Data Curation Network Special Report
Results of the Librarian-Focused Rankings of the Data Curation Activities
Release Date: May 19, 2017
Authors: Lisa R Johnston (PI), Jake Carlson, Cynthia Hudson-Vitale, Heidi Imker, Wendy Kozlowski, Robert Olendorf, and Claire Stewart.

The Data Curation Network project began the first planning phase 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. The results presented here represent one activity of the DCN's first year to seek input from researchers to better understand how data curation services fit into their research workflow and data management needs.

Introduction

During the phase planning of the Data Curation Network project, our team surveyed researchers on the importance of various data curation activities for their data. Our goal was to identify highly rated activities that researchers felt were important but were either not currently happening for their data, or not happening particularly well. The results of our work in this area are presented here (link to summary) and in this full report (Summary Researcher Engagement Events). The team had several opportunities to run the rating exercise on the importance of Data Curation Activities with various librarian groups. The Librarian perspectives are presented here and compared with our findings from researchers.

Methodology

The rating activity was run in two different ways for librarian groups. First, in a training session for 17 library staff attending the SHARE "Data Curators" User Meeting held in Atlanta, GA on January 24, 2017, the PI of the grant lead the same card ranking exercise used in all 6 of the researcher engagement sessions (full methodology described here). Second, using a different approach, the team captured ratings of importance in a survey instrument facilitated as part of the annual American Research Libraries SPEC Kit program. Our survey on Data Curation services (SPEC Kit No. 354) was run January 3, 2017 through January 30, 2017 and collected responses from 80 of the 124 ARL member libraries (65%). In this instrument, the first question branched the survey (see Table 1). Therefore based on their answer to the question "Does your institution currently provide research data curation services?" the tool would jump according to the following rules:

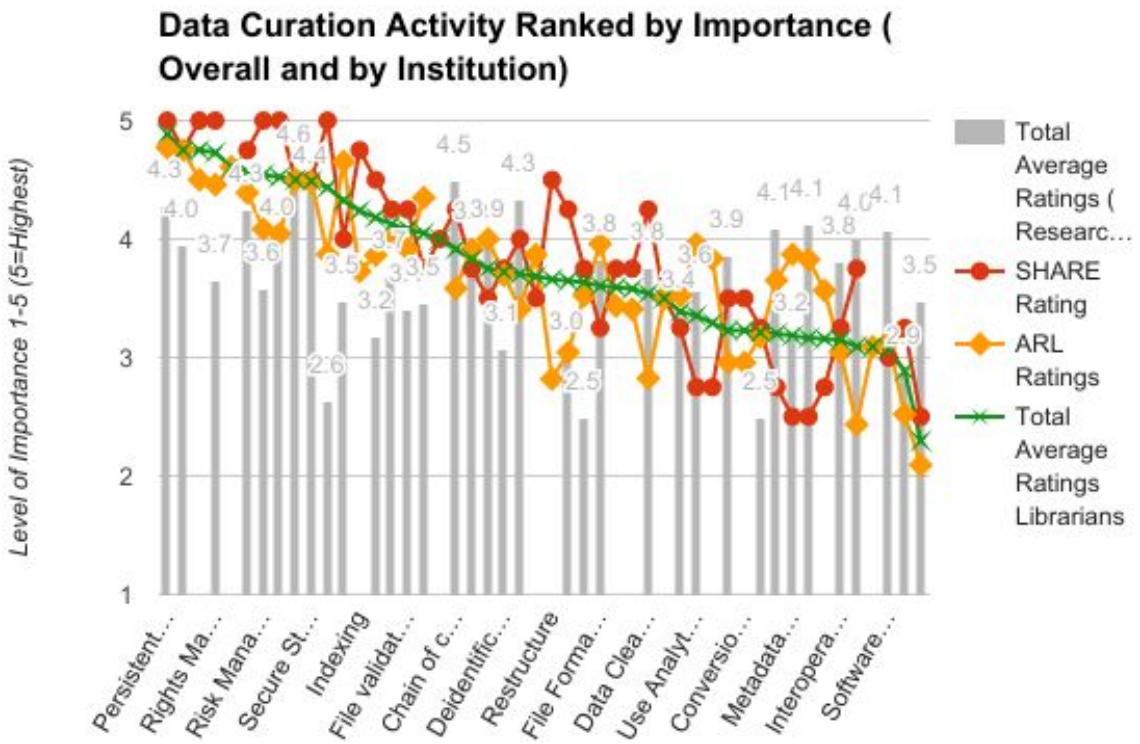
- "Yes" responses: the survey tool directed the 51 respondent to the longer version of the survey which asked detailed questions about their services and their levels of support for these activities.
- "No" or "In process" responses: the survey directed 29 respondents to our questions about the importance of data curation activities.

Table 1: ARL SPEC Kit Survey N0. 354 Responses to "Does your institution currently provide research data curation services?"

Yes	51	64%
No	16	20%
In process	13	16%

Results

The comparison of librarian ratings and researcher rating of the importance of different data curation activities is presented in Table 2.



Comparison Key: Comparing the rank percentiles for each activity provides a divergence label system.

Well Aligned = Both groups agree on the level of importance ranking, Deviation < +/- 0.3	Diverged = The groups diverge on the level of importance ranking, Deviation +/- 0.3-0.6	Very Diverged = The groups strongly diverge, Ranking Deviation > +/- 0.6

Table 2: Rating of Importance from Librarians and Researchers surveyed by the Data Curation Network

	Librarians			*Researchers*			*Comparison*	
Rank (Lib.)	**Data Curation Activity**	**Total Count (Libs.)**	**Ave Rating (Libs.)**	**Total Count (Res.)**	**Ave Rating (Res.)**	**Rank (Res.)**	**Alignment (% the Libs. Rank Deviated from Res.)**	
Librarian Ranking = 5 out of 5 Level of Importance "Essential"								
#1	Persistent Identifier	39	4.9	91	4.3	#5	.12	Well Aligned
#2	Metadata	41	4.8	80	4.0	#12	.30	Diverged
#3	Deposit agreement	41	4.8	0	n/a			
#4	Rights Management	41	4.7	62	3.7	#20	.50	Diverged
#5	File download	23	4.6	0	n/a			
#6	Discovery Services	40	4.6	11	4.3	#6	.04	Well Aligned

#7	Risk Management	41	4.5	80	3.6	#21	.46	Diverged
#8	File Audit	40	4.5	49	4.0	#11	.15	Well Aligned
#9	Documentation	41	4.5	91	4.6	#1	-.17	Well Aligned
#10	Secure Storage	40	4.5	60	4.4	#3	-.13	Well Aligned
#11	Restricted Access	40	4.4	24	2.6	#33	.73	Very Diverged
#12	Terms of Use	40	4.3	62	3.5	#24	.44	Diverged
#13	Indexing	39	4.2	0	n/a			
#14	File Inventory or Manifest	39	4.2	62	3.2	#28	.52	Diverged
#15	Embargo	40	4.1	91	3.7	#19	.23	Well Aligned
#16	File validation	41	4.1	67	3.4	#26	.42	Diverged
#17	Data Citation	40	4.0	67	3.5	#25	.37	Diverged
#18	Disk Image	17	4.0	0	n/a			
Librarian Ranking = 4 out of 5 Level of Importance “Very Important”								
#19	Chain of custody	41	3.9	11	4.5	#2	-.35	Diverged
#20	Contextualize	40	3.8	91	3.9	#14	-.02	Well Aligned
#21	Versioning	40	3.8	91	3.9	#13	-.07	Well Aligned
#22	Deidentification	40	3.7	62	3.1	#30	.41	Diverged
#23	Quality Assurance	39	3.7	73	4.3	#4	-.38	Diverged
#24	Succession Planning	40	3.7	0	n/a			
#25	Restructure	39	3.7	0	n/a			
#26	Repository Certification	40	3.6	18	3.0	#31	.35	Diverged
#27	Full-Text Indexing	40	3.6	13	2.5	#35	.45	Diverged
#28	File Format Transformations	40	3.6	73	3.8	#16	-.13	Well Aligned
#29	Cease Data Curation	40	3.6	0	n/a			
#30	Transcoding	39	3.6	0	n/a			
#31	Data Cleaning	40	3.5	24	3.8	#18	-.14	Well Aligned
#32	Selection	41	3.5	0	n/a			
#33	Migration	40	3.4	29	3.4	#27	.08	Well Aligned
#34	Use Analytics	40	3.4	91	3.6	#22	-.08	Well Aligned
#35	Authentication	41	3.3	0	n/a			
#36	Code review	40	3.2	91	3.9	#15	-.33	Diverged
#37	Conversion (Analog)	40	3.2	0	n/a			
#38	Contact Information	40	3.2	18	2.5	#34	.18	Well Aligned
#39	Tech. Monitoring and Refresh	40	3.2	18	4.1	#8	-.60	Very Diverged
#40	Metadata Brokerage	40	3.2	80	3.2	#29	-.01	Well Aligned
#41	Curation Log	40	3.2	11	4.1	#7	-.68	Very Diverged
#42	Arrangement and Description	40	3.2	0	n/a			
#43	Interoperability	39	3.1	24	3.8	#17	-.42	Diverged
#44	Data Visualization	40	3.1	24	4.0	#10	-.65	Very Diverged
#45	File renaming	22	3.1	0	n/a			
#46	Software Registry	39	3.1	29	4.1	#9	-.72	Very Diverged

Librarian Ranking = 3 out of 5 Level of Importance “Important”								
#47	Emulation	40	2.9	29	2.9	#32	-.07	Well Aligned
#48	Peer-review	39	2.3	42	3.5	#23	-.35	Diverged
Librarian Ranking = 2 out of 5 Level of Importance “Less Important”								
Librarian Ranking = 1 out of 5 Level of Importance “Not Important”								

Note: n/a indicates that this activity was not rated by this group.

Discussion

Librarians and researchers in our sample were well aligned, divergent, or poorly divergent with how they rate the importance of data curation activities.

Not surprising was the alignment in how they rated the “very important” activities of minting Persistent Identifiers for data (#1 Libs, #5 Res.), creating necessary Documentation (#9 Libs, #1 Res.) and obtaining Secure Storage (#10 Libs, #3 Res.). Perhaps more surprising is the agreement that Discovery Services for research data (#6 Libs, #6 Res.) and File Audits (#8 Libs, #11 Res.) are equally very important.

Where librarians and researchers diverged or greatly in their rankings shows areas where more mutual understanding is needed. Perhaps unsurprisingly, researchers more greatly valued building a Software Registry (#9 Res.), obtaining Data Visualization (#10 Res.) services for their data, Technology Monitoring and Refresh (#8 Res.) activities, ensuring the Interoperability of datasets, preserving the Chain of custody (#2 Res.) for a dataset, and data curation services such as Quality Assurance (#4 Res.), Peer-review and Code review. More surprising was that the Curation Log activity which is described as “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.” was more highly ranked by researchers than by librarians. On the other hand, librarians more greatly valued Restricted Access mechanisms, creating a File Inventory or Manifest, both Risk and Rights Management, and data repository services such as Full-Text Indexing, Terms of Use, File validation, Data Citation, and obtaining Repository Certification. Key elements that librarians slightly more highly valued were (unsurprisingly) Metadata creation and (surprisingly) Deidentification services.

Conclusions

Librarians and researchers both face similar challenges when it comes to curating research data. However the importance placed on certain data curation activities may highlight each group's collective’ strengths and weaknesses. Working together to address common data curation goals would aid in each groups better understanding and support for long-term data findability, access, interoperability, and reuse.

Supplementary Data available