This expert from the book *Curating Research Data Volume Two: A Handbook of Current Practice* by Lisa R. Johnston for ACRL released January 2017. The book is available from the American Library Association (www.alastore.ala.org) and Amazon.com in print and as an open access e-book.

Data Curation Handbook Steps

By Lisa R. Johnston

Preliminary Step 0: Establish Your Data Curation Service: Repository data curation services should be sustained through appropriate staffing and business models.

- 0.1 Mission: As an organization, acknowledge the institutional commitment (e.g., resources, staff) to providing data curation services at a level appropriate for your goals.
- 0.2 Policies: Define the scope of what data will be curated and establish written policies (e.g., criteria for acceptance and rejection).
- 0.3 Target Audience: Perform a market analysis (e.g., user needs assessment, gap analysis of existing services, stakeholder analyses) to better understand the target audience needs and motivations for using the curation service.
- 0.4 Costs: Understand and plan for the costs involved with staffing, resourcing, launching, and maintaining the curation service.
- 0.5 Staffing: Define and allocate the organizational infrastructure to provide the data curation services at the level appropriate for your organization.
- 0.6 Technological Infrastructure: Develop (or secure from a third party) the needed repository infrastructure to securely house and store the data.

Step 1.0: Receive the Data: Repository data curation services should enable data creators to submit their data for deposit into a data repository (institutional or disciplinary).

- 1.1 Recruit Data for Your Service: Potential submitters learn of your services through successful recruitment strategies and a robust communications plan that engages your target audiences.
- 1.2 Negotiate Deposit: The repository and the submitter come to a clear understanding and agreement of the terms of deposit for data into the repository (e.g., policies and conditions for access and reuse are known and well communicated).
- 1.3 Obtain Author Deposit Agreements: A rights transfer agreement is signed by the data author (or authorized submitter), and any conditions involving legally protected or restricted use data are well understood.
- 1.4 Facilitate Transfer of the Data: The data files are transferred to the repository in a secure manner that protects the integrity and authenticity of the data.

1.5 Obtain Metadata and Documentation: The repository collects the author-generated metadata and supporting documentation necessary to use and understand the data. This information will be included in the curation process as part of the data submission.

1.6 Receive Notification of Data Arrival: The appropriate repository staff are alerted that a new data submission was received and is ready for curation.

Step 2.0: Appraise and Select: Repository data curation services should include a review of data submissions to allow for selection and rejection of data that does not meet collection policy and mitigate known risk inherent to digital data.

2.1 Appraise the Files: Determine that the repository is the appropriate home for this data (e.g., the data meets all collection policies) and, with appropriate curation, there is a potential long-term value for reuse.

2.2 Consider Any Risk Factors: The repository has clear understanding of the types of data (e.g., federally protected data, sensitive information, copyright violations) that should not be accepted and has protocols in place to reject or facilitate remediation of data that should not have been transferred to the repository.

2.3 Inventory the Submission: The data submission is inspected and the number, file types, and file sizes of the data are understood and documented. Identify any missing, duplicate, or corrupt (e.g., unable to open) files. Capture the organization of the files and any technical metadata (e.g., date last modified). Request more information from the author if necessary.

2.4 Select: The data submission is accepted or rejected based on the above actions. If accepted, determine if any additional information or files need to be acquired from the author before moving to the next step. This step may include the deselection and removal of any duplicate or unnecessary files.

2.5 Assign the Submission: Accept the data submission for inclusion in the repository and assign curation responsibility to the appropriate data curator based on subject and format expertise required.

Step 3.0: Processing and Treatment Actions for Data: Repository data curation services should include processing actions for all data deposited in order to best arrange, transform, and prepare the data according to established procedures.

3.1 Secure the Files: Create a working copy of the files in order to protect the originals when making any changes or additions to the submission during Step 3.

3.2 Start a Curation Log: Track any changes to the data in a curation log in order to keep a record of the correspondence (e.g., e-mails) between the repository and the submitter.

3.3 Inspect the File Representation and Organization: Understand the directory structures, file relationships, and any naming conventions used. Preserve any relationships in the files or generate documentation to help others understand how the files relate.

3.4 Inspect the Data: Review the content of the data files (e.g., open and run the files). Check for quality and usability issues such as missing data, ambiguous headings, code execution failures, and data presentation concerns. Try to detect and extract any "hidden documentation" inherent to the data files that may facilitate reuse. Generate a list of questions for the data author to fix any errors or issues.

3.5 Work with the Author to Enhance the Data Submission: Verify all metadata provided by the author and review the available documentation. Determine if this description of the data is sufficient for a user with similar qualifications to the author's to understand and reuse the data. If not, seek out or create additional documentation (e.g., use a readme.txt template).

3.6 Consider File Formats: Identify specialized file formats and their restrictions (e.g., Is the software freely available? Link to it or archive it alongside the data). Verify technical metadata (min resolution, audio/video codec) that would optimize the files for reuse. Transform files into open, nonproprietary file formats that broaden the potential audience for reuse and ensure that preservation actions might be taken by the repository in later steps. Retain original files if data transfer is not perfect.

3.7 Arrangement and Description: Organize and rename the files to optimize their meaning, and display them in a way that might facilitate reuse (e.g., Which files are the primary object and which are supplementary? Too many files, consider repackaging for display.)

Step 4.0: Ingest and Store Data in Your Repository: Repository data curation services should ingest and store data in secure locations using appropriate repository infrastructure.

4.1 Ingest the Data Files: Transfer the processed data files to the repository while maintaining integrity and verifying fixity throughout the process (e.g., generate file checksums).

4.2 Store the Assets Securely: Add the ingested files to a well-configured (in terms of hardware and software) archival storage environment. Perform routine checks and provide disaster recovery capabilities as needed.

4.3 Develop Trust in Your Repository: Become a trusted digital repository for data by applying for accreditation and growing your reputation locally and beyond.

Step 5.0: Descriptive Metadata: Repository data curation services should apply appropriate descriptive metadata and enhance existing metadata that best facilitates discovery.

5.1 Create and Apply Descriptive Metadata: Structure author-generated metadata into the metadata schema used by your repository in order to maximize search and discovery functionality. Create and apply new metadata for the data record, including technical and provenance metadata.

5.2 Consider Metadata Standards for Disciplinary Data: When appropriate, structure and present metadata in multiple schemas to facilitate discovery and future integration into other systems.

Step 6.0: Access: Repository data curation services should facilitate access through discovery, dissemination, retrieval, and download functionality.

6.1 Determine Appropriate Access Conditions: Determine what level of access is required and how target audiences for reuse may be affected by this condition (e.g., a terms of use).

6.2 Apply the Terms of Use and Any Relevant Licenses and Copyright Notices for the Data: Work with the author to choose to apply any specific reuse conditions (e.g., a terms of use agreement), and if appropriate, apply a license, such as Creative Commons Licenses.

6.3 Contextualize the Data: The discovery and access environment for the data should convey the possibilities of reusing the data. This can be done by visualizing the data, showcasing the primary data file contents to convey meaning, and linking to articles and projects that successfully reused the data.

6.4 Enhance the Submission to Increase Exposure and Discovery: Work directly and indirectly with third-party indexers to disseminate your repository holdings. For example, web search engines and services, such as Google, Bing, and Yahoo, rely heavily on text-based information to facilitate discovery. If possible, enhance the data submission for discovery purposes by generating search-engine-optimized formats of the data (e.g., full-text index).

6.5 Apply Any Necessary Access Controls: Depending on the conditions for access and reuse, place access restrictions on some or all data files.

6.6 Ensure Persistent Access: Generate a persistent identifier (e.g., a DOI) to help facilitate reliable long-term access to the data (e.g., via bibliographic citation).

6.7 Release Data for Access and Notify Author: Finalize the data submission by allowing the metadata of the data set to "go live" for discovery and access. Notify the author of this event and any further obligations that might be required of them (e.g., responding to requests for access).

Step 7.0: Preservation of Data for the Long Term: Repository data curation services should preserve data in many forms for as long as the data is useful and adhere to policy-driven decisions for how long the data must be retained.

7.1 Plan for Long-Term Reuse: The delivery and use of the data will rely on long-term preservation planning that anticipates format obsolescence and storage failures.

7.2 Monitor Preservation Needs and Take Action: Actively monitor the integrity and reusability of the data files using appropriate software, and apply digital preservation strategies.

Step 8.0: Reuse: Repository data curation services evaluate the impact or value of the data and determine whether to keep or dispose.

8.1 Monitor Data Rese: Track any requests for access, file downloads, data set citations, and other factors that might indicate the reuse value of data over time.

8.2 Consider Post-Publication Review Techniques: Allow others, public and experts, to provide feedback on the data in order to provide additional post-publication quality control. Consider peer-review mechanisms to track input and provide quality measures for data housed in your repository.

8.3 Provide Ongoing Support as Long as Necessary: Provide services that meet the evolving needs of the data over the anticipated life or usefulness of the data, such as new versions, supplemental file additions, and user-generated documentation.

8.4 Cease Data Curation: Plan for any contingencies that will ultimately terminate access to the data, such as loss of funding for the repository. For example, how will you respond to takedown requests and deselection (e.g., provide tombstones)?