Alfred P. Sloan FOUNDATION

The Data Curation Network will enable academic institutions to better support researchers that are faced with a growing number of requirements to ethically share their research data.

Data Curation Network

A network of expertise model for curating research data in digital repositories

Lisa R. Johnston, (PI) University of Minnesota
Jake Carlson, University of Michigan
Cynthia Hudson-Vitale, Washington University in St. Louis
Heidi Imker, University of Illinois at Urbana—Champaign
Wendy Kozlowski, Cornell University
Robert Olendorf, Pennsylvania State University
Claire Stewart, University of Minnesota

The State of Data Curation in ARL Libraries: Looking Forward – Growth & Challenges

Background

A survey of American Research Libraries (ARL) was released in 2017 as Spec Kit #354: Data Curation. The purpose was to uncover the current infrastructure (policy and technical) at ARL member institutions for data curation, explore the current level of demand for data curation services, and discover any challenges that institutions currently face when providing these services.

Email invitations to complete the survey were sent to the SPEC survey liaisons at 124 ARL institutions. A total of 80 responses were collected from Jan 3 – Jan 30 using SurveyMonkey®. In addition to quantitative data, the survey generated comments regarding the extent of curation activities taking place at each campus.

Question

What are common barriers that institutions face when implementing robust curation treatments for research data?

Table 1: Current state of curation activities at responding ARL institutions.

Curation Step	Treatment Activities	Currently Providing at least one activity	Would like to provide at least one activity, but currently unable to
Ingest	authentication; chain of custody; deposit agreement; documentation; file validation; metadata	46 of 49	23 of 49
Appraisal	rights management; risk management; selection	34 of 49	23 of 49
Processing & Review	arrangement and description; code review; contextualize; conversion; curation log; data cleaning; de-identification; file format transformations; file inventory; file renaming; indexing; interoperability; peer-review; persistent identifier; quality assurance; restructure; software registry; transcoding	43 of 48	45 of 48
Access	contact information; data citation; data visualization; discovery services; embargo; file download; full-text indexing; metadata brokerage; restricted access; terms of use; use analytics	43 of 48	34 of 48
Preservation	cease data curation; emulation; file audit; migration; repository certification; secure storage; succession planning; technology monitoring and refresh; versioning	39 of 49	39 of 48

Methods

1. The number of institutions who answered they 'currently provide' (#1) at least one curation treatment per curation step and the number of institutions who answered they 'would like to provide, but unable to at this time' (#3) at least one curation treatment per curation step were calculated to show extent of interest in service development (Table 1). *Note that institutions could be currently be providing one treatment activity and would also like to provide another activity.*
2. In order to identify barriers to providing curation activities, we coded 67 free text comments for six questions (28-33) that directly addressed current and aspirational data curation treatments. (Table 2)
3. We applied a grounded theory method of qualitative data coding and analysis across each comment – specifically coding for barriers to implementing curation treatments. *Note that responses could have more than one variable attributed to them.*
4. Coding variables were independently validated by two team members to ensure accurate interpretation and consistency.
5. Distribution of variables was evaluated for each themed curation activity. (Figure 1)

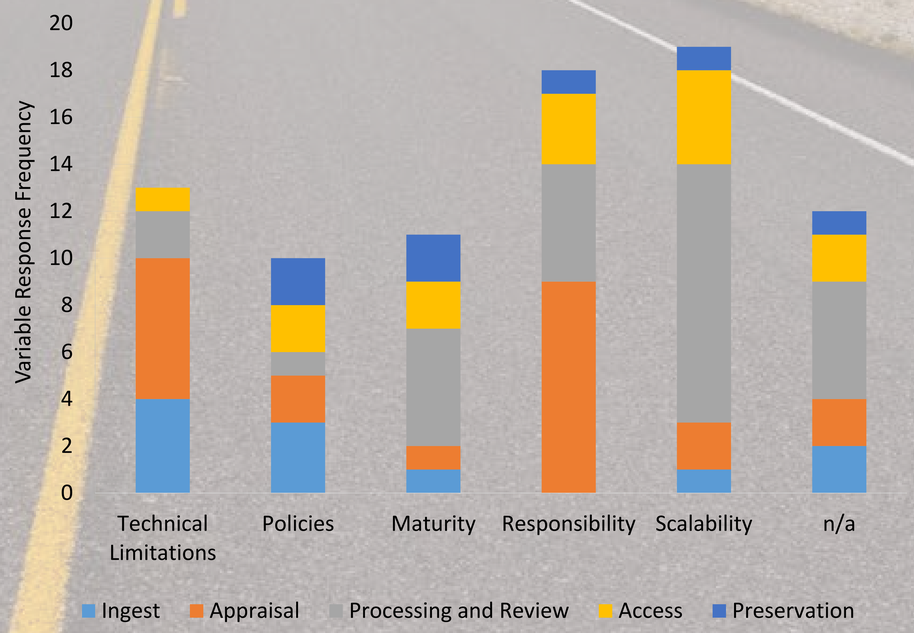
Results

- Most comment barriers varied by activity, with responsibility (n=18) and scaling (n=19) occurring more frequently than technical limitations (n=13), policies (n=10) or maturity (n=11) (Figure 1).

Table 2: Six distinct variables describe comments on institutional barriers for providing data curation services

Barrier	Definition	Examples
Not applicable (n/a)	The comment did not reference a challenge that was preventing them from conducting the activity.	
Technical limitations	The comment indicated that they were unable to complete this activity because of limitations in the repository or other technical equipment/expertise.	• Current platform analytics has limited capability and functionality at this time. • Our Dataverse is self-deposit.
Policies	The comment indicated that they were unable to complete this activity because of policy limitations or changes.	• Policies are currently under review. • Terms of use are in review by the university counsel's office.
Maturity	The comment indicated that they were unable to complete this activity because services or policies are not mature not enough.	• Multiple internal studies are currently underway looking at support for these date curation issues. • Most of these services are provided ad-hoc… however, we do not yet have an established service for data sets.
Responsibility	The comment indicated that they were unable to complete this activity because it is the responsibility of the submitter, or is carried out by another unit on campus.	• Currently, "risk management" responsibility is placed on the submitter. • We rely on the researcher to comply with stated deposit agreement. • Transcoding is done as needed by a unit outside of Data Curation.
Scaling	The comment indicated that they were unable to complete this activity because they require partnerships, additional staff or additional funding.	• Authentication and chain of custody are not done [to] the level described here… • There can be significant costs associated with the reprocessing of information.

Figure 1: Barriers attributed to curation activities



Conclusions

The state of data curation activities across ARL institutions is fairly robust, with many reporting that they provide some level of service, which may or may not be automated by the technology they use.

For those who would like to provide a more robust level of service it was surprising to learn that responsibility was one of the most significant barriers across all activities. This suggests that while libraries would like these activities to be improved, they feel it should be in partnership with the researchers or other campus entities.

The importance of staffing and scaling also supports anecdotal evidence that many institutions require additional funding to fully curate and preserve the data assets they steward.

Subsequent research should conduct a similar assessment of curation practices outside of ARL institutions to determine similarities and differences.

Learn more | http://publications.arl.org/Data-Curation-SPEC-Kit-354/ | https://sites.google.com/site/datacurationnetwork | #DataCurationNetwork