Bureau of Transportation of Statistics
Data Services & Data Curation

Leighton L Christiansen 0000-0002-0534-4268
Data Curator, National Transportation Library
ntldatcurator@dot.gov
National Transportation Library

Data Services

Policies, Practices, Standards, & Workflows

Data Description

Data Access

Data Appraisal

Data Curation & Preservation

Data Format Migration

Data Management Education

Archival Retention

U.S. Department of Transportation
Some Key Terms

• Data Management
  – deliberate planning, creation, storage, access and preservation of data produced from a given investigation

• Data Curation
  – enables data discovery and retrieval, maintains data quality, adds value, and provides for re-use over time

• Data Science
  – drawing useful conclusions from large and diverse data sets through exploration, prediction, and inference
Linked Processes

Data Management \(\in\) Data Curation

Data Curation \(\Rightarrow\) Data Science
U.S. DOT Public Access Requirements

Researchers and authors are subject to U.S. DOT Public Access requirements if:

• U.S. DOT funded scientific research;
• Newly funded or extended on or after January 1, 2016.

Researchers and authors will need to submit:

• ORCiDs
• 2 – 3 page Data Management Plan (with proposal);
• Final peer-reviewed manuscript;
• Final Digital Datasets; and,
• Any other written outputs (final reports, technical reports, tech summaries, etc.)

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

http://ntl.bts.gov/publicaccess/creatingaDMP_extramural.html

Data Services:

Education & Cheerleading

- PA Compliance Guidance
- Data Management Planning training & consulting
- DMP Sufficiency Checklist and Training
- Data Citation Recommendations
- Best Practice sharing

Future Actions

- Data Curation & Preservation
- FAIR Data Principles
- Repository

5 Sections in DMP

- Data Description
- Standards Used
- Access Policies
- Re-Use, Redistribution, and Derivative Products Policies
- Archiving and Preservation Plans
In-House Training

DMP Self-Assessment

- Gives visual feedback on current state of DMP best practices
- Tested with each BTS office
- Used to begin DMP process for 2 BTS datasets
  - Transborder Freight
  - State Transportation Stats

---

**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: ____________________________________________

Evaluator: ________________________________________________

Date: ____________________________________________________

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

Based on NTL’s “Creating Data Management Plans” [http://ntl.bts.gov/publicaccess/creatingDMPlnput](http://ntl.bts.gov/publicaccess/creatingDMPlnput) v01, 2016-08-11
2. Data Description:

<table>
<thead>
<tr>
<th>Data Description Narrative Evaluation Prompts</th>
<th>Explained Fully</th>
<th>Partially Explained</th>
<th>No information</th>
<th>Not applicable</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.01 The DMP names the data, data collection project, or data producing program.</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>2.02 The DMP describes the purpose of the research or data collection.</td>
<td>☒</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>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.).</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>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.).</td>
<td>☒</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>2.05 The DMP discusses the period of time data will be collected and frequency of update.</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>2.06 The DMP describes the relationship between the new data collected for this effort and any existing data also used.</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>2.07 The DMP lists potential users of the data.</td>
<td>☒</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>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.</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>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.</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>2.10 The DMP indicates the party responsible for managing the data.</td>
<td>☒</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>2.11 The DMP describes how project leads will check for adherence to this data management plan.</td>
<td>☒</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
</tbody>
</table>

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.
Data on the Web
Best Practices

Data Description & Standards
- Metadata
- Data Licenses
- Data Provenance
- Data Quality
- Data Versioning
- Data Identifiers

Access Policies
- Data Formats
- Data Vocabularies
- Data Access
- Data Access APIs

Data Reuse & Derivatives
- Feedback
- Data Enrichment
- Republication

Archiving & Preservation
- Data Preservation

Benefits
- Reuse
- Trust
- Comprehension
- Access
- Linkability
- Interoperability
- Discoverability
- Processability

https://www.w3.org/TR/dwbp/
NTL’s Repository & Open Science Access Portal (ROSA P)

Immediate Steps
• Launch in April 2017
• Based on the CDC Public Access platform, which hosts the CDC Stacks and NOAA Institutional Repository digital libraries
• AWS cloud storage
• Publications & Data
• DOIs

Future Steps
• Trusted repository status
Data Seal of Approval – World Data Systems
Repository Audit and Certification

1. The data can be found on the Internet
2. The data are accessible (clear rights and licences)
3. The data are in a usable format
4. The data are reliable
5. The data are identified in a unique and persistent way so that they can be referred to
Data Curation Lifecycle Model

- Standards & Workflow Development & Implementation
- Complete Description
  - Metadata Standards
- Preservation Planning
  - DM Training
- Curation & Preservation
  - Access & Reuse
  - Repository Ingest
  - Preservation
  - Migration
  - Disposition
Benefits of Data Curation

- Protects Unique Data from Loss
- Improves Data Search & Retrieval
- Enables Reuse
- Facilitates Longitudinal and/or Meta Analyses
- Avoids Duplication of Effort & Spending
- Increases Verifiability
- Opens New Lines of Scientific Discovery
- Satisfies Public Access & Open Government & Legal Requirements
Questions?

ntldatacurator@dot.gov
References

2. Graduate School of Library and Information Science at the University of Illinois at Urbana-Champaign. “Specialization in Data Curation,” 2013. http://www.lis.illinois.edu/academics/programs/specializations/data_curation