to increase the effectiveness of these data curation activities for the researchers who engage in them:

- 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 data (only 12.5% satisfied), and
- creating and or applying metadata (only 29% satisfied).



Data Curation Activity	Documentation	Secure Storage	Chain of custody	Metadata	Data Visualization	Quality Assurance	Software Registry	Persistent Identifier	Technology Monitoring	Discovery Services	File Audit
Importance Rating (1-5)	4.6	4.4	4.5	4	4	4.3	4.1	4.3	4.1	4.3	4
Yes, this happens	80%	75%	64%	63%	58%	52%	41%	37%	33%	18%	16%
Yes	26%	38%	27%	29%	13%	14%	14%	19%	0%	0%	2%

Figure 7. Percent of Satisfaction for the Data Curation Activities rated Very Important where light grey represents "Yes this happens" and dark grey represents "Yes, this happens and I'm Satisfied" on a 100% scale.

CONCLUSION

The results of our focus groups with researchers provided a number of key findings that were used to build evidence for the specific activities a collaboratively staffed Data Curation Network might focus on in the future. But we also learned several things that could inform the development of better academic library data curation services more generally. Our focus groups revealed that while researchers were actively engaged in a variety of data curation activities for their data, none of these activities were happening in a satisfactory way for the majority of our group. Second, discussions with researchers revealed the various ways in which researchers engaged in some data curation activities as well as their barriers for not doing so for others, including time constraints and the lack of clear standards. We suggest, therefore, that research libraries stand to benefit their users by emphasizing, investing in, and/or heavily promoting the highly valued services that may not be happening for many researchers, namely minting and managing persistent identifiers, maintaining a software registry, providing tools and support for auditing file integrity, creating and managing metadata that places data within a context of related publication sources, and providing code-review services. Similarly, libraries might support better tools and/or best

practices to increase the levels of satisfaction for these commonly occurring data curation activities that are falling short of expectations, including maintaining up-to-date data documentation templates that could be used by a variety of researchers, providing best practices for secure storage, creating quality assurance checklists and review procedures for a variety of data formats and types, recommending best practices or tools for data visualization, promoting better adoption of metadata standards across disciplines, recommending tools and file-naming schemas for versioning data sets, and being more transparent about the conditions and procedures for file format transformations.

REFERENCES

Bardyn, T. P., Resnick, T., & Camina, S. K. (2012). Translational researchers' perceptions of data management practices and data curation needs: Findings from a focus group in an academic health sciences library. Journal of Web Librarianship, 6(4), 274–287. https://doi.org/10.1080/19322909.2012.730375

Bowden, H., Lee, C., & Tibbo, H. (2011, June 28). Closing the digital curation gap focus groups report. Digital Curation Exchange Year 2 Advisory Board Meeting, London, UK. http://digitalcurationexchange.org/cdcg/sites/default/files/CDCG_FocusGroupReport.pdf

Carlson, Jake. (2017, January 9). Data curation priorities and activities: A report from a researcher engagement event at the University of Michigan." Deep Blue. http://hdl.handle.net/2027.42/136229

Carlson, J., & Johnston, L. R. (Eds.). (2015). Data information literacy: Librarians, data, and the education of a new generation of researchers. West Lafayette, IN: Purdue University Press. https://doi.org/10.26530/OAPEN_626975

Carlson, J., Fosmire, M., Miller, C. C., & Nelson, M. S. (2011). Determining data information literacy needs: A study of students and research faculty. portal: Libraries and the Academy, 11(2), 629–657. https://doi.org/10.1353/pla.2011.0022

Center for Research Libraries (CRL) and Online Computer Library Center (OCLC). (2007, February). Trustworthy Repositories Audit and Certification (TRAC): Criteria and checklist. Version 1.0. Chicago: CRL and Dublin, OH: OCLC. http://www.crl.edu/sites/default/files/d6/attachments/pages/trac_0.pdf

Digital Curation Center (DCC). (n.d.) DCC curation lifecycle model. Retrieved May 2, 2017, from http://www.dcc.ac.uk/resources/curation-lifecycle-model

Higgins, S. (2008). The DCC curation lifecycle model. International Journal of Digital Curation, 3(1), 134–140. https://doi.org/10.2218/ijdc.v3i1.48

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. https://www.clir.org/pubs/reports/pub154/pub154.pdf

Johnston, L. R. (2017). Summary of the "Data Curation Handbook Steps" from Curating research data volume two: A handbook of current practice. American College & Research Libraries. Retrieved from the University of Minnesota Digital Conservancy, http://hdl.handle.net/11299/183502

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. https://doi.org/10.2218/ijdc.v3i2.62

Kouper, I., Fear, K., Ishida, M., Kollen, C., & Williams, S. C. (2017). Research data services maturity in academic libraries. In L. R. Johnston (Ed.), Curating research data, volume one: Practical strategies for your digital repository, 153–170. Chicago: Association of College and Research Libraries.

Lee, C. (2009, June 17). Matrix of digital curation knowledge and competencies (overview), version 13, DigCCurr Project. Retrieved from https://ils.unc.edu/digccurr/digccurr-matrix.html

Madrid, M. M. (2013). A Study of digital curator competences: A survey of experts. The International Information & Library Review, 45(3), 149–156. https://doi.org/10.1016/j.iilr.2013.09.001

McLure, M., Level, A. V., Cranston, C. L., Oehlerts, B., & Culbertson, M. (2014). Data curation: A study of researcher practices and needs. portal: Libraries and the Academy, 14(2), 139–164. https://doi.org/10.1353/pla.2014.0009

Parham, S. W., Bodnar, J., & Fuchs, S. (2012). Supporting tomorrow's research: Assessing faculty data curation needs at Georgia Tech. College & Research Libraries News, 73(1), 10–13. https://doi.org/10.5860/crln.73.1.8686

Perrier, L., Blondal, E., Ayala, A. P., Dearborn, D., Kenny, T., Lightfoot, D., . . . & MacDonald, H. (2017). Research data management in academic institutions: A scoping review. PloS ONE, 12(5), e0178261. https://doi.org/10.1371/journal.pone.0178261

Research Data Alliance Data Foundation and Terminology Work Group (RDA). (2016, December 22). Data Foundation and Terminology Work Group Products. https://doi.org/10.15497/06825049-8CA4-40BD-BCAF-DE9F0EA2FADF

Schmidt, B., & Shearer, K. (2016, June). Librarians' competencies profile for research data management. Joint Task Force on Librarians' Competencies for E-Research and Scholarly Communication. Retrieved from https://www.coar-repositories.org/files/Competencies-for-RDM_June-2016.pdf

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

Tenopir, C., Birch, B., & Allard, S. (2012, June). Academic libraries and research data services: Current practices and plans for the future; an Association of College and Research Libraries white paper. Retrieved from http://www.ala.org/acrl/sites/ala.org.acrl/files/content/publications/whitepapers/Tenopir_Birch_Allard.pdf

Witt, M., Carlson, J., Brandt, D. S., & Cragin, M. H. (2009). Constructing data curation profiles. International Journal of Digital Curation, 4(3), 93–103. https://doi.org/10.2218/ijdc.v4i3.117

APPENDIX

Definitions of Data Curation Activities and Ranking in Our Focus Groups

Definitions were written by the authors by consulting the following sources: The CASRAI Dictionary (http://dictionary.casrai.org/Main_Page), the Research Data Alliance (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 handbook by Lisa R. Johnston (ALA/ACRL Press) *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).

Data Curation Activity	Definition	Focus Group Rank
Arrangement and Description	The reorganization of files (e.g., new folder directory structure) in a data set 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 Rated
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 Rated
Ceasing 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 Rated
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 is 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
Contextualizing	Use metadata to link the data set to related publications, dissertations, and/or projects that provide added context for how the data were generated and why.	34
Conversion (Analog)	In an 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., PDFs) 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 Rated
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 data set 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 and 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 information (e.g., sensitive geographic locations) from a data set prior to sharing with third parties.	30
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 Rated
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 Rated
File Format Transformations	Transform files into open, nonproprietary 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., unopenable) files are discovered.	28
File Renaming	To rename files in a data set, often to standardize and/or reflect important metadata.	Not Rated

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 whether 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 Rated
Interoperability	Formatting the data using a disciplinary standard for better integration with other data sets 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) and/or specific elements inherent to data sets (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 who meet predefined criteria.	33
Restructure	Organize and/or reformate poorly structured data files to clarify their meaning and importance.	Not Rated
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., deidentification 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 is deemed appropriate for acceptance and ingest into the repository according to local collection policy and practice.	Not Rated
Software Registry	Maintain copies of modern and obsolete versions of software (and any relevant code libraries) so that data may be opened/used over time.	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 Rated
Technology Monitoring and Refreshing	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 Rated
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 over time that includes metadata describing the version history and any changes made for each version.	13