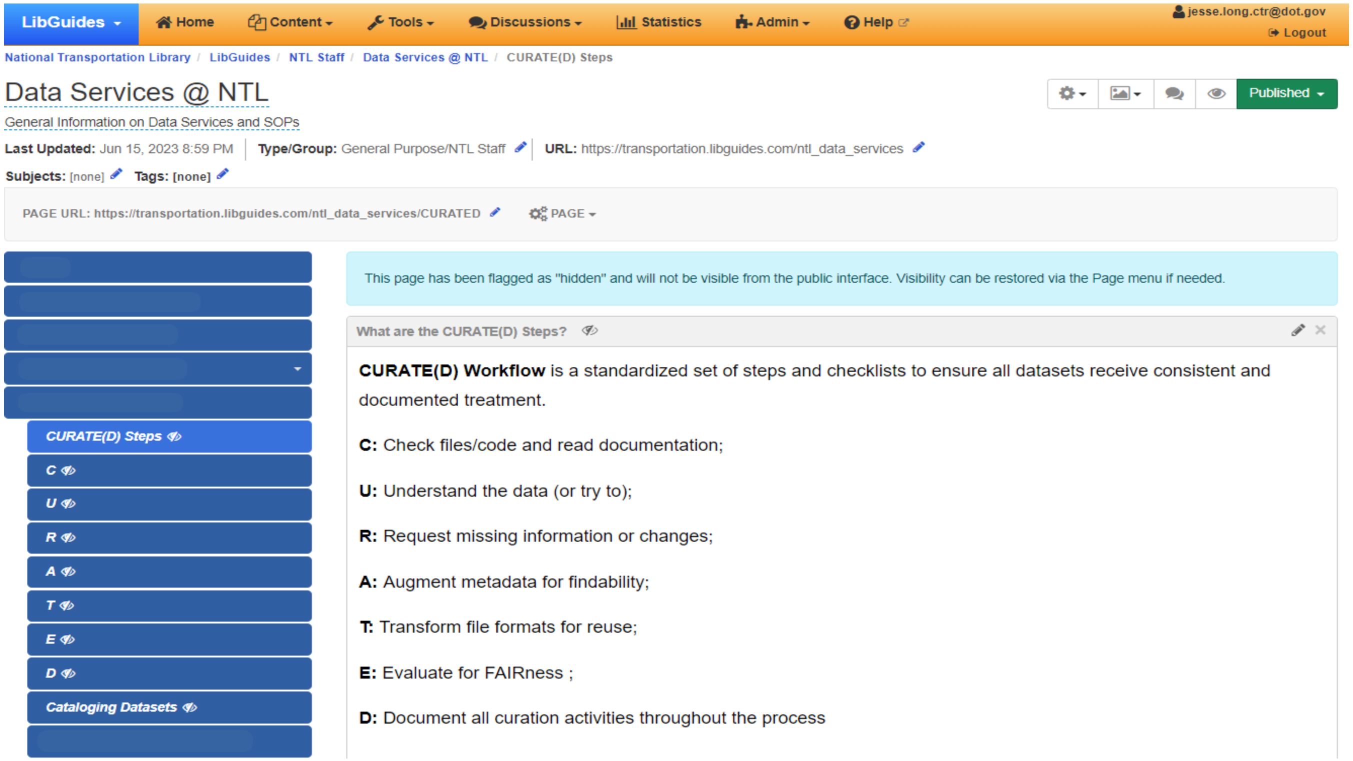
Integrating the DCN CURATE(D) Steps into the National Transportation Library's (NTL) Workflow.

Table of Contents

Note on Structure

My organization, The National Transportation Library (NTL), uses the software, LibGuides, which is a content management system that is primarily used by "librarians to curate knowledge and share information, organize class and subject specific resources, and to create and manage websites." This is software is used at NTL to create both internal staff guides and external guides for researchers. The below deliverable was originally created within the LibGuides software and then transferred into a word document, since it is an internal staff guide that would not be shared outside of the organization. As a result, there may be some inconsistencies in the language and layout, because the information exists across 9 pages in LibGuides, which allows for better navigation and version control. Additionally, certain links in the document won't work because they reference other internal staff guides. I have included an image to provide a visual understand of the guide's structure.



Data Services @ NTL

General Information on Data Services and SOPs

Last Updated: Jun 15, 2023 8:59 PM | **Type/Group:** General Purpose/NTL Staff | **URL:** https://transportation.libguides.com/ntl_data_services

Subjects: [none] **Tags:** [none]

PAGE URL: https://transportation.libguides.com/ntl_data_services/CURATED

This page has been flagged as "hidden" and will not be visible from the public interface. Visibility can be restored via the Page menu if needed.

What are the CURATE(D) Steps?

CURATE(D) Workflow is a standardized set of steps and checklists to ensure all datasets receive consistent and documented treatment.

C: Check files/code and read documentation;

U: Understand the data (or try to);

R: Request missing information or changes;

A: Augment metadata for findability;

T: Transform file formats for reuse;

E: Evaluate for FAIRness ;

D: Document all curation activities throughout the process

You can see from the image on the left, that each step in the CURATE(D) Process receives its own page dedicated to it. The 9 pages that are in the internal staff LibGuide corresponds to the 9 sections (excluding this note and references) that are found in the table of contents on the previous page.

CURATE(D) Steps

What are the CURATE(D) Steps?

CURATE(D) Workflow is a standardized set of steps and checklists to ensure all datasets receive consistent and documented treatment.

C: Check files/code and read documentation;

U: Understand the data (or try to);

R: Request missing information or changes;

A: Augment metadata for findability;

T: Transform file formats for reuse;

E: Evaluate for FAIRness;

D: Document all curation activities throughout the process

The CURATE(D) Steps were developed by the Data Curation Network (DCN).
https://datacurationnetwork.org/outputs/workflows/

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

NTL abides and operates according to the USDOT Public Access Plan, titled: "Plan to Increase Public Access to the Results of Federally-Funded Scientific Research."

We are required:

- by law to share the metadata about any funded datasets
- to make public as much of the data as possible
- to only release data that is not sensitive and does not disclose sensitive personal, business, or government information.

Resources

Data Curation Network CURATE(D) Training Modules:

The goal of the CURATE(D) Training is to offer an introduction to applied data curation. This training is designed for those completely new to data curation, those hoping to refresh their data curation skills, or those looking to apply data curation knowledge to the management of their own research data. This training and associated CURATE(D) model are teaching and research tools that are presented and best understood sequentially.

It is recommended that any new staff complete the training as part of their onboarding process. Additionally, this training can serve as a good refresher and source of information for the CURATE(D) Workflow.

Other Resources provided / recommended by the DCN:

- Curation Glossary - 50 curation activities defined
- Institutional Data Curation Survey Tool - Self-assessment tool to benchmark your data curation
- Curating Research Data: A handbook of current practice (Johnston, 2017) - a free ebook
- A Toolbox for Curating and Archiving Research Software for Data Management Specialists (John Hopkins Libraries) - an online, self-paced training modules

Starting Process: Current

If you are new to the Data Services Team you will need to ensure that you have access to the NTL Data Curator Inbox.

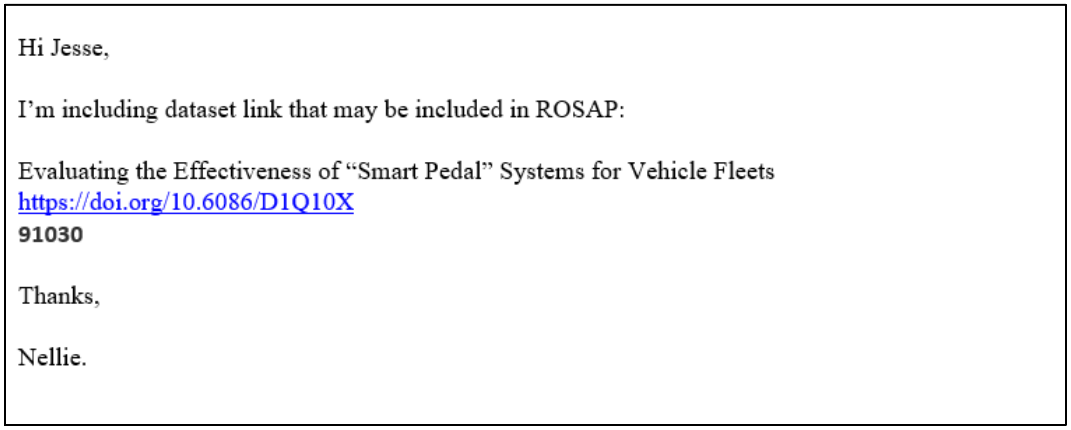
- If you have yet to receive access to this Inbox, please request access by: Asking a team member who already has access to the box to request your access through the OCIO Client (Help Desk). Once OCIO processes the request you will receive a confirmation email explaining how to successfully add the Inbox into Outlook.

New datasets that require cataloging will be forwarded to the NTL Data Curator Inbox (NTLDataCurator@dot.gov) from the NTL Digital Submissions Inbox (NTLDigitalSubmissions@dot.gov). This will be done by Nellie or Shawn after they have cataloged the associated report (if one was present), this way we can ensure the report and data are linked in Workroom and ROSA P.

The email will include the original contents from the submitter and additional information provided by the cataloger, both are necessary to ensure the CURATE(D) Steps are completed correctly. Below is an example of an email and the information need to pull for the CURATE log and subsequent cataloging of the dataset.

Email from Cataloger:

- This email will provide the link to the dataset that needs to be cataloged and, if applicable, the Workroom ID of the cataloged report that is associated with the data.



Hi Jesse,

I'm including dataset link that may be included in ROSAP:

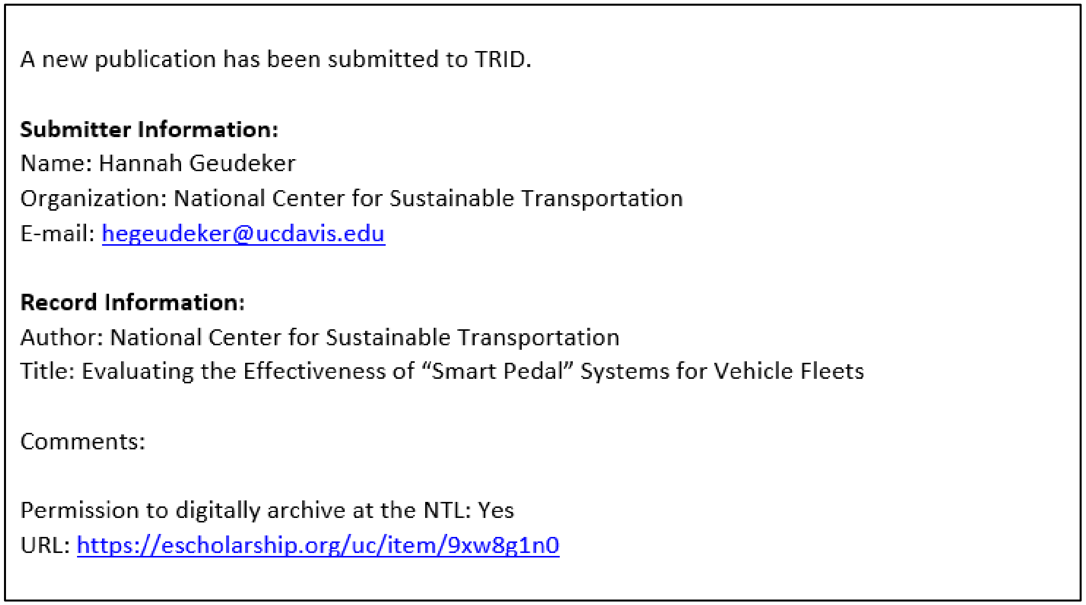
Evaluating the Effectiveness of "Smart Pedal" Systems for Vehicle Fleets
https://doi.org/10.6086/D1Q10X
91030

Thanks,

Nellie.

Email from Original Submission:

- This email will provide the submitter's information and original links, title, and author of the submitted item (this will be the report, dataset, or both)



A new publication has been submitted to TRID.

Submitter Information:
Name: Hannah Geudeker
Organization: National Center for Sustainable Transportation
E-mail: hegeudeker@ucdavis.edu

Record Information:
Author: National Center for Sustainable Transportation
Title: Evaluating the Effectiveness of “Smart Pedal” Systems for Vehicle Fleets

Comments:

Permission to digitally archive at the NTL: Yes
URL: https://escholarship.org/uc/item/9xw8g1n0

After receiving this email, please notify other members of the Data Services team that you will being the process of curating the dataset. After notifying your intention, log in to Workroom and select "Add New Record," and create the cataloging record for the dataset. You will not be completing the cataloging process at this point, just creating the record to serve as a placeholder and begin the CURATE(D) process. When the new record has popped up you will put in the following information:

- Title of the Dataset**:** (i.e. Evaluating the Effectiveness of "Smart Pedal" Systems for Vehicle Fleets [supporting datasets])
- Staff Notes: "Related to Workroom record ..." (i.e. Related to Workroom 91030)
- Status: Change to InProcess
- Status Comment: "Record created by [your initials] [today's date]" (i.e. Record created by JAL 20230612)
- Click Finish

Once the record is completed note the Workroom record number that has been provided for the dataset record, which is the final piece of information you need to start the CURATE log and begin the CURATE(D) Process.

Go to Next Step/Page "C"

Starting Process: Future

In the coming future, NTL will release a new web-based submission form for submitters to increase cataloging efficiency. Once this form is implemented the following process will become the standard for starting the CURATE(D) steps workflow.

If you are new to the Data Services Team you will need to ensure that you have access to the NTL Data Curator Inbox.

- If you have yet to receive access to this Inbox, please request access by: Asking a team member who already has access to the box to request your access through the OCIO Client (Help Desk). Once OCIO processes the request you will receive a confirmation email explaining how to successfully add the Inbox into Outlook.

Once submitters complete the submission form, the NTL Data Curator Inbox (NTLDataCurator@dot.gov) will be notified. If you intend to begin the CURATE process for the latest submission, please notify other members of the Data Services team that you will being the process of curating the dataset. After notifying your intention, log in to the submission form database to view the submission.

[insert image, when available]

Once the submission is located you can begin the CURATE(D) Process.

Go to Next Step/Page "C"

C: Check

Terms to Know

Submission Information Package (SIP): Items that have been submitted by the depositor.

Archival Information Package (AIP): A package that contains data that will be stored within a digital archive.

Dissemination Information Package (DIP): A package created from the Archival Information Package (AIP) to distribute digital content to users.

File Inventory: The list of files in the submission information package (SIP).

File Organization: The act of structuring files in a hierarchical way to ensure findability.

README File: A file that is usually a text file (.txt) or a rich text format file (.rtf) or markdown (.md) that gives information about the creators of the data, where the data was created, methods used to produce the data, sharing privileges, and so on.

Metadata: Structured information that describes attributes of the dataset. Metadata can include the author, file size, the date the document was created and keywords to describe the document.

Saving Data to P: Drive

When receiving a new dataset, the Submission Information Package (SIP), we need to save the data to the P: Drive in order to complete the CURATE(D) Steps and prepare the Archival Information Package (AIP) and Dissemination Information Package (DIP).

After downloading the data, go to the P: Drive location, P:\NTL\data_management_curation\Data_to_CURATE, and select the appropriate folder to store the data. The folders reflect the different transportation modes we receive data from and externally funded research. After selecting the correct folder for the data, open it and make a new folder within it and name it according to the dataset's Workroom ID.

Once created, enter into the dataset's folder and create two additional folders entitled "SIP" and "Working Folder." You will then place a copy of the data into each. However, the SIP folder will serve as the MASTER copy and will not be touched moving forward, while the copy in the Working Folder will be used to complete the CURATE(D) Steps.

Folder Hierarchy Example:

- Data to CURATE
  - BTS
    - 91030
      - SIP
      - Working Folder

Additionally, once you have completed the above steps make note of the file location for the data, which will need to be recorded in the next step with the CURATE Log. An example of the file location for the above example would be: P:\NTL\data_management_curation\Data_to_CURATE\BTS\91030.

How the different data packages interact:

- The working copy of the SIP will become an AIP through the process of curation and will contain our CURATE log for the dataset, and a version of the AIP, without our CURATE log included, will become the DIP to be retrieved by a user.

CURATE Log

You can locate the CURATE Log in the P: Drive at the following location and file name.

- P:\NTL\data_management_curation\Data_to_CURATE

There are two files you will be using to document the CURATE(D) Process. First, you will open up the file named NTL_CURATED_LOG.xlsx. This file is a living document that will track all the datasets that either have been CURATED or are in the process. As a result, it will be used by all members of the Data Services, so please be aware if you get the message that someone else is currently working in the document. We want to ensure all the information stays organized and in a single location.

Upon opening this document navigate to the first sheet titled "Overview." In this sheet, you will record the new dataset that you will be working on. At this time you will fill in the following columns with the information your currently have:

- A: Workroom ID
- B: Title

- L: Submitter
- M: Date Submitted
- N: Assigned to
- O: Files (s) Location
- P: Rows with Headers *(Rows describing what data was collected)*
- Q: Data Rows *(Rows contain data collected in the research project)*
- S: Curation Level (*According to CoreTrustSeal's Levels of Curation)*

The other columns will be updated as you perform each step of CURATE(D). After adding the information to the Overview Sheet in NTL_CURATED_LOG.xlsx, save and close the file.

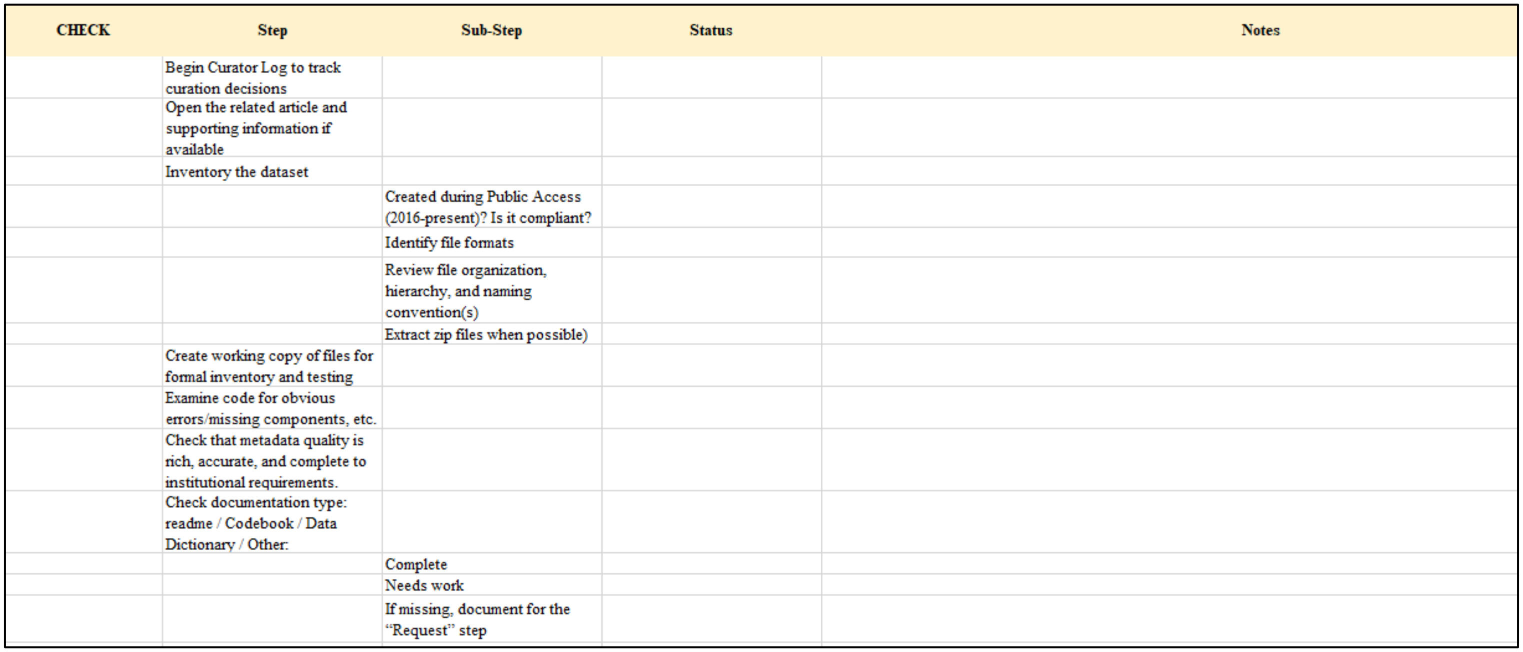
Next, open the file CURATE_LOG_TEMPLATE_Single_Record.xlsx. This is the CURATE log that you will create for the specific dataset that you are working with. Upon opening the file the first thing you want to do is go to "File" and "Save As," in order to create a new copy and not overwrite the original template. When saving this file make sure to follow the file naming structure defined below and in the same location as the data.

- File Naming Structure: WorkroomID_CURATE_LOG_StartDate.xlsx
    - Example: 91030_CURATE_LOG_20230612.xlsx

After saving the file make sure to rename the sheet to match the Workroom ID and fill in the information you already have regarding the dataset in rows 1-9 (this will be the same information you inserted into the Overview Sheet on the NTL_CURATED_LOG.xlsx file).

Below is a sample image of the CURATE log, which will document your work during the CURATE(D) Steps. There are 5 columns:

- Column 1: Shows the overall step you are currently working on. (i.e. Check, Understand, Request, etc.)
- Columns 2 and 3: Reveals the specific descriptions for what is being evaluated, which are broken down into steps in column 2 and sub-steps in column 3.
- Column 4: Is the status of each point or subpoint. In this column, there is a drop-down menu in each cell, so each should contain either "Checked", "Not Checked," or "Not Applicable."
- Column 5: Is where you will document notes on your findings while completing the evaluation of each step or sub-step.



CHECK	Step	Sub-Step	Status	Notes
	Begin Curator Log to track curation decisions			
	Open the related article and supporting information if available			
	Inventory the dataset			
		Created during Public Access (2016-present)? Is it compliant?		
		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: readme / Codebook / Data Dictionary / Other:			
		Complete		
		Needs work		
		If missing, document for the "Request" step		

Once you have completed the CURATE(D) Steps you will make a copy of this file and make it its own individual sheet in the living document NTL_CURATED_LOG.xlsx. This will be explained in the final Step "D."

Check Step

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

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
  - NTL's Collection Development Policy
- 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

Workflow

Now that all the preliminary steps are complete it is time to begin working through the CURATE(D) Steps. Each step should be evaluated and documented inside the dataset's specific CURATE log (WorkroomID_CURATE_LOG_StartDate.xlsx).

Below is a table to help guide you through the "Check" process and provide NTL-specific considerations.

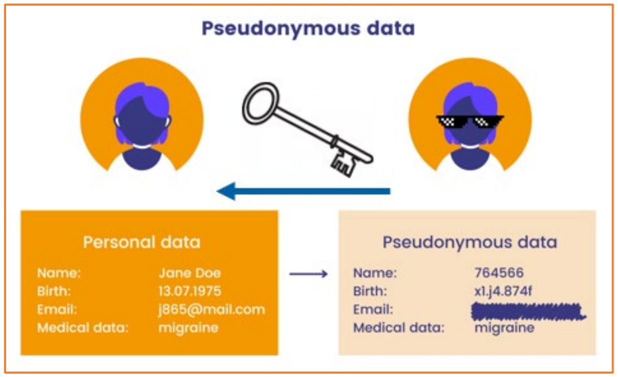
What you are checking:	NTL-Specific:	What to Document:
Begin CURATE Log to track curation decisions.	Follow the naming convention listed in CURATE Log Section.	Date CURATE Log started and store in data's folder in P: Drive.
Open the related article and supporting information if available.	Most data submitted to NTL will have an associated report that is cataloged by the cataloging team.	The Workroom ID, ROSA P URL, and DOI (if provided) of the related report.
Inventory the dataset. (Public Access, files in submission, file formats, folder hierarchy, naming structure)	Was the data created under the USDOT Public Access Plan and if so, is it compliant? Do you have the required software to open the data files? If not does another member of the Data Services team? Does the hierarchy or naming structure follow NTL preferences?	- Compliance with the Public Access Plan - List all files submitted (full names) - file formats - Notes on the organization of files
Create a working copy of files for formal inventory and testing.	This should have already been completed during the previous section on downloading and storing data in P: Drive.	File location in P: Drive.
Examine code for obvious errors/missing components, etc.	NTL does not currently receive a lot of code, but we do get one once in a while. If you don't have the software or know how to	Note if there are errors/missing components.

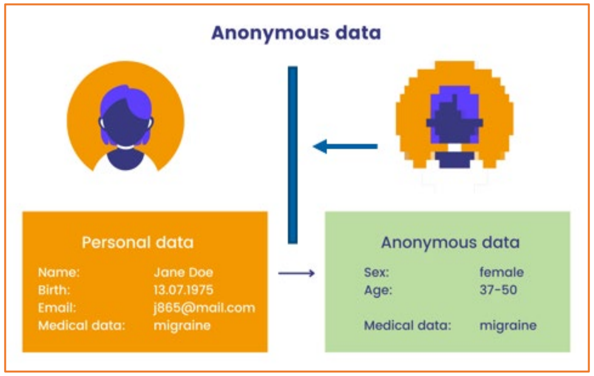
	check the code, please bring it up during the team meeting.	
Check that metadata quality is rich, accurate, and complete to institutional requirements.	NTL will need to create the government-mandated metadata file following DCAT-US Schema in addition to any discipline-specific metadata provided by the submitter.	Notes on review of metadata submitted by the reviewer.
Check documentation type: readme / Codebook / Data Dictionary / Data Management Plan / Other.	In some cases, depending on the submitter, NTL will take care of the necessary README file. However, a Codebook/Data Dictionary/Methodology should be provided.	What files were included in the submission? Were any missing?
Check whether human subject data (data about humans regardless of IRB determination) is present.	If there is human subject data, you can utilize the resources provided below to accomplish this step, as needed.	Note if there is or isn't human subject data. If so, express any potential concerns regarding identification and potential solutions. (An example of this type of situation will be included when addressing the Request step). Also if human subject data is found there should be a consent for more participation agreement listed in the documentation files.
Check the accessibility of all files.	NTL has a LibGuide on Accessibility for reference, as needed.	Notes on Compliance
Check whether any visualization(s) of data are easily accessible.	NTL has a LibGuide on Accessibility for reference, as needed.	Notes on Compliance

Go to Next Step/Page "U"

Resources

Pseudonymisation vs. Anonymisation





Assessing Reasonable Likelihood of Re-identification

- No unique entries, e.g not only one person above 90 years
  - Size and distribution of the data set are relevant
- Consider general privacy and security measures, such as data access restrictions or encryption
  - What are likely threat scenarios?

Read More: Luk Arbuckle and Khaled El Emam, *Building an Anonymization Pipeline: Creating Safe Data* (2020) & *Anonymizing Health Data: Case Studies and Methods to Get You Started* (2013)

Strategies for Anonymization

- Delete or don’t collect (in-)direct identifiers
- Aggregation/ reduction of information (e.g. age intervals instead of precise age)
- Randomization (e.g. exchanging values or noise addition)
- Synthetic data (creating data set with similar statistical patterns as an original dataset)

Read More: Article 29 Data Protectio Working Party. 10.04.2014. *Opinion 05/2014 on Anonymisation Techniques*
Open Data Institute. 2019. ‘Anonymisation and open data: An introduction to managing the risk of re-identification

U: Understand

Terms to Know

Absolute/Relative Path: The path refers to the location of a file in the directory structure of where it is stored.

- Absolute paths provide a full list of all of the folders from the beginning (or "root") of the storage unit. On most unix-based systems, the root directory is "\." On Windows systems, the root directory usually begins with a drive letter such as "C:\".
- Relative paths provide a list of folders that begin at a designated folder (usually the initial folder of a project). In the case of curating for secondary use, relative paths are preferred for long-term preservation since it is generally easier to share and preserve the initial project folder and all subsequent folders.

Codebook: A codebook provides a description of all the items contained in the data collection with information about individual files, measures, and codes used to represent those files. A codebook will often provide additional context about the data collection process, assumptions, requirements, and descriptive statistics that enable a secondary user to understand the context for the collection while also validating the integrity of the data collection.

Commented Code: Documentation embedded in the computer programming code that is ignored by the interpreter or compiler when the computer program executes.

Data Dictionary: A list of the elements contained in a dataset and their position in the data file. Each file in a data submission may have its own data dictionary.

Delimited file: A data file in which each data element is separated by a common character. Comma-separated values (.csv) files are very popular, but tab-separated values (.tsv) files and pipe delimited (|) files are also used in many data projects.

File dependency: Software code that requires the presence of certain files (file dependency) or software libraries for the program to execute. Some dependencies may require a particular version of a software code for execution.

Understand Step

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

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)

Format/Domain Specific Questions

Currently, in the CURATE log there are checklists for:

- Tabular Data (e.g, .CSV, Microsoft Excel. etc)
- Database(s)
- Geodatabase(s)
- Code

These are the most likely formats that data will be submitted in to NTL. However, at times or in the future we may receive more of a variety of formats. As a result, additional checklist tasks for the CURATE log based on specific file formats and subject domains are listed below:

- Acrobat PDF Primer
- ATLAS.ti Primer
- Confocal Microscopy Image 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

All primers created by the DCN can be found in the University of Minnesota's Libraries Digital Conservancy in the digital collection "Data Curation Network Primers." Persistent link to this collection: https://hdl.handle.net/11299/202810

Interactive primers available for download and derivatives at: https://github.com/DataCurationNetwork/data-primers

Workflow

The Check and Understand steps are deeply linked, and you may find you are performing them at the same time. That is ok.

Below is a table to help guide you through the "Understand" process and provide NTL-specific considerations.

Do you Understand:	NTL-Specific:	What to Document:
Examine files, organization, and documentation more thoroughly. Are there changes that could enhance the dataset?	What is the data's total file size? (if it is larger than 1 GB it can't be downloaded from ROSA P, so if it is larger we need to discuss options for it and should be brought up at the team meeting).	Notes on data, if there is anything missing, rescue and reproducibility, does documentation support understanding of data.
Data (e.g, Microsoft Excel) Questions:	This is the most frequent data type that you will evaluate at NTL, which is why this format's questions are already listed within the CURATE log (along with two lesser common formats, i.e. database and code). If you encounter a different data format, please use the additional references listed in the Format/Doman Specific Questions in the above section to select the one that fits	Notes on whether the data meets qualifications outlined in format/domain-specific questions.

	the data, add those questions into the CURATE log and evaluate them.	

Go to Next Step/Page "R"

R: Request

Request Step

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

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 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.

Communication Tips and Examples

1. Establish Rapport with the Researcher (It may take more than one email)
    - Start with a “thank you” and let them know that the NTL is happy to be able to share their research with our users.
    - It is easier to ask for supplemental information once rapport is established; the initial email should be as brief and concise as possible.
    - Offer to schedule a meeting if needed.
    - A little empathy goes a long way. Consider:
        - What are the depositor’s needs in depositing their dataset?
        - This may be an opportunity for them to learn, and the beginning of a long-term relationship as they continue to deposit data in the repository: treat it like a collaboration.
        - Keep in mind that the depositor may be under a tight time frame and may not have the time to make recommended improvements this time around.
2. Let the depositor know what changes and additions need to be made to the data to make it as complete and understandable as possible.
    - Formulate your request for the depositor to respond. Explain what you need from the depositor, and why.
        - Do you need additional information or permission from the depositor to take action? If so, make sure this ask is part of your request.
        - Make your request as specific as possible. Provide examples and resources when relevant.

- Prioritize your requests! Ask for information that is most important for the dataset quality first.
    - Essential: What information is critical to improving the quality of the dataset? (for example: missing README or metadata, files that won't open or are missing extension information, errors found in the data set).
    - Important: What information is useful to improve the quality of the dataset? (for example, incomplete README or metadata, changes to file names).
    - Supplemental: What information is helpful to improve the quality of the dataset? (for example: small edits recommended to improve clarity or increase FAIRness of data like ORCIDs, funding information, keywords).
- Make it easy for depositors to respond.
    - Limit to four asks.
    - Be specific, but concise.
    - Keep it simple; ideally, depositors could respond with a yes or no.
    - Provide resources where useful.
    - If possible, offer to make changes yourself and ask for approval.

3. Encourage the depositor to supply missing information.
    - Tell the depositor how they should get you the information.
    - Offer to help, for example: include a README template that you have started populating with their information and ask them to complete it.

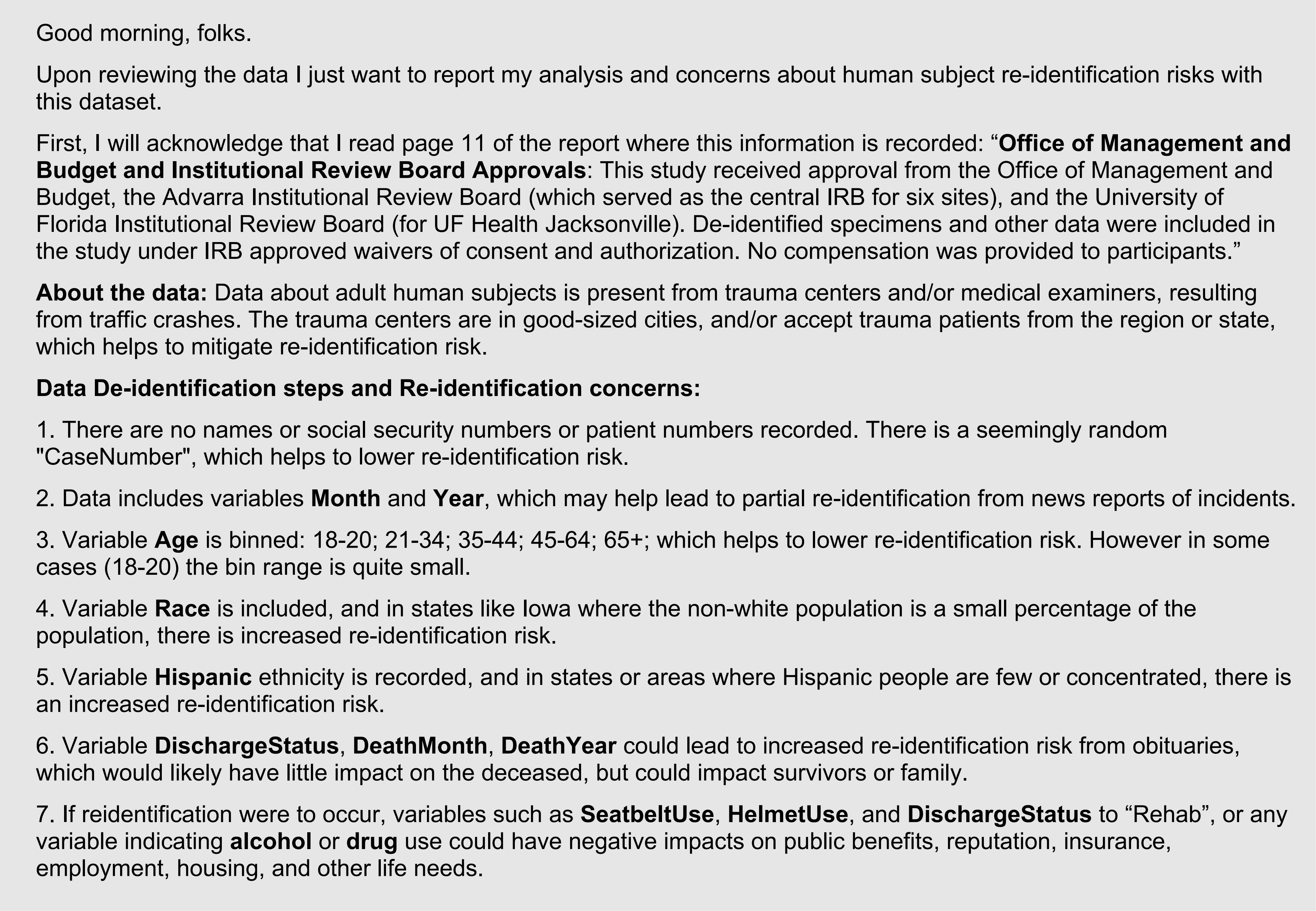
4. Get permission to make any recommended changes to improve the quality of the dataset.
    - If your request includes changes, you would be able to make yourself (for example, changes to file names), ask for the researcher's approval before making the changes.
    - Keep it simple. Explain what changes you recommend and allow researchers to respond with a "yes" or "no".

Above Recommendations from DCN's Training Module on the "R Step: Requesting Missing Information."

Sample email from previous dataset CURATE(D) by NTL:

For this example, I would like to provide the additional information that the Data Services team met with members of the submitting organization prior to receiving this specific dataset.

Initial email after completing "C" and "U" Steps:



Good morning, folks.

Upon reviewing the data I just want to report my analysis and concerns about human subject re-identification risks with this dataset.

First, I will acknowledge that I read page 11 of the report where this information is recorded: "**Office of Management and Budget and Institutional Review Board Approvals**: This study received approval from the Office of Management and Budget, the Advarra Institutional Review Board (which served as the central IRB for six sites), and the University of Florida Institutional Review Board (for UF Health Jacksonville). De-identified specimens and other data were included in the study under IRB approved waivers of consent and authorization. No compensation was provided to participants."

About the data: Data about adult human subjects is present from trauma centers and/or medical examiners, resulting from traffic crashes. The trauma centers are in good-sized cities, and/or accept trauma patients from the region or state, which helps to mitigate re-identification risk.

Data De-identification steps and Re-identification concerns:

1. There are no names or social security numbers or patient numbers recorded. There is a seemingly random "CaseNumber", which helps to lower re-identification risk.

2. Data includes variables **Month** and **Year**, which may help lead to partial re-identification from news reports of incidents.

3. Variable **Age** is binned: 18-20; 21-34; 35-44; 45-64; 65+; which helps to lower re-identification risk. However in some cases (18-20) the bin range is quite small.

4. Variable **Race** is included, and in states like Iowa where the non-white population is a small percentage of the population, there is increased re-identification risk.

5. Variable **Hispanic** ethnicity is recorded, and in states or areas where Hispanic people are few or concentrated, there is an increased re-identification risk.

6. Variable **DischargeStatus**, **DeathMonth**, **DeathYear** could lead to increased re-identification risk from obituaries, which would likely have little impact on the deceased, but could impact survivors or family.

7. If reidentification were to occur, variables such as **SeatbeltUse**, **HelmetUse**, and **DischargeStatus** to "Rehab", or any variable indicating **alcohol** or **drug** use could have negative impacts on public benefits, reputation, insurance, employment, housing, and other life needs.

Recommendations?

I do not have any specific recommendations on how to further decrease any potential re-identification risk, with the exception of the very small bin for 18-20 years old. And now that the analysis is done, that may be a correction that cannot be made. If we had caught that during data collection planning, maybe we could have suggested an alternative. A bad actor with an enough time and access to news reports of vehicle crashes and obituaries might be able to come close to guesses about the identity of some folks. But I would say the risk is low.

Actions

If you all are comfortable with making this data public as-is, then we can go ahead.
If you have concerns, we can pause for a moment.

Please advise,"

<u>Upon further review of the data, additional recommendations were sent to the submitter:</u>

"Hi folks.

Last email, I promise. My apologies, I have been think about this all day Tuesday and overnight.

My final recommendation for helping to prevent re-identification of human subject in this dataset are these:

1. Remove these variables, as they raise re-identification risk (when paired with other public data and information such as news and accident reports), and are not used to support the conclusions:
    a. Year
    b. Month
    c. Race
    d. Hispanic
    e. Comorbidities
    f. Complications
2. Ask contract to resubmit .sav and .csv datafiles, as well as updated Codebook with those variables removed.
3. NTL will add a note to the README text and data management plan that these variables were removed before data sharing.

Please let me know your thoughts."

Workflow

Below is a table to help guide you through the "Request" process and provide NTL-specific considerations.

Requesting:	NTL-Specific:	What to Document:
Ask about additional data contributors, beyond publication authors, or contributor roles.		Not the requests you make to the submitter.
Summarize conversations / outreach efforts in Curator Log	If an issue arises that we haven't previously dealt with or addressed consider adding a new example to this workflow/guide so we can remain informed for the future.	Document all communication between submitters and dates.

Go to Next Step/Page "A"

A: Augment

Terms to Know

Metadata: Data describing the context, content and structure of records and their management through time. Metadata is: information; about some other form of communication; in a structured format; designed to serve a particular purpose; and which may serve in some circumstances as a surrogate for the original communication.

Metadata Standards: Metadata is made up of a number of elements which can be categorised into the different functions they support. A metadata standard will normally support a number of defined functions, and will specify elements which make these possible. A metadata standard may support some or all of the following functions: descriptive metadata, technical metadata, administrative metadata, use metadata, and preservation metadata.

Persistent Identifiers (PIDs): A PID is a long-lasting digital reference to an object, contributor, or organization, “a code which remains constant as a means of identifying a digital object regardless of changes to its location on the internet." An “identifier” is “an association between a string (a sequence of characters) and an information resource." Web URLs are an example of a common identifier. The term "persistent" refers to the need for an identifier to provide continued access to and provenance for the object it refers to for years to come.

For more information on PIDs consult NTL's LibGuide on the topic.

- DOI: DOIs are persistent unique identifiers designed for research objects, such as articles, books and book chapters, conference proceedings, data sets, etc. The DOI system is designed to identify objects wherever they are located on the web, unlike a URL that points to a specific location on the web which may change or disappear over time. DOIs alleviate the problem of dead links or link rot.
- ORCID: Open Researcher and Contributor IDs (ORCID iDs) enable identification, linking, and discovery between researchers. ORCID provides a registry where individuals may obtain a unique PID, which can be used in connection with their research and scholarly workflows.

README: A file that is usually a text file (.txt) or a rich text format file (.rtf) or markdown (.md) that gives information about the creators of the data, where the data was created, methods used to produce the data, sharing privileges, and so on.

Augment Step

Augment: metadata for findability (DOIs, metadata standards, discoverability)

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.

Workflow

Below is a table to help guide you through the "Augment" process and provide NTL-specific considerations.

Augmenting the Data:	**NTL-Specific:**	**What to Document:**
Reserve DOI (if needed)	DOIs are required in the USDOT Public Access Plan, ideally, the data will already have a DOI associated with it. Either through NTL or another source. However, if there isn't one it needs to be completed here to ensure it is included within all the needed documentation files.	The DOI associated with the data.
Review information received from the researcher from the initial deposit and all subsequent conversations		Record any new or changed documents that were obtained from the submitter during the Request step.
Updates to the Data: Metadata, Documentation files (README, data dictionary, etc.), Replacement files, Organization &	This is where you will create the required DCAT-US Schema metadata file and README file (if needed).	Document any file creations (including the file names) and changes to current files.

Arrangement, Organization, and naming structure.	Additionally, it is time to make final changes that have been identified in previous steps to the documentation files and organization structure prior to cataloging.	
Discoverability tasks: Add related links, Provide additional descriptions of files as appropriate for external indexing or other purposes, and Determine ROSA P Collection.	When reviewing ROSA P Collections and determining where the data will live in the repository remember to review the related report's Workroom record to see what collection it was placed into. If it isn't in one, or you feel it has been incorrectly placed bring the topic up at the next team meeting, so we can ensure the report and data live in the same collection and are correctly placed in ROSA P.	Provide notes on the tasks or actions taken for this step and its sub-steps When Collection is determined go ahead and document it within the Workroom Record
Ensure keywords are sufficient and representative	In Workroom, we collect metadata for TRT (Transportation Research Thesaurus) Terms and Subject keywords. Resources and information on how best to select these terms are provided in the NTL Procedures for Submission and Workroom LibGuide Under: • TRT Terms • Subject Keywords When deciding these terms for the dataset you can go ahead and complete this task in the Workroom record. Additionally, if there is already a report record that the data is associated with, please compare/review the terms the cataloger already selected in its Workroom record.	Document the day you added the keyword to the Workroom record.
Address suggestions to improve the accessibility of content (e.g., alt-text or additional descriptions; color contrast; etc).	Make any needed changes to improve the accessibility of the data and supporting documentation. Following NTL recommendations from the Accessibility LibGuide. These	Document any changes made to increase accessibility.

	suggestions should have come from your analysis of the files in the Check step.	

Go to Next Step/Page "T"

T: Transform

Terms to Know

Conversion or Transformation: The migration of information from one file format to another, usually for purposes of preservation or access.

Access: The act of making information available. To increase ease of access, data should be made available in a convenient and modifiable form.

Accessibility: Content that is accessible is designed and developed so that people with disabilities can use it. For data curators, accessibility can include technical requirements that facilitate access for people with a diverse range of hearing, movement, sight, and cognitive ability (e.g. formatting that is compatible with screen readers), as well as requirements that facilitate user interactions (e.g. understandable instructions remove barriers to access, understanding and reuse of the data). Curating with accessibility in mind can improve data for all future users.

Interoperability: Data formatted using a disciplinary standard for better integration with other datasets and/or systems.

Preservation: Ensuring that data remain intact, accessible and understandable over time. This requires preserving the integrity of digital files themselves, and can be very complicated. Preservation actions may include preserving the software required to interact with the data or emulating older systems, migrating data to new formats and new media, and ensuring there is sufficient metadata to understand, interpret, manage and preserve the data.

Proprietary Format: A proprietary format is a file format of a company, organization, or individual that contains data that is ordered and stored according to a particular encoding-scheme, designed by the company or organization to be secret, such that the decoding and interpretation of this stored data is easily accomplished only with particular software or hardware that the company itself has developed. (Wikipedia)

Open Format or Non-proprietary Format: An Open File Format is a file format that is published and freely available for anyone to use. An open file format is licensed with an open license. For example, an open format can be implemented by both proprietary and free and open-source software, using the typical software licenses used by each. Open file formats are often recommended for preservation purposes because they typically do not require special software to open.

Transform Step

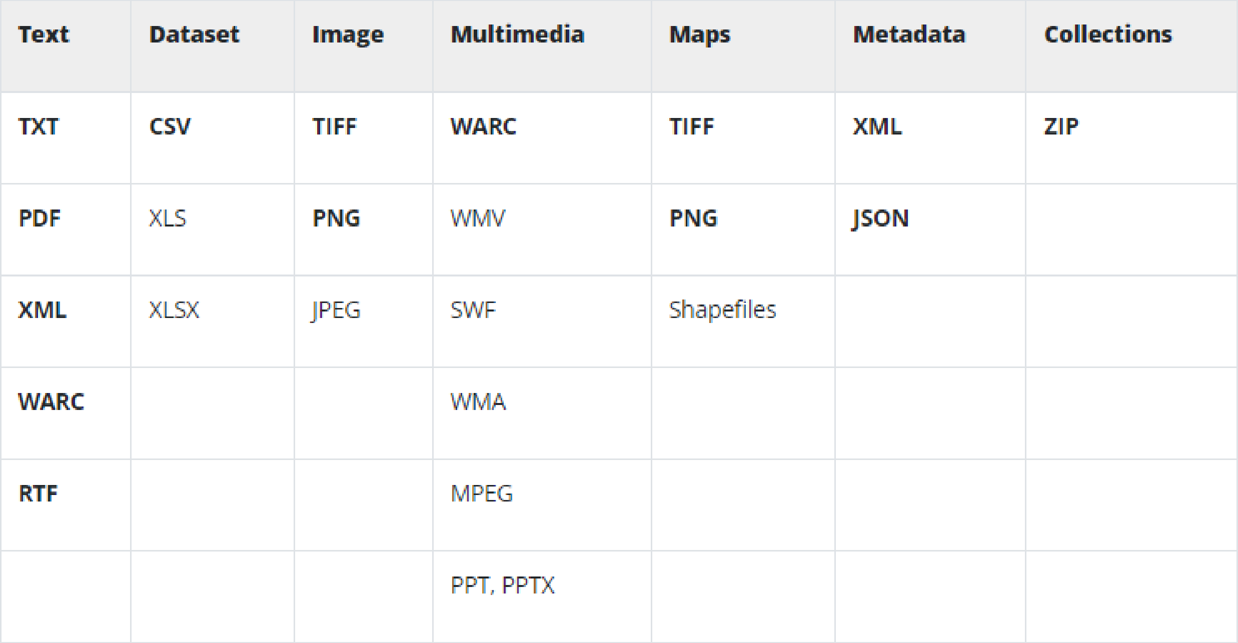
Transform: file formats for reuse (data preservation, conversion tools, data viz)

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.

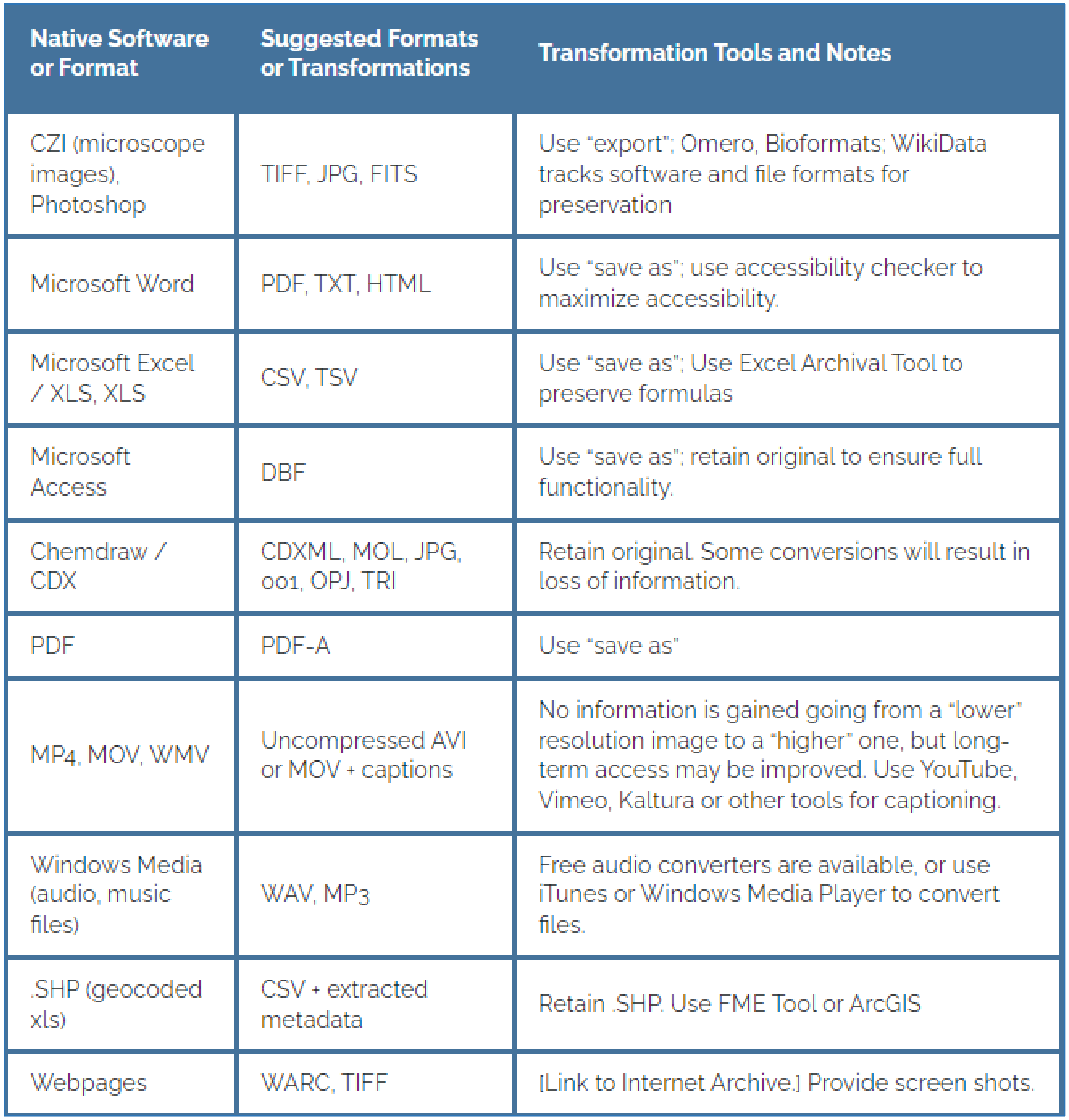
Resources

NTL has recommended formats for digital objects submitted to ROSAP (https://www.bts.gov/ntl/submitting-content). Digital objects provided by the content creators or owners in one of the following file formats (preferably one in **bold**).



Text	**Dataset**	**Image**	**Multimedia**	**Maps**	**Metadata**	**Collections**
TXT	**CSV**	**TIFF**	**WARC**	**TIFF**	**XML**	**ZIP**
PDF	XLS	**PNG**	WMV	**PNG**	**JSON**	
XML	XLSX	JPEG	SWF	Shapefiles		
WARC			WMA			
RTF			MPEG			
			PPT, PPTX			

NTL supports and advocates for the use of open access formats, but we do except proprietary formats because getting a copy of the data is our top priority. However, when able we will transform the proprietary format into an open access format to increase the data's usability. It is in these cases that the transform step will be relevant. If data is provided initially in an open format the transform step will not be necessary.



Native Software or Format	Suggested Formats or Transformations	Transformation Tools and Notes
CZI (microscope images), Photoshop	TIFF, JPG, FITS	Use "export"; Omero, Bioformats; WikiData tracks software and file formats for preservation
Microsoft Word	PDF, TXT, HTML	Use "save as"; use accessibility checker to maximize accessibility.
Microsoft Excel / XLS, XLS	CSV, TSV	Use "save as"; Use Excel Archival Tool to preserve formulas
Microsoft Access	DBF	Use "save as"; retain original to ensure full functionality.
Chemdraw / CDX	CDXML, MOL, JPG, 001, OPJ, TRI	Retain original. Some conversions will result in loss of information.
PDF	PDF-A	Use "save as"
MP4, MOV, WMV	Uncompressed AVI or MOV + captions	No information is gained going from a "lower" resolution image to a "higher" one, but long-term access may be improved. Use YouTube, Vimeo, Kaltura or other tools for captioning.
Windows Media (audio, music files)	WAV, MP3	Free audio converters are available, or use iTunes or Windows Media Player to convert files.
.SHP (geocoded xls)	CSV + extracted metadata	Retain .SHP. Use FME Tool or ArcGIS
Webpages	WARC, TIFF	[Link to Internet Archive.] Provide screen shots.

Sometimes we will need to transform the data into a public access format for release, although we will preserve both the proprietary and open format in the Archival Information Package (AIP). To the Left is a Common Transformations Table that DCN created to assist curators with the most likely transformations that they might encounter on the job. When you determine that a transformation is needed this is a good place to start.

Additional Resources:

- Data Curation Software Tools List (DCN). This list is a good reference when trying to determine what software is needed to open a specific file format.
- NTL File Format List. This list was created while working with external datasets (through an old workflow) and has been preserved to provide insight into the common formats that staff has encountered with datasets. If you come across a new format during your work, please add your findings to this list.
- File Extensions: If you do come across a file format that is new to you this is a great resource to figure out what it is and what software is needed to open it. (This is the site that was used when creating NTL's File Format List).
- Cornell University Libraries preservation format recommendations. This is a list of the best recognized and accepted formats when it comes to long-term preservation.

Workflow

Transformation may be recommended for several reasons including increasing access, interoperability, and likelihood of long-term preservation. If a file transformation has taken place, it is NTL's preference that both the original file and the transformed file be included in the AIP and DIP. An exception may arise due to file size, presence of executables, etc. Some file types are better for long-term preservation, but in the act of transformation, some of the proprietary file functionality is lost. Therefore it is recommended to share both versions with users whenever possible.

Below is a table to help guide you through the "Transform" process and provide NTL-specific considerations.

Transforming the Data:	**NTL-Specific:**	**What to Document:**
Check whether preferred file formats are in use.	NTL recommends the use of open access file formats. Above you will find the recommended/preferred file formats that NTL excepts for submission into ROSA P.	Are the file formats open access? If not, is the reason for the proprietary format documented (including software needed to view the data)? Can we have both an open access and propriety format if the researcher feels the proprietary format is necessary?

Check whether the software needed is readily available.		Document what software is needed to open each file type, and link to the software if possible.
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).	NTL has a LibGuide on Accessibility for reference, as needed.	Note all files that are converted and comply with accessibility guidelines.
Reorganize files as appropriate		What reorganization, if any, took place and why.
Standardize file names	If renaming files is needed or takes place for another determined reason, beware and check to ensure supporting documentation files reflect this change.	Document why file names were changed and that supporting documentation files were reviewed. Otherwise N/A for no change.

Go to Next Step/Page "E"

E: Evaluate

Terms to Know

FAIR: The FAIR Principles were developed by a set of diverse stakeholders that outline how scientific data should be shared. They stand for Findable, Accessible, Interoperable, and Reusable.

Having FAIR data is an important end goal of data sharing and data curation.

Findable
The first step in (re)using data is to find them. Metadata and data should be easy to find for both humans and computers. Machine-readable metadata are essential for automatic discovery of datasets and services, so this is an essential component of the FAIRification process.

F1. (Meta)data are assigned a globally unique and persistent identifier

F2. Data are described with rich metadata (defined by R1 below)

F3. Metadata clearly and explicitly include the identifier of the data they describe

F4. (Meta)data are registered or indexed in a searchable resource

Accessible
Once the user finds the required data, she/he/they need to know how they can be accessed, possibly including authentication and authorisation.

A1. (Meta)data are retrievable by their identifier using a standardised communications protocol

A1.1 The protocol is open, free, and universally implementable

A1.2 The protocol allows for an authentication and authorisation procedure, where necessary

A2. Metadata are accessible, even when the data are no longer available

Interoperable

The data usually need to be integrated with other data. In addition, the data need to interoperate with applications or workflows for analysis, storage, and processing.

I1. (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.

I2. (Meta)data use vocabularies that follow FAIR principles

I3. (Meta)data include qualified references to other (meta)data

Reusable

The ultimate goal of FAIR is to optimise the reuse of data. To achieve this, metadata and data should be well-described so that they can be replicated and/or combined in different settings.

R1. (Meta)data are richly described with a plurality of accurate and relevant attributes

R1.1. (Meta)data are released with a clear and accessible data usage license

R1.2. (Meta)data are associated with detailed provenance

R1.3. (Meta)data meet domain-relevant community standards

The principles refer to three types of entities: data (or any digital object), metadata (information about that digital object), and infrastructure. For instance, principle F4 defines that both metadata and data are registered or indexed in a searchable resource (the infrastructure component).

Evaluate Step

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

In this step, review the dataset and companion data record against international standards, including FAIR, CARE, and FATE. 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

Currently, at NTL and in the CURATE log we are only including questions regarding FAIR. However, this step was designed by DCN in preparation for other principles as the emerge and become nationally or globally recognized. To of these upcoming principles, identified by DCN are CARE and FATE. Although NTL has not adapted the evaluating into the current iteration of the CURATE(D) Workflow the resources for CARE and FATE are included and listed below because they have the potential to be useful and integrated in the future.

Resources

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

Curation is a partnership between:

- the curator and the researcher
- the researcher and the repository system
- the curator and the repository system

1: Researcher / Curator Relationship:

- Was communication with the researcher successful? Did they make/accept the recommended modifications to the dataset?
- Did the expertise of the curator allow you to effectively work with the researcher's data?
- Did the researcher value the curation process?

2: Researcher / Repository Relationship:

- Do the features of the system/platform facilitate making the data FAIR (i.e., minting PIDs, assigning licenses, structured metadata, etc.)?
- Is the technology well supported and maintained?
- What standards and best practices does the repository follow? (i.e., digital preservation, etc.)

3: Curator / Repository Relationship:

- As a depositor/user, do I trust this repository?
- Will this repository ensure my data are FAIR? How?

- Is there transparency with what actions will be taken with my data?

Workflow

Below is a table to help guide you through the "Evaluate" process and provide NTL-specific considerations.

Evaluating the Data:	**NTL-Specific:**	**What to Document:**
Test that files successfully download		Document Success or if there are issues what they are.
Check that any transformations didn't introduce problems	If any transformations were performed in the previous step, was information lost in the conversion? If so, we will need to discuss at the team meeting how we want to limit information loss.	Document information loss, if there was any.
Review the final state of data and record with the researcher before publication		Document Communication between researcher (dates email sent and their response)
Findable: • Metadata exceeds researcher/ title/ date • There is a unique Persistent ID (DOI, Handle, PURL, etc.) • The data/record is discoverable via web search engines.		Document if the dataset fails any points and if so plan to discuss with team.
Accessible:		Document if the dataset fails any points and if so plan to discuss with team.

• The Data/record is retrievable via a standard protocol (e.g., HTTP). • The Data/ record is 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).		Document if the dataset fails any points and if so plan to discuss with team.
Reuseable: • 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 if the dataset fails any points and if so plan to discuss with team.

Go to Next Step/Page "D"

D: Document

Terms to Know

Accession: The record of deposit made upon addition to the repository collection with relevant cataloging details such as dates and depositor names.

Catalog: The descriptive metadata associated with the deposited files and dataset covering accession, findability, contributors, and content overviews.

Archival Information Package: In the context of digital preservation, refers to the inclusion of all files related to dataset preservation, including documentation, in a single bundled format such as .TAR or .Zip, often including checksum metadata for fixity checks, such as MD5 files.

Dissemination Information Package (DIP): A package created from the Archival Information Package (AIP) to distribute digital content to users.

Provenance: In the context of digital preservation, refers to the history of a digital object including its origins (e.g. depositor name) and any transformations applied to that original object before release.

Checksum: A unique digital code with a separate reference copy applied to, and preserved with, a given digital object or bundle. Any change in bits checked periodically against the reference copy indicates corruption of the preserved original files, signaling the need to replace it with a backup.

Terms of use: Generally a set of metadata fields covering user's obligations when downloading files, including licensing, such as Creative Commons, citation requirements, confidentiality declarations, and other conditions.

Document Step

Document: your curation activities (Curator Log, correspondence)

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)

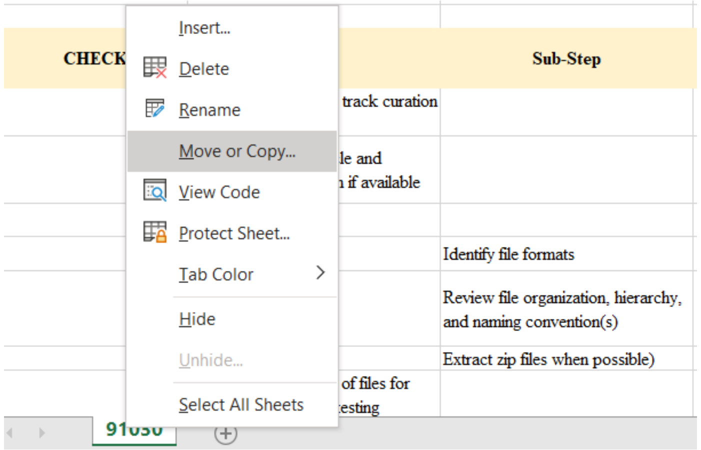
Below is a table to help guide you through the "Document" process and provide NTL-specific considerations.

Documenting CURATE(D):	**NTL-Specific:**	**What to Document:**
Complete Cataloging Process in Workroom	You can see information on cataloging datasets in Workroom on the "Cataloging Datasets" in this LibGuide(including specifics for both internal and external datasets). Also, you can find the items that should be included in both the AIP and DIP. **Reminder:** your CURATE Log does NOT belong in the DIP. Only the DIP should be ingested in the Workroom record. The SIP and AIP should not, as there is a not-Zero risk that the files could accidently be made public through human error or Workroom software glitch. Additional information on cataloging in Workroom is in the LibGuide "NTL Procedures for Submissions and Workroom."	Data Cataloging is completed in Workroom
Ensure the following information is captured in the CURATE Log: • Activities taken during the CURATE process. • Accessioning & deposit records (Names, dates, contact information, submission agreements, etc.). • Repository collection metadata.	Most of these items should have been completed within the CURATE Log as you complete each CURATE(D) Step, but please review and ensure all documentation is complete	Confirm you reviewed the dataset's full CURATE Log and that everything has been documented descriptively.

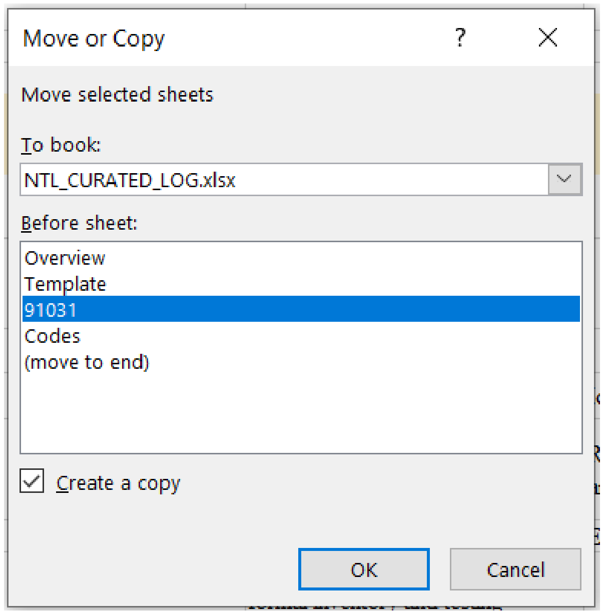
• Provenance logs (change by curators in the Transform step). • Service workflow. • Correspondences and other interactions. • Preservation packaging. • Any additional requirements at your institution.		

Upon Completion of the CURATE(D) Process, you need to make a copy of the dataset's completed CURATE log into the NTL_CURATED_LOG.xlsx.

1. Open both the NTL_CURATED_LOG.xlsx and the dataset's CURATE Log (i.e. WorkroomID_CURATE_LOG_StartDate.xlsx)
2. Make sure the sheet name in the dataset's CURATE Log is not Template, (which is what it was originally when you made the initial copy), but the Workroom ID.
    3. Right Click on the sheet name at the bottom and select "Move or Copy"



4. After selecting "Move or Copy" a pop-up window will appear. To ensure a successful transfer you will need to:
   - Under "To book" select NTL_CURATED_LOG.xlsx.
   - Under "Before sheet" select the sheet that comes after Template, to maintain the order of the latest CURATE(D) dataset being first within the document.
   - Make sure the "Create a copy" box is checked.



Move selected sheets

To book:
NTL_CURATED_LOG.xlsx

Before sheet:
Overview
Template
91031
Codes
(move to end)

☑ Create a copy

5. Click "OK" and the NTL_CURATED_LOG.xlsx document should show up with the dataset's CURATE log being copied into the correct location.
6. Save and close both files.

You have successfully CURATED a dataset! Take a well-deserved break.

Cataloging Datasets

General Cataloging Tips

Generally, the Data Curation team will follow the cataloging steps laid out in the Workroom Quick Guide: https://transportation.libguides.com/workroom

Preparing AIP and DIP Prior to Cataloging

Archival Information Package (AIP): A package that contains data that will be stored within a digital archive.

Contains:

- Complete data files (may contain copies of both open access and propriety file formats) (.ZIP).
- Additional supporting documentation (as applicable): Codebook, Data Dictionary, Code, etc.
- README file (.txt)
- Metadata file (.json)
- DMP (.pdf)
- CURATE Log (.xlsx)

Dissemination Information Package (DIP): A package created from the Archival Information Package (AIP) to distribute digital content to users.

Contains:

- Complete data files (ideally all open access file formats) (.ZIP).
- Additional supporting documentation (as applicable): Codebook, data dictionary, code, etc.
- README file (.txt)
- Metadata file (.json)
- DMP (.pdf)

General Cataloging Datasets Workflow

Cataloging is completed as part of the "D" Step in CURATE(D).

It’s helpful to open the record of the final report which the dataset supports in Workroom to use as a template.

- Enter the title of the dataset in the new Workroom record:
    - Format: Title of Final Report [supporting dataset(s)]
    - Example: Black Cat Napping Impacts [supporting datasets]
- **Data Entry Tab 1**: Use the final report as a template for all fields.
- **Data Entry Tab 2**: Use the final report as a template for **Corp. Publisher(s)**, and **Geographical Coverage**
    - **Publication Date**: If a date cannot be identified from the file names or the metadata file, use the final report publication date.
    - **Alternate Title(s)**: Report Number [supporting dataset]
        - Example: DC33 [supporting dataset]
    - **Format**: ZIP
        - Format(s) of the datasets and related files, not the format of the final report. Workroom only allows one format. For now, the best format will usually be ZIP.
    - **Resource Type**: Dataset

    - **Public Note**: There at 2 versions of the public note that can be used depending on the level of curation work taken during this workflow. These levels of curation are taken directly from CoreTrustSeal's guidance for levels of curation. The public note also describes whether the dataset is within USDOT control and when it was last successfully accessed by the cataloger. This information is very important for external datasets of which we have little control over.
        - For external datasets where minimal curation was done, only an evaluation and the creation of a DCAT-US metadata file, this would be curation level "C." Use the following text:
            - "National Transportation Library (NTL) Curation Note: As this dataset is preserved in a repository outside U.S. DOT control, as allowed by the U.S. DOT’s Public Access Plan (https://doi.org/10.21949/1503647) Section 7.4.2 Data, the NTL staff has performed NO additional curation actions on this dataset. This dataset has been curated to CoreTrustSeal's curation level "C. Initial Curation." To find out more information on CoreTrustSeal's curation levels, please consult their "Curation & Preservation Levels" CoreTrustSeal Discussion Paper" (https://doi.org/10.5281/zenodo.8083359). NTL staff last accessed this dataset at its repository URL on **[YYYY-MM-DD]**. If, in the future, you have trouble accessing this dataset at the host repository, please email NTLDataCurator@dot.gov describing your problem. NTL staff will do its best to assist you at that time.

        - For internal datasets, the curation process is much more extensive, requiring transformation, editing and creating support documentation, and ensuring the highest level of accessibility possible within our power. For this level of curation, this is known as a level "B." Use the following text:
            - "National Transportation Library (NTL) Curation Note: This dataset has been curated by the NTL Data Services Team. This dataset has been curated to CoreTrustSeal's curation level "B. Logical-Technical Curation." To find out more information on CoreTrustSeal's curation levels, please consult their "Curation & Preservation Levels" CoreTrustSeal Discussion Paper" (https://doi.org/10.5281/zenodo.8083359). NTL staff last accessed this dataset on **[YYYY-MM-DD]**. If, in the future, you have trouble accessing this dataset, please email NTLDataCurator@dot.gov describing your problem. NTL staff will do its best to assist you at that time.

    - **Abstract**: Add the abstract from the final report, followed by a paragraph about the data and all documentation files within the zip file, including: Final Report Title, ROSA P link, file formats within the zip, zip size in MB, how to unzip the files, how to open the file formats
        - Example of zip paragraph: The supporting zip file contains case study datasets, a README.txt with a data dictionary, and a .json metadata file in Project Open Data format, for Black Cat Napping Impacts, https://rosap.ntl.bts.gov/view/dot/4039. The data files are in comma-separated value (.csv) format. The compressed zip file is 84.15 MB. These files can be unzipped using any zip decompression software. The .csv files can be read with any basic text editor. The README.txt file is a plain-text file and can be opened with any basic text editor. The metadata file is in .json format and can be opened with any basic text editor, but is better viewed in a metadata editor or more advanced text editor, such as Notepad++. PDF files can be opened by web browsers or with PDF readers.
- **Data Entry Tab 3**: Use the final report as a template for all fields.
    - If a DOI was provided for the dataset it will be different from the final report, do not include the DOI, if there is one, that is listed in the metadata record for the report.
    - In Staff Notes Include: Related to Workroom [Workroom number for the report].”
    - Change Status to “InProcess”
- **Go back to Data Entry Tab 1:**
    - Click **Ingestion**
    - Using the **Choose File** Button you are going to add the below files
        - README file .pdf
        - Metadata file .json

        - Data Files .zip
    - Once all the files are added, make sure only the README file .pdf is listed first, to ensure it is the display document that will show in the viewer within ROSA P.
    - When the files have been added and ordered correct, click **Finish**.
- Hit Modify Again and go back to **Data Entry Tab 3**
    - Change status to “Complete”
    - Add “Completed XXXX-XX-XX Initials” to Status Comment
        - Ex: Completed 2022-03-23 JL
- Click **Finish**.

Cataloging External Datasets

Cataloging is completed as part of the "D" Step in CURATE(D).

NTL is not often involved in the data management of USDOT-funded external research data. As a result, our CURATE(D) Steps may not be as complete as with BTS or DOT-created research data.

Cataloging the External Dataset:

It’s helpful to open the record of the final report which the dataset supports in Workroom to use as a template, the workroom record for the report was provided in the initial email from the cataloging team.

- Enter the title of the dataset in the new Workroom record:
    - Format: Title of Final Report [supporting dataset(s)]
    - Example: Black Cat Napping Impacts [supporting datasets]
- **Data Entry Tab 1**: Use the final report as a template for all fields.
- **Data Entry Tab 2**: Use the final report as a template for **Corp. Publisher(s)**, and **Geographical Coverage**
    - **Publication Date**: If a date was not provided within the dataset’s record on the repository’s site, use the final report publication date.
    - **Alternate Title(s)**: If the title of the dataset on the repository’s site differs from the report title include it here.
    - **Format**: ZIP
        - Format(s) of the datasets and related files, not the format of the final report. Workroom only allows one format. For now, the best format will usually be ZIP.

- **Resource Type**: Dataset
- **Public Note:** Add the below curation note in the public note field and ensure you update the [bracketed] information in the below note to reflect the dataset you are cataloging.
    - **Outside Repository Public Note:** National Transportation Library (NTL) Curation Note: As this dataset is preserved in a repository outside U.S. DOT control, as allowed by the U.S. DOT's Public Access Plan (https://doi.org/10.21949/1503647) Section 7.4.2 Data, the NTL staff has performed NO additional curation actions on this dataset. This dataset has been curated to CoreTrustSeal's curation level "C. Initial Curation." To find out more information on CoreTrustSeal's curation levels, please consult their "Curation & Preservation Levels" CoreTrustSeal Discussion Paper" (https://doi.org/10.5281/zenodo.8083359). NTL staff last accessed this dataset on **[YYYY-MM-DD]**. If, in the future, you have trouble accessing this dataset at the host repository, please email NTLDataCurator@dot.gov describing your problem. NTL staff will do its best to assist you at that time.
    - **As Is Public Note:** National Transportation Library (NTL) Curation Note: This dataset was submitted from an external researcher through a USDOT-funded grant, in accordance with U.S. DOT's Public Access Plan (https://doi.org/10.21949/1503647) Section 7.4.2 Data, the NTL staff has performed NO additional curation actions on this dataset. The dataset was ingested, as is into the digital repository ROSA P. This dataset has been curated to CoreTrustSeal's curation level "C. Initial Curation." To find out more information on CoreTrustSeal's curation levels, please consult their "Curation & Preservation Levels" CoreTrustSeal Discussion Paper" (https://doi.org/10.5281/zenodo.8083359). NTL staff last accessed this dataset at on **[YYYY-MM-DD]**."
- **Abstract**: Add the abstract/description that was provided for the dataset from the outside repository's site. If one was not listed use the same abstract from the final report.
    - Additionally, write a paragraph about the files included in this dataset.
        - Example: The total size of the described zip file is 8.5MB. The ZIP file for this dataset contains files with the following files extensions: File extension .gitignore is associated with Git, a version control system developed by Linus Torvalds for various platforms that can run on local machine also as server app. These .gitignore files are text configuration files used by Git used to determine which files and directories to ignore, before user make a commit (for more information on .gitignore files and software, please visit https://www.file-extensions.org/gitignore-file-extension). File extension .md is used in creating GitHub Issues, and can be opened in a basic text editor. Files with the extension .PNG are image files that

can be opened with any basic photo viewer or editor. The .csv, Comma Separated Value, file is a simple format that is designed for a database table and supported by many applications. The .csv file is often used for moving tabular data between two different computer programs, due to its open format. Any text editor or spreadsheet program will open .csv files. PDF are used to display text and images and can be opened with any PDF reader or editor. Files with the .xlsx extension are Microsoft Excel spreadsheet files. These can be opened in Excel or open source spreadsheet programs. .py files are written in the Python programming language, and can be opened in basic text editors or in Python programming environments. Reading Python files will take special knowledge of programming. Files with the extension .bat traditionally used for batch files. A batch file is a text file that contains a sequence of commands for a computer operating system. These can be opened in an editor or in command line. This will take special knowledge or appropriate software to operate.

- **Data Entry Tab 3**: Use the final report as a template for all fields.
    - If a DOI was provided for the dataset it will be different from the final report, do not include the DOI, if there is one, that is listed in the metadata record for the report.
    - In Staff Notes Include “Related to Workroom [Workroom number for report].”
    - Change Status to “InProcess”
- **Go back to Data Entry Tab 1:**
    - Click **Ingestion**
    - Using the **Choose File** Button you are going to add the below files
        - External Metadata file .json
        - External Data
    - Once all the files are added, make sure only the External Metadata file .json is checked. With externals, we want to point users to where the data lives in its external repository while still keeping a copy of the data for ROSA P if it is ever needed.
- When the files have been added and only the 1 file selected, click **Finish**.
- Hit Modify Again and go back to **Data Entry Tab 3**
    - Change status to “Complete”
    - Add “Completed XXXX-XX-XX Initials” to Status Comment
        - Ex: Completed 2022-03-23 JL
    - Click **Finish**

References

Article 29 Data Protection Working Party. 10.04.2014. *Opinion 05/2014 on Anonymisation Techniques.* Open Data Institute. 2019. ‘Anonymisation and open data: An introduction to managing the risk of re-identification’

Association of Research Libraries. (n.d.). *Institutional Data Curation Survey Tool*. SPEC Survey on Data Curation. https://drive.google.com/open?id=0B-OrOOY8nJYjX0M4N0tCUEluV0k

Blake, Mara; Borda, Susan; Carlson, Jake; Darragh, Jennifer; Fearon, David; Hadley, Hannah; Herndon, Joel; Johnston, Lisa; Kalt, Marley; Kozlowski, Wendy; Hess, Sophia Lafferty; Moore, Jennifer; Narlock, Mikala; Scott, Dorris; Vitale, Cynthia Hudson; Wham, Briana Ezray; Wright, Sarah (2022), *Data Curation Network: CURATED Training.* https://datacurationnetwork.github.io/CURATED/

Global Indigenous Data Alliance. (n.d.). *CARE principles*. https://www.gida-global.org/care

Data Curation Network (DCN). (2018). *The DCN CURATE(D) steps*. Data Curation Network. https://datacurationnetwork.org/outputs/workflows/

Data Curation Network (DCN). (n.d.-a). *Data Curation Activities: Curation Glossary.* Data Curation Network. https://datacurationnetwork.org/data-curation-activities/

Data Curation Network (DCN). (n.d.-b). *Data Curation Tools List.* Data Curation Network. https://docs.google.com/spreadsheets/d/1X1MNEDn20pGTsuPQaWf5ZZDpQ6Gzt6TXLK0eU9iqOMU/edit?usp=sharing

Data Curation Network (DCN). (n.d.-c). *Data Curation Network Primers*. University of Minnesota’s Libraries Digital Conservancy. https://hdl.handle.net/11299/202810

Data.gov Program. (n.d.). *DCAT-US schema V1.1 (Project Open Data Metadata schema)*. A Repository of Federal Enterprise Data Resources. https://resources.data.gov/resources/dcat-us/

GO FAIR International Support & Coordination Office (GFISCO). (2022, January 21). *FAIR Principles*. GO FAIR. https://www.go-fair.org/fair-principles/

FATE. (2023, February 16). *FATE: Fairness, Accountability, Transparency, and Ethics in AI*. Microsoft Research. https://www.microsoft.com/en-us/research/theme/fate/

John Hopkins Libraries. (n.d.). *A Toolbox for Curating and Archiving Research Software for Data Management Specialists* https://dmsdata.jhu.edu/wp-content/uploads/SoftwareArchiving_v3-Storyline%20output/story_html5.html

Johnston, L. (2017). *Curating Research Data. A Handbook of Current Practice*. Association of College and Research Libraries, a division of the American Library Association.

Cornell University Library. (n.d.). *Recommended File Formats*. ECommons: Cornell's Digital Repository. http://guides.library.cornell.edu/ecommons/formats

Luk Arbuckle and Khaled El Emam, *Building an Anonymization Pipeline: Creating Safe Data* (2020) & *Anonymizing Health Data: Case Studies and Methods to Get You Started* (2013)

National Transportation Library. (n.d.-a). *NTL's Collection Development Policy*. Bureau of Transportation Statistics. https://ntl.bts.gov/ntl/policies/collection-development

National Transportation Library. (n.d.-b). *Submitting Content*. Bureau of Transportation Statistics. https://www.bts.gov/ntl/submitting-content

Statice. (n.d.). *Pseudonymization vs anonymization: Differences under the GDPR*. https://www.statice.ai/post/pseudonymization-vs-anonymization

The Source for File Extensions Information. (n.d.). https://www.file-extensions.org/

Springshare. (n.d.). *LibGuides - Content Management and Curation Platform for Libraries*. LibGuides. https://springshare.com/libguides/

United States. Department of Transportation. (2015, December 16). *Plan to increase public access to the results of federally-funded scientific research results*. ROSA P. https://rosap.ntl.bts.gov/view/dot/29637

Wikimedia Foundation. (2022, March 24). *Proprietary Format*. Wikipedia. https://en.wikipedia.org/wiki/Proprietary_format