U.S. Department of Transportation
Office of the Secretary of Transportation

Data Management Strategies for the National Transportation Data Archive: Dealing with Legacy Data

Transportation Research Board 99th Annual Meeting
Washington, D.C., January 12-16, 2020
Poster: P20-20652

Abstract

The National Transportation Library (NTL) is currently working to expand the National Transportation Data Archive (NTDA). Recently, this work has focused on successfully implementing data management strategies for legacy datasets, such as the Omnibus Household Surveys (OHS). Improper data management took place at the time that OHS data was being collected, resulting in a variety of challenges associated with finding, managing, and preserving the data now. Data management is important at all stages of a data project. However, managing legacy data long after collection presents added challenges. These challenges include: sorting through files to locating the data and relevant documentation; deciphering file names; obtaining software to open files; and, migrating files into open access formats. Additionally, other companion documentation files need to be created, such as a data management plan (DMP), Readme file, and metadata file. Finally, datasets need to be assigned persistent identifiers. After all issues are addressed a data package can be created for each dataset. Working with legacy data has reinforced NTL's goal of developing and implementing a standard data management protocol to ensure that steps or proper steps are taken when the data is being created and not after the fact. This poster will review the challenges of managing legacy data after the fact, highlighting our efforts within the NTDA, and offer best practices for – and the benefits of – data management during the lifecycle of the data collection project.

USGS Data Lifecycle Model

Data curators see data as having a "lifecycle." There are several definitions of the data lifecycle. To paraphrase many of them, the data lifecycle is "All the phases of data's existence from planning to collection, through preservation, to reuse and potential destruction" Additionally, there are several models of the data lifecycle, below is the USGS Data Lifecycle. This is one of the simpler versions, which NTL views as more beneficial, because the very first action is planning. A crucial step, and if not completed can result in great difficulties down the road, especially with legacy data.

National Transportation Data Archive

The National Transportation Data Archive (NTDA) is intended to preserve and provide access to high-value, discrete datasets, created or aggregated the by Bureau of Transportation Statistics (BTS). The NTDA also seeks to preserve other transportation research or statistical data sets, created by non-US DOT entities, within the borders of the United States, its territories, and possessions. https://doi.org/10.21949/1504517

Omnibus Survey Program

The Omnibus Surveys were designed as a convenient way to get very quick input on transportation issues, as well as to gauge public satisfaction with the transportation system and government programs. The Omnibus Surveys consisted of two vehicles: (1) a recurring household survey of 1,000 households that collected data on core questions about general travel experiences, satisfaction with the system, and some demographic data; and (2) a series of targeted surveys to address specific transportation issues or domains. The OHS program ran from 2000 to 2009.

Working with Legacy Data

Since the Omnibus Surveys are a legacy dataset that is 10 to 20 years old, and did not have proper data management best practices implemented at the time, this project started with a limited understanding of the survey and files associated with it. This meant that the first step would be to assess what BTS had stored on the survey program.

	A	B	C	D	E
1	Data	Data	Documentation	Supporting Files Found	Data Package Complete
2	2000-08	Y	Y	Summary Tables (also in documentation doc)	Yes
3	2000-09	Y	Y	Summary Tables (also in documentation doc)	Yes
4	2000-10	Y	Y	Summary Tables (also in documentation doc)	Yes
5	2000-11	Y	Y	Summary Tables (also in documentation doc)	Yes
6	2000-12	Y	Y	Summary Tables (also in documentation doc)	Yes
7	2001-01	Y	Y	Summary Tables (also in documentation doc), Data Collection Report	Yes
8	2001-02	Y	Y	Summary Tables (also in documentation doc), Data Collection Report	Yes
9	2001-03	Y	Y	Summary Tables (also in documentation doc), Data Collection Report	Yes
10					
11	2001-07	Y	Y	data dictionary (.xls), survey table results (.xls, .pdf, .doc, .txt), SASLabels (.txt), SASFormat Library (.txt)	Yes
12	2001-08	Y	Y	data dictionary (.xls), survey table results (.xls, .pdf, .doc, .txt), SASLabels (.txt), SASFormat Library (.txt)	Yes
13	2001-10	Y	Y	data dictionary (.xls), survey table results (.xls, .pdf, .doc, .txt), SASLabels (.txt), SASFormat Library (.txt)	Yes
14	2001-11	Y	Y	data dictionary (.xls), survey table results (.xls, .pdf, .doc, .txt), SASLabels (.txt), SASFormat Library (.txt)	Yes
15	2001-12	Y	Y	data dictionary (.xls), survey table results (.xls, .pdf, .doc, .txt), SASLabels (.txt), SASFormat Library (.txt)	Yes
16	2002-01	Y	Y	data dictionary (.xls), survey table results (.xls, .pdf, .doc, .txt), SASLabels (.txt), SASFormat Library (.txt)	Yes
17	2002-02	Y	Y	data dictionary (.xls), survey table results (.xls, .pdf, .doc, .txt), SASLabels (.txt), SASFormat Library (.txt)	Yes
18	2002-03	Y	Y	data dictionary (.xls), survey table results (.xls, .pdf, .doc, .txt), SASLabels (.txt), SASFormat Library (.txt)	Yes
19	2002-04	Y	Y	data dictionary (.xls), survey table results (.xls, .pdf, .doc, .txt), SASLabels (.txt), SASFormat Library (.txt)	Yes
20	2002-05	Y	Y	data dictionary (.xls), survey table results (.xls, .pdf, .doc, .txt), SASLabels (.txt), SASFormat Library (.txt)	Yes
21	2002-06	Y	Y	data dictionary (.xls), survey table results (.xls, .pdf, .doc, .txt), SASLabels (.txt), SASFormat Library (.txt)	Yes
22	2002-07	Y	Y	data dictionary (.xls), survey table results (.xls, .pdf, .doc, .txt), SASLabels (.txt), SASFormat Library (.txt)	Yes
23	2002-08	Y	Y	data dictionary (.xls), survey table results (.xls, .pdf, .doc, .txt), SASLabels (.txt), SASFormat Library (.txt)	Yes
24	2002-09	Y	Y	data dictionary (.xls), survey table results (.xls, .pdf, .doc, .txt), SASLabels (.txt), SASFormat Library (.txt)	Yes
25	2002-10	Y	Y	data dictionary (.xls), survey table results (.xls, .pdf, .doc, .txt), SASLabels (.txt), SASFormat Library (.txt)	Yes
26	2002-12	Y	Y	data dictionary (.xls), survey table results (.xls, .pdf, .doc, .txt), SASLabels (.txt), SASFormat Library (.txt)	Yes
27	2003-02	Y	Y	data dictionary (.xls), survey table results (.xls, .pdf, .doc, .txt), SASLabels (.txt), SASFormat Library (.txt)	Yes
28	2003-04	Y	Y	data dictionary (.xls), survey table results (.xls, .pdf, .doc, .txt), SASLabels (.txt), SASFormat Library (.txt)	Yes
29	2003-06	Y	Y	data dictionary (.xls), survey table results (.xls, .pdf, .doc, .txt), SASLabels (.txt), SASFormat Library (.txt)	Yes
30	2003-08	Y	Y	data dictionary (.xls), survey table results (.xls, .pdf, .doc, .txt), SASLabels (.txt), SASFormat Library (.txt)	Yes
31	2003-10	Y	Y	data dictionary (.xls), survey table results (.xls, .pdf, .doc, .txt), SASLabels (.txt), SASFormat Library (.txt)	Yes
32					
33	2004-12	N	N	Nothing found	
34	2005-10	Y	N	data dictionary (.xls), survey table results (.txt), SASLabels (.txt), SASFormat Library (.txt), List of Variables to BTS (.xls)	
35	2006-11	Y	Y	data dictionary (.xls), survey results/frequency tables-also in stats format (.txt, .xls), SASFormat Library (.txt), SAS Labels (.txt), BTS Special Report for OHS 2006-2007 data (.pdf)	
36	2007-11	N	Y	data dictionary (.doc),data collection plan (.doc), stats for response rates (.xls), BTS Special Report for OHS 2006-2007 data (.pdf), BTS Special Report for OHS 2007-2008 data (.pdf)	
37	2008-11	Y?	Y	report tables (.xls), data dictionary (.xls), BTS Special Report for OHS 2007-2008 data (.pdf), BTS Special Report for OHS 2008 data (.pdf)	
38	2009-10	Y	Y	Survey results/Frequency tables for both NAT and MSA data (.xls)	Yes

Common Issues

Inconsistent or unclear file names:

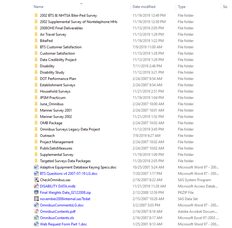
+ A common issue found with legacy data is an undefined and simplistic naming structure. I have found, in multiple cases that names tend to be both general and repeatedly used. For example, a file that I came across for OHS was simply named "disposition." Do to this basic name and limited documentation I was unable to understand the purpose of the data within this file, and was further confused since the word "disposition" was used in the documentation to reference various variables. In the end, I spent wasted time comparing the file with the data dictionary and main dataset, until I was finally able to piece together a purpose.

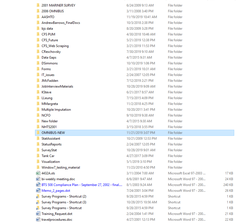
Various locations of files:

+ P:\Omnibus Surveys
+ P:\Survey Programs\OMNIBUS-NEW
+ P:\pchandhok
+ P:\MSchiro\backup\bts\programs\omnibus_surveys\household_survey\2009\original_files
+ P:\Web Team Data\WebFarmRITASiteArchive\FromAsteroid\wwwrita\bts\public\programs\omnibus_surveys

Limited Documentation:

+ Not only does the lack of documentation affect our current understanding of the data, but can also lead to the other issues mentioned above. Having robust documentation is major part of ensuring data management and preservation can be continued successfully for years in the future.





Plan → Acquire → Process → Analyze → Preserve → Publish/Share

Authors

Jesse A Long
https://orcid.org/0000-0002-4962-1380.
Data Management and Data Curation Fellow, National Transportation Library
Jesse.long.ctr@dot.gov

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

Recommended Citation

Long, Jesse A. https://orcid.org/0000-0002-4962-1380. 2020. "Data Management Strategies for the National Transportation Data Archive: Dealing with Legacy Data." Transportation Research Board 99th Annual Meeting. Washington, D.C., USA. https://doi.org/10.21949/1506098

Data Package

Key Elements

1. Dataset
2. Readme.txt
3. Metadata file in Project Open Data .json
4. Data Management Plan (DMP)

Optional

5. Code or scripts used in data analysis
6. Supporting files, tables, etc.

For more information on data packages please check out a past poster at TRB, which can be found in ROSAP:

Christiansen, Leighton L. http://orcid.org/0000-0002-0543-4268. 2018. "Delivering Data Packages for Discovery, Analysis, and Preservation." Transportation Research Board 97th Annual Meeting. Washington, D.C., USA https://doi.org/10.21949/1500456

Preventing these Issues

By implementing data management and sharing plans and embedding data curators into data collection teams we can ensure proper data packages are created efficiently and comprehensive, eliminating the time consuming issues and gaps that have been identified with legacy data.

Data Management and Sharing Plans:

Planning before data collection is arguably the most important step, and arguably why it sits as the first step in the USGS Data Lifecycle. During the Planning step many things needed to be determined, such as:

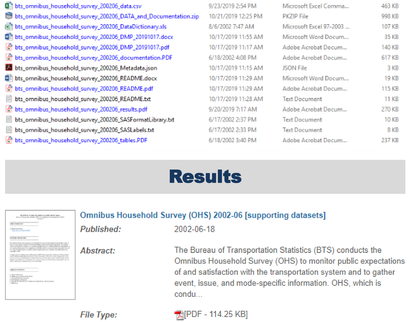
- What data is going to be collected?
- How data will be collected?
- What types of data will be collected?
- What file types of data can be expected?
- How will the data be organized? Who will be responsible for data?
- When will backups occur? Where will backups reside?
- What size of data is expected?
- Will there be sensitive data collected and if so how will it be handled?
- Whether and how much data will be shared?

Only after all of these questions are answered, can data start being collected. By creating a Data Management or Sharing Plan all following steps in the lifecycle will progress successfully and attribute to data package creation. Additionally, this plan should be thought of as a living document and updated as needed throughout the project.

Embedded Data Curators:

It is NTL's goal to have embedded data curators to implement suggestions and achieve greater transparency when it comes to BTS data. NTL believes if the goal is to gather a team of professionals best able to carry out a data collection project, analyze data into statistics, and document, preserve, and share the statistics, the team deserves a trained, professional data curator.

- Data curators possess technical and research skills other team members won't have, but will contribute directly to data and statistical transparency.
- Data curators, can serve as fresh eyes on repeated data collection projects and make explicit knowledge that is implicit and "obvious" to the team.
- Data curators work under the assumption that data should be shared, while remaining aware of data sensitivity.
- Data curation practices will improve team efficiency around sharing, preservation, and transparency, by default.
- Curators take a lifecycle view of the data, and can relieve other team members of that duty.
- Data curators can also plan for end of data lifecycle events and disposition, in ways consistent with established best practices.



Results

Omnibus Household Survey (OHS) 2002-06 [supporting datasets]

Published: 2002-06-18

Abstract: The Bureau of Transportation Statistics (BTS) conducts the Omnibus Household Survey (OHS) to monitor public expectations of and satisfaction with the transportation system and to gather event, issue, and mode-specific information. OHS, which is condu...

File Type: PDF - 114.25 KB