Delivering Data Packages for Discovery, Analysis, and Preservation

U.S. Department of Transportation
National Transportation Library

Research Data Alliance
11th Plenary
Berlin, Germany,
March 21-23, 2018

The United States Department of Transportation (USDOT) Plan to Increase Public Access to the Results of Federally-Funded Scientific Research (PA) requires, in part, "digitally formatted scientific data resulting from unclassified research supported wholly or in part by Federal funding to be stored and publicly accessible for search, retrieval, and analysis." The PA goes on to require that researchers deliver a Data Management Plan (DMP) which identifies practices they will employ to ensure the long-term access and preservation of the project data.

The National Transportation Library (NTL), of USDOT's Bureau of Transportation Statistics (BTS), is applying the standards and practices required by the PA to new datasets created by BTS. Further, in order to ensure the greatest possible longevity of discovery and preservation, as well as encouraging data interoperability and reuse, NTL is helping BTS staff to create even more robust documentation for datasets. This documentation is collectively known as a "data package." In addition to the final dataset, and the DMP, a data package includes other documentation, which as defined by NTL, is "needed to contextualize the dataset for any and all users."

This poster will explore the various elements of data packages, and look at their initial use within BTS and NTL, using the 2016 National Census of Ferry Operators dataset as an example.

What is a "Data Package"?

A "Data Package" is the dataset, the data management plan (DMP), and all other documentation needed to contextualize the dataset for any and all users and re-users.

At the National Transportation Library, a data packages is comprised of 4 required files - or elements - and 2 (or more) optional elements. Each plays a different role in contextualizing a dataset.

Look in the back of truck for the inventory of Data Package Elements.

NTL Dataset Data Package Elements

1) **Dataset**
   - .csv or other open format
2) **Readme.txt**
   - Includes Data Dictionary
   - Notes standards used
   - Defining Zero, Null, and Unknown
   - FAQs and other notes
3) **Metadata file**
   - in Project Open Data .json
4) **Data Management Plan (DMP)**
5) *Code or scripts* used in data analysis
6) *Supporting files, tables, etc.*

(**Bold** = Required; *Italics* = Optional, or Required if Applicable)

Element: Data

bts_osp_national_census_ferry_operators_2016_Segment_2017_10_26.csv - Notepad

File Edit Format View Help

```
SEGMENT_ID,SEGMENT_NAME,SEG_TERMINAL1_ID,SEG_TERMINAL2_ID,SEG_TYPE,SERVES_NPS,SURVEYYEAR
222,Ketchikan Terminal - Prince Rupert,45,1,3,0,2016
223,Anacortes - Lopez,613,534,1,0,2016
224,Friday Harbor - Orcas,862,541,1,0,2016
226,Lopez - Shaw,534,550,1,0,2016
227,Orcas - Anacortes,541,613,1,0,2016
228,Shaw - Orcas,550,541,1,0,2016
229,Anacortes - Friday Harbor,613,862,1,0,2016
237,Sombra ONT - Marine City,11,295,3,0,2016
239,Victoria Inner Harbor - Black Ball Ferry Line,4,542,3,0,2016
264,"Road Town, Tortola - W. Blyden Marine Terminal",19,492,1,0,2016
265,"West End, Tortola - W. Blyden Marine Terminal",20,492,1,0,2016
276,Enighed Pond - Red Hook,494,495,1,0,2016
312,Catano - Old San Juan Pier Two,447,452,1,0,2016
```

Best data file format for long-term preservation of tabular data: .CSV

- Open, non-proprietary file format
- Backward and forward compatible
- Long-term preservable

Best Practice Tip:

In order for data to be machine readable and interoperable, it should have only two types of rows: 1) a single header row, and 2) all others rows are data.

Element: Metadata

Project Open Data Metadata V1.1

Machine-readable; Required by DOT open data policy; Required by data.gov

bts_osp_national_census_ferry_operators_2016_Metadata_2017_10_26.json - Notepad

File Edit Format View Help

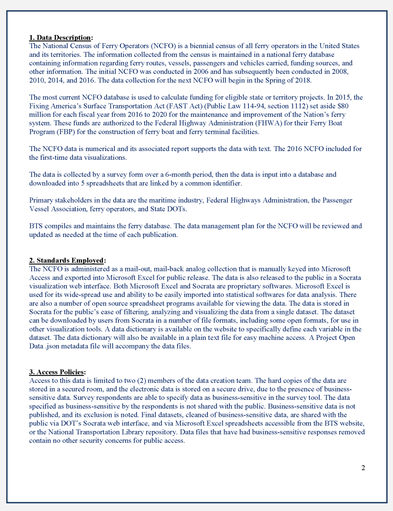
```
[
        {"conformsTo": "https://project-open-data.cio.gov/v1.1/schema"},
        {
                "title":"2016 National Census of Ferry Operators (NCFO)",
                "description":"The 2016 NCFO dataset is comprised of the responses of all operators who completed the 2016 N
                "keyword":"passenger, freight, transit, travel, maritime, terminals, vessels, ferry",
                "modified":"2017-10-06",
                "publisher":"U.S. Department of Transportation, Bureau of Transportation Statistics, Office of Survey Progra
                "contactPoint": {
                                "@type": "vcard:Contact",
                                "fn": "Janine McFadden",
                                "hasEmail": "janine.mcfadden@dot.gov"
                                },
                "identifier":null,
                "accessLevel":"public",
                "accessLevelComment":null,
                "bureauCode":"021:04",
                "programCode":"021:053",
                "accessURL":"https://data.transportation.gov/Public-Transit/2016-National-Census-of-Ferry-Operators-NCFO-/5k
                "webService":null,
                "format":"csv",
                "license":"Public Domain",
                "spatial":"United States",
                "temporal":"2015",
                "accrualPeriodicity":"R/P2Y",
                "landingPage":"https://www.bts.gov/product/highlights-2016-national-census-ferry-operators-ncfo",
```

Element: Data Management Plan

A data management plan (DMP) is a succinct narrative document which describes the deliberate planning, creation, storage, access, and preservation of data produced from a given investigation.

A robust DMP contains 5 sections:

1. **Data Description**
2. **Standards Employed**
3. **Access Policies**
4. **Re-Use, Redistribution, and Derivative Products Policies**
5. **Archiving and Preservation Plans**



1. Data Description:
The National Census of Ferry Operators (NCFO) is a biennial census of all ferry operators in the United States and its territories. The information collected from the census is maintained in a national ferry database containing information regarding ferry routes, vessels, passengers and vehicles carried, funding sources, and other information. The initial NCFO was conducted in 2006 and has subsequently been conducted in 2008, 2010, 2014, and 2016. The data collection for the next NCFO will begin in the Spring of 2018.

The most current NCFO database is used to calculate funding for eligible state or territory projects. In 2015, the Fixing America's Surface Transportation Act (FAST Act) (Public Law 114-94, section 1112) set aside $80 million for each fiscal year from 2016 to 2020 for the maintenance and improvement of the Nation's ferry system. These funds are authorized to the Federal Highway Administration (FHWA) for their Ferry Boat Program (FHP) for the construction of ferry boat and ferry terminal facilities.

The NCFO data is numerical and its associated report supports the data with text. The 2016 NCFO included for the first-time data visualizations.

The data is collected by a survey form over a 6-month period, then the data is input into a database and downloaded into 5 spreadsheets that are linked by a common identifier.

Primary stakeholders in the data are the maritime industry, Federal Highways Administration, the Passenger Vessel Association, ferry operators, and State DOTs.

BTS compiles and maintains the ferry database. The data management plan for the NCFO will be reviewed and updated as needed at the time of each publication.

2. Standards Employed:
The NCFO is administered as a mail-out, mail-back analog collection that is manually keyed into Microsoft Access and exported into Microsoft Excel for public release. The data is also released to the public in a Socrata visualization web interface. Both Microsoft Excel and Socrata are proprietary softwares. Microsoft Excel is used for its wide-spread use and ability to be easily imported into statistical softwares for data analysis. There are also a number of open source spreadsheet programs available for viewing the data. The data is stored in Socrata for the public's ease of filtering, analyzing and visualizing the data from a single dataset. The dataset can be downloaded by users from Socrata in a number of file formats, including some open formats, for use in other visualization tools. A data dictionary is available on the website to specifically define each variable in the dataset. The data dictionary will also be available in a plain text file for easy machine access. A Project Open Data .json metadata file will accompany the data files.

3. Access Policies:
Access to this data is limited to two (2) members of the data creation team. The hard copies of the data are stored in a secured room, and the electronic data is stored on a secure drive, due to the presence of business-sensitive data. Survey respondents are able to specify data as business-sensitive in the survey tool. The data specified as business-sensitive by the respondents is not shared with the public. Business-sensitive data is not published, and its exclusion is noted. Final datasets, cleaned of business-sensitive data, are shared with the public via DOT's Socrata web interface, and via Microsoft Excel spreadsheets accessible from the BTS website, or the National Transportation Library repository. Data files that have had business-sensitive responses removed contain no other security concerns for public access.

2

Best Practice Tip:

To write a DMP that conforms to the USDOT Public Access Plan, go to the NTL Public Access Guidance website at: https://ntl.bts.gov/publicaccess/

Element: Code Book & Supporting Files

Include code books, scripts used during analysis, auxiliary tables, and other supporting files that were created or used to collect, process, clean, or analyze the data.

Best Practice Tip:

Be as complete as possible. The goals of a robust data package include fully documenting your processes to allow a naïve user to replicate your results and understand the full context of the data.

This will enable them to decide intelligently whether your data meets their reuse needs.

Element: Readme.txt

The purpose of the README document is to provide all of that contextual information that could not be included in the data file because it would break the data.

Data Dictionary Example:

AN. Name: SURVEYYEAR
Full name: Survey Year
Variable Definition or Description: Year of survey.
Data format: YYYY
Field length: 4
Representation of Null values: No Null values present; Not Applicable.

NTL ReadMe.txt Template Contents

A. General Information
B. Sharing/Access Information
C. Data and Related File Overview
D. Methodological Information
E. Data-Specific Information, Data dictionary
F. Appendices

Best Practice Tip:

In order for you and future users to know which of the dozens of "README.txt" files on your desktop go with which datasets, use clarifying, human-readable files names, such as:

bts_osp_national_census_ferry_operators_2016_README_2017_10_26.txt

A Data Package is...

The dataset and all documentation needed to contextualize the dataset for any user.

bts_osp_national_census_ferry_operators_2016_DATA_FlatFile_2017_10_26.csv
bts_osp_national_census_ferry_operators_2016_DATA_Operator_2017_10_26.csv
bts_osp_national_census_ferry_operators_2016_DATA_OperatorSegment_2017_10_26.csv
bts_osp_national_census_ferry_operators_2016_DATA_Segment_2017_10_26.csv
bts_osp_national_census_ferry_operators_2016_DATA_Terminal_2017_10_26.csv
bts_osp_national_census_ferry_operators_2016_DATA_Vessel_2017_10_26.csv
bts_osp_national_census_ferry_operators_2016_DMP_2017_10_26.pdf
bts_osp_national_census_ferry_operators_2016_DMP_2017_10_26.txt
bts_osp_national_census_ferry_operators_2016_METADATA_2017_10_26.json
bts_osp_national_census_ferry_operators_2016_README_2017_10_26.txt

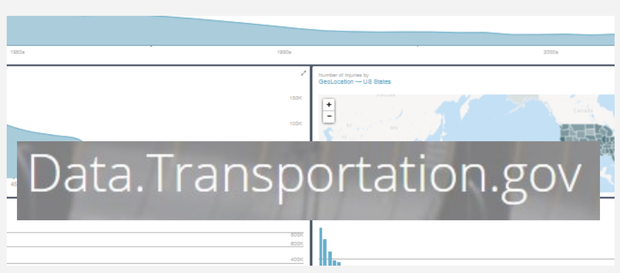
Best Practice Tip:

Use consistent, descriptive, human-readable file names, with date and timestamps to help you keep your data package well organized and up to date.

Where does NTL Deliver Data Package?

National Transportation Library
Repository & Open Science Access Portal
COME SEE
rosap

https://rosap.ntl.bts.gov/



https://data.transportation.gov/

Leighton L Christiansen http://orcid.org/0000-0002-0543-4268

Data Curator, NTL leighton.christiansen@dot.gov @purpleleighton

Recommended Citation: Christiansen, Leighton L http://orcid.org/0000-0002-0543-4268 . 2018. "Delivering Packages for Discovery, Analysis, and Preservation." Research Data Alliance 11th Plenary. Berlin, Germany.

Acknowledgements: The author thanks Alpha Wingfield for design assistance, & the National Transportation Library staff for their input, support, and teamwork.