



U.S. Department of Transportation



Considerations on Data Retention

<https://doi.org/10.21949/1526875>

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

Presented to: Open Mobility Foundation's Privacy, Security, and Transparency
Committee, on 2022-08-17

Slide Title: Considerations on Data Retention

[Speaker Script]

I want to thank Stephanie Dock, Pamela Lee, and Angela Giaccheti for inviting me to speak with you today.

[Next slide]

Slide Text not read:

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

Presented to: Open Mobility Foundation's Privacy, Security, and Transparency
Committee, on 2022-07-20

Introduction



Slide Title: Introduction

[Speaker Script]

Just two quick notes.

First, any opinions that I express today, especially during the discussion period, are mine alone and do not necessarily represent those of the United States Department of Transportation.

Second, you have been given or will receive a handout version of these slides. The handout includes enhanced and extended text that I will not be able to say today in the interest of time. You may read that extended text at your leisure.

[Next slide]

About the National Transportation Library (NTL)



Established in 1998 by the Transportation Equity Act for the 21st Century (TEA-21), the National Transportation Library (NTL) provides access to:

- Digital collections
- Data services
- Reference and research services
- Knowledge Networks

NTL is an open access digital repository. Most items are in the public domain and available for reuse without restriction.

Important Links

- Home Page: <https://ntl.bts.gov/ntl>
- Repository & Open Science Access Portal ((ROSA P): <https://rosap.ntl.bts.gov/>
- NTL LibGuides: <https://transportation.libguides.com/>

3

Slide Title: About the National Transportation Library (NTL)

[Speaker Script]: Skipping in the interest of time.

[Next Slide]

[Slide Text not read]

Established in 1998 by the Transportation Equity Act for the 21st Century (TEA-21), the National Transportation Library (NTL) provides access to:

Digital collections

Data services

Reference and research services

Knowledge Networks

NTL is an open access digital repository. Most items are in the public domain and available for reuse without restriction.

Important Links

Home Page: <https://ntl.bts.gov/ntl>

Repository & Open Science Access Portal ((ROSA P): <https://rosap.ntl.bts.gov/>

NTL LibGuides: <https://transportation.libguides.com/>

NTL's Legislative Mandate



49 U.S.C. 6304(a)(8), Moving Ahead for Progress in the 21st Century Act, 2012

To support the information management and decision making needs of transportation officials at the Federal, State, and local levels, there is established in the Bureau a National Transportation Library which shall coordinate efforts among, and cooperate with, transportation libraries, information providers, and technical assistance centers, in conjunction with private industry and other transportation library and information centers, with the goal of developing a comprehensive transportation information and knowledge network that supports the activities described in section 6302(b)(3)(B)(vi).

Important Links

- Home Page: <https://ntl.bts.gov/ntl>
- Repository & Open Science Access Portal ((ROSA P): <https://rosap.ntl.bts.gov/>
- NTL LibGuides: <https://transportation.libguides.com/>

4

Slide Title: NTL's Legislative Mandate

[Speaker Script]: Skipping in the interest of time.

[Next Slide]

[Slide Text not read]

The National Transportation Library's mandate was expanded by the MAP-21 legislation in 2012. 49 U.S.C. 6304(a)(8), Moving Ahead for Progress in the 21st Century Act, 2012

In short, the NTL is tasked:

To support the information management and decision making needs of transportation officials at the Federal, State, and local levels, there is established in the Bureau a National Transportation Library which shall coordinate efforts among, and cooperate with, transportation libraries, information providers, and technical assistance centers, in conjunction with private industry and other transportation library and information centers, with the goal of developing a comprehensive transportation information and knowledge network that supports the activities described in section 6302(b)(3)(B)(vi).

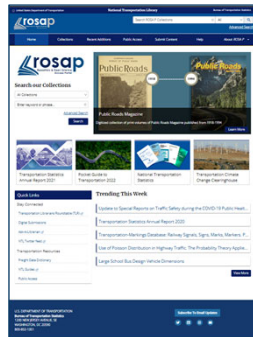
Important Links

Home Page: <https://ntl.bts.gov/ntl>

Repository & Open Science Access Portal ((ROSA P): <https://rosap.ntl.bts.gov/>

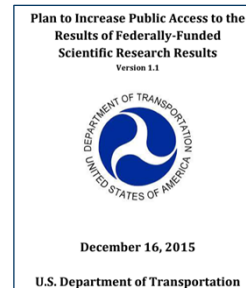
NTL LibGuides: <https://transportation.libguides.com/>

My Roles



Data Curator

- Collect, Catalog, Manage, and Share U.S. DOT-created or -funded statistical and research datasets;
- Especially from the Bureau of Transportation Statistics (BTS) and the University Transportation Centers Program;
- In our Repository & Open Science Access Portal (ROSA P) <https://rosap.ntl.bts.gov/>



Public Access Policy

- Lead implementation of DOT Public Access efforts <https://doi.org/10.21949/1503647>
- Train DOT employees and contracted researchers on data management best practices
- Work on federal, national, and international open science and research data sharing bodies

5

Slide Title: My Roles

[Speaker Script]:

I perform two main duties at the U.S. Department of Transportation, both related to data.

As mentioned, I am the Data Curator at the National Transportation Library. So what does that mean? It means that I:

- **Collect, Catalog, Manage, and Share U.S. DOT-created or -funded statistical and research datasets;**

Secondly, I help DOT implement our Public Access Policy

<https://doi.org/10.21949/1503647>, that says, in part, that all non-sensitive research data generated by DOT or its paid researchers, must be made available to the general public for their use.

Those caveats aside, let us talk about data retention.

[Next Slide]

[Extended Slide Text not read]

I perform two main duties at the U.S. Department of Transportation, both related to data.

As mentioned, I am the Data Curator at the National Transportation Library. So what does that mean? It means that I:

- Collect, Catalog, Manage, and Share U.S. DOT-created or -funded statistical and research datasets;
- Especially those from the Bureau of Transportation Statistics (BTS) and the University Transportation Centers (UTC) Program;
- In our Repository & Open Science Access Portal (ROSA P)
<https://rosap.ntl.bts.gov/>

Secondly, I help DOT implement our Public Access Policy, that says, in part, that all non-sensitive research data generated by DOT or its paid researchers, must be made available to the general public for their use. This means that I:

- Lead implementation of DOT Public Access efforts
<https://doi.org/10.21949/1503647>;
- Train DOT employees and contracted researchers on data management best practices; and,
- Work on federal, national, and international open science and research data sharing bodies.

My degree is a Masters of Library and Information Science (MLIS), with a Certificate of Study focused on Data Curation, from the University of Illinois at Urbana-Champaign.

And in the life I lead before becoming a data curator, I was, among many other things, an over the road truck driver, and have 1.5 millions miles, and the back ache to prove it.

[Next slide]

Data Retention



Slide Title: Data Retention

The Question:

How long should we plan to retain our data?

Slide Title: The Question:

[Speaker Script]

I was asked today to try to answer the following question: “How long should we plan to retain our data?”

[Next slide]

The Short Answer:

It depends.

Slide Title: The Short Answer:

[Speaker Script]

It depends.

I know that answer is not sufficient. So let us go deeper.

[Next slide]

The Longer Answer:

The period of time to retain any dataset is a case-by-case business decision. **There is no “one-answer-fits-all-cases” solution.**

Data retention depends on **a number of factors**, which include, but are not limited to:

1. Laws, Policies, Contracts, Funder Agreements, and Licenses
2. Organizational and/or Repository Resources and Infrastructure
3. Characteristics of Data Collection
4. Potential Data Uses and Reuses
5. Data Management and Preservation Planning

9

Slide Title: The Longer Answer:

[Speaker Script]

The period of time to retain any dataset is a case-by-case business decision. There is no “one-answer-fits-all-cases” solution.

Data retention depends on a number of factors, which include, but are not limited to:

- 1. Laws, Policies, Contracts, Funder Agreements, and Licenses**
- 2. Organizational and/or Repository Resources and Infrastructure**
- 3. Characteristics of Data Collection**
- 4. Potential Data Uses and Reuses**
- 5. Data Management and Preservation Planning**

So, with that quick preview, let us go a little deeper into each of these factors.

[Next slide]

1. Laws, Policies, Agreements & Licenses



Your data may already be subject to a retention period.

Data retention periods may be written into legal or organizational documents. Do you know yours? They are not always easy to find.

1. **Federal Law:** National Archives and Records Administration “NARA Record Schedule” <https://www.archives.gov/files/about/records-schedule/nara-records-schedule-list.pdf> is 328 pages long! Retention varies by type of record.
2. **City Law:** DC “Data and Records Retention Policy” <https://octo.dc.gov/publication/data-and-records-retention-policy> : “...agenc[ies] must create a Records Retention Schedule (“Schedule”) which has been approved by the District’s Office of Public Records...”
3. **Funder Agreements:** National Institutes of Health (NIH) “Data Management & Sharing Plan” <https://sharing.nih.gov/data-management-and-sharing-policy/planning-and-budgeting-DMS/writing-a-data-management-and-sharing-plan#after> : “... make ... data available for **as long as ... useful** for the larger research community...”

10

Slide Title: 1. Laws, Policies, Agreements & Licenses

[Speaker Script]

Your data may already be subject to a retention period.

Data retention periods may be written into legal or organizational documents.

The documents might include:

- **Laws,**
- **Policies,**
- **Contracts,**
- **Funder Agreements, and,**
- **Data use or re-use Licenses**

Do you know which laws and policies affect the retention period of the data you are collecting? These documents can go by many names and are not always easy to locate.

Let us compare two of three examples on the slide.

Example 1, the National Archives and Records Administration “NARA Record Schedule” is 328 pages long, with specific retention periods for certain types of

federal records.

On the other hand, Example 3, the National Institutes of Health (NIH) “Data Management & Sharing Plan”: calls for researchers to make data available for “as long as they anticipate it being useful for the larger research community, institutions, and/or the broader public.”

Take away: there is a broad range of answers to the data retention question. Some of these answer depend on organizational infrastructure.

[Next slide]

[Extended Slide Text not read]

Your data may already be subject to a retention period.

Data retention periods may be written into legal or organizational documents. The documents might include:

- Laws,
- Policies,
- Contracts,
- Funder Agreements, and,
- Data use or re-use Licenses

Do you know which laws and policies affect the retention period of the data you are collecting? These documents can go by many names and are not always easy to locate.

Here are three examples. Let’s highlight just 1 at the moment:

1. **Federal Law:** National Archives and Records Administration “NARA Record Schedule” <https://www.archives.gov/files/about/records-schedule/nara-records-schedule-list.pdf> is 328 pages long! Retention varies by type of record.
 1. The length of the NARA schedule should give you a sense of why there is no one-size-fits-all answer.
2. **City Law:** DC “Data and Records Retention Policy” <https://octo.dc.gov/publication/data-and-records-retention-policy> : “...agenc[ies] must create a Records Retention Schedule (“Schedule”) which has been approved by the District’s Office of Public Records...”
 1. The DC City Laws pushes the responsibility down to individual agencies to create policies consistent with the general guidelines of the law.
3. **Funder Agreements:** National Institutes of Health (NIH) “Data Management & Sharing Plan”: NIH also encourages researchers to make scientific data available for as long as they anticipate it being useful for the larger research community, institutions, and/or the broader public.
 1. The NIH policy is written to cover as many general cases as possible, so it’s retention language is vague on purpose.

This slide was about legal impacts on data retention. But data retention is also impacted by organizations.

2. Organizational and/or Repository Resources



Data retention can depend your **organization's resources and infrastructure**. Or that of the **repository** where you keep data.

Does your organization or repository have:

- 1. Secure funding to retain and preserve data?**
- 2. Data curation staff to backup and migrate data?**
- 3. Robust Organizational Data Governance?**
- 4. A good Backup Strategy?**

Backup Best Practice: 3-2-1 Backup Strategy

- 3 Copies of the Data**
- 2 Different Geographical Regions**
- 1 Other Type Storage Media**

Desirable Characteristics of Data Repositories for Federally Funded Research. National Science and Technology Council (NSTC) Subcommittee on Open Science (SOS), 2022.
<https://rosap.ntl.bts.gov/view/dot/62310>

11

Slide Title: 2. Organizational and/or Repository Resources

[Speaker Script]

Data retention can depend your organization's resources and infrastructure. Or that of the repository where you keep data.

There are a number of considerations to make, and I will highlight 2 of the 4 on the slide:

- 1. Does your organization or repository have secure funding to retain and preserve data?**
- 4. Does your organization or repository have a good Backup Strategy?**

I want to take a moment highlight a Backup Best Practice: The 3-2-1 Backup Strategy.

- Keep 3 Copies of the Data,**
- In 2 Different Geographical Regions, and**
- On at least 1 Other Type of Storage Media**

This simple practice can protect you from all sorts of data loss scenarios, and can be used in any data collection activity.

[Next slide]

[Extended Slide Text not read]

Data retention can depend your organization's resources and infrastructure. Or that of the repository where you keep data.

There are a number of considerations to make, and I include just 4 below:

1. Does your organization or repository have secure funding to retain and preserve data?
2. Does your organization or repository have data curation staff to backup and migrate data?
3. Does your organization or repository have robust Organizational Data Governance?
 1. By Data Governance, I mean in the way that the Data Management Association (DAMA) defines "data governance" as the "planning, oversight, and control over management of data and the use of data and data-related resources." <https://dama.org/sites/default/files/download/DAMA-DMBOK2-Framework-V2-20140317-FINAL.pdf> . Data governance, as described above, is a managerial process to ensure data is managed, and is not the day-to-day managing of data by data creators and data stewards. John Ladley (Data Governance, 2012, page 11) encourages us "not to confuse the management of data with ensuring data is managed."
4. Does your organization or repository have a good Backup Strategy?

I want to take a moment highlight a Backup Best Practice: The 3-2-1 Backup Strategy.

- Keep 3 Copies of the Data,
- In 2 Different Geographical Regions, and
- On at least 1 Other Type of Storage Media

This simple practice can protect you from all sorts of data loss scenarios.

If you are thinking about contracting with a third-party repository for data preservation, I suggest you take a moment to review this short document from the National Science and Technology Council (NSTC) Subcommittee on Open Science (SOS):

Desirable Characteristics of Data Repositories for Federally Funded Research. National Science and Technology Council (NSTC) Subcommittee on Open Science (SOS), 2022. <https://rosap.ntl.bts.gov/view/dot/62310>

I was one of many co-authors.

We have just talked about the characteristics of organizations. Now let us turn the characteristics of gathering data.

[Next slide]

3. Characteristics of Data Collection

Data retention can also depend on characteristics of the **data collection activity**. Is the data collection activity, or the data:

1. Subject to Laws and Policies?
2. Used as Basis for Laws or Policies?
3. **Unique or Rare Natural or Cultural Phenomenon?**
4. **Costly to Create or Re-Create?**
5. **Important to Discipline, Entity, Heritage, Society, or Business Needs?**
6. From a long time ago? (Is the data out of date?)
7. Bound by Data Use and Re-use Licenses and Agreements?

“Lost” Apollo 11 Moon Landing Video Data

In 2006, NASA, prepping for an Apollo 11 anniversary, discovered they could not locate original 1969 Apollo 11 video recordings.

After months searching, NASA “determined that they had most likely been erased and used again, which was standard practice at the time.” However, no data was actually lost, because all footage had been transferred from original tapes and backed up elsewhere. <https://www.nasa.gov/feature/not-unsolved-mysteries-the-lost-apollo-11-tapes>

There may not have been a written data retention plan for the tapes.

Data Collection Characteristics

- Uniqueness or Rarity
- Costly to Create
- Important to Discipline and Society

12

Slide Title: 3. Characteristics of Data Collection

[Speaker Script]

Data retention can also depend on characteristics of the data collection activity. Is the data collection activity, or the data:

(of the 7 questions below, I will highlight just three)

3. Is the data collection activity a Unique or Rare Natural or Cultural Phenomenon?

4. Is the data collection activity costly to create or re-create?

5. Is the data collection activity important to Discipline, Entity, Heritage, Society, or Business Needs?

So, let’s look at an example where the question of data retention made international news.

You may remember from about 15 years back the media uproar about the so-called “Lost” Apollo 11 Moon Landing Video Data.

In 2006, NASA, prepping for the Apollo 11 40th Anniversary, discovered they could not easily locate the original 1-inch wide, 14-track tapes that included 1969 Apollo 11 slow scan video recordings.

After months searching, NASA “determined that [the tapes] had most likely been erased and used again, which was standard practice at the time.” The media jumped on this story.

However, at the end of the search, it was determined that no data was actually lost, because all footage had been transferred from original tapes and backed up elsewhere.

So, all ended well.

However, I did not find any discussion of a written data retention plan for the tapes. Maybe there was, maybe there wasn't.

If you had been on the Apollo team, what characteristics might you have considered when thinking about data retention?

In my mind the obvious Data Collection Characteristics are:

- **Uniqueness or Rarity**
 - You only land on the moon for the first time once.
- **Costly to Create**
 - From 1960 to 1973, NASA spent \$28 billion on the lunar program.
- **Important to Discipline and Society**
 - Data important to future missions.

So we just heard about data tapes, getting reused, what about data re-use?

[Next slide]

[Extended Slide Text not read]

Data retention can also depend on characteristics of the data collection activity. Is the data collection activity, or the data:

(of the 7 questions below, I will highlight just three)

1. Is the data collection activity or the resulting data subject to laws and policies?
2. Is the data collection activity or the resulting data used as basis for laws or policies?
3. **Is the data collection activity a Unique or Rare Natural or Cultural Phenomenon?**
4. **Is the data collection activity costly to create or re-create?**
5. **Is the data collection activity important to Discipline, Entity, Heritage, Society, or Business Needs?**
6. Is the data collection activity from a long time ago? (Is the data very old or out of date?)
7. Is the data collection activity bound by Data Use and Re-use Licenses and Agreements?

So, let's look at an example where the question of data retention made international news.

You may remember from about 15 years back the media uproar about the so-called “Lost” Apollo 11 Moon Landing Video Data.

In 2006, NASA, prepping for an Apollo 11 anniversary, discovered they could not easily locate the original 1-inch wide, 14-track tapes that included 1969 Apollo 11 slow scan video recordings.

After months searching, NASA “determined that [the tapes] had most likely been erased and used again, which was standard practice at the time.” The media jumped on this story.

However, at the end of the search, it was determined that no data was actually lost, because all footage had been transferred from original tapes and backed up elsewhere.

The recordings have since been digitally enhanced and are available in their entirety.

<https://www.nasa.gov/feature/not-unsolved-mysteries-the-lost-apollo-11-tapes>

So, all ended well.

However, I did not find any discussion of a written data retention plan for the tapes.

Maybe there was, maybe there wasn't.

If you had been on the Apollo team, what characteristics might you have considered when thinking about data retention?

In my mind the obvious Data Collection Characteristics are:

- Uniqueness or Rarity
 - You only land on the moon for the first time once.
- Costly to Create
 - From 1960 to 1973, NASA spent \$28 billion on the lunar program.
- Important to Discipline and Society
 - Data important to future missions.
 - Plus, the space race and the moon landing were central “science and engineering superiority” bragging points in the Cold War between the United States and the Soviet Union at the time.

Now we have seen how data must be retained in order to be re-used, let us take a look at how data re-use might impact data retention.

[Next slide]

4. Potential Data Uses and Reuses



Data retention can depend on **potential data uses and reuses**.

Considerations:

1. Data collected for 1 purpose may be useful in other settings.
- 2. Data can be combined with other datasets to reveal new knowledge.**
3. Transportation data may hold information related to economics, climate, health, population groups, etc.
4. Data may be re-used in innovative ways or to create new tools and applications.
- 5. You may not be best person to determine reuse. It may take fresh eyes to see other uses.**

Transportation Research Data Re-Use Example:

Tewari, Sanjay. **Combined Effect of Sea-Level Rise and Coastal Land Subsidence – Identification of Critical Transportation Infrastructure At-Risk in Coastal SPTC Region: Part I - Louisiana**. Louisiana Tech University, Ruston; Southern Plains Transportation Center; Office of the Assistant Secretary for Research and Technology, 2019. <https://rosap.ntl.bts.gov/view/dot/62266>

13

Slide Title: 4. Potential Data Uses and Reuses

[Speaker Script]

Data retention can depend on potential data uses and reuses.

Of the five considerations on the slide, I will highlight just two:

2. Data can be combined with other datasets to reveal new knowledge.

5. You may not be best person to determine reuse. It may take fresh eyes to see other uses.

And for a transportation example this time, we can look Dr Sanjay Tewari's 2019 research Combined Effect of Sea-Level Rise and Coastal Land Subsidence – Identification of Critical Transportation Infrastructure At-Risk in Coastal SPTC Region: Part I - Louisiana.

In the project Dr Tewari used NOAA's National Geodetic Survey (NGS) data of coastal Louisiana to create GIS tools that would identify transportation infrastructure imperiled by sea level rise and land subsidence.

By retaining and making NGS data public, NOAA made it easier for the researcher to identify at-risk infrastructure, and will give authorities an opportunity to direct resources.

In order to be able to share data, the NOAA team had to plan to manage the data.
[Next slide]

[Extended Slide Text not read]

Data retention can depend on potential data uses and reuses.

Considerations:

1. Data collected for 1 purpose may be useful in other settings.
2. Data can be combined with other datasets to reveal new knowledge.
3. Transportation data may hold information related to economics, climate, health, population groups, etc.
4. Data may be re-used in innovative ways or to create new tools and applications.
5. You may not be best person to determine reuse. It may take fresh eyes to see other uses.

Transportation Research Data Re-Use Example:

Tewari, Sanjay. Combined Effect of Sea-Level Rise and Coastal Land Subsidence – Identification of Critical Transportation Infrastructure At-Risk in Coastal SPTC Region: Part I - Louisiana. Louisiana Tech University, Ruston; Southern Plains Transportation Center; Office of the Assistant Secretary for Research and Technology, 2019. <https://rosap.ntl.bts.gov/view/dot/62266>

In this study, Dr Sanjay Tewari re-used the National Oceanic and Atmospheric Administration’s (NOAA) National Geodetic Survey (NGS) data <https://www.ngs.noaa.gov/> of coastal Louisiana to create GIS tools that would identify transportation infrastructure imperiled by sea level rise and land subsidence. By retaining and making NGS data public, NOAA made it easier for the researcher to identify at-risk infrastructure, and this will give the state of Louisiana and the U.S. DOT more time act, as well as knowing where to target resources.

You can greatly improve the re-use of data for yourself and others by writing and following a data management plan, so let us look at Data Retention and DMPs next.

[Next slide]

5. Data Management and Preservation Planning



Data retention policies impact data management and data preservation planning.

1. Organizational data retention policy and/or records retention schedules should guide retention decisions for individual datasets.
2. Record retention period in DMP.
3. Make sure current team and future team members know the retention period.

U.S. Department of Transportation, Bureau of Transportation Statistics. (1997). **Vehicle Inventory and Use Survey (VIUS) 1997 [datasets]**. <https://doi.org/10.21949/1521118>

5. Archiving and Preservation Plans:

The dataset will be archived in the National Transportation Library Repository and Open Science Access Portal (ROSA P).

The dataset will be **retained in perpetuity**.

14

Slide Title: 5. Data Management and Preservation Planning

[Speaker Script]

Data retention policies impact data management and data preservation planning. Highlighting just one of the three points, as a best practice, you should Record retention periods in DMPs.

Do you want to know what that looks like in a real DMP, Then check out one of NTL's DMPs. Look at the 1997 Vehicle Inventory and Use Survey (VIUS) [datasets].

Section 5 Archiving and Preservation Plans of that DMP says: The dataset will be archived in the National Transportation Library Repository and Open Science Access Portal (ROSA P)...The dataset will be retained in perpetuity.

Or, as close to “forever” as we can, depending, on funding, the eventual collapse of the power grid, other natural and social phenomenon, etc.

Let us take a closer look at DMPs.

[Next slide]

[Extended Slide Text not read]

Data retention policies impact data management and data preservation planning.
Record retention periods in DMPs.

1. Organizational data retention policy and/or records retention schedules should guide retention decisions for individual datasets.
2. Record retention period in DMP.
3. Make sure current team and future team members know the retention period.

For an example of how this is recorded in a DMP, let us look at the U.S. Department of Transportation, Bureau of Transportation Statistics, (1997), **Vehicle Inventory and Use Survey (VIUS) 1997 [datasets]**. <https://doi.org/10.21949/1521118>

The current DOT DMP template has a specific section where we record retention information, section: **5. Archiving and Preservation Plans:**

For the VIUS dataset the text reads, in part: The dataset will be archived in the National Transportation Library Repository and Open Science Access Portal (ROSA P)...The dataset will be retained in perpetuity.

This means NTL is committing to retaining the dataset for as close to “forever” as we can. That will all depend, of course, on funding, the eventual collapse of the power grid, other natural and social phenomenon, etc.

Speaking of DMPs, let us take a little closer look at DMPs next.

[Next slide]

More on DMPs

1. A Data Management Plan or DMP describes the deliberate planning, creation, storage, access, and preservation of data produced from a given investigation.
2. A DMP should be created during the project planning phase. A DMP is designed to help you think through all of your data externalizes and dependencies, as well as plan for access, storage, sharing, and preservation.
3. Although initially created during the planning phase, a DMP should be thought of as a living document and reviewed as frequently as necessary, and, updated to capture every project change that is made.

DMPs should capture:

- Project Title & Information
- Data Description
- Roles & Responsibilities
- Access & Sharing Policies
- **Privacy & Sensitive Data Policies**
- **Retention & Preservation Plans**
- Applicable Laws & Policies

National Transportation Library
Research Data Management LibGuide
<https://transportation.libguides.com/researchdatamanagement>

United States. Department of Transportation
Creating Data Management Plans (DMPs)
<https://doi.org/10.21949/1520562>

15

Slide Title: More on DMPs

[Speaker Script]

Of my three tips on Data management plans on the slide, I want you all to remember point 3: Although initially created during the planning phase, a DMP should be thought of as a living document and reviewed as frequently as necessary, and, updated to capture every project change that is made.

DMPs should capture a number of elements, as shown on the slide, including for our discussion:

- **Privacy & Sensitive Data Policies**
- **Retention & Preservation Plans**

The links at the bottom of the slide lead you to DOT resources for writing robust data management plans.

Let us now look at other elements of the data lifecycle.

[Next slide]

[Extended Slide Text not read]

I want to give three key thoughts on data management plans:

1. A Data Management Plan or DMP describes the deliberate planning, creation, storage, access, and preservation of data produced from a given investigation.
2. A DMP should be created during the project planning phase. A DMP is designed to help you think through all of your data externalizes and dependencies, as well as plan for access, storage, sharing, and preservation.
3. Although initially created during the planning phase, a DMP should be thought of as a living document and reviewed as frequently as necessary, and, updated to capture every project change that is made.

DMPs should capture:

- Project Title & Information
- Data Description
- Roles & Responsibilities
- Access & Sharing Policies
- **Privacy & Sensitive Data Policies**
- **Retention & Preservation Plans**
- Applicable Laws & Policies

If you would like to know more about how US DOT and NTL create DMPs or how we encourage our researchers to create DMPs, please see these two resources:

1. National Transportation Library, Research Data Management LibGuide available at <https://transportation.libguides.com/researchdatamanagement>
2. United States. Department of Transportation, Creating Data Management Plans (DMPs) available at <https://doi.org/10.21949/1520562>

Now that we've talk about DMPs, let us take a look at other aspects of data retention during the data lifecycle.

[Next slide]

How can we make Data Retention easier?

Plan for Data Retention throughout the Data Lifecycle.



1. **Planning:**
 1. **Plan for Long-Term Retention**
 2. Plan for Sharing
2. **Acquisition:**
 1. Long-term retention-friendly data formats
 2. Document! Data dictionary; Methodology;
 3. Update DMP
3. **Processing:**
 1. Document cleaning and modifications;
 2. Update DMP
4. **Analysis:**
 1. Record & publish analytical models and tools
5. **Publication & Sharing:**
 1. Choose well-run repository
 2. Use persistent identifiers for data
 3. Plan for sharing from the beginning: re-used data is data worth retaining
6. **Preservation & Disposition:**
 1. Preservation-friendly Data Formats
 2. **Retain or Dispose based on Policy or Schedule**
 3. "Automated" deletion should have a human in the review loop.

16

Slide Title: How can we make Data Retention easier?

[Speaker Script]

You should be planning for Data Retention throughout the Data Lifecycle. This slide shows some specific actions that you can take during each stage of the data lifecycle to aid retention.

So as you will note from my crude data lifecycle diagram, I get paid for data curation, not graphic design.

That aside, a simple, generic data lifecycle contains these six stages:

1. **Planning for data collection**
2. **Acquisition of data**
3. **Processing and Cleaning of data**
4. **Analysis of data**
5. **Publication & Sharing of data**
6. **Preservation & Disposition of data**

During each of these stages there are actions you can take make data retention easier and more impactful.

Let us highlight just two of the many:

At the Planning stage, you should plan for long-term retention, especially longer than you might think the data will be of interest.

And at the Preservation and Disposition stage, be sure to use preservation friendly data formats, and to retain or dispose of data based on policy or retention schedule.

Wait! Did I just say you can delete data?

[Next slide]

[Extended Slide Text not read]

You should be planning for Data Retention throughout the Data Lifecycle. This slide shows some specific actions that you can take during each stage of the data lifecycle to aid retention.

So as you will note from my crude data lifecycle diagram, I get paid for data curation, not graphic design.

That aside, a simple, generic data lifecycle contains these six stages:

1. Planning for data collection
2. Acquisition of data
3. Processing and Cleaning of data
4. Analysis of data
5. Publication & Sharing of data
6. Preservation & Disposition of data

Now during each of these stages there are actions you can take make data retention easier and more impactful. This list is far from exhaustive.

1. Planning:

1. Long-Term Retention
2. Sharing

2. Acquisition:

1. Long-term retention-friendly data formats
2. Document! Data dictionary; Methodology;
3. Update DMP

3. Processing:

1. Document cleaning and modifications;
2. Update DMP

4. Analysis:

1. Record & publish analytical models and tools

5. Publication & Sharing:

1. Choose a well-run repository
2. Use persistent identifiers for data
3. Plan for sharing from the beginning: re-used data is data worth retaining

1. Preservation & Disposition:

1. Preservation-friendly Data Formats
2. Retain or Dispose based on Policy or Schedule
3. “Automated” deletion should have a human in the review loop

So, some of you may have noticed that the sixth stage of the lifecycle is “Preservation & Disposition.”

Does that mean I am saying you can delete data?

The answer can be yes. Let us talk about that in the next slide.

[Next slide]

Is it Ever OK to Delete Data?

Yes. Not all data has to be kept.
And data that you keep today does not have to be kept forever.

Deletion Considerations:

1. Is the data an official record or covered by specific law or policy?
2. How old is the data?
- 3. Is the data obsolete or superseded?**
4. How large is the data?
5. Is limited access, cold storage an option?
6. Is there another agency willing to house the data?
- 7. Has the data ever been re-used?**

Preservation & Disposition Actions:

- 1. Periodically review data access, re-use, and citation statistics.**
2. Contact outside stakeholders and the community of interest.
3. Document your review.
4. Delete if appropriate.
- 5. “Tombstone” deleted data.**
 1. Update the dataset landing page to let the public know that the data did exist and that it now no longer does.

17

Slide Title: Is it Ever OK to Delete Data?

[Speaker Script]

Yes. Not all data has to be kept. And data that you keep today does not have to be kept forever

Let us look at two of the considerations for deleting data:

3. Is the data obsolete or superseded?

Delete it.

7. Has the data ever been re-used?

No? Then will anyone miss it?

If you are going to delete the data, there are a number of Preservation & Disposition Actions you should take as best practice, including:

1. Periodically review data access, re-use, and citation statistics.

5. “Tombstone” deleted data.

Update the dataset landing page to let the public know that the data did exist and that it now no longer does.

Let us have a quick review...

[Next slide]

[Extended Slide Text not read]

Yes. Not all data has to be kept. And data that you keep today does not have to be kept forever.

Let us look at some considerations on why you might choose to delete some data:

1. Is the data an official record or covered by specific law or policy that dictates a retention schedule or disposition timeline?
 1. If so, follow that policy.
2. How old is the data?
 1. If the data is very old, is it still relevant? Is it historical? Was it just overlooked?
3. Is the data obsolete or superseded?
 1. Delete it.
4. How large is the data?
 1. Is retaining this data using a lot of resources that could be better spent?
5. Is limited access, cold storage an option?
 1. This will allow you to hold on to the data, but at less cost.
6. Is there another agency willing to house the data?
 1. Give the data a new home.
7. Has the data ever been re-used?
 1. No? Then will anyone miss it?

If you are going to delete the data, there are a number of Preservation & Disposition Actions you should take as best practice:

1. Periodically review data access, re-use, and citation statistics.
2. Contact outside stakeholders and the community of interest.
3. Document your review.
4. Delete if appropriate.
5. “Tombstone” deleted data.
 1. Update the dataset landing page to let the public know that the data did exist and that it now no longer does.

Let us have a quick review....

[Next slide]

In Review:

Data retention depends on **a number of factors**, such as:

1. Laws, Policies, Contracts, Funder Agreements, and Licenses
2. Organizational and/or Repository Resources and Infrastructure
3. Characteristics of Data Collection
4. Potential Data Uses and Reuses
5. Data Management and Preservation Planning

That said, robust data retention planning and implementation can have beneficial impacts on other data goals.

18

Slide Title: In Review:

[Speaker Script]

Data retention depends on a number of factors, such as:

- 1. Laws, Policies, Contracts, Funder Agreements, and Licenses**
- 2. Organizational and/or Repository Resources and Infrastructure**
- 3. Characteristics of Data Collection**
- 4. Potential Data Uses and Reuses**
- 5. Data Management and Preservation Planning**

That said, robust data retention planning and implementation can have beneficial impacts on other data goals.

With that I want to go to one last slide of tools and Resources,...

[Next slide]

Resources



Research Data Management LibGuide. National Transportation Library.
<https://transportation.libguides.com/researchdatamanagement>

Creating Data Management Plans (DMPs). United States. Department of Transportation.
<https://doi.org/10.21949/1520562>

Desirable Characteristics of Data Repositories for Federally Funded Research. National Science and Technology Council (NSTC) Subcommittee on Open Science (SOS), 2022.
<https://rosap.ntl.bts.gov/view/dot/62310>

Data Management of Researchers: Organize, Maintain and Share Your Data for Research Success. Kristin Briney. (2015). Pelagic Publishing. 9781784270117.

Repository and Open Science Access Portal. Collection: "Public Access Resources." National Transportation Library. https://rosap.ntl.bts.gov/collection_par

Repository and Open Science Access Portal. Collection: "US DOT Public Access Data Management Plans." National Transportation Library. https://rosap.ntl.bts.gov/collection_pa_dmp

19

Slide Title: Resources

[Speaker Script]

....is a list of a few resources to help you implement data retention best practices.

I hope you will find them helpful.

Thank you for your time and attention today.

[Next slide]

[Extended Slide Text not read]

Research Data Management LibGuide. National Transportation Library.
<https://transportation.libguides.com/researchdatamanagement>

Creating Data Management Plans (DMPs). United States. Department of Transportation. <https://doi.org/10.21949/1520562>

Desirable Characteristics of Data Repositories for Federally Funded Research. National Science and Technology Council (NSTC) Subcommittee on Open Science (SOS), 2022. <https://rosap.ntl.bts.gov/view/dot/62310>

Data Management of Researchers: Organize, Maintain and Share Your Data for Research Success. Kristin Briney. (2015). Pelagic Publishing. 9781784270117.

Repository and Open Science Access Portal. Collection: “Public Access Resources.” National Transportation Library. https://rosap.ntl.bts.gov/collection_par

Repository and Open Science Access Portal. Collection: “US DOT Public Access Data Management Plans.” National Transportation Library. https://rosap.ntl.bts.gov/collection_pa_dmp

Questions?



20

Slide Title: Questions?

[No text]

[Next Slide]



U.S. Department of Transportation



Contact

Leighton L Christiansen

Data Curator

National Transportation Library

leighton.christiansen@dot.gov