DATA CURATION NETWORK

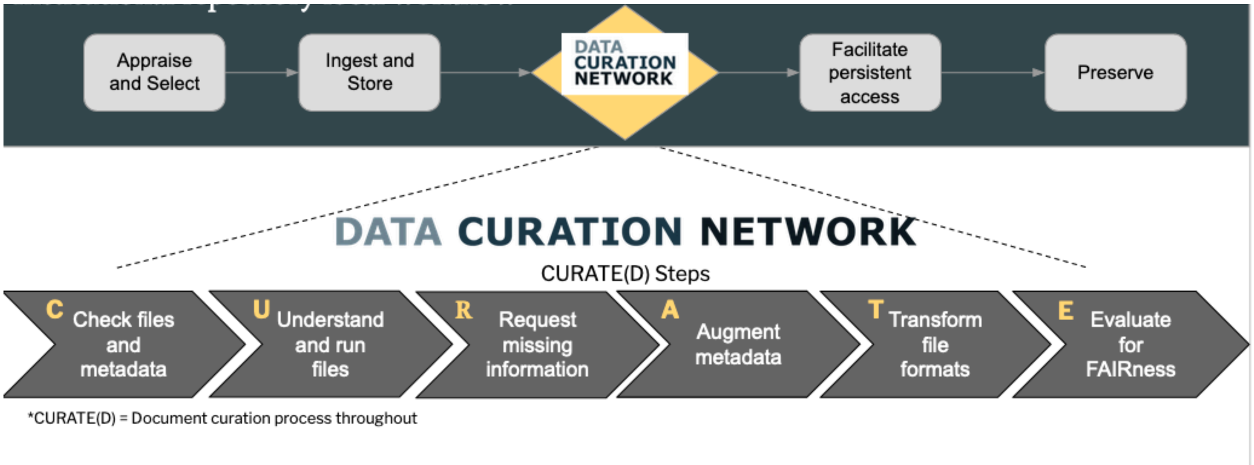
CURATE(D): Checklist for Data Curation

The CURATE(D) steps are a teaching and representation tool. This model is useful for onboarding data curators and orienting researchers preparing to share their data. It serves as a demonstration for the type of work involved in robust data curation, and was created to fit within institution-specific data repository workflows. Curation may not follow this exact workflow every time, and procedures may differ slightly depending on data needs and institutional processes. Moreover, the CURATE(D) process, while presented sequentially, is not necessarily linear. curation can jump between steps and repeat actions as necessary. Curators using this checklist should review all steps before reaching out to dataset creators in the "R–Request" step.

This checklist also flows from a preceding series of critical appraisal decisions. This includes deciding which data to keep (selection), and where to share data (e.g., is this data in-scope for a particular repository based on a range of factors including local policy and potential for reuse?).

Should the data be shared?
Data curators analyze content to assess near and long-term impacts of data sharing, which is especially critical when evaluating for ethical concerns in data derived from human participants. To learn more about this, review:

- Human Participants Data Essentials primer
- Curation of Data Collected by Informed Consent
- CARE Principles for Indigenous Data Governance
- Principles for Advancing Equitable Data Practice



CC-BY Suggested Citation: Data Curation Network (2022). "CURATE(D) Steps and Checklist for Data Curation, version 2" http://z.umn.edu/curate.

Checklist of CURATE(D) Steps Performed by the DCN

C **Check** files and read documentation (risk mitigation, file inventory, appraisal/selection)

U **Understand** the data (or try to), if not… (run files/environment, QA/QC issues, readme)

R **Request** missing information or changes (tracking provenance of any changes and why)

A **Augment** metadata for findability (DOIs, metadata standards, discoverability)

T **Transform** file formats for reuse (data preservation, conversion tools, data viz)

E **Evaluate** for FAIRness (licenses, responsibility standards, metrics for tracking use)

D **Document** your curation activities (Curator Log, correspondence)

CHECK Step

Check data files/code and read documentation

In this step we secure the dataset by inventorying and reviewing the contents, applying local appraisal and selection criteria. Common CHECK steps include:

- Review to ensure data is in scope for the repository
- Inventory the contents of the data files (e.g., open and sample the files or code)
- Verify all metadata provided by the researcher; check available documentation

Key Ethical Considerations

- Review participant agreement and data use agreements; examine potential impacts of sharing this data. Consider:
    - Individuals and communities represented
    - Representativeness of diverse human populations
    - Protection or endangerment status of species
    - Geographic locations (e.g., contested boundaries, historical and current political situations)

    - ○ Animal research ethics and approval
- Is it possible that the dataset may impact a specific group?
- Does this dataset follow compliance & institutional policy?

Essential Tasks

- ☐ Begin Curator Log to track curation decisions
- ☐ Open the related article and supporting information if available
- ☐ Inventory the dataset
    - ☐ Identify file formats
    - ☐ Review file organization, hierarchy, and naming convention(s)
    - ☐ Extract zip files when possible
- ☐ Create working copy of files for formal inventory and testing
- ☐ Examine code for obvious errors/missing components, etc.
- ☐ Check that metadata quality is rich, accurate, and complete to institutional requirements.
- ☐ Check documentation type (circle)
  readme / Codebook / Data Dictionary / Other: ______________________
    - ☐ Complete
    - ☐ Needs work
    - ☐ If missing, document for the “Request” step
- ☐ Check whether human subject data (data about humans regardless of IRB determination) is present. If so,
    - ☐ Request consent form / participation agreement if not present
    - ☐ If the data are not de-identified, document for the "Request" step.
- ☐ Check the accessibility of all files
    - ☐ Ensure there are robust descriptions in plain text of data files and any images.
- ☐ Check whether any visualization(s) of data are easily accessible
    - ☐ Review alt-text and visualization descriptions. Ensure these describe, but do not interpret, associated visualizations.
    - ☐ Check data visualizations follow accessible color contrast guidelines
    - ☐ Recommend graphical representation ___________
    - ☐ Recommend web-accessible surrogate _______________

UNDERSTAND Step

Understand the data (or try to)

In this step, examine the dataset closely to understand what it is, how the files interrelate, and what information is needed to reuse. Common UNDERSTAND steps include:

- Check for quality assurance 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 or expose unintended information
- Determine if the documentation of the data is sufficient for a user with similar qualifications to the researcher's to understand and reuse the data. If not, recommend or create additional documentation (e.g., a readme.txt template)

Key Ethical Considerations

- If working with human data, is this research done *with* and not *on* communities and populations involved? (You may wish to review data sources, researchers, and their connections to the communities and subjects they are serving to facilitate further conversation with researcher(s).)
  - Are there authoritative group representatives who should be contacted in the next (request) step?
- Are there labels or other descriptive indicators that could be applied to better represent or protect an identified group of people impacted by this dataset? (Example: TK labels)

Essential Tasks

- □ Examine files, organization, and documentation more thoroughly. Are there changes that could enhance the dataset?
  - □ Are there missing data?
  - □ Could a user with similar qualifications to the author's understand and reuse these data and reproduce the results?
  - □ Are the data, documentation and/or metadata presented in a way that aids in interpretation? (e.g., readme Example)

- □ Record all questions and concerns in Curation Log.

Tasks vary based on file formats and subject domain. Sample tasks based on format:

Tabular Data (e.g, Microsoft Excel) Questions:

- □ Check the organization of the data–is it well-structured?
- □ Are headers/codes clearly defined?
- □ Is quality control clearly defined?
- □ Is methodology clear and sufficient?

Database(s) Questions:

- □ Is there documentation on tables, relationships, queries, etc?
- □ Can the data be exported (to CSV(s), TXT or other) easily?
- □ Which tables or queries are the relevant ones used in a publication?

Code Questions:

- □ Does the provided code execute without errors?
- □ Is the code commented, i.e., did the author provide descriptive information on sections of code?
- □ Is data for input missing? Are environmental conditions and parameters noted? Is it clear which language(s) and version(s) are used?
- □ Does the code use absolute paths or relative paths? If absolute paths, is this documented in the readme?
- □ Are packages or additional libraries used? Is so, is this noted with clear use instructions?
- □ Are any data organized consistently for access by the code ?
- □ Is there an indication of whether the depositor intends reusers to be able to run the code and reproduce results, or just see the process used?

To view additional UNDERSTAND steps based on format, view the following primers:

- Acrobat PDF Primer
- ATLAS.ti Primer
- Confocal Microscopy Image Primer
- Geodatabase Primer
- GeoJSON Primer
- Jupyter Notebook Primer
- Microsoft Access Primer
- Microsoft Excel Primer

- netCDF Primer and Tutorial using NCAR dataset
- SPSS Primer
- STL Primer
- R Primer
- Tableau Primer
- (Twitter primer?)
- Wordpress.com Primer

REQUEST Step

Request missing information or changes

In this step, generate a list of questions to help the researcher fix any errors or issues and enrich the usability of the data. Common REQUEST steps include:

- Triage and prioritize issues. Identify and highlight those with the highest data reuse implications
- Convey a sense of urgency, as there it becomes more difficult to get responses from researchers as time passes.
- Collaborate with the researcher(s) to make necessary changes
- Communicate any changes you, the curator, will make on their behalf
- Pause and consider how best to frame and communicate requests. This should be the start of a conversation.

Key Ethical Considerations

- Consider asking researchers if their participants will be notified that their data (in addition to published results) are being shared.
- If you feel uncomfortable about sharing the data in its current state and/or it does not meet your institution's requirements, reserve the right not to publish.
- Consider asking researcher(s) if there are limitations to how data could/should be used to include in documentation. (Based on, e.g., representativeness of sample).

Essential Tasks

- ☐ Ask about additional data contributors, beyond publication authors. Consider using the Contributor Roles Taxonomy to communicate this: https://casrai.org/credit/
- ☐ Summarize conversations / outreach efforts in Curator Log

Sample email to researcher:

Dear [*name of the data set author or contact*],

Thank you for depositing your data set, [*title of the data set*], to [*name of repository*].

After we receive a data set, we review it to ensure that the data we host are as complete and understandable as possible. We have reviewed your data set and have the following recommendations for you:

- Recommendation #1
- Recommendation #2
- Recommendation #3
- Recommendation #4

We look forward to hearing your response.

Please let us know if you have any questions about our recommendations. We would be happy to talk with you or meet in person to discuss our review of your data, if you would like

Sincerely,

[*Name of Curator*]

AUGMENT Step

Augment the dataset

In this step we ensure metadata conforms to repository and/or appropriate discipline standards; adjust metadata to improve findability and accessibility; and improve documentation to make data more understandable, interoperable and reusable.

Common AUGMENT steps include:

- Enhance metadata to best facilitate discoverability, such as by ensuring datasets have a persistent identifier.
- Create and apply metadata for the data record, including descriptive keywords
- When appropriate, structure and present metadata in domain-specific schemas to facilitate interoperability with other systems
- Implement any other agreed-on enhancements to metadata or documentation following discussion with researcher

Key Ethical Considerations

- Make sure bibliographic information reflects correct author attribution.
- Ensure any augmentation by the depositor to resolve ethical questions from previous steps is completed.

Essential Tasks

- □ Review information received from the researcher from initial deposit and all subsequent conversations
- □ Update, as appropriate:
    - □ Metadata
    - □ Documentation (readme, Codebook, Data Dictionary, Other)
    - □ Replacement files
    - □ Organization and Arrangement of files
    - □ Documentation of file organization, hierarchy, and naming convention(s)
- □ Facilitate discoverability:
    - □ Add links to related publications, grants, reports, source data, etc.
    - □ Provide additional description of files as appropriate for external indexing or other purposes.
    - □ Add subject terms
- □ Ensure keywords are sufficient and representative
- □ Record all changes in the Curation Log
- □ Provide suggestions to improve accessibility of content (e.g., alt-text or additional descriptions; color contrast; etc)

TRANSFORM Step

Transform file formats

In this step, consider the file formats in the dataset to make them more interoperable, reusable, preservation friendly, and non-proprietary when possible.[1] Common TRANSFORM steps include:

- Identify specialized file formats and their restrictions (e.g., Is the software freely available? If so, link to it or archive it alongside the data)
- Propose open source or more reusable formats when appropriate
- Retain original file formats

[1] See Cornell's list of preservation format recommendations: http://guides.library.cornell.edu/ecommons/formats

Key Ethical Considerations

- Consider how best to navigate researcher bandwidth limitations and ownership of data with repository commitments to reducing barriers to reuse.
- Decide how to balance the potential benefits of transformation with the risks of mistakes and loss of content/context, especially if the curator or repository will be performing transformation. Document the decision.

Essential Tasks

- ☐ Check whether preferred file formats are in use
  - ☐ If not, recommend conversion
  - ☐ Retain original formats
- ☐ Check whether software needed is readily available
  - ☐ Suggest open source options, if applicable and appropriate
  - ☐ Ensure software and software version is documented
- ☐ Convert any data visualization(s) that are not accessible (e.g., R visualizations, which need to be converted for screen reader use, or visualizations that do not meet color contrast guidelines)
- ☐ Reorganize files as appropriate
- ☐ Standardize file names
- ☐ Record any transformations in Curator Log

EVALUATE Step

Evaluate and rate the dataset

In this step, review the dataset and companion data record against international standards, including FAIR,[2] CARE,[3] and FATE.[4] Common EVALUATE steps:

- Score the dataset and recommend ways to increase the FAIRness of the data
- Review data for ethical concerns in line with CARE and FATE

[2] Rubric evaluating the FAIR principles are based on the scoring matrix by Dunning, de Smaele, & Böhmer (2017).
[3] CARE principles: https://www.gida-global.org/care
[4] FATE in AI: https://www.microsoft.com/en-us/research/theme/fate/

Key Ethical Considerations

- Final review--remember it is not too late to surface any ethical concerns.
- Verify the words/language being used are not racist/harmful.
- Remind the submitter of their responsibility, if they choose to ignore requests for de-identification or similar concerns.

Essential Tasks

- ☐ Test that files successfully download
- ☐ Check that any transformations didn't introduce problems
- ☐ Review final state of data and record with researcher before publication
- ☐ Add any final changes to Curator Log

This is a sample checklist for evaluating datasets against a set of principles.

FAIR evaluation

Findable:

- ☐ Metadata exceeds researcher/ title/ date.
- ☐ There is a unique Persistent ID (DOI, Handle, PURL, etc.).
- ☐ Data/record is discoverable via web search engines.

Accessible:

- ☐ Data/ record are retrievable via a standard protocol (e.g., HTTP).
- ☐ Data/ record are free, open (e.g., via a download link).

Interoperable:

- □ Metadata is formatted in a standard schema (e.g., Dublin Core).
- □ Metadata is provided in machine-readable format (OAI feed).

Reusable:

- □ Data include sufficient metadata and supporting documentation about the data characteristics for reuse.
- □ A way to contact the researcher directly for further questions is provided
- □ There are clear indicators of who created, owns, and stewards the data.
- □ Data are released with clear data usage terms (e.g., a CC License).

DOCUMENT Step

Document curation activities throughout

In the Curator Log mentioned throughout this guide, record the significant treatments or actions applied to the dataset. This is for your archival record keeping (distinct from documentation the researcher(s) created to accompany their own datasets).
DOCUMENT requires:

- Recording all information relevant to the tracking and administration of the deposit, about who did what to the dataset and when
- Tracking communication with the researcher(s)

Key Ethical Considerations

- Document that disclosure risk review has taken place. State if changes from original data have been made, but do not give enough detail on changes to reverse-engineer any anonymization.
- Include consent (or waiver) and/or IRB approval of sharing with administrative documentation.
- Consider collecting contributor demographics.

Essential Tasks

- □ Ensure the following information is captured in the Curator Log:
    - □ Activities taken during the CURATE process

□ Accessioning & deposit records (Names, dates, contact information, submission agreements, etc.)
□ Repository collection metadata
□ Provenance logs (changes by curators in the Transform step)
□ Service workflow
□ Correspondences and other interactions
□ Preservation packaging
□ Any additional requirements at your institution