National Transportation Data Preservation Network (NTDPN) Workshop 2

Detailed Proceedings

February 2020

U.S. Department of Transportation
John A. Volpe National Transportation Systems Center

Volpe

National Transportation Data Preservation Network (NTDPN) Workshop 2
Held at the Transportation Research Board (TRB) 99th Annual Meeting, Washington D.C.

Walter E. Washington Convention Center
Conference Room 305
Thursday, January 16, 2020
1:00 PM – 2:30 PM

Detailed Proceedings

Participants

In-Person

- Angela Berthaume, *USDOT (Volpe)*
- Sandy Brady, *Louisiana Transportation Research Center*
- Leighton Christensen, *USDOT (OST)*
- Bobbi deMontigny, *Montana Department of Transportation*
- Allison Gwinup, *Tennessee Department of Transportation*
- Dianne Kresich, *Arizona Department of Transportation Research Center*
- Kendra Levine, *UC Berkeley ITS Library*
- Jesse Long, *USDOT (OST)*
- Adam McCormick, *Booz Allen Hamilton*
- Courtney Mumma, *Texas Digital Library*
- Hilary Nixon, *San Jose State University*
- Sue Sillick, *Montana Department of Transportation*
- Brendan Williams, *Portland State University*
- Chris Baglin, *Xentity Corporation*

Virtual

- Charles Ducker, *USDOT (OST)*
- Dana May, *Montana State University*
- Emily Chan, *San Jose State University*
- Jacob Carlson, *University of Michigan*
- Lisa Kay Schweyer, *Carnegie Mellon University*
- Maria Praetzellis, *California Digital Library*
- Megan O'Donnell, *Iowa State University*
- Michael Witt, *Purdue University (PURR)*
- Nate Pulliam, *Georgia Department of Transportation*
- Roger Weaver, *Missouri University of Science and Technology*
- Sandi Caldrone, *Purdue University (PURR)*

Meeting Notes

Network Overview and Development Efforts

- Data archiving at the Bureau of Transportation Statistics (BTS) initially focused on small datasets from Department of Transportation (DOT) surveys, but has since evolved to include a more diverse collection of data types, including university research data. This change introduced the need to work with, store, and archive huge databases, overburdening the limited BTS and National Transportation Library (NTL) staff. As such, the BTS and NTL looked to develop networks among other institutions that can house large datasets.

- While the introduction of the internet facilitated making data collected by public agencies publicly available at no cost, this service is not free for the provider. Storing, managing, and providing data costs money. The BTS and NTL have identified a need to leverage programs to provide data efficiently and effectively, and find funding to make data Findable, Accessible, Interoperable, and Reusable (FAIR) in order to support researchers' and data users' needs.
- In January 2017, the NTL Data Curator began working with several academic libraries and university transportation centers (UTCs) to help bring their institutional repositories into conformance with USDOT recommendations. During these discussions, the NTL realized that while setting up a repository was not possible for some research organizations, some large research universities may be able to provide contracted repository services at reasonable costs.
- In April 2019, the BTS and NTL sponsored a first of its kind workshop to investigate the possibility of building a National Transportation Data Preservation Network (NTDPN) that would provide unified access and discovery for transportation research data, where data are clearly described and defined. As a result, researchers would be able to combine and reuse data, opening up new paths of inquiries that leverage, rather than duplicate, earlier efforts. This first workshop was held at the Research Data Alliance 13th Plenary, in Philadelphia, Pennsylvania, and brought together a core group of stakeholders to take the first steps supporting this effort.
- The vision for the NTDPN is outlined by two key goals:
  - To help users find transportation-related data currently housed in numerous organizational and institutional repositories and archives; and
  - To help researchers find reliable homes for the digital data if their organization does not have a repository of its own.
- The primary goal of this first workshop was to bring together repository managers with large transportation research data collections to discuss creating a repository network or consortium. During this first workshop, interested parties discussed priorities and considerations for transportation data management, the feasibility of building a network of trustworthy repositories for transportation research data, and the rewards and challenges to building such a network.
  - The priorities and considerations included:
    - Defining relevant nodes and stakeholders along with their needs and contributions;
    - Funding mechanisms and costs associated with data storage and management;
    - Measuring and documenting research impacts to demonstrate the value of research funding;
    - Research scoping, data misrepresentation/misuse, and confidentiality; and,
    - Partnering opportunities and potential network model examples.
  - Some rewards and challenges identified included:
    - Preservation, metadata, and repository workflows and needs;
    - How a network of transportation research repositories could work together;
    - The successes and failures of similar networks in other disciplines;
    - Contracting and service models for smaller research units who become members of the network; and,
    - Preparing for next steps in improving long-term preservation of Federally funded transportation research data.
- The meeting minutes for the April 2019 workshop are archived and available in the NTL's Repository and Open Science Access Portal (ROSAP) at: https://doi.org/10.21949/1506118.
- Since this first workshop, the group has continued to meet regularly during 2019 to continue developing the Network's focus and model moving forward. For this second workshop, the group has identified potential next steps for the Network to take.

- *Discussion* – There was a question on the purpose of the Network in regards to researchers. The Network is intended to help research organizations that don't have repositories of their own find a reliable and trustworthy repository for their data. Some universities or repositories can offer these services for little or no costs, or can set up pricing schemes for contracts based on the data size and amount of curation needed to store and archive the data. The goal is to find long-term partners to provide services to those without access to their own repository.

Network Next Steps Discussions
- The group is looking for more stakeholders and more individuals to be involved. Up to this point, there has been a core working group of 13 people to determine if developing this Network was worth pursuing. The group decided it was, and is ready to move forward to build this Network and advance next steps.
- Next steps the Network previously identified include:
  - USDOT Public Access Plan update – The USDOT Public Access Plan is currently four years old. Laws have been updated and added, clarifications are necessary, and there are lessons learned to incorporate. The update is expected sometime summer 2020.
  - Stakeholder recruitment – The recruitment efforts at the Transportation Research Board (TRB) poster session earlier this week on Tuesday, January 14, 2020 identified 21 new potential stakeholders.
  - Capabilities assessment – Conduct a capabilities assessment of the Network, its nodes, and its members to determine what each group can or cannot provide, define any gaps in the Network infrastructure, and determine who is best suited for each element to inform a plan moving forward.
  - Use cases development – Develop use cases that clearly identify the value of the Network for different users to encourage stakeholder participation and help develop the Network.
  - Network model establishment – Establish a network model to define how the Network will operate. Considerations include having a governing body, requiring membership fees, or adopting existing models such as the National Science Foundation (NSF) spoke-and-hub network connecting users to data and information.
  - Confidentiality considerations – Alleviate confidentiality concerns.
  - Research Data Management (RDM) outreach and education – While there is still a need for training, there are many folks implementing transportation data preservation who hold expertise in data management. Take advantage of these opportunities.
- One member asked about the data management requirements for data contributors versus what the Network is providing, and if the Network is assuming contributors have a repository that can be linked to. Currently, the Network has no assumptions about contributors other than that their data is compliant with the USDOT Public Access Plan. If the Network grows and is sustained, perhaps it can be expanded to other data. Additionally, different repositories have different terms of service, so the data contributor should provide guidance on the level of curation needed – for example, the data may require archival management rather than full curation.
- The Network is technology agnostic and brings repositories together. This Network is not trying to create a domain specific repository for transportation.
- One member is looking for the Network to provide guidance on making data findable and on how to best set up metadata so their data gets captured in searches. This guidance would be important for researchers and data contributors, many of who have not been instructed on how to correctly set up their metadata.
  - One participant noted that this is a service many libraries are offering their repositories.

- o Another participant clarified that the discoverability of the data is in the metadata which is up to the repository, not the researcher. It is the repository's responsibility to establish appropriate metadata for the data to be discoverable.
- One participant noted it is difficult to track DOT-funded research because the unique identifiers are grant numbers rather than a digital object identifier (DOI), so the research is not in Crossref. The Research Organization Registry (ROR) works with Crossref and is not dependent on journal publications, so is useful for researchers developing technical reports as opposed to journal articles. ROR is community driven and has no user fees.
- One participant suggested the Network provide a metadata catalog for repositories. Another participant confirmed many researchers are interested in using a metadata catalog. This is a goal worth pursuing through the Network that would both add value to the field and be appreciated by researchers. All the metadata elements are not yet available since they are limited to the data.
  - o Transport Research International Documentation (TRID) is working on a new record type for datasets. TRID functions as an index and does not host any data itself. TRID could be a central search engine for transportation datasets that search across all repositories.
    - **The Network could support the Transportation Research Board (TRB) and TRID/Transportation Research Information Services (TRIS) efforts to further the development of a metadata catalog.**
    - **One participant suggested reaching out to Bill McLeod, TRIS Manager, to pursue this effort.**
  - o A metadata catalog may be useful for researchers and repository managers since one of the challenges repository managers face is connecting the data hosted in their repositories to the communities that need the data.
  - o Another benefit of a metadata catalog is that it is search engine agnostic.
  - o The group discussed registries and the desire to promote findability. Some examples include Schema.org and FAIRsharing.org.
    - **The Network can identify data registries and determine how they can be used by the Network.**
- One member asked about the requirements for joining the Network. Data stored on a website does not count as a repository since the goal is for the Network to provide long-term repository solutions.
  - o This discussion highlighted the need for developing use cases. This workshop brought together a variety of stakeholders from researchers to data managers to research funders to State DOTs to institutional repository managers. Developing used cases was a previously identified need that has now materialized as an immediate need for the Network.
  - o **The Network will develop use cases from a variety of perspectives to help recruit members by clarifying stakeholder roles and outlining benefits of joining the Network, and help get executive buy-in.**
    - **If anyone is interested in helping develop use cases, let Leighton know.**
- The group discussed the CoreTrustSeal (CTS). While not specifically required in the USDOT Public Access Plan, having the CTS is encouraged and is a way to have data become reliable. Repositories having the CTS is a Network aspiration rather than requirement.
  - o One member noted CTS takes a while to achieve and is a five-year goal for their agency. Another noted that they have CTS, and it took 14 months.
  - o One member agreed requiring the CTS is likely too high of an entry requirement into the Network, but suggested the group identify what a repository would need to do or have in place in order to participate in the Network to ensure data reliability.
    - **The Network should establish criteria or guidance for repositories to join or participate in the Network to ensure data reliability.**

- While different network models have been suggested, no specific model has been identified for the Network to follow and the working group has not determined what type of model will be most appropriate. As the group grows with the addition of new stakeholders, the group can decide which model will be best to emulate.
  - Chronopolis is a good example. It is based in the University of California, San Diego, and provides dark, long-term storage across the community of non-commercial nodes. They have been around for 16 years, so are stable. Chronopolis has a different structure but the same goal as LOCKSS (Lots of Copies Keeps Stuff Safe) based in Stanford University.
  - One participant suggested connecting with the Airport Cooperative Research Program (ACRP) and other cooperative research programs in the future to see how these institutions ask researchers keep their data.
  - **If anyone is interested in leading or joining the effort to establish the Network model, let Leighton know.**
    - **Courtney Mumma is interested in being involved in these efforts.**
- One member suggested encouraging the American Association of State Highway and Transportation Officials (AASHTO) to assist in growing this Network and look for additional opportunities as the Network engages more stakeholders.

Joining the Network
- Everyone can learn from each other based on their different roles, perspectives, and involvement with data. The Network needs to bring together diverse viewpoints to help achieve its goals and more effectively manage the rich data generated through research so it can be used again. The Network will only improve is more people are engaged and ask questions guiding its development.
- From an academic perspective, researchers both create data through their own work and use data from other sources to further their work. A benefit of the Network for researchers is providing recognition to research contributions by making data more discoverable and also providing a mechanism for citing that data. A benefit of the Network for institutional repositories is being able to provide and offer access to a more complete set of information enriching what is known about transportation and facilitating further research.
- Other benefits of the Network include technology transfer, data and technology reuse, discoverability, data availability, and not duplicating prior efforts.
- The intent is to grow this Network to be a national repository for all transportation data, not just USDOT funded data. The Network is starting with this subset of data since the USDOT is a large funder of transportation research, but acknowledges that the USDOT is not the only funder or source for transportation data. The hope is the Network will eventually grow and engage more stakeholders. However, first, this group needs to demonstrate the value of the Network with respect to data usage and research.
- There is no financial commitment to join the Network. The group is looking for folks' time and expertise.
- **The group will reconvene virtually in March.**

Appendix A: National Transportation Data Preservation Network (NTDPN) Workshop 2 Informational Handout

National Transportation Data Preservation Network (NTDPN) Workshop 2

Held at the Transportation Research Board (TRB) 99th Annual Meeting, Washington D.C.

Walter E. Washington Convention Center

Conference Room 305
Thursday, January 16, 2020
1:00 PM – 2:30 PM

Virtual Connection Information
Phone: 877-336-1839; Access Code: 3995486
Web-room: https://connectdot.connectsolutions.com/ntldatacurator/

Agenda

1:00-1:15 PM	Welcome and Introductions
1:15-1:30 PM	Network Overview and Development Efforts
1:30-2:15 PM	Network Next Steps
2:15-2:30 PM	Joining the Network

April Meeting Minutes, archived in the National Transportation Library's repository:
https://doi.org/10.21949/1506118

Building a National Transportation Data Preservation Network Workshop [poster]:
https://doi.org/10.21949/1506103

Invitation to Join the Network

Even with discussions on benefits of "big data," the transportation research community and industry are still in their infancy of using data to drive informed decision making and setting better policy. Simultaneously, we are on the cusp of a transportation revolution with new technologies and concepts for alternative modes of transportation. From geospatial maps and transit routes, to safety data collected about pedestrians and bicyclists along a particular route, transportation-related data can be used for multiple purposes.

With the guidance of the National Transportation Library of the U.S. Department of Transportation, several academic and public institutions are working to design and establish a data preservation network capable of providing all the benefits and services the transportation field needs, such as curating, preserving, and sharing transportation-related data generated during federal, state, industry, and academic research.

So, what do we need to make this happen? You! This is our invitation to you to join us in creating – from the ground up – the data preservation network transportation needs as we move into a more data-centric era.

Please join us in this effort and help us to make transportation data curation, preservation, and access best practices our common practice!

Benefits of a Data Preservation Network

Today, transportation data in the U.S. is stored and managed in disparate ways across numerous platforms. This can make finding data a challenge. Further, as best practices for data preparation, documentation, and preservation are well established, these practices are unevenly implemented across the transportation research domain. This can have a negative impact on long-term data accessibility.

Making data findable, accessible, interoperable, and reusable (FAIR), produces many benefits: researchers creating data receive credit through data citation; greater ease in identifying data gaps; and improved findability of previously collected data for reuse.

Why a Data Preservation Network is Important and Key Benefits of a NTDPN

1. Improve tracking and coordination of data as there currently is no central repository for U.S. transportation research data
2. Identify partners willing to help archive and preserve collected data even without access to an institutional repository; and
3. Encourage adoption of repository standards which makes data easier to find, for people and computers, through the use of robust metadata among other benefits.

Benefits of Joining a Data Preservation Network

- Help establish repository standards for transportation data preservation;
- Contribute to developing research norms for accessing and citing data;
- Improve data availability for the widest possible use; and
- Meet like-minded individuals who care about transportation data preservation.

Network Vision

Data Access & Preservation

The network will aggregate transportation data from across the U.S. to be easily discovered across various repositories and institutions. Network participants will adhere to FAIR Principles ensuring data will be uniformly described and identified to allow researchers to easily find useful and related data, with an understanding of the quality and utility of the data. Long-term preservation and access will be guaranteed through CORETRUSTSEAL certification.

Networking Infrastructure

The network will use a hub-and-spoke architecture (like the NSF model), where member nodes feed into the discovery and access points. This model extends existing repositories and efforts, allowing for participation by researchers regardless of their institutional affiliation. The network will also feed into other established discovery tools for data and transportation.

Community of Practice (CoP)

The network will support a CoP for transportation research, with collaboration through data management, shared governance, education, compliance, and reuse.

**Appendix B: National Transportation Data Preservation Network (NTDPN) Workshop 2
Presentation Slides**

National Transportation Data Preservation Network (NTDPN) Workshop 2

Held at the Transportation Research Board (TRB) 99th Annual Meeting, Washington D.C.
Walter E. Washington Convention Center
Conference Room 305
Thursday, January 16, 2020
1:00 PM – 2:30 PM

Agenda

1:00-1:15: **Welcome and Introductions**
    Angela Berthaume
1:15-1:30 **Network Overview and Development Efforts**
    Jesse Long
1:30-2:15: **Network Next Steps**
    Leighton Christiansen
2:15-2:30: **Joining the Network**
    Hilary Nixon

Welcome and Introductions

Angela Berthaume, MSCE, MPPA
Technology Policy Analyst | Policy Analysis and Strategic Planning V-322
Volpe, The National Transportation Systems Center
U.S. Department of Transportation

Network Overview and Development Efforts

Jesse Long https://orcid.org/0000-0002-4962-1380
Data Curation & Data Management Fellow,
National Transportation Library, Bureau of Transportation Statistics, OST-R
US Department of Transportation
jesse.long.ctr@dot.gov

NTDPN - Overview

Initial Meeting: Building a National Transportation Data Perseveration Network Held at RDA's 13th Plenary, Philadelphia, PA

Key Goals:
- To help searchers find transportation-related data in the numerous organizational and institutional repositories and archives where it now resides.
- Help researchers find reliable homes for the digital data if their organization does not have a repository of its own.

Jesse Long
https://orcid.org/0000-0002-4962-1380
Data Curation & Data Management Fellow,
National Transportation Library,
Bureau of Transportation Statistics,
OST-R, US Department of Transportation
jesse.long.ctr@dot.gov

A National Transportation Data Preservation Network would provide unified access and discovery for transportation research data, where data are clearly described and defined. As a result, researchers would be able to combine and reuse data, opening up new paths of inquiries that leverage (rather than duplicate) earlier efforts.

Goals:

- help searchers find transportation-related data in the numerous organizational and institutional repositories and archives where it now resides
- help researchers find reliable homes for the digital data if their organization does not have a repository of its own.

NTDPN - Overview

Background
- BTS shift from small datasets to a larger and diverse collection of data types.
- New Issues
- Identifying a need and looking for a solution

Workshop 1
- Bring stakeholders together
- Things discussed:
  - Priorities and considerations for transportation data management
  - Rewards and challenges
  - Next steps and outcomes
- April 2019 Meeting Minutes: https://doi.org/10.21949/1506118.

Data archiving at BTS initially focused on small datasets from Department of Transportation

(DOT) surveys, but has evolved to include a more diverse collection of data types, including university research data. This change has introduced the need to work with, store, and archive huge databases, resulting in the inability for the limited BTS and NTL staff to tackle these new issues by themselves. So, they look to develop networks among other institutions that can house large datasets, and determine how to divide challenges in a consistent manner. The introduction of the internet has made it easier to make data collected by public agencies available to the public at no cost, but it is not free for the provider. Storing and providing data costs money. BTS and NTL has identified a need to figure out how to afford what everyone wants to accomplish and leverage programs to provide data efficiently and effectively, and find funding to make data Findable, Accessible, Interoperable, and Reusable (FAIR).

Starting in January 2017, the NTL Data Curator started to working with several academic libraries and university transportation centers (UTCs) to help them bring their institutional repositories into conformance with the USDOT recommendations. During these discussions NTL realized that setting up a repository was not possiblefor some research organizations, while large research universities might be able to provide contracted repository services at reasonable costs.

Workshop 1, at the Research Data Alliance 13th Plenary, in Philadelphia, brought a core of stakeholders together to take the first steps.

The major goal of this first workshop was to bring together repository managers with large transportation research data collections to discuss creating a repository network or consortium. During this first workshop, interested parties discussed feasibility, rewards, and challenges to building such a network of trustworthy repositories for transportation research data.

The group discussed several topics related to priorities and considerations for transportation data management, as well as the rewards and challenges that come with building such a network.

The Priorities and considerations include:

• Defining relevant nodes and stakeholders, their needs and contributions;

• Funding mechanisms and costs associated with data storage and management;

• Measuring and documenting research impacts to demonstrate the value of research funding;

• Research scooping, data misrepresentation/misuse, and confidentiality; and,

• Partnering opportunities and potential network model examples.

Then some rewards and challenges that were identified are:

• preservation, metadata, and repository workflows and needs;

• how a network of transportation research repositories could work together;

• the success and failures of similar networks in other disciplines;

• contracting and service models for smaller research units who become members of the network; and,

• preparing for next steps in improving long-term preservation of federal funded transportation research data

To learn more about the April 2019 workshop the complete minutes are available in ROSA P at: https://doi.org/10.21949/1506118.

NTDPN - Overview

- Other meetings:
    - August 13, 2019
    - October 8, 2019
    - November 13, 2019

- Next Steps

Since the workshop the group has continued to meet multiple times during 2019, to continue developing the network's focus and model moving forward.

Now for this second workshop we have identified the next steps the network is ready to take, which Leighton will discuss.

Network Next Steps

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

Network Next Steps Discussion

Identified

- Update DOT Public Access Plan
- Recruit New Stakeholders
- Conduct Capabilities Assessment
- Develop Use-cases
- Establish Network Model
- Alleviate Confidentiality Concerns
- RDM Outreach & Ed.

For Discussion

- Other steps?
- Priorities?
- Name?
- Volunteers?

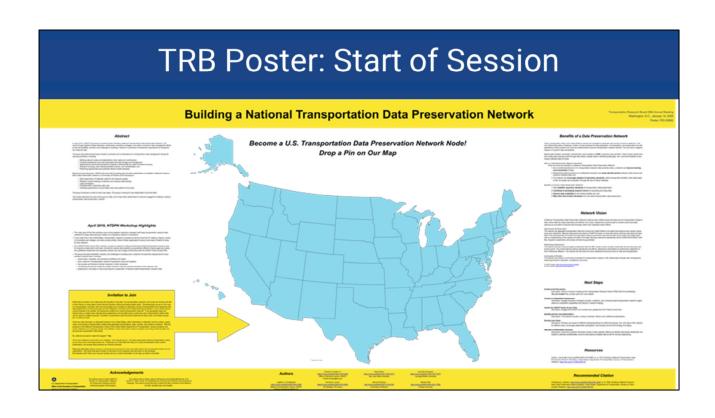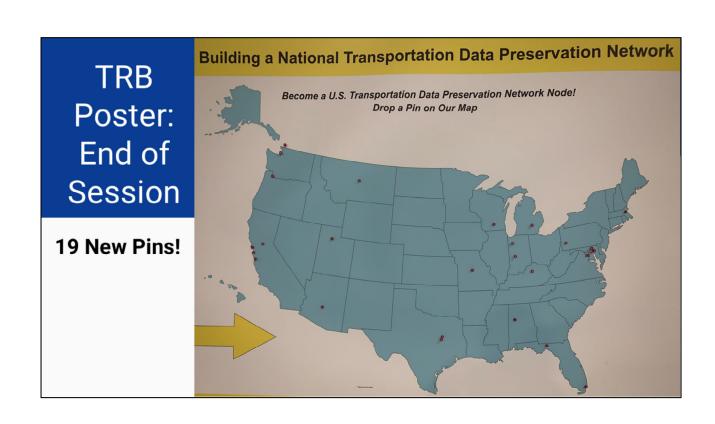See October 8 meeting notes

Joining the Network

Hilary Nixon, Ph.D.
Deputy Executive Director
Mineta Transportation Institute
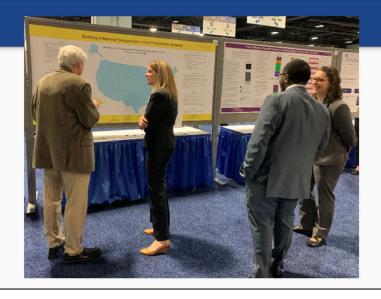San Jose State University

Joining the Network

How to Get Involved and Why Diverse Perspectives are Important – An Academic's View

- Have data?
- Get data?
- Learn about data?
- Bring it all together
- Recognize contributions
- Your voice matters → will make the network better

TRB Poster: Start of Session

TRB Poster: End of Session

19 New Pins!

Joining the Network Can Make you Smile!

Thank you