National Transportation Library Digital Curation Policy for the Repository and Open Science Access Portal (ROSA P)

Version 2.0 October 2024

Background and Scope

The National Transportation Library (NTL) is a digital library that provides national and international access to transportation information, coordinates information creation and dissemination, and provides reference services for DOT employees and public stakeholders. Established in 1998 by the Transportation Equity Act for the 21st Century (TEA-21; P.L. 105-178) (https://doi.org/10.21949/1522467), NTL's authorized role was expanded in 2012's Moving Ahead for Progress in the 21st Century (MAP-21; P.L. 112- 141) (https://doi.org/10.21949/1522466). In these legislative documents one key role for NTL is the creation of a repository to provide open access to federally funded transportation research, which has resulted in the creation of the the Repository & Open Science Access Portal (ROSA P). As a result, the staff of NTL and ROSA P are synonymous and are designed to support the continued development of ROSA P. NTL has no other library collection apart from the digital items located in ROSA P.

Specifically, MAP-21 mandates that NTL:

- acquire, preserve, and manage transportation information and information products and services for use by DOT, other Federal agencies, and the public;
- provide reference and research services;
- serve as a central digital repository for DOT research results and technical publications
- become a central clearinghouse for transportation data and information of the Federal Government;
- serve as coordinator and policy lead for transportation information access;
- coordinate among and cooperate with multiple external parties to develop the comprehensive set of transportation statistics on the performance and impacts of the national transportation system, including statistics on the eleven topics required in 49 U.S.C. § 6302(b)(3)(B)(vi); and,
- publicize, facilitate and promote access to information products and services.

NTL's primary products and services are the Repository & Open Science Access Portal (ROSA P; https://rosap.ntl.bts.gov/); Ask-a-Librarian virtual reference desk and knowledge base (https://transportation.libanswers.com/); and coordination of the National Transportation Knowledge Network (NTKN) (https://transportation.libguides.com/ntkn).

Since 2016, NTL has led the implementation of the Official DOT Public Access Plan (https://doi.org/10.21949/1503647) issued in response to the February 22, 2013, Office of Science and Technology Policy (OSTP) Memorandum for the Heads of Executive Departments and Agencies entitled "Increasing Access to the Results of Federally Funded Scientific Research" (https://doi.org/10.21949/1528360). NTL will be updating its practices and workflows to reflect the August 25, 2022, OSTP memo "Ensuring Free, Immediate, and Equitable Access to Federally

Funded Research" (https://doi.org/10.21949/1528361) as soon as the new version of the USDOT's Public Access Plan is published.

Founded as an all-digital library, ROSA P's collections include full-text digital publications, datasets, and other resources. Legacy print materials that have been digitized are collected if they have historic, technical, or national significance. All items acquired by NTL are ingested and preserved in ROSA P. Collections in ROSA P are available without restriction to transportation researchers, statistical organizations, the media, and the general public. All research funded by the Department of Transportation is required to be submitted to NTL for long-term preservation.

As the national resource for US Transportation Research, the National Transportation Library ensures equitable access of transportation research and data to all through robust curation practices and extensive preservation work. We as an organization have become a leader in the field of transportation research, active contributors to new transportation research data management practices, and strong advocates for the future of open science. We connect people and organizations with transportation research, making us a go-to source for public and private sector individuals who seek the newest innovations in transportation.

Technology and Integrations

The National Transportation Library utilizes a custom, in-house built cataloging system, branded Workroom. Workroom has been in operation since 2000. NTL uses an AWS Relational Database Service to manage its cataloging environment. The storage environment consists of 3 sub-environments: production, disaster recovery, and staging.

1. Beginning in 2017, NTL contracted with the Centers for Disease Control and Prevention (CDC) Library for the CDC-developed repository software STACKS (https://stacks.cdc.gov/). The NTL STACKS implementation is branded as the Repository & Open Science Access Portal (ROSA P).
2. CDC supplies the technical infrastructure needed for NTL to share datasets and textual outputs.
    A. The contract between NTL and CDC is renewed regularly.
    B. CDC uses checksums to ensure file integrity and prevent duplication of records. Using checksums allows the CDC to monitor and prevent data degradation and bit rot for each file.
    C. SWAT only exists in the staging environment, and it allows curators to interface with the content and make changes without interfering with the production environment/content.
    D. The DOT software infrastructure consists of a combination of open source and custom software developed and maintained by the CDC (Centers for Disease Control and Prevention). Multiple different open-source software stacks are integrated and combined with custom code to produce the final environment. The following open-source software is used:
        - Apache Web Server
        - Apache Tomcat
        - Apache SOLR
        - MySQL

- Fedora Commons
- Image Magic
- Samvera (Hyku)

Curation Process

NTL's goal is to preserve all digital information at the bit level, at a minimum. This means that the NTL will protect digital information from bit rot and media failure, ensuring future devices will be able to faithfully reproduce the sequence of bits encoded in a digital information object. To achieve this goal, NTL employs the following practices.

All items submitted to NTL (datasets, text-based, images, etc.) are assessed by NTL staff and checked for collection development scope, metadata, 508 compliance, and other criteria. No content is "distributed as submitted." NTL systems are all staff-mediated and direct deposit into the repository is not allowed.

Data Curation

For datasets, code, websites, GIS Geodatabases, software, and other digital data products. Data Services preforms robust data curation, this includes evaluating the dataset and determining the appropriate curation level. Each dataset goes through the same curation workflow to ensure data quality and transparency.

Levels of Curation

NTL Performs curation at the following levels. These curation levels are taken from the "Curation and Preservation Levels: CoreTrustSeal Position Paper" document version 3 (https://doi.org/10.5281/zenodo.11476980).

- D. Deposit Compliance - Data content and supporting metadata deposited are checked for compliance with defined criteria, e.g. data formats, metadata elements, and compliance with legal and ethical norms. Digital objects that do not meet these criteria may be rejected or moved forward to initial curation. ROSA P staff does check submission to ensure data was submitted and complies with repository scope found in the collection development policy.

- C. Initial Curation - The digital objects are curated by the repository to meet defined criteria, which may exceed those defined for Deposit Compliance. This initial curation for access and use may include, e.g., the correction or enhancement of metadata and/or data content, or the creation of dissemination formats. ROSA P staff's action include the creation of DCAT-US (https://resources.data.gov/resources/dcat-us/) metadata files, authoring of a robust public note (including software requirements and decencies), and an evaluation of the dataset's FAIRness using NTL's adaptation of the DCN's CURATE(D) Steps (https://doi.org/10.21949/1530073) for ROSA P.

- A. Active Preservation - In addition to D and/or C above, the repository takes long-term responsibility for ensuring that the data and metadata can be understood and rendered as required by the designated community for reuse. The preservation actions can be aimed at logical-technical, semantic, or quality aspects of the (meta)data, for example in response to the threat of technological obsolescence, to accommodate changing needs of the Designated Community, or in response to other considerations such as security or legal concerns.

> Logical-technical measures include updating hard and software environments, archival and dissemination formats of digital objects and metadata. Semantic measures include updating the content of metadata elements and other semantic artefacts such as controlled vocabularies and ontologies if necessary. It may include responsibility for editing the structure and content of deposited data. ROSA P staff converts data to open formats if necessary, while still preserving the original file format. Additionally, staff will write documentation including READMEs, data dictionaries, and codebooks. In the special cases staff will use OCR software (ABBYY FineReader) to transform, extract, and clean data when taking data from PDF to tabular formats for preservation and reuse. Staff will always check data submission for Personally Identifiable Information (PII), deidentify if necessary, and mint a PID for the data if it does not already have one.

Data Curation Workflow

Summer 2023, NTL Data Services staff adapted the Data Curation Network's (DCN) "CURATE(D)" workflow (https://datacurationnetwork.org/outputs/workflows/). The CURATE(D) steps track a dataset from deposit to ingestion in ROSA P. The use of CURATE(D) for a dataset is documented throughout the process and results in a complete CURATE log for each individual dataset so any changes to data, metadata, documentation, or other information are stored and preserved. In addition to each dataset having its own CURATE log, NTL tracks all datasets through a centralized log entitled "CURATED Log". The centralized log is a living document that will track all the datasets that either have been through the NTL CURATE(D) workflow or are in process. Curation level is reported to the public through the public note metadata field. The CURATE(D) workflow ensures that all datasets, code, geodatabases, and other supporting data products are all handled uniformly and are as accessible and reusable as possible. A published version of NTL's CURATE(D) Workflow can be found at (https://doi.org/10.21949/1530073).

Metadata

Digital resources in ROSA P are fully described to fostering discovery, access, and re-use. Creating quality metadata is an essential part of cataloging and records' quality assurance. All metadata is created using strict and clear procedures by cataloging staff. In addition to our cataloging procedures, NTL follows the global guidelines for FAIR Principles to make all items their metadata findable, accessible, interoperable, and reusable (https://www.go-fair.org/fair-principles/).

Dublin Core Metadata

All records in ROSA P are cataloged using the Dublin Core Metadata Initiative (DCMI) schema (https://www.dublincore.org/specifications/dublin-core/dcmi-terms/). This schema is used for every record published in ROSA P. While the majority of our terms are derived from this schema, NTL uses terms unique to our organization for many of our records. Some of these unique terms include: Report Number, Contract or Cooperative Agreement Number, Transportation Research Thesaurus Subject Terms, ResearchHub Display ID, and more. The unique terms allow our cataloging term to capture information that is specific to USDOT practices and funding.

Dataset Metadata

By Federal guidance (https://resources.data.gov/resources/dcat-us/) and by the U.S. DOT Public Access Plan (https://doi.org/10.21949/1503647), all datasets should be accompanied by a DCAT-US Schema version 1.1 .json metadata file to help ensure long-term preservation and discovery.

NTL staff may supply the submitter with a DCAT-US template, a link to create it using NTL's DCAT-US Version 1.1 JSON Generator (https://transportation.libguides.com/researchdatamanagement/datapackages), or may create the DCAT-US metadata during the curation process.

Persistent Identifiers

USDOT's "Plan to Increase Public Access to the Results of Federally-Funded Scientific Research Results" (https://doi.org/10.21949/1503646) requires researchers to have ORCIDs and assign a persistent object identifier for all publications and datasets. Additionally,
NTL assigns a digital object identifier (https://www.doi.org/), or DOI, as a persistent identifier to all information and data resource landing pages in *ROSA P*. This globally unique link will always lead users to a landing page containing the resource and its metadata, or a description documenting the alteration and/or destruction of the information source, if applicable.

By incorporating DOIs and ORCID iDs, NTL is continuing its mission to make data available to the public and help increase data sharing in the community.

DOIs

- Beginning in 2016, NTL joined the federal Digital Object Identifier (DOI) consortium managed by the United States Department of Energy (DOE) Office of Science and Technical Information (OSTI) (https://www.osti.gov/).
- In 2024, OSTI announced the discontinuation of their Persistent Identifier Interagency DOI Service API and encouraged the transition of its users to the DataCite API and platform directly. This change allows for more robust metadata, complex automation, and the use of the DataCite metadata schema. The DataCite metadata schema is a more widely used schema for DOIs, bringing NTL into a more widely used standard.
- The contract between NTL and OSTI is renewed annually.

ORCiD IDs

- Beginning in 2023, NTL joined the federal ORCID consortium (https://www.osti.gov/pids/orcid-services/us-gov-orcid-consortium) managed by the United States Department of Energy (DOE) Office of Science and Technical Information (OSTI) (https://www.osti.gov/).
- The consortium provides NTL reduced-cost access to ORCID APIs and systems integration.
- NTL currently captures ORCIDs and is beginning the process of reporting ORCID through the STACKS/ROSA P system.
- NTL plans for software integration between Workroom and ORCID during 2025.

ROR IDs

- In 2024, NTL began to incorporate ROR IDs into their data curation workflows. ROR IDs are currently listed and used when listing researchers and their affiliated organizations in dataset README documents. They are also used for authoring and funding organizations in these same READMEs. While NTL has not integrated ROR IDs into the cataloging process and does not have the ability to display that information through CDC STACKS software, NTL does advocate for their use by researchers when possible.

- While a paid membership is currently not a feature of the ROR community, NTL staff participates in ROR community events and webinars.
- NTL is advocating and planning integration of ROR with ROSA P for organizations and affiliations.
- For DOI metadata requests using the DataCite schema and API, NTL staff captures organizations' ROR information

Preservation and Backups

To protect digital information and data from loss, NTL employs the "3-2-1" backup rule[2].

NTL maintains

- three (3) copies of the electronic files;
- stored on two (2) different kinds of storage media;
- with at least one (1) copy stored in a different geographic and geologic region.

Currently, NTL maintains a copy of its repository content and metadata in the following locations:

- USDOT- managed AWS cloud environment
- Backups on the USDOT-managed AWS cloud environment are in the disaster recovery site, in a different geographical area than USDOT headquarters.
- CDC Public Access Platform (Amazon Web Services cloud environment)
- Removable media (external drive)
    - The external drive housing *ROSA P* content is housed at USDOT headquarters in Washington, DC.
    - This drive is updated quarterly.

The AWS, Amazon Web Services, environment consists of three sub-environments, each acting as a separate archive. There's a production environment, where all public users can reach published content. A disaster recovery environment, where a slower but functional archive is ready in case of disaster. And finally, a staging environment, where all content can be managed by curators and stage content prior to production. S3 is AWS managed and backed up regularly. Other Backups are taken using AWS backup. In using cloud storage, we are already mitigating risks associated with standard disk storage by backing up data regularly in case of failure.

Backups on the CDC Public Access Platform are in the disaster recovery (DR) site on the US West Coast, a different geographic area than CDC headquarters. The DR is updated daily. All daily backups of the staging server and weekly backups of the production servers are kept for 45 days.

Note on other digital versions: Because of the ease of copying electronic files, and the desire of NTL to share information and data freely, information objects housed in *ROSA P* may also be represented in other digital repositories. The NTL staff cannot guarantee the authenticity or provenance of files accessed from repositories outside our control. Whenever a question of authenticity or provenance arises, users should refer to the information item's landing page in *ROSA P* for the authentic version of an item. Further, the NTL staff will not be able to inspect electronic copies of information and data in other repositories for bit rot or corruption. However, if corrupted files are found in other repositories, NTL will work with these repositories to replace corrupted files with authentic copies of the original files.

Format Migration

Per NTL's Collection Development Policy (https://doi.org/10.21949/1530598), format preferences for content submitted to NTL are non-proprietary and open electronic file formats, such as .txt, .csv, .pdf, .tiff, as well as others as described by the Library of Congress in *Sustainability of Digital Formats* (https://www.loc.gov/preservation/digital/formats/index.shtml).
When content is migrated from one format to another, NTL will:

- Record the event in metadata
- Provide a description on the landing page
- Keep one (1) copy in the original format
- Maintain access to all versions.
- Curatorial activities include migrating data from one format into another when earlier formats or devices become obsolete, and as NTL resources permit.

Content Alteration or Removal

Alteration or removal of resources, including publications and datasets, may be required if they contain data that is not publicly accessible. For example, material may be under copyright, may contain confidential information, or may compromise privacy or national security information. (See section 6 of the Selection Statement in the NTL Collection Development and Maintenance Policy for further detail: https://doi.org/10.21949/1530598) Any such changes made to a resource will be noted on the landing page or record in the database, which will remain accessible even if the data is no longer available.

Trusted Digital Repository Status

To further its commitment to preservation and long-term access of items in its collection, NTL supports standards for repository trustworthiness. Assessment for trustworthiness is based on the International Standards Organization (ISO) *Reference Model for an Open Archival Information System (OAIS)* and ISO Standard 16363:2012 *Audit and Certification of Trustworthy Digital Repositories.* NTL is currently undergoing the application for a Trustworthy Digital Repository identified by CoreTrustSeal (https://www.coretrustseal.org/why- certification/requirements/). The results of this will be updated in this policy when the process is complete.

Review Cycle

This policy is subject to a five-year review cycle. The policy may be reviewed, altered, and reissued as needed to meet changing needs and practices of the community, federal regulations, and global and national standards. Policy changes will be noted in the "Digital Curation Policy Update Log" section at the end of this document and will be designated by incrementing the version number and updating the policy date.

Digital Curation Policy Update Log
Version 1.0 published May 2017
Version 1.1 updated October 2017
Version 1.2, updated December 2017
Version 1.3, updated January 2018
Version 2.0, updated October 2024