Data Curation Network Results
Title: Rating of Importance for Data Curation Activities
Release Date: 12-15-2016
Authors: Lisa R Johnston (PI), Jake Carlson, Cynthia Hudson-Vitale, Heidi Imker, Wendy Kozlowski, Robert Olendorf, and Claire Stewart.

Ratings of Importance for Data Curation Activities

The Data Curation Network held six focus groups in the fall of 2016 and asked 91 researchers from a wide variety of disciplines: how would you rate the "importance" of a variety of key data curation activities. Our preliminary results for the importance ranking of 35 data curation activities are presented in the table below along with a brief summary of our methodology and attendance numbers by discipline.

Key finding: The five most important data curation activities indicated in our focus groups were creating documentation for data, preserving the "chain of custody" of a dataset, securely storing data, providing quality assurance for data, and minting a persistent identifier for a data set.

Figure 1: Results of the Average Ranking of Importance for Activities that were ranked by the Data Curation Network Focus Groups (5= highest importance, 1 = not important)

"Most Important" *Average Ranking of 4.0 - 4.9*	"Important" *Average* Ranking of 3.0 - 3.9	"Less Important" *Average Ranking of 2.0 - 2.9*	"Not Important" *Average Ranking 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	

Table 1: Rating of Importance by Participants in the Data Curation Network Focus Groups, Fall 2016

(sorted by average rating on a scale of 1-5 with 5 = highest)

Data Curation Activity	Data Curation Network Definition	Average Rating	Institutional Deviation +/-
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).	4.6	0.54
Chain of custody	Intentional recording of provenance metadata of the files (e.g., metadata about who created the file, when it was last edited, etc.) in order to preserve file authenticity when data are transferred to third-parties.	4.5	n/a
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.	4.4	0.47
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).	4.3	n/a
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/blank values, or unclear acronyms.	4.3	0.29
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.	4.3	0.27
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.	4.1	n/a
Technology Monitoring and Refresh	Formal, periodic review and assessment to ensure responsiveness to technological developments and evolving requirements of the digital infrastructure and hardware storing the data.	4.1	n/a
Software Registry	Maintain copies of modern and obsolete versions of software (and any relevant code libraries) so that data may be opened/used overtime.	4.1	0.27
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.	4.0	0
File Audit	Periodic review of the digital integrity of the data files and taking action when needed to protect	4.0	0.41

	data from digital erosion (e.g., bitrot) and/or hardware failure.		
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).	4.0	0.4
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.	3.9	0.54
Contextualization	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.	3.9	0.4
Conversion (Analog)	In effort to increase the usability of a data set, the information is transferred into digital file formats (e.g., analog data keyed into a database). Note: digital conversion is also used to convert fixed data (e.g., PDF formats) into machine-readable formats.	3.9	0.45
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.	3.9	0.47
Data Cleaning	A process used to improve data quality by detecting and correcting (or removing) defects & errors in data.	3.8	0.35
Interoperability	Formatting the data using a disciplinary standard for better integration with other datasets and/or systems.	3.8	0.8
File Format Transformations	Transform files into open, non-proprietary file formats that broaden the potential for long-term reuse and ensure that additional preservation actions might be taken in the future. Note: Retention of the original file formats may be necessary if data transfer is not perfect.	3.8	0.59
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).	3.7	0.51
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.	3.7	0.32
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.	3.6	0.35
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.	3.6	0.46
Peer-review	The review of a data set by an expert with similar credentials and subject knowledge as the	3.5	1.13

	data creator for the purposes of validating the soundness and trustworthiness of the file contents.		
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).	3.5	0.12
Data Citation	Display of a recommended bibliographic citation for a dataset to enable appropriate attribution by third-party users in order to formally incorporate data reuse as part of the scholarly ecosystem.	3.5	0.58
Migration	Monitor and anticipate file format obsolescence and, as needed, transform obsolete file formats to new formats as standards and use dictate.	3.4	0.74
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.	3.4	0.45
File Inventory or Manifest	The data files are inspected periodically and the number, file types (extensions), and file sizes of the data are understood and documented. Any missing, duplicate, or corrupt (e.g., unable to open) files are discovered.	3.2	0.31
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.	3.2	0.55
Deidentification	Redacting or removing personally identifiable or protected information (e.g., sensitive geographic locations) from a dataset prior to sharing with third-parties.	3.1	0.99
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).	3.0	n/a
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)	2.9	0.35
Restricted Access	In order to maintain the privacy of research subjects without losing integral components of the data, some data access will be protected and/or mediated to individuals that meet predefined criteria.	2.6	0.35
Full-Text Indexing	Enhance the data for discovery purposes by generating search-engine-optimized formats of the text inherent to the data.	2.5	n/a
Contextualization	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.	3.9	n/a

Methodology of the Data Curation Activities Ratings Exercise

First, the Data Curation Network (DCN) team defined 48 data curation activities by consulting the following sources:

- CASRAI Dictionary (http://dictionary.casrai.org/Main_Page),
- Research Data Alliance (RDA) Terms Definition Tool (http://smw-rda.esc.rzg.mpg.de/index.php/Main_Page),
- Digital Curation Center (DCC) Glossary (http://www.dcc.ac.uk/digital-curation/glossary),
- Data Curation Steps from the 2017 book "Curating Research Data, Volume Two: A Handbook of Current Practice" (http://hdl.handle.net/11299/183502),
- ICPSR Glossary of Social Science Terms (http://www.icpsr.umich.edu/icpsrweb/ICPSR/support/glossary),
- Research Data Canada Glossary (https://www.rdc-drc.ca/glossary/),
- Digital Preservation Coalition Glossary (http://handbook.dpconline.org/glossary), and
- Society of American Archivists Terms Glossary (http://www2.archivists.org/glossary/terms).

Next, the DCN team held six focus groups between October 21, 2016 and November 18, 2016 that we termed as “Data Curation Roundtable” sessions at the following academic institutions: Cornell University, Penn State University, University of Illinois at Urbana-Champaign, University of Michigan, University of Minnesota, and Washington University in St. Louis. The participants represented a good mix of experience (faculty, graduate student, post-doc) and discipline (see table 2). Each session lasted 1 ½ hours over lunch, which was provided by the DCN project in exchange for their participation.

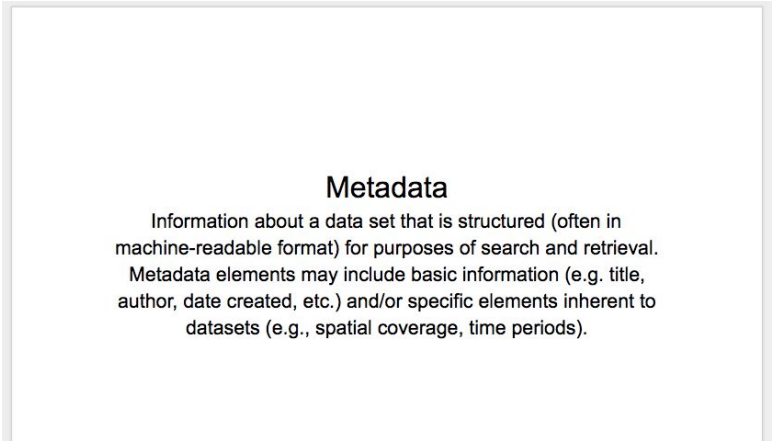
The DCN team asked researchers to rate the importance of 35 data curation activities as one exercise in the focus group session. Not all the activities were presented at each of the six session and it was up to the local DCN team member to select the subset of activities to focus on. To keep the exercise 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. The researcher was given 2-4 cards at a time and asked to read the definition and then rank that activity’s importance from 1 (lowest) to 5 (highest). Once each card in their hand was ranked the researchers were asked to exchange their cards with another researcher in the room and repeat the ranking exercise for the new card. This was repeated for four rounds. Since 2 or 3 copies of the same card circulated around the room, researchers were advised to trade with those who had cards they had not ranked previously. After a quick total of all four rounds, this exercise provided our team with a priority list of data curation activities that were used as the focus of the group discussion through the session.

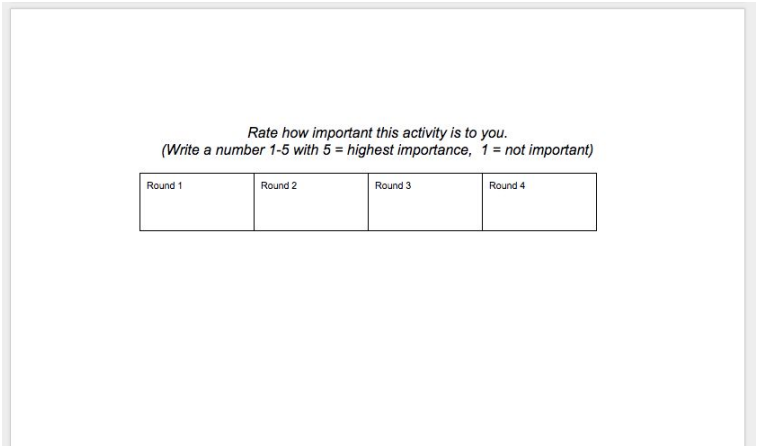
Table 2: Disciplinary representation at the six researcher engagement sessions

Institution	Cornell	Wash U	Illinois	Penn State	Minnesota	Michigan	Totals
Date of Session	10-21-2016	10-25-2016	10-27-2016	11-4-2016	11-14-2016	11-18-2016	**6**
Sciences & Engineering	9	6	10	5	11	8	**49**
Social Sciences	6	1	2	1	1	4	**15**
Humanities	0	1	1	1	0	2	**5**
Staff/Service Providers**	5	3	5	4	1	0	**16**
Medical	0	0	0	0	0	4	**4**
Total	20	11	18	11	13	16	**91**

*Service providers, such as campus-based IT staff and library staff, as well as library and information science faculty were grouped into this category.

Figure 2: The front and back of an example card used in the card ranking activity.





Appendix: Data curation activities that were not ranked in the focus group sessions

Arrangement and Description	The re-organization of files (e.g., new folder directory structure) in a dataset that may also involve the creation of new file names, file descriptions, and the recording of technical metadata inherent to the files (e.g., date last modified).
Authentication	The process of confirming the identity of a person, generally the depositor, who is contributing data to the data repository. (e.g., password authentication or authorization via digital signature). Used for tracking provenance of the data files.
Cease Data Curation	Plan for any contingencies that will ultimately terminate access to the data. For example, providing tombstones or metadata records for data that have been deselected and removed from stewardship.
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.
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.
File download	Allow access to the data materials by authorized third parties.
File renaming	To rename files in a dataset, often to standardize and/or reflect important metadata.
Indexing	Verify all metadata provided by the author and crosswalk to descriptive and administrative metadata compliant with a standard format for repository interoperability.
Restructure	Organize and/or reformate poorly structured data files to clarify their meaning and importance.
Selection	The result of a successful appraisal. The data are determined appropriate for acceptance and ingest into the repository according to local collection policy and practice.
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.
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).