Bureau of Transportation of Statistics Data Services & Data Curation
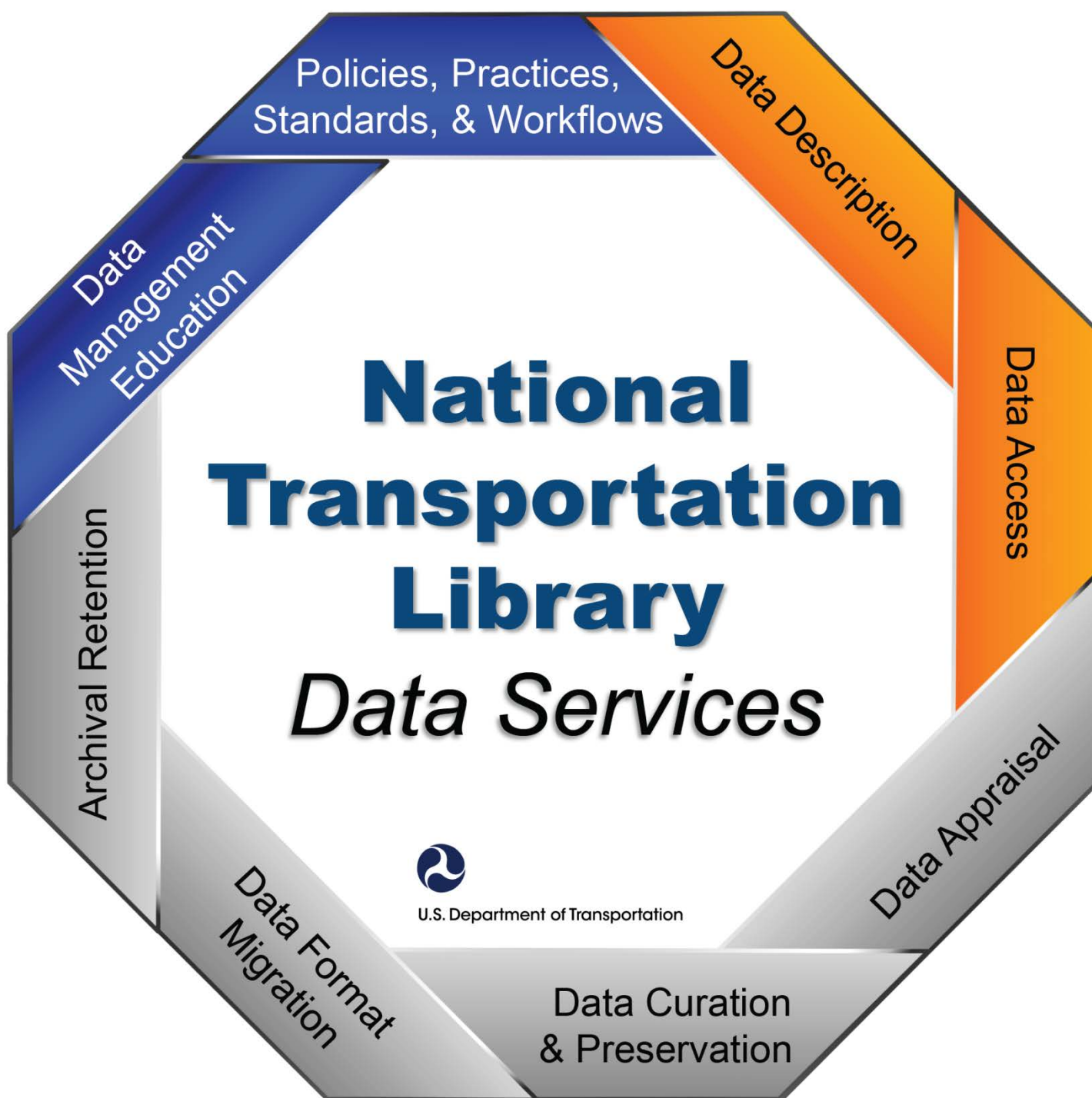
Leighton L Christiansen 0000-0002-0534-4268

Data Curator, National Transportation Library

ntldatacurator@dot.gov

Presentation to:
DOT TRCC
2017-03-23

U.S. Department of Transportation
National Transportation Library

National Transportation Library
Data Services

Policies, Practices, Standards, & Workflows

Data Description

Data Access

Data Appraisal

Data Curation & Preservation

Data Format Migration

Archival Retention

Data Management Education

U.S. Department of Transportation

Some Key Terms

- ## Data Management[1]
  - deliberate planning, creation, storage, access and preservation of data produced from a given investigation[1]

- ## Data Curation[2]
  - enables data discovery and retrieval, maintains data quality, adds value, and provides for re-use over time[2]

- ## Data Science[3]
  - drawing useful conclusions from large and diverse data sets through exploration, prediction, and inference[3]

Linked Processes

Data Management $\in$ Data Curation

Data Curation $\Rightarrow$ Data Science

U.S. DOT Public Access Requirements

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

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.)

U.S. Department of Transportation
National Transportation Library

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**

U.S. Department of Transportation
National Transportation Library

6

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**

In-House Training

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.

DMP Sufficiency Evaluation

- **Gives quantitative feedback to bolster subjective judgement**
- **Precise feedback to researcher on what is missing**
- **Better DMPs improve likelihood of long-term data preservation for public re-use**

U.S. Department of Transportation
National Transportation Library

Data on the Web Best Practices[4]

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

https://www.w3.org/TR/dwbp/

Benefits

R — Reuse

T — Trust

C — Comprehension

A — Access

L — Linkability

I — Interoperability

D — Discoverability

P — Processability

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**

Trusted Repository Status & Data Curation[5]

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[6]
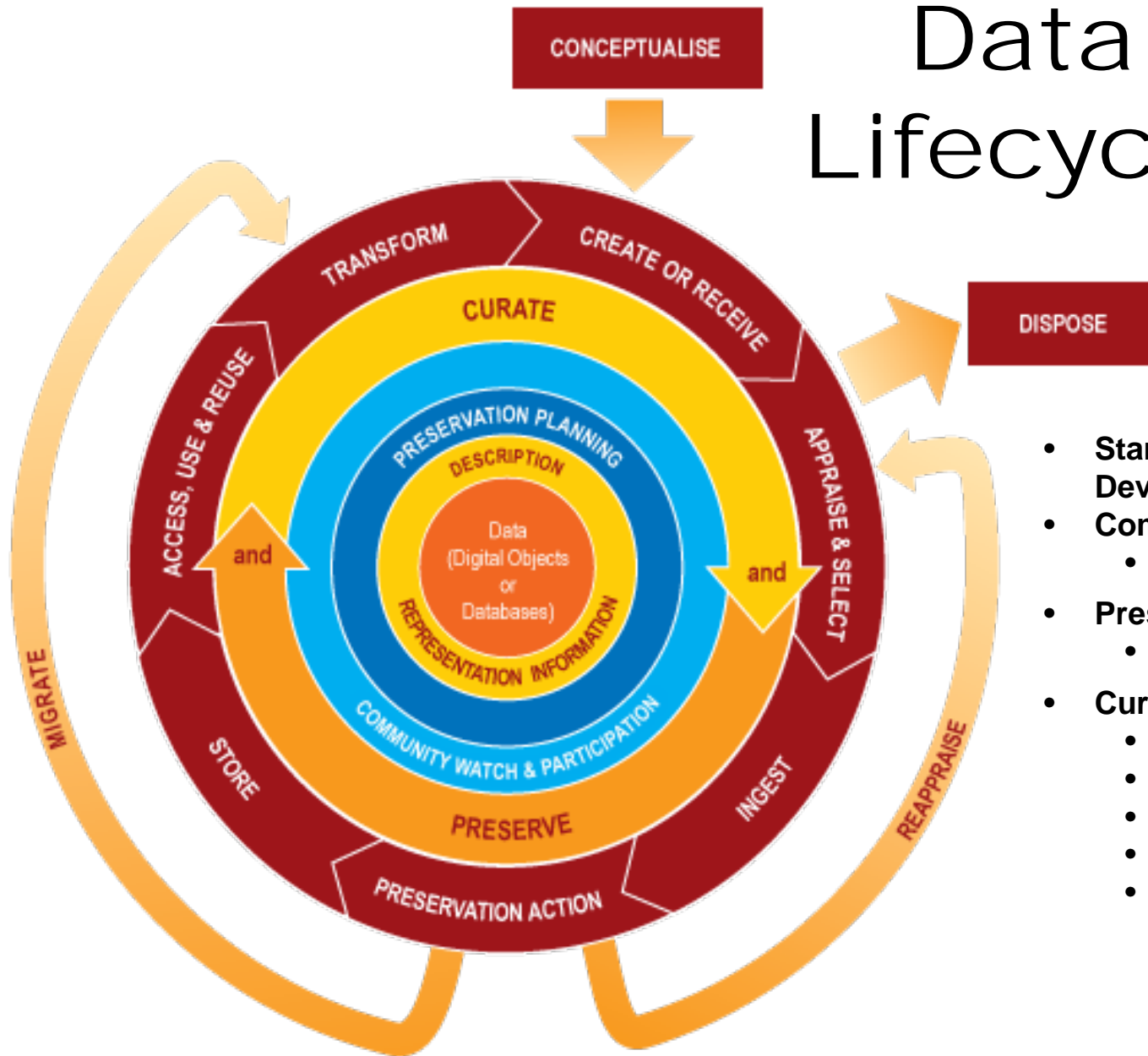
- **Standards & Workflow Development & Implementation**
- **Complete Description**
  - **Metadata Standards**
- **Preservation Planning**
  - **DM Training**
- **Curation & Preservation**
  - **Access & Reuse**
  - **Repository Ingest**
  - **Preservation**
  - **Migration**
  - **Disposition**

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

1. University Library, Texas A&M University. "Data Management Defined- Research Data Management - Guides at Texas A&M University." Research Data Management, October 1, 2013. http://guides.library.tamu.edu/DataManagement
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
3. Definition based on Ani Adhikari and John DeNero, "The Foundations of Data Science" http://www.inferentialthinking.com/index.html "What is Data Science" http://www.inferentialthinking.com/chapter1/what-is-data-science.html
4. Adapted from: W3C. 2016. "Data on the Web Best Practices." https://www.w3.org/TR/dwbp/
5. Repository Audit and Certification DSA–WDS Partnership WG Recommendations https://www.rd-alliance.org/group/repository-audit-and-certification-dsa%E2%80%93wds-partnership-wg/outcomes/dsa-wds-partnership
6. Digital Curation Centre. Data Curation Lifecycle Model. http://www.dcc.ac.uk/resources/curation-lifecycle-model