2017-01-09

Data Curation Priorities and Activities: A Report from a Researcher Engagement Event at the University of Michigan

Carlson, Jake

Data Curation Priorities and Activities: A Report from a Researcher Engagement Event at the University of Michigan

Jake Carlson
January 9, 2017

Facilitation: Jake Carlson, Amy Neeser and Lisa Johnston (University of Minnesota)
Date: 2016-11-18
Time: 11:30 - 1:00pm
Location: Clark Library Presentation Space

Overview

The Data Curation Network (DCN) is an effort to develop a "network of expertise" model for libraries providing data curation services to connect and collaborate with each other to increase member capabilities in curating research data beyond what any one institution could provide. Currently, the DCN is supported by a one year planning grant from the Sloan foundation and includes six institutions: The University of Minnesota (lead), the University of Michigan, the University of Illinois, Cornell University, Penn State University, Washington University in St. Louis. More information about the DCN can be found on the project's website.

Our activities for the year include hosting an event at each member institution to engage with researchers to understand their needs and priorities in data curation. This report describes the event that was held at the University of Michigan (U-M) on November 18, 2016 and what was learned from it.

Attendees

Invitations for the event were sent to faculty, research staff, graduate students, lab managers and others at U-M with roles in supporting the management and curation of research data. Many of the invitations were sent to those with whom we had prior contact and who have expressed some interest in data management or curation issues. However, we wanted to insure that the perspective of different disciplines were well represented at this event and so several library liaisons were contacted and asked to recommend people to invite.

Invitations were sent out beginning in late October. If no response was received after a week a second message was sent. Several of those invited who could not attend recommended others to attend in their place. Others who could not attend expressed an interest in attending similar events in the future. Of the 50 invitations that were ultimately sent only five people did not respond at all.

Twenty one people accepted the invitation of which 18 showed up for the event, though 2 arrived late and one had to leave early. More information about attendees is listed in Table 1.

Table 1 - U-M DCN Event Attendees

Count	Departmental Affiliation	Role	Data Producer or Manager	Discipline Category
1	Climate and Space Sciences	Professor	Producer	Engineering
2	Chemical Engineering	Lab Manager	Manager	Engineering
3	Climate and Space Sciences	IT Manager	Manager	Engineering
4	Naval Architecture and Marine Engineering	Graduate Student	Producer	Engineering
5	Humanities Based Research Center	Director	Manager	Humanities
6	Kelsey Museum of Archeology	Research Scientist	Producer	Humanities
7	Medical School	IT - Data Architect	Manager	Medical
8	Psychiatry	Research Specialist	Producer	Medical
9	Bio-Medical Engineering	Lab Manager	Manager	Medical
10	Microbiology and Immunology	Lab Manager	Manager	Medical
11	Astronomy	Professor	Producer	Science
12	Ecology and Evolutionary Biology	Professor	Producer	Science
13	Earth & Environmental Science	Professor	Producer	Science
14	Earth & Environmental Science	Professor	Producer	Science
15	Public Policy	Research Fellow	Producer	Social Science
16	Social Science Based Research Center (1)	Data Manager	Manager	Social Science
17	Social Science Based Research Center (1)	Technical Director	Manager	Social Science
18	Social Science Based Research Center (2)	Managing Director	Manager	Social Science

Composition of the Event

The researcher engagement event consisted of two exercises. For the first exercise, we asked attendees to rate specific curation activities according to their perceived importance. In the second exercise, we asked attendees to indicate if they or a 3rd party performed these data

curation activities, and if so to indicate how satisfied they were with the results. The results of each of the exercises were then used to launch small table discussions.

Prior to these events, the DCN partner institutions developed an extensive list of activities that related to data curation in some way. Each of the six DCN partners selected the data curation activities of particular interest to them from this list to use in the two exercises described above. Some of the data curation activities from the larger list were selected by all six partners, some by a majority of partners, and other data curation activities were only selected by one or two of the DCN institutions for use in their particular researcher engagement event.

The definitions of each of the 20 curation activities selected for the U-M event are included as Appendix 1 in this report.

Exercise 1: Ranking Data Curation Activities

In the first exercise, we gave each of the attendees a card that had a definition of a particular data curation activity printed on the front of it. We asked the attendees to read the definition and then to rate the importance of the data curation activity for their data on a scale of 1 (not important) to 5 (very important). Attendees wrote their ranking on the back of the card and then handed the card to another attendee. We asked each attendee to do this for four different curation activities. Once completed, the scores were added up and the average score was calculated.

We had expected 21 people to attend this event, however only 17 were present for the first activity. We removed three of the cards from the possible 20 so that none of the attendees had to focus on more than one card at a time.

Table 2 displays curation activities in rank order according to attendees at the U-M event and compares this to the average rank order assigned by all attendees across the six DCN partner institutions. Table 2 also indicates how many of the six participating institutions included the specific data curation activity in their own researcher engagement event.

Table 2 - Curation activities as ranked by UM researchers and compared with all DCN institutions

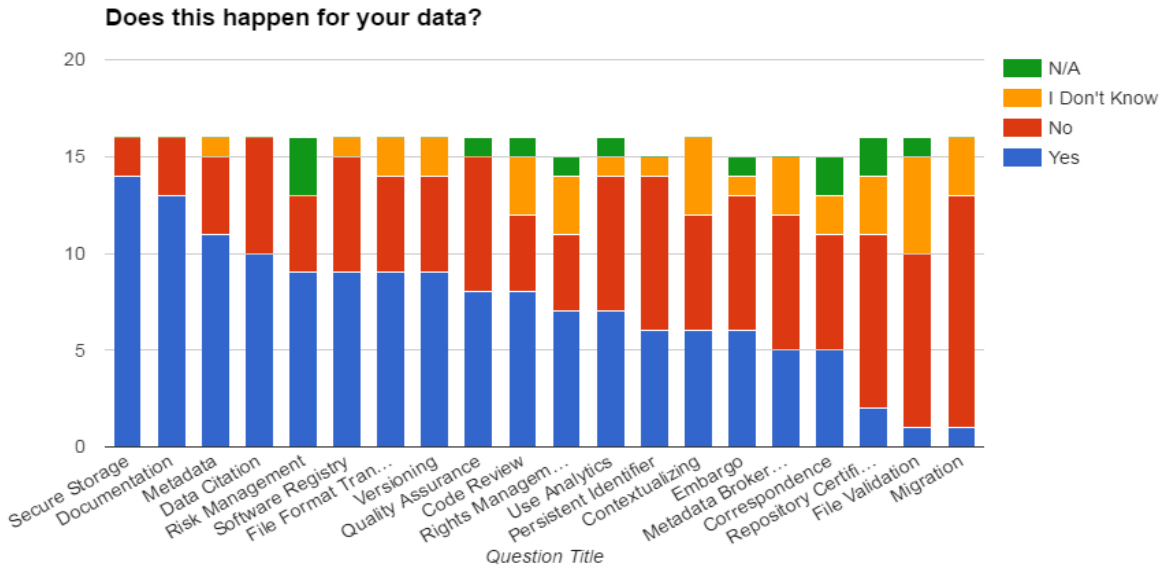
Data Curation Activity U-M Cohort (n=17)	Total Average at U-M	U-M Ranking (n=17)	Average from all partners	Overall Ranking of all partners	Standard Deviation of Overall Ranking	# of Institutions who Included the Activity
Secure Storage	5.0	1	4.4	3	0.47	4
Documentation	4.9	2	4.6	1	0.54	6
Software Registry	4.3	3	4.1	5	0.27	2
File validation	4.0	4	4.3	4	0.27	4
Persistent Identifier	4.0	5	3.4	12	0.45	6

Code review	3.9	6	3.9	7	0.54	6
Risk Management	3.9	7	3.6	10	0.35	5
Rights Management	3.8	8	3.7	9	0.51	4
Embargo	3.6	9	3.7	9	0.32	6
Metadata	3.4	10	4.0	6	0.47	5
Versioning	3.4	11	3.9	7	0.40	6
Contextualize	3.3	12	3.9	7	0.45	6
Data Citation	3.3	13	3.5	11	0.58	4
Metadata Brokerage	3.0	14	3.6	10	0.46	5
Use Analytics	3.0	15	3.2	13	0.55	6
Migration	2.8	16	3.4	12	0.74	2
Correspondence	2.5	17	2.5	18		1
Chain of Custody	(not asked)	(not asked)	4.5	2		1

Exercise 2: Engagement and Satisfaction with Data Curation Activities

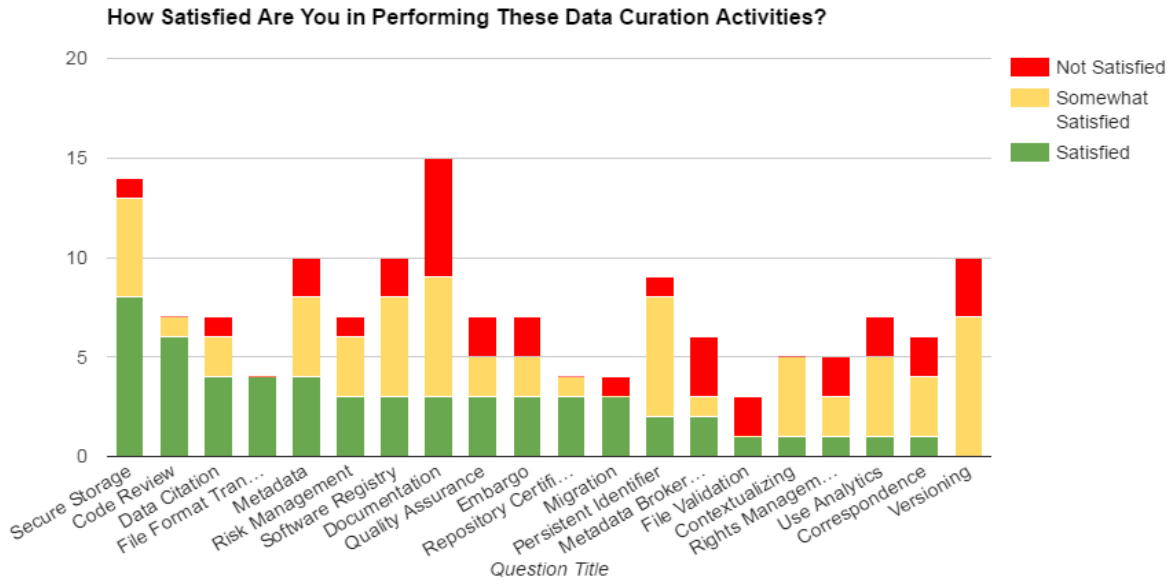
In the second exercise, we asked attendees to fill out a worksheet that listed the 20 different data curation activities selected for this event. For each activity we asked them to indicate if this was an activity that they or a 3rd party performed themselves by selecting “yes”. If this activity is not done by themselves or a 3rd party we asked them to select “no”. Attendees could also select “I don’t know” if they were not sure if the activity was done or not, or “N/A” if the curation activity is not relevant to their data. 16 attendees fill out the worksheet though only 15 filled out some of the questions. Their responses are displayed in Chart 1.

Chart 1 - Are these Data Curation Activities performed? (n=16)



For the attendees that selected “yes”, we then asked them how satisfied they were with the outcomes. Results are displayed in Chart 2 below. Some of the attendees answered the satisfaction question even though they did not select "yes" to the first question.

Chart 2 - If Performed, are Attendees Satisfied with Data Curation Activities?



Discussion

After the two activities were completed we held small group discussions with attendees. Attendees were seated across three tables. Roughly six attendees were seated at each table. Each table had a moderator to help guide the discussion.

The discussion varied across the tables but several important points emerged:

- Attendees expressed a desire for more services and resources to help support their work in managing, sharing and curating research data. They expressed disappointment in the lack of support they receive from campus, the NSF and other funding agencies.

- There needs to be more incentives in place for researchers to manage, share and curate data in ways that support open access.
    - Citations to data and software are a good start but not sufficient in and of themselves. There’s really no way to track who is using your data and how, other than citation, which is limited and doesn’t always happen. Measures of impact for data need to be developed and applied to the promotion and tenure process.
    - There also needs to be a way to track the return on investment for the time and effort spent on working with data beyond satisfying the immediate needs of researchers.

- There is a need for standardization to facilitate sharing, access and curation; however even when standards exist, they often do not mesh well with individual lab practices in generating data for researcher's specific purposes. There are often good reasons for these individual practices which need to be respected. However, this is a barrier towards adopting shared systems (databases), standards, protocols or procedures with data.
    - E-lab notebooks may be helpful in provided a structured workspace, and they can be configured to promote good practice, but they are not a complete answer.
    - There are standards that are emerging that are demonstrating value in helping researchers describe and share their data in ways that benefit their communities. The Brain Imaging Data Structure (BIDS) standard in neuroscience was cited as an example. However there are still challenges of scale and funding to support the needed resources for adoption and use of the standard.

- There is a lot of variation in how research is supported and conducted between disciplines, sub-disciplines, fields of study and even within research centers and local labs. Variation includes availability of funding, resources, storage, costs and other aspects of research that effect data management and curation. These variations also make it difficult to create shared systems to administer and work with data.

- A challenge for researchers is the high number of people coming into or going out of their research groups. Many labs are highly dependent on students which can be problematic as some students do a better job than others and all will leave the lab once they graduate. There is interest in partnering with campus organizations to help with this problem.

- The data transfer process from researcher to data manager does not always go as smoothly as it should. Data managers do not always receive as much information about the data (metadata, documentation) as they need. The areas of responsibility between data producers and those who manage the data are not always clear. Considering the purpose of the data set and the eventual use of the data early on to help identify the steps needed to enable the data to serve its purpose and ensure they get done is important.

- There was a definite interest in generating protocols and agreements on how the data should be developed and managed. However, the amount of time needed to develop the protocols, the education required for all to follow and make use of them, and the high rate of turnover in labs were identified as barriers.

- There were questions and concerns over the lack of defined ownership over research data which limited what could be done with the data.

Observations

Although attendees at this event constitute a very small fraction of the U-M community and may not be representative of the full range of needs across the university, having the opportunity to engage them for an hour and a half was a useful exercise and provided us with a better understanding of the pressures and challenges they face in curating their data.

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

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

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

Recommendations

The U-M Library is seen by the researchers and data managers attending this event as a knowledgeable 3rd party in data management and curation issues. Attendees were open to working with 3rd parties, including libraries, to address curation concerns and issues, and several were already doing so with varying degrees of satisfaction. Several attendees expressed a strong interest in working with the U-M Library to better understand data curation issues and to work with the library in developing approaches that would address their particular needs. The library should take advantage of opportunities to work with researchers and data

managers on curation issues. The potential to build strong relationships with researchers through helping them solve real-world problems with their data would enable the U-M library to demonstrate further the value of our librarians as information experts to the U-M community.

Many of the data curation activities we asked our attendees to respond to were seen as important. Attendees also indicated that they would benefit from more support and guidance in carrying out these activities. There were a number of data curation activities that were highly ranked that are either not being done currently, or not being done satisfactorily, indicating that there may be an opportunity for the library to further develop and provide services in these areas. They include:

- **Documentation:** Ranked 2nd overall in importance and done by 13 of 15 attendees who responded. However, six attendees indicated they were not satisfied with the results and another six indicated only partial satisfaction.
- **Software Registry:** Ranked 3rd overall. Of the ten attendees who answered the satisfaction question, two stated they were not satisfied and five indicated only partial satisfaction.
- **File Validation:** Ranked 4th overall, with nine attendees indicating it is not done and another five responding that they did not know if it was done or not.
- **Persistent Identifiers:** Ranked 5th overall, with eight attendees indicating they are not yet assigned to their data and six attendees indicating they are only somewhat satisfied with the results.
- **Rights Management**: Ranked 8th overall, with four attendees indicating it does not happen for their data and another three indicating they did not know if it happened. Of the seven who indicated that it was done, two attendees were not satisfied with the results and two were only somewhat satisfied.

Documentation in particular was seen as an important focus and one in which librarians could make a positive impact. In addition to documenting the data itself, researchers who work with students or as a part of a group would benefit from coming to a consensus as to how data should be managed and curated and then documenting these decisions in writing as policy and practices. Without this written documentation to refer to, researchers, support staff and students are left to make their own decisions in working with data; decisions that may impact data quality and reuse over time. Making time to have the discussions needed to reach consensus on data and then to document decisions in actionable ways is a challenge as described earlier, however the library could develop tools and models to make the process easier. For example, the library could potentially provide a template of questions to ask or checklists of issues to consider. Librarians could also provide assistance in moderating discussions or walking researchers through questions and then helping them draw up a data policy for their lab or team.

Though the opportunities to have an impact on data practices through providing much needed support at U-M were apparent from this event, there are many challenges that the library will have to address to realize the benefits. Time is a precious commodity to researchers. Although our attendees recognized the need and potential benefits of good data curation practices, the

immediate pressures of publishing and presenting, securing funding and resources, and mentoring graduate students, have traditionally limited the amount of attention paid to data management and curation. It's likely that researchers will have a limited window when they are able to focus on data management and curation issues and the library will need to be ready to act quickly when those moments arrive. The library will also need to recognize that saving the time of the researcher should be included as a goal of the services we provide.

A second challenge for the library in providing research data services is the desirability of standardization versus the need to recognize and accommodate individual data practices. The library has been and will continue to raise awareness about relevant data and metadata standards. However, to help researchers apply these standards effectively librarians will need to acquire an understanding of the data being generated and the current practices surrounding their management, documentation and use. In cases where there are no relevant standards or they cannot be applied for whatever reason, librarians should possess an understanding of general good practices for documenting data and how they could be applied.

Appendix 1 - Definitions of Data Curation Activities

Data Curation Activity	Definition
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.
Contact Information	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.
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.
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.
Data Citation	Provide third-party users with a recommended bibliographic citation for a dataset to enable appropriate attribution and help formally incorporate data reuse as part of the scholarly ecosystem.
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.
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.
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.
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).
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.
Software Registry	Maintain copies of modern and obsolete versions of software (and any relevant code libraries) so that data may be opened/used overtime.
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.