Opening Transportation Data for the Public Good: Getting Our Bits in a Row

Denver, CO, September 14-17, 2016
RDA 8th Plenary & International Data Forum

Abstract

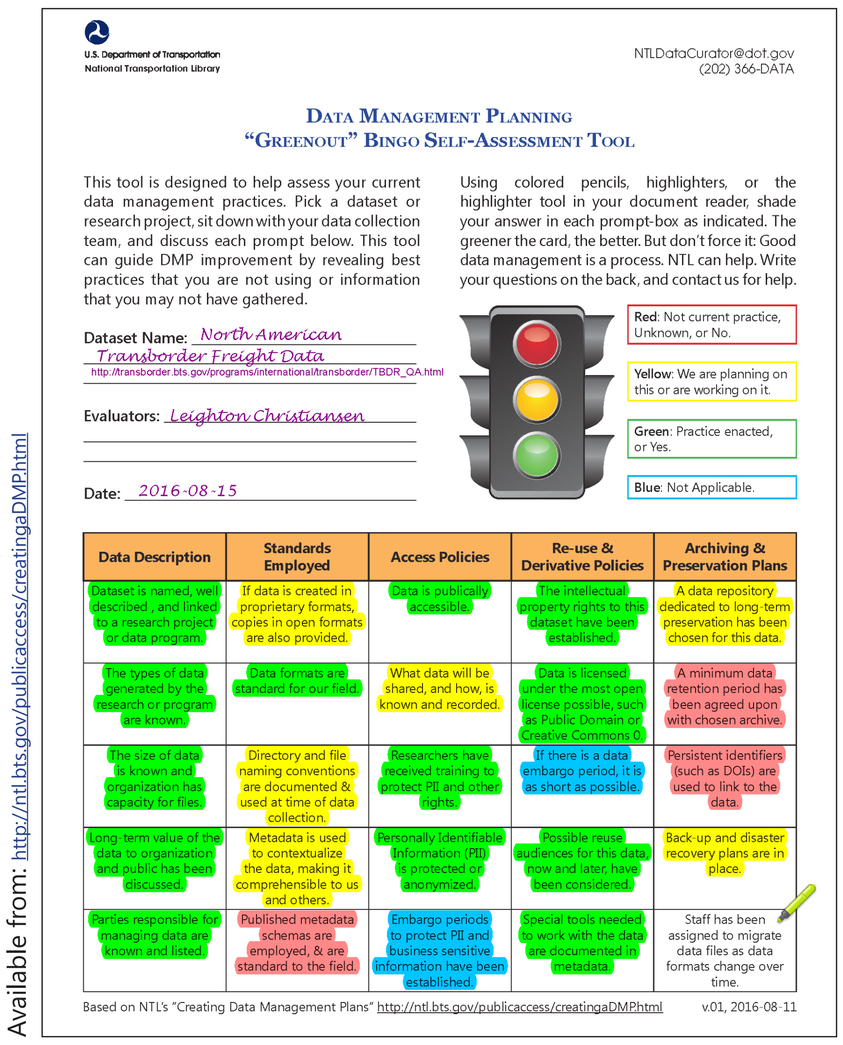
The U.S. Department of Transportation (USDOT) Public Access Plan was issued in November 2015, in response to the February 22, 2013 Office of Science and Technology Policy (OSTP) Memorandum for the Heads of Executive Departments and Agencies entitled "Increasing Access to the Results of Federally Funded Scientific Research ." Through this memorandum, OSTP directed all Executive Departments with greater than $100 million in yearly research and development expenditures to prepare a plan for improving the public's access to the results of federally funded research.

The USDOT plan establishes objectives to ensure public access to **publications** and **digital datasets** arising from USDOT-managed research and development (R&D) programs. USDOT already provides access to intramural and extramural research in progress, technical reports, and many final publications. USDOT's Open Data Plan and Data Governance Policy already require openness for data as well as listing on the Enterprise Data Inventory. Coupled with the implementation of the USDOT Public Access Plan, both extramural and intramural research programs are taking comprehensive steps to increase data sharing and management.

The USDOT plan requires that intramural researchers and extramural awardees submit publications and metadata for datasets resulting from research projects that fall within the scope of the public access plan to the National Transportation Library's (NTL) digital repository. As USDOT's solution for public access, NTL's systems –which are built on international standards and protocols for interoperability, information exchange, federated searching, and metadata sharing– maximize the potential for creative reuse to enhance value to all stakeholders. As the lead for implementation, NTL is also the source for guidance, best practices, and training for compliance with the USDOT Public Access Plan.

NTL's legislative mandate is a primary reason the library serves a central role in public access implementation at USDOT. The mandate includes a requirement to maintain a repository; instruction to be a portal to federal transportation data; and direction to partner with other federal, state, local, and industry organizations to facilitate access to and use of transportation information. These activities enable NTL to both maximize the impact of the Federal research investment in transportation and foster private-public collaboration. This poster highlights a few tools created by NTL to help intramural and extramural data creators employ better data management practices and come into compliance with the Public Access Plan.

Defining Data Management & Playing Bingo!: BTS Roadshows



U.S. Department of Transportation
National Transportation Library

NTLDataCurator@dot.gov
(202) 366-DATA

Data Management Planning
"Greenout" Bingo Self-Assessment Tool

This tool is designed to help assess your current data management practices. Pick a dataset or research project, sit down with your data collection team, and discuss each prompt below. This tool can guide DMP improvement by revealing best practices that you are not using or information that you may not have gathered.

Using colored pencils, highlighters, or the highlighter tool in your document reader, shade your answer in each prompt-box as indicated. The greener the card, the better. But don't force it: Good data management is a process. NTL can help. Write your questions on the back, and contact us for help.

Dataset Name: North American Transborder Freight Data
http://transborder.bts.gov/programs/international/transborder/TBDR_QA.html

Evaluators: Leighton Christiansen

Date: 2016-08-15

Red: Not current practice, Unknown, or No.

Yellow: We are planning on this or are working on it.

Green: Practice enacted, or Yes.

Blue: Not Applicable.

Data Description	Standards Employed	Access Policies	Re-use & Derivative Policies	Archiving & Preservation Plans
Dataset is named, well described , and linked to a research project or data program.	If data is created in proprietary formats, copies in open formats are also provided.	Data is publically accessible.	The intellectual property rights to this dataset have been established.	A data repository dedicated to long-term preservation has been chosen for this data.
The types of data generated by the research or program are known.	Data formats are standard for our field.	What data will be shared, and how, is known and recorded.	Data is licensed under the most open license possible, such as Public Domain or Creative Commons 0.	A minimum data retention period has been agreed upon with chosen archive.
The size of data is known and organization has capacity for files.	Directory and file naming conventions are documented & used at time of data collection.	Researchers have received training to protect PII and other rights.	If there is a data embargo period, it is as short as possible.	Persistent identifiers (such as DOIs) are used to link to the data.
Long-term value of the data to organization and public has been discussed.	Metadata is used to contextualize the data, making it comprehensible to us and others.	Personally Identifiable Information (PII) is protected or anonymized.	Possible reuse audiences for this data, now and later, have been considered.	Back-up and disaster recovery plans are in place.
Parties responsible for managing data are known and listed.	Published metadata schemas are employed, & are standard to the field.	Embargo periods to protect PII and business sensitive information have been established.	Special tools needed to work with the data are documented in metadata.	Staff has been assigned to migrate data files as data formats change over time.

Based on NTL's "Creating Data Management Plans" http://ntl.bts.gov/publicaccess/creatingaDMPhtml v.01, 2016-08-11

Available from: http://ntl.bts.gov/publicaccess/creatingaDMPhtml

The National Transportation Library (NTL) is part of USDOT's Bureau of Transportation Statistics (BTS). BTS was created in 1991 to administer data collection, analysis, and reporting of transportation statistical information.

BTS creates nearly 100 transportation databases and reports annually. BTS datasets have always been publically available, first as print publications, and then through the BTS website and other tools.

However, data management practices are uneven across BTS. Further, there has been little planning for the long-tem preservation and re-use of the data, which is unsurprising as BTS data programs' focus has been on creation, compilation, and dissemination.

NTL staff are now embedding with BTS data program offices to educate, train, and consult on DMP best practices.

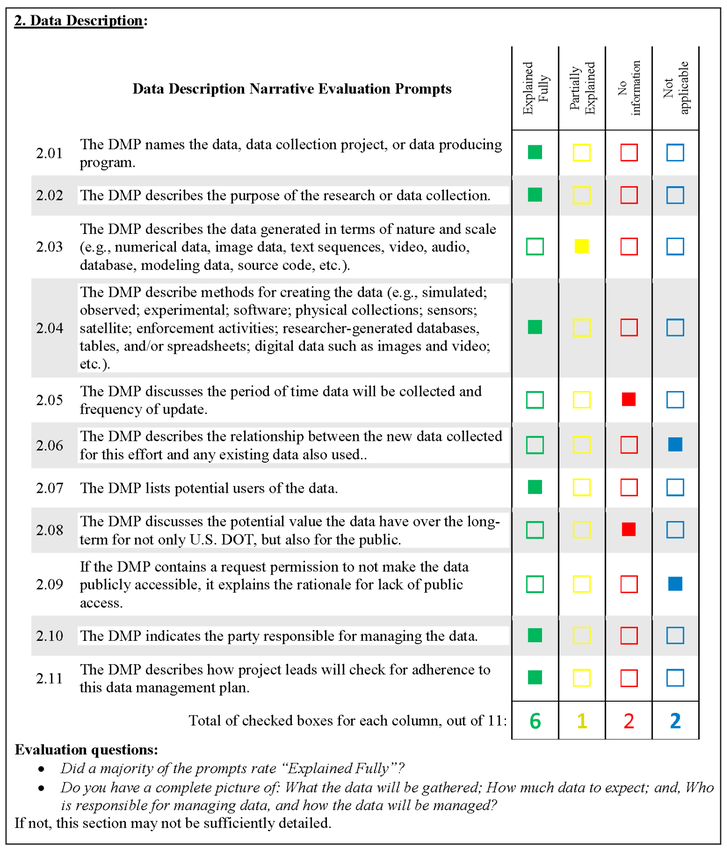
NTL's Data Curator has presented an overview of best practices to each BTS office, called "Data Management Roadshows."

One tool NTL has created for the sessions is the "Data Management Planning 'Greenout' Bingo Self-Assessment Tool." By honestly assessing each prompt, users create a visual representation of data management strengths and weaknesses around a specific dataset.

During the hour-long Roadshows, the Data Curator walks the audience through using the Bingo tool, using a dataset they produce as an example.

NTL's Data Curator can then help data creators implement missing data management practices, improving the preservation and reuse potential of these publicly accessible datasets.

NTL Sufficiency Checklist for DMP Evaluation



2. Data Description:

	Data Description Narrative Evaluation Prompts	Explained Fully	Partially Explained	No information	Not applicable
2.01	The DMP names the data, data collection project, or data producing program.	■			
2.02	The DMP describes the purpose of the research or data collection.	■			
2.03	The DMP describes the data generated in terms of nature and scale (e.g., numerical data, image data, text sequences, video, audio, database, modeling data, source code, etc.).		■		
2.04	The DMP describe methods for creating the data (e.g., simulated; observed; experimental; software; physical collections; sensors; satellite; enforcement activities; researcher-generated databases, tables, and/or spreadsheets; digital data such as images and video; etc.).	■			
2.05	The DMP discusses the period of time data will be collected and frequency of update.			■	
2.06	The DMP describes the relationship between the new data collected for this effort and any existing data also used..				■
2.07	The DMP lists potential users of the data.	■			
2.08	The DMP discusses the potential value the data have over the long-term for not only U.S. DOT, but also for the public.			■	
2.09	If the DMP contains a request permission to not make the data publicly accessible, it explains the rationale for lack of public access.				■
2.10	The DMP indicates the party responsible for managing the data.	■			
2.11	The DMP describes how project leads will check for adherence to this data management plan.	■			
	Total of checked boxes for each column, out of 11:	6	1	2	2

Evaluation questions:
- *Did a majority of the prompts rate "Explained Fully"?*
- *Do you have a complete picture of: What the data will be gathered; How much data to expect; and, Who is responsible for managing data, and how the data will be managed?*

If not, this section may not be sufficiently detailed.

USDOT's Public Access Plan requires DMPs submitted to address the following areas:
- Data Description;
- Standards Used;
- Access Policies;
- Re-Use, Redistribution, and Derivative Products Policies; and,
- Archiving and Preservation Plans.

USDOT and NTL staff created guidance pages for intramural and extramural data creators, which include prompts for the types of information that could be supplied in each section listed above. These include:
- List in what format(s) the data will be collected, & if they are open or proprietary.
- Describe what data will be publicly shared, & how data files will be shared.

Using the information prompts from these guidance pages, NTL staff has created a DMP Sufficiency Checklist in order to aid USDOT staff in assessment of submitted data management plans. Example pages of the checklist are shown in the images.

The DMP Sufficiency Checklist is designed to assist evaluators assessing the sufficiency of the DMPs required of research projects funded by USDOT.

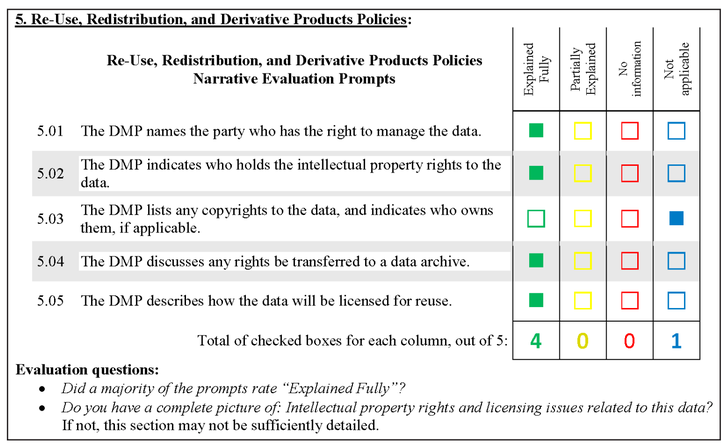
A sufficiently detailed DMP, like a well-developed research methodology, is one vital component of a research proposal for extramural researchers and the research project plan for intramural researchers.

After completing NTL's checklist, non-data management professionals should be able to determine whether or not researchers have thought through all facets of data management planning. NTL's Data Curator can also help.

While USDOT evaluators will still need to make a subjective assessment of a proposed DMP, the checklist supports decision making by providing a quantitative measure of DMPs.

For USDOT intramural research projects, the checklist will guide USDOT research programs toward better data management by highlighting missing information, workflows, and/or practices.

NTL staff will make secondary use of the DMP Sufficiency Checklist to evaluate and update our DMP guidance webpages and training.



5. Re-Use, Redistribution, and Derivative Products Policies:

	Re-Use, Redistribution, and Derivative Products Policies Narrative Evaluation Prompts	Explained Fully	Partially Explained	No information	Not applicable
5.01	The DMP names the party who has the right to manage the data.	■			
5.02	The DMP indicates who holds the intellectual property rights to the data.	■			
5.03	The DMP lists any copyrights to the data, and indicates who owns them, if applicable.				■
5.04	The DMP discusses any rights be transferred to a data archive.	■			
5.05	The DMP describes how the data will be licensed for reuse.	■			
	Total of checked boxes for each column, out of 5:	4	0	0	1

Evaluation questions:
- *Did a majority of the prompts rate "Explained Fully"?*
- *Do you have a complete picture of: Intellectual property rights and licensing issues related to this data?*

If not, this section may not be sufficiently detailed.

Data Citation Recommendations

Properly citing datasets is a hot topic wherever public and open access are discussed. Many data creators feel datasets should be given the same weight as publications, and would like to receive credit for the work they do. NTL staff share the view that datasets should be cited in the same manner as other information resources. As USDOT is implementing a new dataset management system, NTL felt it was important to set a data citation recommendation before its launch. Other federal scientific and technical information (STI) agencies with which NTL partners share the same desire to promulgate a data citation standard for datasets produced by their agencies as part of a data management program.

While an international standard has not yet been set, NTL staff and intern Nicole Strayhorn consulted existing recommendations from data repositories (i.e., Dryad, Harvard Dataverse, Digital Curation Centre, Zenodo) and style guides (i.e., Chicago, MLA, and APA), as well as the required citation information for bibliographic items in the U.S. Government Printing Office Style Manual, 2008. The result, is as follows:

Work by U.S. DOT Office:
Corporate Author. (Release Date/Year of publication, Update Frequency or Revision/Modified Date). Title of dataset, Version number/Edition number. Subset information. [Data file type]. *Archive/Source*. Accessed date from Persistent identifier

U.S. Department of Transportation, Federal Highway Administration. (2015). Public Road and Street Mileage by Functional System(a) 1990-2013. Table 1-5: U.S. [statistical table]. *National Transportation Statistics*. Accessed 2016-07-15 from http://www.rita.dot.gov/bts/sites/rita.dot.gov.bts/files/publications/national_transportation_statistics/html/table_01_05.html

Work by named author:
Author(s)/Principle Investigator(s) ORCID iDs. (Release Date/Year of publication, Update Frequency or Revision/Modified Date). Title of dataset, Version number/Edition number. [Data file type]. *Archive/Source*. Accessed date from Persistent identifier or URL

Moore, Jason K. http://orcid.org/xxxx-xxxx-xxxx-xxxx , Koojiman, J. D. G. http://orcid.org/xxxx-xxxx-xxxx-xxxx , & Schwab, Arend L. http://orcid.org/xxxx-xxxx-xxxx-xxxx . (2015-06-23). Delft Instrumented Bicycle Data and Videos. [dataset]. *Zenodo*. Accessed 2016-07-19 from 10.5281/zenodo.18862

Work by U.S. DOT Office, with subset information:
Corporate Author. (Release Date/Year of publication, Update Frequency or Revision/Modified Date). Title of dataset, Version number/Edition number. Subset information. [Data file type]. *Archive/Source*. Accessed date from Persistent identifier

U.S. Census Bureau. (Date not Given). 2010-2014 American Community Survey 5-Year Estimates. Subset: DP05 - ACS Demographic and Housing Estimates. [statistical table]. *American FactFinder*. Accessed 2016-07-27 from http://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=ACS_14_5YR_DP05&src=pt

USDOT & NTL Public Access Links

USDOT Public Access Plan:
https://www.transportation.gov/mission/open/official-dot-public-access-plan-v11

USDOT Open Government site:
https://www.transportation.gov/mission/open/open-government

NTL Public Access Guidance site:
http://ntl.bts.gov/publicaccess/

NTL "Creating DMPs" guidance:
http://ntl.bts.gov/publicaccess/creatingaDMP.html

NTL Public Access FAQs:
http://ntl.bts.gov/publicaccess/FAQs.html

NTL Digital Repository:
http://ntl.bts.gov/

USDOT Research Hub:
http://ntlsearch.bts.gov/researchhub/index.do

Future Actions

- Work with Bureau of Transportation Statistics (BTS) offices to create Data Management Plans (DMPs) for each dataset created by BTS.
- Help BTS data creators obtain ORCID iDs.
- Integrate ORCID iDs into repository search.
- Train USDOT staff on use of DMP Sufficiency Evaluation Tool (in-person and online module formats).
- Include DMPs in required data package for NTL's digital repository, to enable DMP searching and analysis.
- Work with other federal repositories towards trusted repository status.
- Migrate to a cloud-based digital repository solution, using CDC's Public Access Platform: Introducing NTL's Repository & Open Science Access Portal (ROSA P), the digital library for transportation data and information.

U.S. Department of Transportation
National Transportation Library

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

Amanda J. Wilson
http://orcid.org/0000-0001-6580-2328
Director, National Transportation Library
amanda.wilson@dot.gov

Recommended Citation
Christiansen, Leighton L http://orcid.org/0000-0002-0543-4268 & Amanda J. Wilson http://orcid.org/0000-0001-6580-2328. (2016). "Opening Transportation Data for the Public Good: Getting Our Bits in a Row." RDA 8th Plenary & International Data Forum, International Data Week 2016. Denver, CO, USA.

Acknowledgements
The authors would like to thank Alpha Wingfield for wonderful design assistance.