U.S. Department of Transportation
Office of the Secretary of Transportation

Bureau of Transportation Statistics
National Transportation Library

Data Curation Practices to Optimize Research Coordination, Preservation, and Reuse

Transportation Research Board 103rd Annual Meeting
Washington, D.C., January 7-11, 2024
Poster: P24-20517
For AJE35 RIIM: Solicited Research Management and Innovation

Abstract

Transportation research organizations must implement a number of funder requirements around **technology transfer**, **data sharing**, and **public access** to research outputs. Further, the practice of scientific research in the digital age is fostering other changes such as the global movement to open science. Tech transfer, data sharing, public access, and **open science** all can lead to exciting and innovative collaborations in global transportation research. However, these many of these practices are new or only recent additions to transportation research program workflows. These new practices require new skills or team members be added to research programs. Program managers are right to wonder: "Which practices will have the greatest impact on furthering research coordination and accelerating tech transfer?"

Data Curation, a lifecycle approach to **research data management (RDM)**, offers many practices to improve research coordination, collaboration, and technology transfer. For example, creating **data management plans (DMP)** at the beginning of the research planning process can greatly enhance research coordination. A robust DMP serves as a living **knowledge management document** for the research team, making it much easier to weather changes in project personnel or technology. Further, a good DMP can satisfy funder requirements!

As another example, the application of **persistent identifiers (PIDs)** to research outputs, researchers, and research organizations, can help to ensure **accurate citation and attribution as** your research is implemented or built on by other research. PIDs can in turn help produce more **accurate reuse or tech transfer metrics**, which can impress funders and promotion boards. Data curation offers many more practices to fuel research innovation.

This poster will:
1. Define Data Curation and describe its research lifecycle approach;
2. Contextualize Technology Transfer within Public Access and Open Science;
3. Describe Data Curation practices transportation research programs can implement to improve research coordination and tech transfer, including tips on choosing a data repository;
4. Discuss how program directors can bring data curation skills into their research programs; and,
5. Highlight data management and curation training the National Transportation Library could provide to your research program.

1. Data Curation & the Data Lifecycle

Just as a research project has a lifecycle, so too does the data generated by that research project. This is not a new revelation. What is new is the expectation that data may have a life independent of the research project that generated it and that data may have more uses than simply as an aid to replicate research findings. A goal of the global **Open Science** movement is to allow data to be reused to generate new knowledge. When data is seen as a long-term investment or asset, this means data must be cared for in new ways, which in turns means new skills, such as those of data curation.

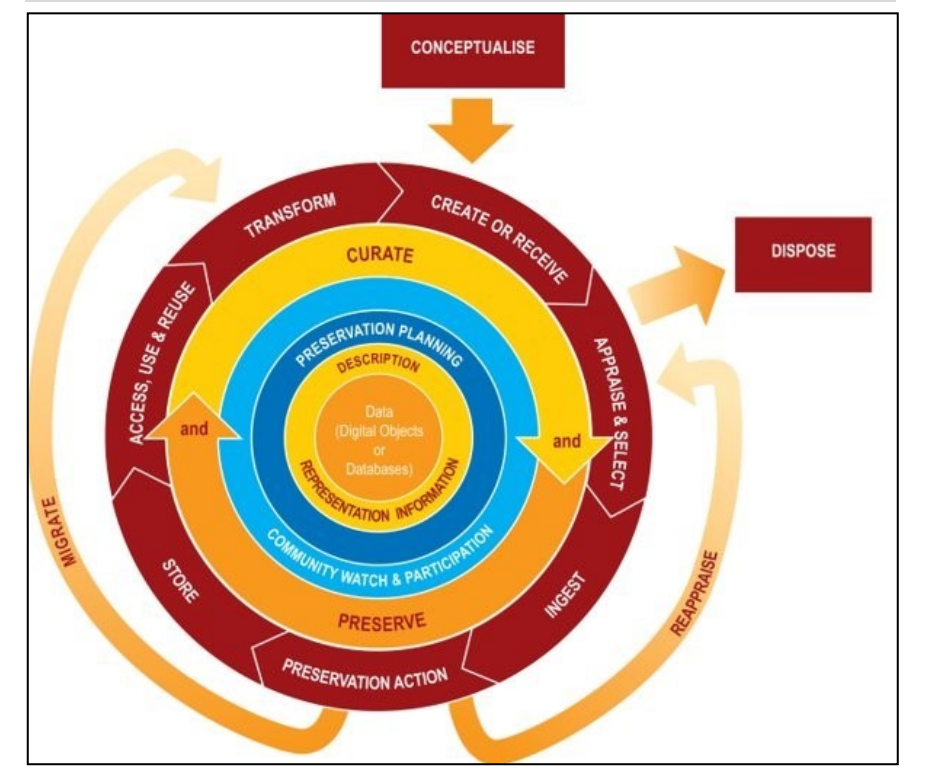
Data Curation is the **active and ongoing** management of data through its lifecycle **of interest and usefulness** to scholarship, science, and education. Data curation enables **data discovery** and **retrieval**, maintains **data quality**, adds **value**, and provides for **reuse over time** through activities including authentication, archiving, management, preservation, and representation. https://tinyurl.com/datacurationuiuc

A primary goal of **Data Curators** is the facilitation of data reuse, sometimes in novel or unforeseen research. This means that **Data Curators** enter into a holistic and ongoing engagement with a dataset, which may outlast that of even the researchers who collected the data.

The **Data Lifecycle** is all the phases of data's existence from planning to collection, through preservation, to reuse and potential destruction. http://dx.doi.org/10.3133/ofr20131265

The Data Curation Centre's **Curation Lifecycle Model** (right) shows the many points before, during, and after the existence of a dataset where a curator may interact.

The Data Curation Centre
Curation Lifecycle Model
[2023] https://www.dcc.ac.uk/guidance/curation-lifecycle-model



During the research planning and data collection phases, a data curator can be a vital team member. Curation tasks can include:
- Team data management **training**;
- Co-authoring a **data management plan (DMP)**;
- Creating **standardized** data collection **workflows**;
- Proposing and implementing **controlled vocabularies** and **metadata standards** to improve data analysis, data comparison, and data discovery & data citation;
- Data **preservation and backup planning** to prevent data loss;
- Create robust **dataset documentation** to improve understanding and data reuse efficiency; and,
- Assignment and recording of **persistent identifiers (PIDs)** which fuel attribution and citation.

A data curator's relationship with a dataset does not end after collection, analysis, and reporting of findings. A data curator may go on to:
- **Catalog and deposit** the dataset into an appropriate **repository**;
- Help set data reuse **access requirements**;
- Perform **preservation** actions;
- **Migrate** datasets into preservation-friendly formats or into new formats as older formats be obsolete; or,
- Plan and implement the secure **disposition**, or deletion, of data that is no longer of interest to science or an organization.

1A. Definitions & Linked Processes

The practices of Data Curation do not exist in a bubble. Rather they are closely linked to or enable other facets of the research and data lifecycles. These linkages will be explored and illustrated below. But first we need to define some terms, specifically **Data Management**; **Data Science**; **Data Stewardship**; and **Data Governance**. Each of these terms has a distinct meaning, and while it can be tempting to use them interchangeably, it is vital to not do so.

1. **Data Curation (DC)** is the **active and ongoing** management of data through its lifecycle **of interest and usefulness** to scholarship, science, and education. Data curation enables **data discovery** and **retrieval**, maintains **data quality**, adds **value**, and provides for **reuse over time** through activities including authentication, archiving, management, preservation, and representation.]The National Transportation Library (NTL) has adopted this definition from the University of Illinois Graduate School of Library and Information Science (2013). https://tinyurl.com/datacurationuiuc]
2. **Data Management (DM)** is the deliberate planning, creation, storage, access, and preservation of data produced from a given investigation. [NTL has adopted this definition from Texas A&M University Libraries (2016). https://tinyurl.com/datamanagementTAM]
3. **Data Science (DS)** is drawing useful conclusions from large and diverse datasets through exploration, prediction, and inference. [NTL has adopted this definition from Ani Adhikari and John DeNero (2016). https://tinyurl.com/adhikariwhatisdatascience]
4. **Data Stewardship** is making connections between researchers, policy makers, software developers, and infrastructure providers to implement the necessary elements that enable researcher to successfully implement RDM [research data management]. [NTL has adopted this definition from Hasani-Mavriqi, et al (2022). https://doi.org/10.3217/p9fvw-rke48]
5. **Data Governance** is the executive or managerial "exercise of authority, control, and shared decision-making over the management of data assets." [NTL has adapted text from the Data Management Association (DAMA) (2017) https://www.dama.org/cpages/body-of-knowledge and John Ladley (2012) https://shop.elsevier.com/books/data-governance/ladley/978-0-12-415829-0 for this definition.]

Again, each role or action above is distinct, although they are all connected, and there is some overlap. It should be kept in mind, however, that each role requires a distinct set of practical skills or organizational position and power. Therefore the roles are not interchangeable. To further blur the lines, a single person may act in one or more roles as a **Data Curator**, **Data Manager**, **Data Scientist**, **Data Steward**, or **Data Governor**. This may be based on the size or structure of a research institution, and on subject matter expertise and training. So confusion is understandable.

For the illustration of linked processes below, we will focus on Data Curation, Data Management, and Data Science.

Linked Processes: Data Management & Data Curation

When we take another look at Curation Lifecycle Model (left) we see that data collection, data description, and preservation planning inhabit the central three rings. These central rings overlap nicely with the definition of data management above. This means that data management (DM) is central to data curation (DC). We also notice that of the large set of data curation practices illustrated in the Lifecycle Model, or described in the lists, data management action make up a subset, but an essential subset.

In fact, if data is not well managed from the beginning, it can incomprehensible or can be lost, and therefore cannot be curated well or at all. So at NTL, we say that:

Data Management (DM) is a Necessary Element of (∊) Data Curation (DC)

This can be illustrated as a formula:

$$\mathbf{DM} \in \mathbf{DC}$$

where **DM** is data management, ∊ is the symbol for "is element of," and **DC** is data curation.

Linked Processes: Data Curation & Data Science

If data science (DS) is defined in part as "drawing useful conclusions from large and diverse datasets," this implies that there are large and diverse datasets available for reuse, repurposing, and combing in novel ways. New information and knowledge cannot be generated unless there is a body of data from which to draw. This means that data scientists can generate new information only if data curation (DC), or the "**active and ongoing** management of data through its lifecycle **of interest and usefulness** to scholarship, science, and education," has taken place. Therefore, at NTL, we say that:

Data Curation (DC) Enables (⇒) Data Science (DS)

This can be illustrated as a formula:

$$\mathbf{DC} \Rightarrow \mathbf{DS}$$

where **DC** is data curation, where we repurpose the symbol for implies ⇒ as "enables," and **DS** is data science.

Linked Processes: Data Management & Data Curation & Data Science

At NTL we combine these formulae in the phrase:

Data Management (DM) is a Necessary Element
of (∊) Data Curation (DC) which Enables (⇒) Data Science (DS), or

$$\mathbf{DM} \in \mathbf{DC} \Rightarrow \mathbf{DS}$$

If data is expected to serve as a useful asset for future research discovery or for organizational operations, then we must understand that conclusions of data science are most trustworthy when drawn from well-curated data. Further, data curation cannot take place if data is not well managed to begin with.

Finally, in order to unlock the full potential of data, **we strongly recommend** adding data curation practices to research programs.

2. Tech Transfer, Public Access, & Open Science

In **Building a Foundation for Effective Technology Transfer through Integration with the Research Process: A Primer**, the U.S. DOT Volpe Center defines **Technology Transfer (T2) Activities** as: "All activities designed to help ensure that technologies created or improved though R&D [research and development] are widely adopted for use outside or within the research-producing organization" (p. 2) [https://rosap.ntl.bts.gov/view/dot/12262]. Or to put it another way, U.S. DOT-funded researchers are expected to implement the research outputs they create **and** to make them available to other transportation organizations for easy and rapid adoption. This means that **sharing research outputs** — concrete aggregate mixtures; new bridge support designs; improved rail signaling devices; aviation collision avoidance software; etc. — is an expected outcome of research.

Sharing research and technology transfer are not new for transportation researchers. **Public Roads Volume 1, Number1** [https://doi.org/10.21949/1523484] includes a bibliography reporting lab test results on gravel, crushed stone, bituminous road materials, and other road building materials at a time when concrete pavement was just being introduced (see page 44). Additional articles discuss road building and maintenance issues and solutions as implemented by various states. As the research progressed through the 20th Century, transportation researchers shared research results through journals and reports. Some reports even included page after page of typeset data tables detailing lab testing results. As we moved in the age of computers, it became easier to send reports and journal articles to transportation colleagues, so that they could rapidly adopt R&D outcomes to improve transportation infrastructure and safety.

So sharing R&D outputs is not new. Even sharing R&D data is not new, and is now easier. What is new is the expectation that research datasets will be shared as openly as possible, and that they may be combined with other datasets for novel or unexpected uses. Where does this expectation come from? The global **Open Science** movement seek to make research data **Findable, Accessible, Interoperable, and Reusable**, or as it is called **FAIR** https://doi.org/10.1038/sdata.2016.18.

The idea of "open science" has been around since 1998, and the phrase was first used as a call for software sharing in the early years of open networking via the Internet. Over the last 25 years "open science" has been refined and re-defined by various bodies in Europe; at the National Academies of Science, Engineering, and Medicine; and, by the United Nations Educational, Scientific, and Cultural Organization (UNESCO). These more recent definitions focus on sharing scientific research, outputs, and data for broad social benefit.

Open Science is "the principle and practice of making research products and processes available to all, while respecting diverse cultures, maintaining security and privacy, and fostering collaborations, reproducibility and equity" [https://doi.org/10.1038/d41586-023-00019-y], according to the White House Office of Science and Technology Policy (OSTP) Subcommittee on Open Science (SOS), a group in which we participate. This definition helps to guide federal development in sharing R&D outputs and data, and now informs policies such as updates to **Public Access**.

U.S. R&D departments first issued **Public Access** plans following a 2013 White House memo [https://doi.org/10.21949/1528360] describing Public Access as the Public being aware or and being able to download and analyze research publications and dataset that are funded in part or wholly by federal agencies. One goal of Public Access is to conduct federally-funded science with more transparency. Another goal is to be able to reuse already-collected scientific data in novel ways to produce new information. Therefore, U.S. -funded researchers are now required to share non-sensitive research dataset at the time that they publish their research results.

Public Access policies for research and data sharing are direct descendants of Open Science. Further, Open Science and Public Access can be seen as formal approaches to the Technology Transfer transportation organizations have been doing for decades.

There is no doubt that sharing research datasets and other outputs through publicly accessible repositories requires new skills and new resources which Technology Transfer organization may not currently possess. However, by adding data curators to research staff, T2 agencies can find it easier to comply with Public Access requirements and to participate in Open Science activities. T2 organizations have been sharing all along, now they themselves need to adopt some new tools to facilitate their long-established Technology Transfer activities in this new research environment.

Authors

Leighton L Christiansen
https://orcid.org/0000-0002-0543-4268
Data Curator,
National Transportation Library
leighton.christiansen@dot.gov

Jesse Ann Long
https://orcid.org/0000-0002-4962-1380
Data Management and Data Curation Fellow,
National Transportation Library
Jesse.long.ctr@dot.gov

Peyton Tvrdy
https://orcid.org/0000-0002-9720-4725
Data Management and Data Curation Fellow,
National Transportation Library
Peyton.Tvrdy.ctr@dot.gov

Scan the QR code above to download and save this poster for future reference.

Recommended Citation

Christiansen, Leighton L; Long, Jesse A.; and Tvrdy, Peyton. "Data Curation Practices to Optimize Research Coordination, Preservation, and Reuse ." [2024]. [Presented at Transportation Research Board 103th Annual Meeting.] National Transportation Library, U.S. Department of Transportation. Washington, D.C., USA. https://doi.org/10.21949/1529901

3. Curation to Improve Coordination

Now that we have spent some time describing **Data Curation** and its benefits, as well as contextualizing **Tech Transfer** within the **Public Access** and **Open Science** landscapes, let us now turn to those curation **practices** which will provide the most immediate benefit to your **research program**.

1. Data Management Planning & Writing DMPs: The most impactful thing you can do right now is include data management planning and writing a DMP in the planning phase of each research project.

A **data management plan (DMP)** is a narrative **Knowledge Management** document created during research proposal writing and planning to capture and record implicit team knowledge into an explicit document. DMPs record team roles and responsibilities that are vital for succession planning as team members change. A robust DMP describes: 1. the team's plan for handling the raw and final dataset(s) generated during research; 2. the variables of interest, how they will be captured, and the file formats they are stored in; and, 3. how the research outputs will be stored, preserved, and shared. A good DMP, along with other documentation such as a data dictionary, helps future users understand the data. And since **the most likely future user of a dataset will be you** or the your organization, writing a DMP is being good to yourself.

For more guidance on DMPs see the NTL Research Data Management LibGuide:
https://transportation.libguides.com/researchdatamanagement/dmp

2. Good Data Collection Practices will help you understand files at a glance, enable long-term preservation and reuse, and protect data from loss. These practices include: 1. Using open, **non-proprietary file formats** when possible; 2. Establish a human-readable **File Naming Structure**; and, 3. Use the **3-2-1 Backup Strategy**: 3 copies of the data, stored in 2 different geographical regions, on at least 1 other type of storage media.

3. Persistently Identifying People, Research Outputs, & Organizations: Persistent Identifiers (PIDs) are alphanumerical strings formatted as URLs, which globally and uniquely identify a person, a dataset or research paper, or a research organization. PIDs are vital as there many people, papers, and entities with the same or similar names in the world. PIDs that you could adopt today include:

a. **ORCID (Open Researcher and Contributor IDentifiers)** https://orcid.org/ : a PID for people

b. **DOI (Digital Object Identifier)** https://www.doi.org/ : a PID for articles and datasets

c. **ROR (Research Organization Registry)** https://ror.org/ : a PID for research organizations & funders

For more guidance on PIDs see the NTL Research Data Management LibGuide:
https://transportation.libguides.com/persistent_identifiers

4. Documentation & Data Packaging: A **Data Package** is the dataset, the data management plan (DMP), and all other documentation needed to contextualize the dataset for any and all users and reusers. A complete Data Package should be submitted to the funder and the chose repository, with these elements:

a. **Research Output(**s): Dataset, Software, Code, Model, Report, etc.;

b. **README.txt** which includes the **data dictionary**;

c. A descriptive **Metadata** file to aid data discovery on the Internet;

d. **Data Management Plan** (DMP); and,

e. Other supporting codes, scripts, or tables.

For more guidance on Data Packages see the NTL Research Data Management LibGuide:
https://transportation.libguides.com/researchdatamanagement/datapackages

5. Decide How Long and Where to Preserve Your Data: Every research program should have policies and guidance for deciding how long to preserve research data (**Data Retention**) and have a list of trustworthy repositories where researchers will deposit the data (**Data Preservation**). Let us look at each concept quickly.

Data Retention is a set of "policies for persistent data and records management for meeting legal and business data archival requirements." https://en.wikipedia.org/wiki/Data_retention. As data retention is meant to be policy driven, data retention is a **data governance** function. Deciding how long to preserve data should be guided by conscious governance, and not left to chance. Factors may include: funder requirements for retention; local or federal legal requirements; the uniqueness of the data; the difficulty in collecting the data again (in the case of loss); among others. The **retention period** should be recorded in the DMP during the project planning phase. Retention policies should guide **data preservation practices**.

Data Preservation is a set of "actions taken to conserve and maintain the safety and integrity of data" https://en.wikipedia.org/wiki/Data_preservation. Notice the word "maintain" in the definition. This means active care of the data over time. Data preservation and **data storage** are NOT synonymous. Data storage is passive activity that may be done once. Preservation actions that may need to take place during the data lifecycle include: **Bit fixity checking** and file replacement; file **format migration** as software develops; and, active **data disposition** decision making. Yes, data can be deleted, under certain circumstances, or as it reached the end of its useful life. But that decision should be guided by policy and carried out by skilled curators. **Good practice:** Choose a repository that actively curates data. How can you know?

6. Choosing a Data Repository: Look for a repository that: 1. allows for easy patron **access**; 2. has explicit **retention policies**; 3. has long-term **organizational sustainability**; 4. uses and records **persistent identifiers** for people, outputs, and organizations; 5. has robust and rich **metadata** describing each dataset; and, 6. has explicit **data security** policies and practices.

To help satisfy federal funder requirements refer to the **Desirable Characteristics of Data Repositories for Federally Funded Research** at https://doi.org/10.5479/10088/113528. You can also reference the NTL list of criteria at https://doi.org/10.21949/1520563 or the list of DOT conformant repositories at https://doi.org/10.21949/1520566. NTL Data Services team members are happy to consult with you as well. Just ask!

The practices describe above can be quickly implemented and will have lasting programmatic impact.

4. Getting DC Practices into Your Program

Now that you are convinced of the benefits of data curation, you are probably asking yourself "So, **how do I** get data curation practices into my program to optimize research coordination, preservation, and reuse?" We are glad you asked! Here are few ideas:

1. **Hire a degreed Data Curator** and **embed** them in each data collection project. Of course this answer seems a little self-serving. However, you wouldn't build a bridge with out a certified Civil Engineer on the project. An expert Data Curator uses their skills to collect, document, and care for the data, freeing other project staff to use their expert skills at greatest efficiency, and at tasks they actually enjoy.

Where do you find data curators? Many Graduate Schools of Library and Information Science have been offering data curation specializations for the past decade. Reach out and talk to the placement office or post positions on the university job board.

2. **Contract curation** through nearby university library or digital repository. Christiansen's position at the Iowa DOT Library was actually contracted through the institute for Transportation (InTrans) at Iowa State University (ISU). This meant Christiansen was an ISU employee working full time in the Iowa DOT building. Check with your nearby **University Library**, as many now offer **research data management** and **data curation** training and support, and **repository services**.

3. Join a "**curation network**." For example, in the United States, there is the Data Curation Network https://datacurationnetwork.org. Networks such as these are currently set up to allow repositories to trade expertise and labor when they are presented with datasets for which they do not have expert staff. Organizational and funding models may vary.

4. **Provide continuing education** and **staff development** time for curation practices. There are many excellent online-based training courses and MOOCs available. Some are available for free and generated by the research and data curation community. You can see some of these at the Earth Science Information Partners (ESIP) Data Management Training Clearinghouse https://dmtclearinghouse.esipfed.org/home. Others can be had from subscription-based for-profit training hubs.

You can also set up trainings with the NTL staff. For an idea of options, see below.

5. Past Trainings from NTL

Here is a sample of the data curation-related trainings available from the NTL Data Services team:

Fulfilling Statistical Policies with Data Curation Practices: https://doi.org/10.21949/1527466

Data Management Plans (DMPs) for Research Proposals: Researchers' Development of a Data Management Plan: Session 1 for Federal Aviation Administration: https://doi.org/10.21949/1524567

Evaluating Data Management Plans (DMPs): Researchers' Development of a Data Management Plan: Session 2 for Federal Aviation Administration: https://doi.org/10.21949/1524568

U.S. Open Science Policy Perspectives & Transportation: Open Science in Transportation: Challenges and Opportunities in a COVID-19 Era: https://doi.org/10.21949/1520725

Introducing the Year of Open Science & Next Steps in Sharing Transportation Research: https://doi.org/10.21949/1529413

How to Share Publications and Datasets under the USDOT Public Data Access Plan: Presentation to the Federal Transit Administration: https://doi.org/10.21949/1524569

Considerations on Data Retention: Presentation to the Open Mobility Foundation's Privacy, Security, and Transparency Committee: https://doi.org/10.21949/1526875

Other topics include: Persistent Identifiers; Research Digitization and Legacy Data Rescue; Documenting Data and Creating Data Packages; Making you Data FAIR; and more. Email us at NTLDataCurator@dot.gov to talk about the options!

6. Conclusions

In this poster we spent a good deal of time introducing you to some practices of **Data Curation** and related data management concepts that we employ at the National Transportation Library. We also contextualized **Technology Transfer**, or research sharing, with **Public Access** and **Open Science**, showing how they support each other in the goal of sharing transportation research for rapid adoption.

One of our goals was to answer a question from research program managers: wonder "Which practices will have the greatest impact on furthering research coordination and accelerating tech transfer?" Section 3 outlined our top 6 practices for immediate impact. We briefly touched on how to get curation skills into your research program and offered some training topics that the NTL Data Services Team provides.

We hope that you will adopt the practices that we shared, and that you will reach out to us if can help in anyway.

Remember to review
NCHRP Report 936
https://www.trb.org/Publications/Blurbs/180230.aspx

NCHRP
RESEARCH REPORT 936
NATIONAL COOPERATIVE HIGHWAY RESEARCH PROGRAM
Guide to Ensuring Access to the Publications and Data of Federally Funded Transportation Research