Authors: Lisa R Johnston (PI), Jake Carlson, Cynthia Hudson-Vitale, Heidi Imker, Wendy Kozlowski, Robert Olendorf, and Claire Stewart.

Data Curation Terms and Activities

The DCN team drafted a dictionary of relevant terms used in data curation by consulting the following sources: SAA, CASRAI, DCC, RDA, DPC, RDC, ICPSR, and steps from ACRL book "Curating Research Data Volume 2: Handbook of current practice" (forthcoming November 2016). The resulting activities will be used in our engagement sessions with researchers at each of the six institutions in the Data Curation Network as well as our 2017 ARL SPEC Kit survey.

Definitions

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.) (RDA).

Data Curation Activities

Data curation activities can be arranged around five steps of data curation life-cycle: Ingest, Appraise (Accept), Curate, Access, and Preserve.

Ingest

- 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.
- 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.
- 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.
- 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).
- 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

---

The Data Curation Network is funded by a generous grant by the Alfred P. Sloan Foundation (2016-2017).

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).

Appraise/Accept

- 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.
- 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.

Curate

- 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).
- 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.
- 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.
- 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.
- 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.
- Data Cleaning: A process used to improve data quality by detecting and correcting (or removing) defects & errors in data.
- Deidentification: Redacting or removing personally identifiable or protected information (e.g., sensitive geographic locations) from a dataset prior to sharing with third-parties.
- 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.

- 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.
- 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.
- Interoperability: Formatting the data using a disciplinary standard for better integration with other datasets and/or systems.
- 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.
- 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.
- Restructure: Organize and/or reformate poorly structured data files to clarify their meaning and importance.
- Software Registry: Maintain copies of modern and obsolete versions of software (and any relevant code libraries) so that data may be opened/used overtime.
- 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).

Access

- 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.
- 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.
- 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.
- 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).
- 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 download: Allow access to the data materials by authorized third parties.
- Full-Text Indexing: Enhance the data for discovery purposes by generating search-engine-optimized formats of the text inherent to the data.
- 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.
- 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.
- 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).
- 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.

Preserve
- 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)
- 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.
- Migration: Monitor and anticipate file format obsolescence and, as needed, transform obsolete file formats to new formats as standards and use dictate.
- 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).
- 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.
- 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.
- 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.
- 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.
- 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.