Integrating the CURATE(D) Steps at the National Transportation Library's (NTL)

By: Jesse Ann Long and Peyton Tvrdy

In June 2023 the National Transportation Library's (NTL) Data Services Team incorporated the Data Curation Network's CURATE(D) workflow into their data cataloging process. NTL is a sub-organization that exists within the Bureau of Transportation Statistics (BTS), which is a part of the United States Department of Transportation (USDOT). Our team was initially introduced to the Data Curation Network's CURATE(D) Steps during their training events in early 2023. Seeing the potential for the CURATE(D) Steps to help NTL, we began the process of structuring the new workflow and documentation. Due to a backlog and minimal staff, the previous process was rather simple in nature, and the CURATE(D) Steps was a perfect tool to build upon our foundational workflow, close noticeable gaps, and ensure documentation throughout the whole process. The introduction of the CURATE(D) Steps additionally allows NTL to take a step forward in data management and data curation practices, thus elevating the level at which we fulfill our mandates and services, such as those outlined in the USDOT Public Access Plan (USDOT, 2015).

Background

The Data Services team is a two-person team dedicated to serving the transportation research community. We provide public access to all federally funded data. We curate datasets for both internal and external research. This includes research that is housed in our repository ROSA P, as well as datasets that are stored in an external repository such as Zenodo, Dryad, etc. We are responsible for all these datasets as they are federally funded. The U.S. DOT Public Access Plan mandates that all datasets must be shared to their fullest extent possible, accounting for Personally Identifiable Information (PII) and sensitive data.

Enter The CURATE(D) Steps

The CURATE(D) Steps are a standardized set of steps and checklists to ensure all submitted datasets receive consistent and documented treatment, with each letter in CURATE(D) standing for one of the steps: C:Check U:Understand, R:Request,

A:Augment, T:Transform, E:Evaluate, and D:Document (DCN, 2018). After deciding to implement the CURATE(D) Steps into NTL's current workflow I began to look through all related resources provided by the DCN and discovered that they had created CURATE(D) Training Modules, which I then proceeded to work through since it had been a couple of months since I completed the initial webinar on the CURATE(D) Steps. In doing so, I was introduced to additional resources that I had made use of within NTL's new CURATE(D) Workflow as well as listing them as a training resource during the onboarding of new team members (Blake, et al.). The biggest issue that needed to be addressed was how the data would be tracked through the CURATE(D) Steps so all members of the Data Services team could remain up to date on each dataset's progress. None of NTL's current systems or software were able to help with this issue. Since it can be challenging to get new software approved, it was decided that going with a simple spreadsheet to track and document the process would be best. The spreadsheet allows us to track each step of the curated process and understand a full count of how many datasets have been processed through this workflow.

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
Workroom ID	**Title**	**C**	**U**	**R**	**A**	**T**	**E**	**D**	**Comments**	**Cataloging Completed Date**	**Submitter**	**Date Submitted**	**Assigned to**	**File(s) Location**	**Rows with Headers**	**Data Rows**	**Approximate Hours of effort**	**CURATION level CoreTrustSea**
96353	Stochastic Rideshare System with Flexible Pickup and Drop-Off Points [supporting dataset]	2024-02-28	2024-02-28	2024-02-28	2024-02-28	2024-02-28	2024-02-28	2024-02-28	External	2024-02-28	Name: Alyssa McDonald Organization: National Center for Sustainable Transportation E-mail: armcdonald@ucdavis.edu	2024-02-23	Peyton	P:\NTL\data_management_curation\Data_to_CURATE\UTC\96353	69462	69458	1.5	C
96419	Utility of Capillary Blood for Gene Expression Studies [supporting dataset]	2024-03-07	2024-03-11	2024-03-11	2024-03-11	2024-03-11	2024-03-11	2024-03-11	Internal	2024-03-11	Tracy, Christopher J (FAA) <Christopher.J.Tracy@faa.gov>	2024-03-06	Peyton	P:\NTL\data_management_curation\Data_to_CURATE\FAA\96419	65279	65268	5	B
96772	Education as a Key Factor in Policy Support: An Evaluation of National Mileage Fee Support as It Varies with Information and Attitudes [supporting dataset]	2024-03-29	2024-03-29	2024-03-29	2024-03-29	2024-03-29	2024-03-29	2024-03-29	External	2024-03-29	Name: Thomas Wong Organization: National Center for Sustainable Transportation E-mail: tmjwong@ucdavis.edu	2024-03-19	Peyton	P:\NTL\data_management_curation\Data_to_CURATE\UTC\96772	2115	2114	1.5	C
96782	Analysis of the Impacts of the COVID-19 Pandemic on Vessel and Cargo Movements in the United States [supporting dataset]	2024-03-29	2024-03-29	2024-03-29	2024-03-29	2024-03-29	2024-03-29	2024-03-29	External	2024-03-29	Name: Amy Shell Organization: Center Manager, MarTREC and MBTC E-mail: martrec@uark.e	2024-03-29	Peyton	P:\NTL\data_management_curation\Data_to_CURATE\UTC\96782	40434	40433	1.5	C
96964	Vehicle to Infrastructure (V2I) and Vehicle to Vehicle (V2V) Passenger and Freight Vehicle Applications to Enhance Safety and Efficiency in Coastal Evacuations [supporting dataset]	2024-04-09	2024-04-09	2024-04-09	2024-04-09	2024-04-09	2024-04-09	2024-04-09	External	2024-04-09	Name: Amy Shell Organization: Center Manager, MarTREC and MBTC E-mail: martrec@uark.e	2024-03-29	Peyton	P:\NTL\data_management_curation\Data_to_CURATE\UTC\96964	50129	50127	1	C

[file name if you would like a better look: curated_log_overview.png]

The first sheet is an overview page that provides a snapshot of every dataset that is currently being CURATE(D) by NTL. Each row has useful information such as the submitter, who is working on the dataset, where it is in the CURATE(D) process, the location of the dataset on the shared storage drive, and the level of curation achieved according to the CoreTrustSeal's levels of curation (CoreTrustSeal Standards and

Certification Board, 2023). The second sheet is a blank template for what would be a specific CURATE Log for each dataset, including general info on the dataset, all seven steps and sub-steps, and space for a thorough documentation of each. Following the blank template will be the sheets for all the datasets that have been CURATE(D) by NTL. This design allowed us to both track each dataset and create its specific CURATE Log all in the same document. Additionally, each specific CURATE Log is downloaded and stored alongside the data in the shared drive. Lastly, we have a final sheet dedicated to our list values. These drop-down list values allow us to standardize parts of the CURATE(D) log and ensure that we are filling out the sheet in the same way. This document lives on the NTL Data Services shared Microsoft Teams channel so everyone can view and edit the document at the same time, if needed. When the current NTL CURATE(D) LOG reaches the NTL accession number of 100,000 sheets/datasets the team stores the file with its name containing the corresponding dates for datasets curated (i.e. NTL_CURATED_Log_XXXX_XX_XX_to_XXXX_XX_XX.xlsx) and create a new one.

Putting the Workflow into Practice

Implementing the CURATE(D) workflow at the NTL has greatly increased the quality of NTL's data services. Each dataset has thorough documentation including a record of actions taken, notes on the dataset's original state, useful metrics on work effort and the time it takes, and all additions we have made to data that we receive. Through the implementation of the CURATE(D) steps, we have made significant changes to the workflow that have improved our processes.

Since implementation we have separated our process into two distinct workflows depending on if the data will be housed in our repository, ROSA P, or will be housed externally on another repository, such as Zenodo, Dataverse, etc. For internally housed datasets by DOT offices, the entire CURATE(D) workflow is applied, including the creation of DCAT-US metadata files, README files, data dictionaries, codebooks, and other supporting documentation needed to aid in the interpretation of the data. For datasets that are housed externally and will only have metadata stored in ROSA P, we enact a lighter approach. For these datasets, we check and understand the submission

and reach out for more details if needed, but no transformations are made to the dataset and its documentation. If a dataset is in a proprietary format or does not include a README or data dictionary, we accept the dataset as is since it is not housed in ROSA P. In this situation, the metadata record in ROSA P points a user to the external repository where the data is stored. This scenario is common for submissions from our University Transportation Centers Program, which awards and administers grants to consortia of colleges and universities across the United States to advance state-of-the-art transportation research and develop the next generation of transportation professionals. Even these datasets that are housed externally receive a DCAT-US metadata file that is stored in ROSA P, which is a core component of compliance with federal regulation for all government data. We process these external datasets through a do have a lighter version of this workflow that includes almost all functions of the CURATE(D) workflow for each dataset we receive.

In addition to NTL's two workflows, there have been significant changes to our CURATE(D) process as more datasets go through it, resulting in new scenarios not covered and reevaluating what is best suited for serving NTL. One significant change that has been made is the addition of specific questions and workflows for the different type of data NTL receives. While the majority of data deposited to NTL is in tabular format, it isn't always the case. NTL also receives data in formats such as geodatabases, code, etc. When these projects need to be CURATE(D), it was important for the data services team to adjust the procedures and ensure the data is evaluated for completeness, accessibility, accuracy, and compliance with federal and international guidelines and standards. For example, there has been a significant increase in the depositing of code, which is externally housed on GitHub. While GitHub is not a preferred data repository, researchers have chosen it for data storage due to its popularity and excellent code versioning. When evaluating the code on GitHub we found it important to check the commits to understand the code's changes and commit messages, which has also led to NTL discovering a new form of README file. Many projects NTL has received in the past have included a README file in .TXT form; however, GitHub encourages the use of Markdown README files so that they render below the code's main branch. This has changed how we understand and evaluate

README files, and it has changed NTL's workflow from creating plain text .TXT READMEs to machine and human readable .MD files. This change might not have occurred without the CURATE(D) workflow encouraging better understanding of NTL's submitted data.

The CURATE(D) workflow has led the data services staff to improve our comfort level and ability to handlewith a variety of software, programming languages, and data types. As a result, NTL has changed the elements that are checked to understand the data, performs in-depth analysis of the code, know coding languages enough to dissect its parts, determine if the code is useable and interoperable, and transform and improve elements of the data package. With the new CURATE(D) workflow, NTL was able to adapt our procedures and the way we approach data submission to the repository. This involved setting new standards in the quality of NTL's cataloging and for our researchers and their submissions.

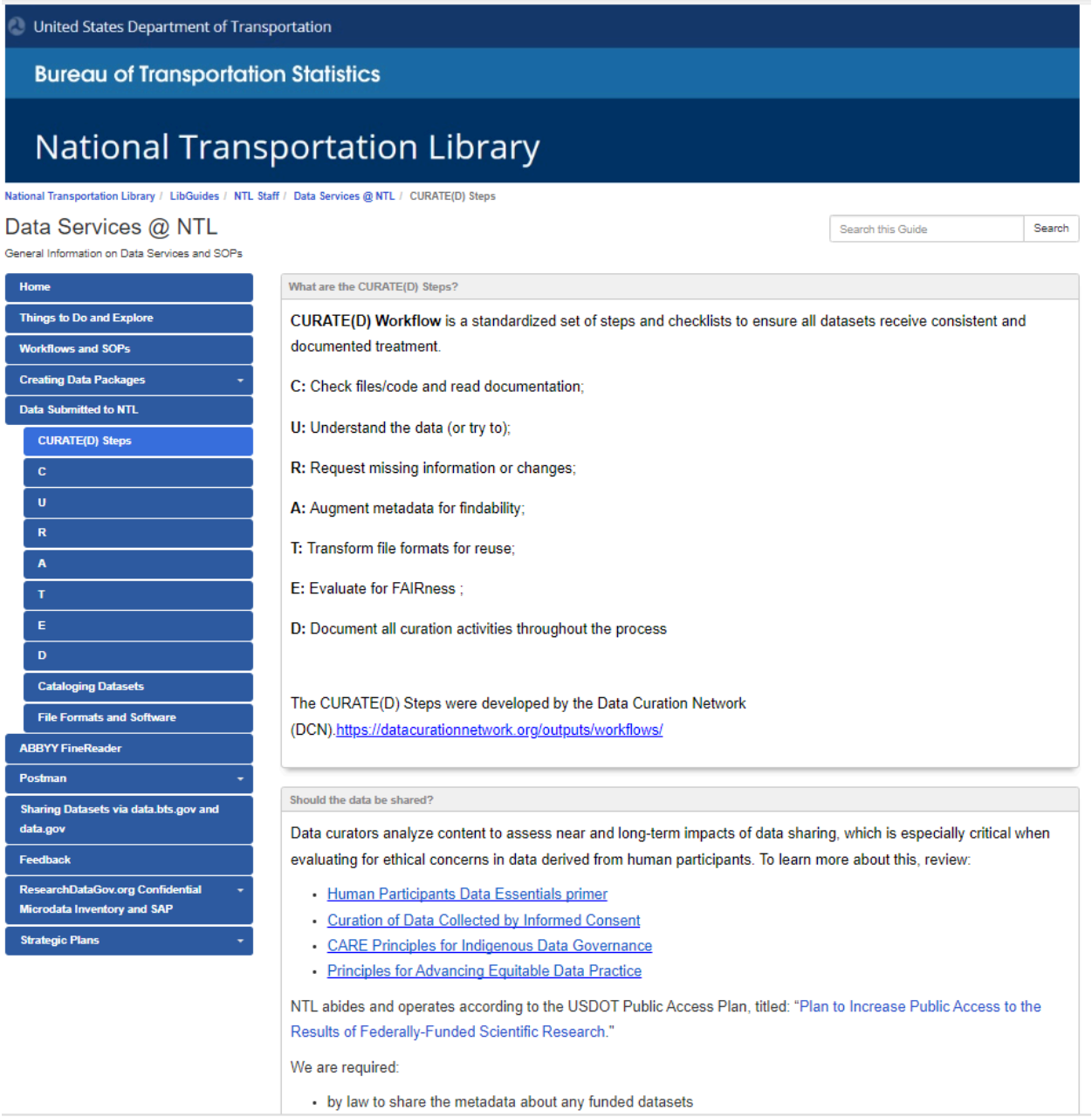
The CURATE(D) workflow has also led to better tracking and metrics of our cataloging efforts as a team. For a small team keeping up with cataloging of all DOT transportation data is no easy feat. However, the metrics we have gained on how many datasets have been CURATE(D), who has taken which actions, and all decisions that have been made when cataloging the data has been a game changer for NTL data services. Since starting this new workflow in Fall 2023, the team has CURATE(D) and cataloged almost 100 datasets for ROSA P. Through the CURATE(D) log, we have metrics, such as externally housed datasets on average take approximately 1.5 hours to complete and internal datasets take roughly 5.25 hours to undergo the entire workflow. These metrics are important not only to measure our accomplishments but also help us better budget time and understand our workload, which is crucial for such a small team.

Transparent Documentation and Description of Workflow

Integrating the CURATE(D) workflow into NTL's data cataloging process was a little daunting due to the number of changes that were needed. NTL's previous workflow was contained to a portion of only one of the seven CURATE(D) steps, so significant

development was needed to ensure that the CURATE(D) Steps were manageable, applied to NTL's policy and standards, and were described adequately.

It was determined that the CURATE(D) workflow was to be stored in the Data Services team internal staff guide where all work-related workflows are stored. This guide (Only accessible by NTL members) was created using the LibGuides software, 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" (Springshare, n.d.). Additionally, by using LibGuides, it keeps all Data Services related workflows in a single place, increases navigation efficiency between CURATE(D) Steps, and groups information and resources that correspond with a particular step on a single page. The CURATE(D) guide has a main page that serves as an intro into the process, with eight sub-pages for each step in CURATE(D) and for the cataloging process (which is a part of step D) according to NTL's systems. Another benefit to using LibGuides to create the workflow is that there are several internal LibGuides used by NTL Staff and linked in the CURATE(D) workflow, making it easier to move between resources that are all within the same platform.

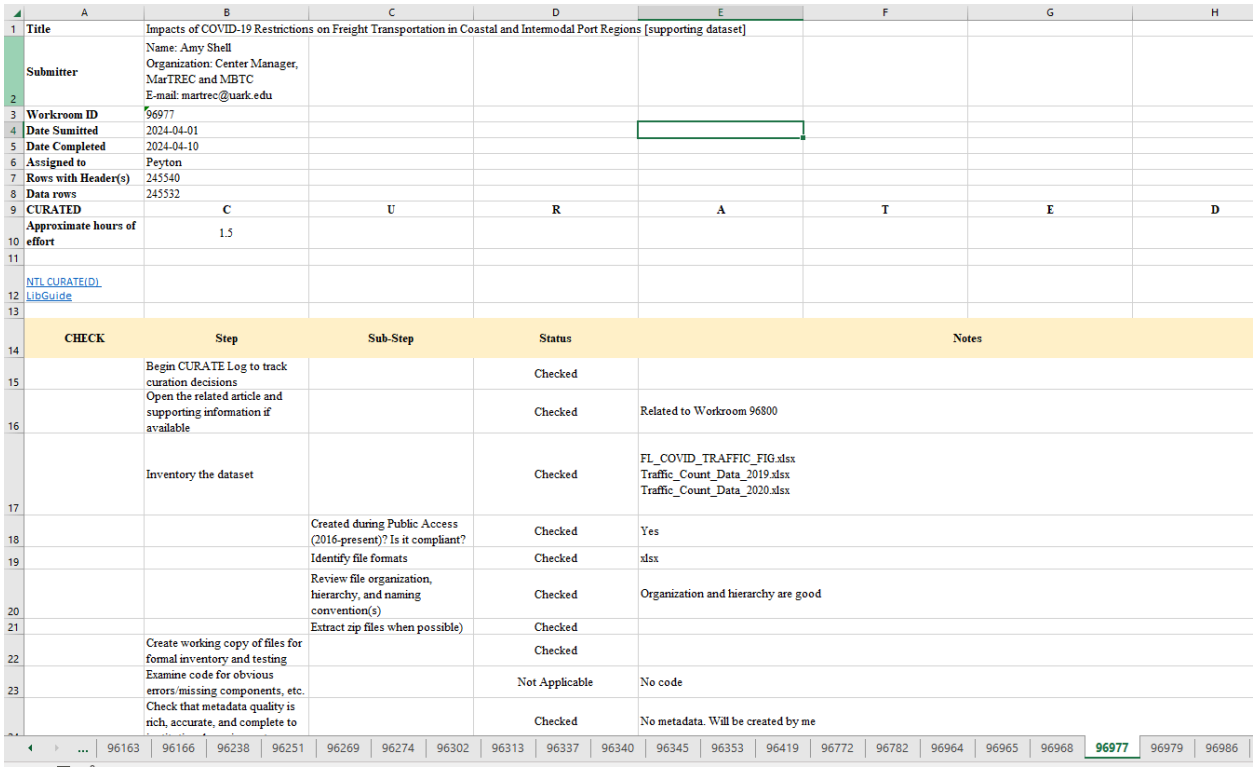


[file name if you would like a better look: libguide.png]

The workflow has been designed to provide support to staff members as they create a curation log for a single dataset (CURATE Log) and complete it according to the CURATE(D) Steps, each page has pretty much the same layout with only a couple of exceptions. This was done to create some consistency across all steps. On any given page a user will find:

- related terms;

- general info about the CURATE(D) Step in question;
- resources and/or examples;
- chart that addresses what topics are covered in the CURATE Log,
- how it related to NTL,
- what to document in the log.



	A	B	C	D	E	F	G	H
1	**Title**	Impacts of COVID-19 Restrictions on Freight Transportation in Coastal and Intermodal Port Regions [supporting dataset]						
2	**Submitter**	Name: Amy Shell Organization: Center Manager, MarTREC and MBTC E-mail: martrec@uark.edu						
3	**Workroom ID**	96977						
4	**Date Sumitted**	2024-04-01						
5	**Date Completed**	2024-04-10						
6	**Assigned to**	Peyton						
7	**Rows with Header(s)**	245540						
8	**Data rows**	245532						
9	**CURATED**	**C**	**U**	**R**	**A**	**T**	**E**	**D**
10	**Approximate hours of effort**	1.5						
11								
12	NTL CURATE(D) LibGuide							
13								
14	**CHECK**	**Step**	**Sub-Step**	**Status**	**Notes**			
15		Begin CURATE Log to track curation decisions		Checked				
16		Open the related article and supporting information if available		Checked	Related to Workroom 96800			
17		Inventory the dataset		Checked	FL_COVID_TRAFFIC_FIG.xlsx Traffic_Count_Data_2019.xlsx Traffic_Count_Data_2020.xlsx			
18			Created during Public Access (2016-present)? Is it compliant?	Checked	Yes			
19			Identify file formats	Checked	xlsx			
20			Review file organization, hierarchy, and naming convention(s)	Checked	Organization and hierarchy are good			
21			Extract zip files when possible)	Checked				
22		Create working copy of files for formal inventory and testing		Checked				
23		Examine code for obvious errors/missing components, etc.		Not Applicable	No code			
		Check that metadata quality is rich, accurate, and complete to		Checked	No metadata. Will be created by me			

... | 96163 | 96166 | 96238 | 96251 | 96269 | 96274 | 96302 | 96313 | 96337 | 96340 | 96345 | 96353 | 96419 | 96772 | 96782 | 96964 | 96965 | 96968 | **96977** | 96979 | 96986

[file name if you would like a better look: curated_log_single_record.png]

The resources on our LibGuide came from DCN CURATE(D) references as well as materials we encountered in the University of Vienna's Data Steward Course. Both of us completed the University of Vienna's Data Steward Course. The workflow was designed with the goal of provide as many resources as possible to help inform the CURATE(D) process. Many of the steps, although not unfamiliar, have not previously been completed by NTL staff and the desire was to create a workflow that sets staff members up for success.

In addition to the adding of vast amounts of information, resources, and structure to NTL's work with datasets, CURATE(D) also led to the reevaluation of standards that were already established. Previously when working with data packages, NTL focused

on singular .zip submissions. However, the CURATE(D) steps address three different types of data packages, the Submission Information Package (SIP), the Archival Information Package (AIP), and the Dissemination Information Package (DIP). Although not a major change in NTL's work process, it is an important step to ensure we are curating each dataset thoroughly and have all the proper documentation and files both for public access and long-term preservation. Similar to the way data is treated, this workflow is designed to be explicit and descriptive as possible so that everything is transparent and well-documented.

We feel that we will see a greater impact of CURATE(D) the more it is used at NTL. We have already identified some key benefits, such as long-term knowledge management, documentation of actions, creation of complete packages for archiving and distribution with robust documentation, and the through tracking of all datasets submitted to NTL. Furthermore, the documentation collected through CURATE(D) can help show and tell success stories and provide evidence that reveals the limitation of what our current systems can manage when curating data. This will hopefully help us get additional resources that are needed to fully support all the data submitted to the repository.

Conclusion

Completing the CURATE(D) workflow for every data submission we receive is no easy task. It takes significant dedication and upkeep every week to manage the workload in addition to our many other responsibilities. Yet, with over 150 hours of total CURATE(D) work under the belts, we have learned to better analyze, improve, and understand the data that we work with and the challenges and needs of our researchers. CURATE(D) has help bridge the gap between our data stewardship efforts upon receiving a dataset and improving the practice of open and FAIR science for DOT's own research. All the lessons we have learned and the knowledge we have gained will go to improving the data of all USDOT funded research. We have made our own workflow (https://doi.org/10.21949/1530073) available on ROSA P. Although specific to NTL, the workflow could easily serve as a starting point for other governmental agencies or institutions to apply to their workflows. Hopefully our

implementation can demonstrate the benefits of this workflow and help others apply it to their work. The CURATE(D) workflow has not only helped NTL to improve data curation and management best practices but expanded our approach to open science, and we hope our experience can help others to do the same.

References

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/

CoreTrustSeal Standards and Certification Board. (2023). Curation & Preservation Levels: CoreTrustSeal Discussion Paper (v02.00). Zenodo. https://doi.org/10.5281/zenodo.8083359

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

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

United States Department of Transportation (USDOT). (2015, December 16). *Plan to Increase Public Access to the Results of Federally-Funded Scientific Research Results.* Welcome to ROSA P. https://rosap.ntl.bts.gov/view/dot/29637