Increasing data reusability through Tidy Data form

2020-12-10

Francisco Juarez
https://orcid.org/0000-0001-5463-9596

Presentation to NTL and BTS

Intro

- Graduate Student
- Library and Information Science
- The University of Illinois at Urbana-Champaign

- Graduate Assistant
  Grainger Engineering Library

About the National Transportation Library

Transportation Equity Act for the 21st Century (TEA-21) (1998)

- “establish and maintain a National Transportation Library, which shall contain a collection of statistical and other information needed for transportation decision making at the Federal, State, and local levels.”

Moving Ahead for Progress in the 21st Century Act (MAP-21) (2012)

- Acquire, preserve and manage transportation information and information products and services for use by DOT, other Federal agencies, and the public;
- Serve as the central repository for DOT research results and technical publications; and,
- Serve as the central clearinghouse for transportation data and information of the Federal Government

White House Office of Science and Technology Policy memo (2103)

- requiring all Executive Departments and Agencies spending more than $100 million/year on R&D to ensure public access to peer-reviewed publications and digital datasets arising from federally-funded scientific research

Foundations for Evidence-Based Policymaking Act of 2018, Title II: The Open, Public, Electronic, and Necessary Government Data Act (OPEN Government Data Act)

Agenda

Government Transportation Financial Statistics(GTFS)

- Series of reports covering transportation-related financial actives of all US government levels

- The government plays an important role in the U.S. transportation system, as a provider of transportation infrastructure and as an administrator and regulator of

Table 1-A

Summary of Federal, State, and Local Transportation Finance by Mode: FY 1985-1999

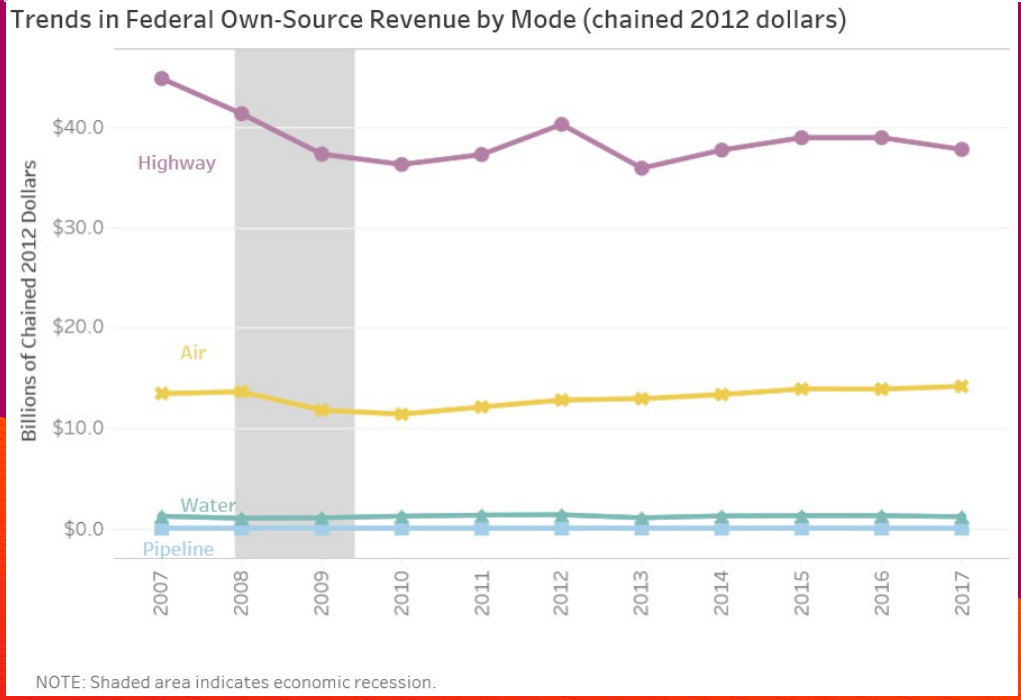
(Current $ millions)

Mode	1985	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999
Revenues															
Total	**52,140**	**54,860**	**58,531**	**62,864**	**67,778**	**69,753**	**77,392**	**80,326**	**85,197**	**87,632**	**93,659**	**96,419**	**100,516**	**111,234**	**126,895**
Highway	38,166	40,230	42,455	46,040	49,457	49,945	53,838	57,780	60,465	62,316	66,743	71,179	71,814	77,299	88,668
Transit	5,636	5,848	6,346	6,428	6,764	7,193	8,778	7,482	8,570	8,948	9,352	10,171	11,417	11,872	13,186
Air	6,711	7,019	7,765	8,190	9,369	10,119	11,924	11,872	12,744	13,101	13,954	11,298	13,544	18,176	21,079
Water	1,626	1,761	1,956	2,198	2,178	2,487	2,840	3,174	3,393	3,242	3,567	3,733	3,704	3,850	3,923
Pipeline	–	–	9	9	10	10	11	14	15	19	35	31	30	29	30
General Support	–	–	–	–	–	–	–	3	10	7	7	7	7	8	8
Expenditures															
Total	**77,230**	**83,856**	**89,457**	**90,612**	**94,766**	**100,629**	**108,284**	**114,587**	**116,461**	**125,882**	**130,542**	**133,359**	**138,361**	**145,659**	**154,845**
Highway	46,604	50,435	54,032	57,361	59,854	62,563	66,526	68,954	69,991	74,531	79,309	81,550	84,212	89,454	95,494
Transit	16,333	17,586	19,321	16,827	17,594	19,261	20,857	22,322	21,279	25,088	26,162	26,346	26,875	28,108	29,027
Rail	1,072	917	817	586	606	541	783	906	819	845	1,043	1,015	1,148	1,099	565
Air	7,903	8,749	9,540	10,422	11,240	12,568	13,974	15,916	17,408	17,941	16,960	17,273	18,776	19,593	21,789
Water	5,124	5,974	5,601	5,245	5,289	5,480	5,847	6,167	6,593	7,046	6,628	6,775	6,996	7,137	7,682
Pipeline	8	4	4	9	15	26	28	32	34	36	43	33	29	32	30
General Support	187	193	143	163	168	191	270	289	337	396	396	367	327	236	258

KEY: "–" = No activity or a value of zero.

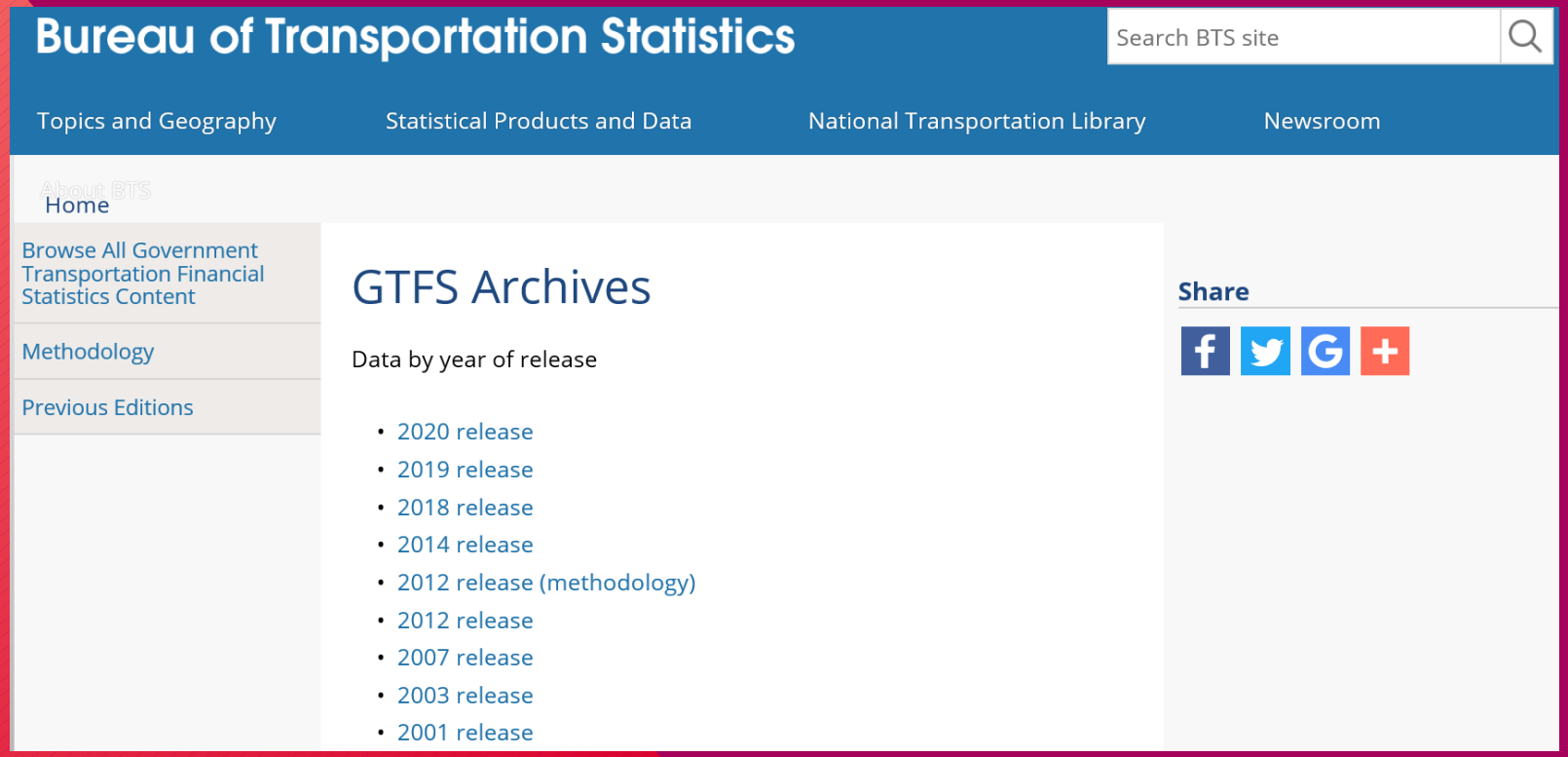
NOTES:
Numbers may not add to totals due to rounding.
For FY 1996 - 1999, state and local pipeline expenditures are not included due to lack of data.

SOURCES:
Federal Revenues:
Highways and Transit:
U.S. Department of Transportation, Federal Highway Administration, *Highway Statistics* (Washington, D.C.: Annual issues), Tables FE-210 (Historical Data).



Useful for viewing trends over the years

Government Transportation Financial Statistics (GTFS) datasets



Goal:

- Archive all but the current GTFS report into ROSA P instead of BTS.gov

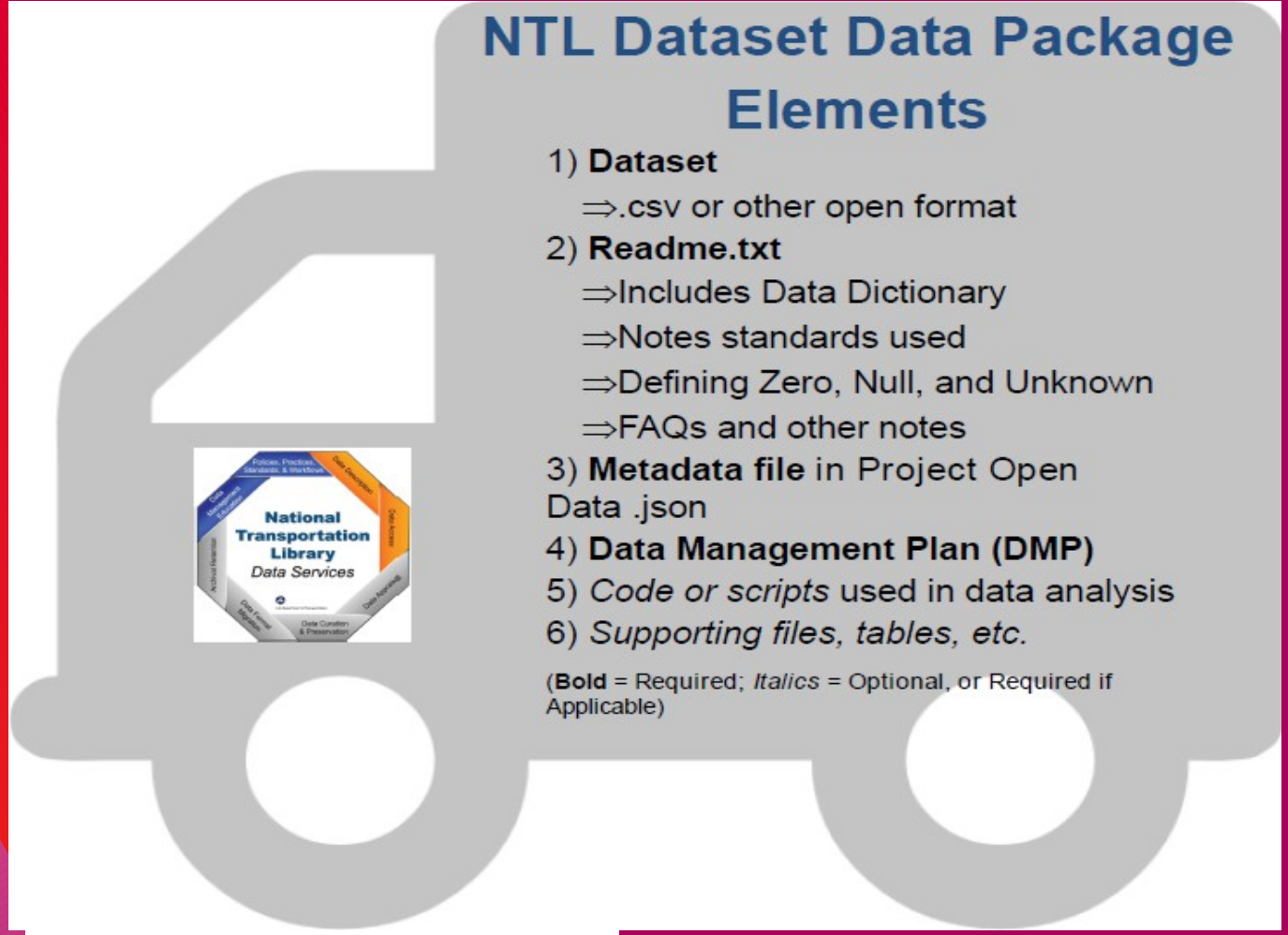
- All archive GTFS datasets to have accompanying data packages

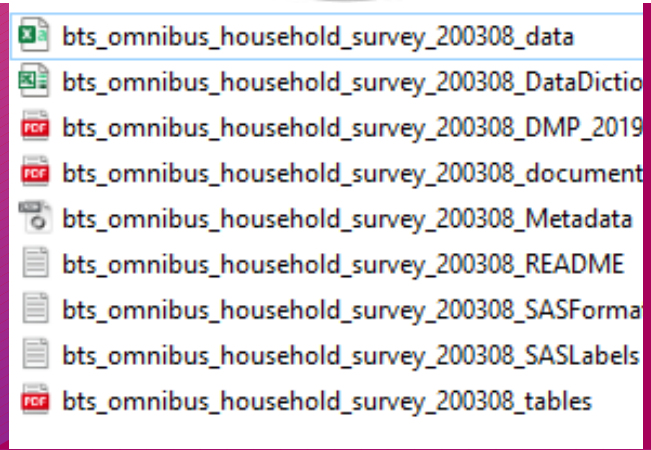
Currently all GTFS report are stored in a bts.gov landing page

Data Package

What is a “Data Package”

- Dataset
- Data Management Plan
- All other documentation needed to contextualize the data set for all users and re-users

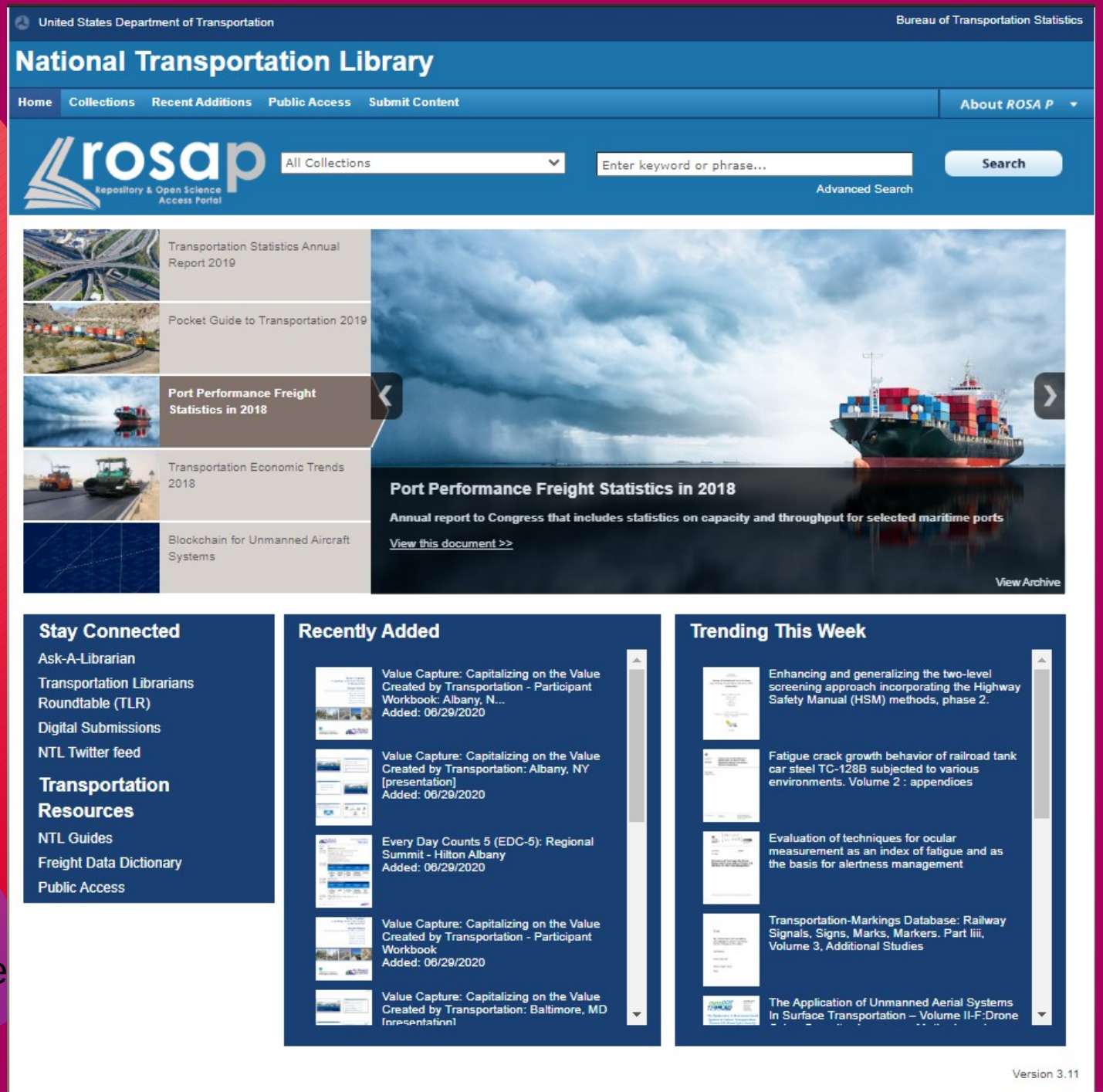




Repository & Open Science Access Portal (ROSA P)

- *ROSA P* is the National Transportation Library's *Repository and Open Science Access Portal*. The name *ROSA P* was chosen to honor the role public transportation plays in the civil rights movement, along with one of the important figures

Visit ROSA P at: https://rosap.ntl.bts.gov/we



Challenges

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Table 3-A A Summary of Transportation Revenue and Expenditure													
2	Thousands of Current Dollars													
3		1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007
4	Revenue Allocated for Transportation, Total	#######	#######	#######	#######	#######	#######	#######	#######	#######	#######	#######	#######	#######
5	Highway	#######	#######	#######	#######	#######	#######	#######	#######	#######	#######	#######	#######	#######
6	Transit	#######	#######	#######	#######	#######	#######	#######	#######	#######	#######	#######	#######	#######
7	Railroads	#######	#######	#######	#######	435,000	759,000	735,000	#######	#######	#######	#######	#######	#######
8	Air	#######	#######	#######	#######	#######	#######	#######	#######	#######	#######	#######	#######	#######
9	Water	#######	#######	#######	#######	#######	#######	#######	#######	#######	#######	#######	#######	#######
10	Pipeline	57,000	65,000	61,000	63,000	64,000	77,000	74,000	100,000	118,000	116,000	123,000	132,000	136,000
11	General Support	430,000	386,000	354,000	265,000	263,000	321,000	325,000	#######	#######	#######	#######	#######	#######
12	Expenditure, Total	#######	#######	#######	#######	#######	#######	#######	#######	#######	#######	#######	#######	#######
13	Highway	#######	#######	#######	#######	#######	#######	#######	#######	#######	#######	#######	#######	#######
14	Transit	#######	#######	#######	#######	#######	#######	#######	#######	#######	#######	#######	#######	#######
15	Railroads	#######	#######	#######	#######	452,900	778,300	753,300	#######	#######	#######	#######	#######	#######
16	Air	#######	#######	#######	#######	#######	#######	#######	#######	#######	#######	#######	#######	#######
17	Water	#######	#######	#######	#######	#######	#######	#######	#######	#######	#######	#######	#######	#######
18	Pipeline	24,000	34,000	33,000	36,000	38,000	46,000	37,000	48,000	65,000	73,000	82,000	91,000	89,000
19	General Support	774,539	715,518	697,602	599,707	631,854	652,820	837,729	#######	#######	#######	#######	#######	#######
20	Notes: Local government receipts and outlays for highway are not included in 2007.													
21	Revenues and expenditures in a given year may differ, due to funding of transportation expenditures out of general funds on the one hand, and the flow of transportation revenues int													
22	Sources:													
23	Executive Office of the President of the United States, Office of Management and Budget, "Budget of the United States Government: Analytical Perspective," Detailed Functional Table													
24	U.S. Department of Transportation (USDOT),													
25	U.S. Department of Transportation (USDOT),													
26														
27	U.S. Department of Commerce, Census													
28	Revenues and expenditures in a given year may differ, due to funding of transportation expenditures out of general funds on the one hand, and the flow of transportation revenues int													
29														

__init__.py | BTS_GTFS_2007_table_03a_intermediate.csv | BTS_GTFS_2007_table_03a.csv | GTFS_Tidy.py

```
Table 3-A   A Summary of Transportation Revenue and Expenditure,,,,,,,,,,,,
Thousands of Current Dollars,,,,,,,,,,,,
,1995,1996,1997,1998,1999,2000,2001,2002,2003,2004,2005,2006,2007
"Revenue Allocated for Transportation, Total","146,955,080 ","152,399,220 ","158,747,257 ","172,201,688 ","198,756,772 ","193,734,554 ","
Highway,"94,249,469 ","101,630,138 ","104,517,620 ","111,277,807 ","127,563,761 ","129,555,925 ","128,209,694 ","134,686,098 ","135,837,9
Transit,"22,894,971 ","23,014,200 ","23,851,278 ","24,547,037 ","32,044,491 ","29,908,966 ","34,306,593 ","34,422,484 ","36,153,416 ","38
Railroads,"1,080,000 ","1,010,000 ","1,137,000 ","1,080,000 ","435,000 ","759,000 ","735,000 ","1,308,000 ","1,219,000 ","1,513,000 ","1,
Air ,"20,961,769 ","19,066,511 ","21,508,074 ","27,547,567 ","30,584,123 ","25,431,972 ","34,303,744 ","37,185,557 ","35,373,368 ","37,57
Water,"7,281,871 ","7,227,371 ","7,318,285 ","7,421,277 ","7,802,396 ","7,680,691 ","8,490,305 ","7,898,825 ","9,544,095 ","10,879,806 ",
Pipeline,"57,000 ","65,000 ","61,000 ","63,000 ","64,000 ","77,000 ","74,000 ","100,000 ","118,000 ","116,000 ","123,000 ","132,000 ","13
General Support,"430,000 ","386,000 ","354,000 ","265,000 ","263,000 ","321,000 ","325,000 ","1,481,000 ","10,038,000 ","1,049,000 ","1,0
"Expenditure, Total","143,255,637 ","149,133,221 ","155,953,603 ","163,543,521 ","182,318,129 ","186,374,184 ","211,179,952 ","223,807,79
Highway,"90,075,441 ","94,745,874 ","98,397,988 ","103,987,699 ","112,258,778 ","119,910,746 ","127,103,943 ","133,672,268 ","138,614,597
Transit,"25,459,587 ","26,113,438 ","27,858,489 ","28,989,888 ","39,169,559 ","34,827,547 ","38,988,965 ","41,603,545 ","41,482,031 ","44
Railroads,"1,048,800 ","1,027,900 ","1,164,100 ","1,100,200 ","452,900 ","778,300 ","753,300 ","1,324,300 ","1,242,000 ","1,533,000 ","1,
Air ,"19,250,284 ","19,769,728 ","20,694,088 ","21,732,265 ","22,066,270 ","22,525,014 ","32,838,620 ","37,025,402 ","34,184,510 ","39,17
Water,"6,622,987 ","6,726,762 ","7,108,336 ","7,097,762 ","7,700,768 ","7,633,757 ","10,620,395 ","8,037,742 ","11,774,611 ","10,904,772
Pipeline,"24,000 ","34,000 ","33,000 ","36,000 ","38,000 ","46,000 ","37,000 ","48,000 ","65,000 ","73,000 ","82,000 ","91,000 ","89,000
General Support,"774,539 ","715,518 ","697,602 ","599,707 ","631,854 ","652,820 ","837,729 ","2,096,538 ","10,729,699 ","1,240,363 ","1,1
Notes: Local government receipts and outlays for highway are not included in 2007.,,,,,,,,,,,,
"Revenues and expenditures in a given year may differ, due to funding of transportation expenditures out of general funds on the one hand
"Sources:
Executive Office of the President of the United States, Office of Management and Budget (OMB), ""Budget of the United States Government:
Executive Office of the President of the United States, Office of Management and Budget (OMB), ""Budget of the United States Government:
"Executive Office of the President of the United States, Office of Management and Budget, ""Budget of the United States Government: Analy
"U.S. Department of Transportation (USDOT), Federal Highway Administration (FHWA), ""Highway Statistics,"" Washington, D.C., Annual issue
```

```python
with open("BTS_GTFS_2007_table_03a.csv", 'r') as f:
    contents = csv.reader(f)
    header = next(contents)

print(header)
```

```
Run:  __init__

['Table 3-A   A Summary of Transportation Revenue and Expenditure', '', '', '', '', '', '', '', '', '', '', '', '', '']

Process finished with exit code 0
```

Journal of Statistical Software
MMMMMM YYYY, Volume VV, Issue II. http://www.jstatsoft.org/

Tidy Data

Hadley Wickham
RStudio

Abstract

A huge amount of effort is spent cleaning data to get it ready for analysis, but there has been little research on how to make data cleaning as easy and effective as possible. This paper tackles a small, but important, component of data cleaning: data tidying. Tidy datasets are easy to manipulate, model and visualise, and have a specific structure: each variable is a column, each observation is a row, and each type of observational unit is a table. This framework makes it easy to tidy messy datasets because only a small set of tools are needed to deal with a wide range of un-tidy datasets. This structure also makes it easier to develop tidy tools for data analysis, tools that both input and output tidy datasets. The advantages of a consistent data structure and matching tools are demonstrated with a case study free from mundane data manipulation chores.

Keywords: data cleaning, data tidying, relational databases, R.

1. Introduction

It is often said that 80% of data analysis is spent on the process of cleaning and preparing the data (Dasu and Johnson 2003). Data preparation is not just a first step, but must be repeated many over the course of analysis as new problems come to light or new data is collected. Despite the amount of time it takes, there has been surprisingly little research on how to clean data well. Part of the challenge is the breadth of activities it encompasses: from outlier checking, to date parsing, to missing value imputation. To get a handle on the problem, this paper focusses on a small, but important, aspect of data cleaning that I call data **tidying**: structuring datasets to facilitate analysis.

The principles of tidy data provide a standard way to organise data values within a dataset. A standard makes initial data cleaning easier because you don't need to start from scratch and reinvent the wheel every time. The tidy data standard has been designed to facilitate initial exploration and analysis of the data, and to simplify the development of data analysis tools that work well together. Current tools often require translation. You have to spend time

Wickham H., (2014). Tidy Data *The Journal of Statistical Software*, vol. 59(10). Retrieved December 9, 2020, from https://vita.had.co.nz/papers/tidy-data.html

Tidying: structuring datasets to facilitate analysis.

An example of a *messy dataset*:

	Treatment A	Treatment B
John Smith	-	2
Jane Doe	16	11
Mary Johnson	3	1

An example of a *tidy dataset*:

Name	Treatment	Result
John Smith	a	-
Jane Doe	a	16
Mary Johnson	a	3
John Smith	b	2
Jane Doe	b	11
Mary Johnson	b	1

Defining Tidy data

Five common problems with messy datasets

row	a	b	c
A	1	4	7
B	2	5	8
C	3	6	9

(a) Raw data

row	column	value
A	a	1
B	a	2
C	a	3
A	b	4
B	b	5
C	b	6
A	c	7
B	c	8
C	c	9

(b) Molten data

Table 5: A simple example of melting. (a) is melted with one colvar, row, yielding the molten dataset (b). The information in each table is exactly the same, just stored in a different way.

Column headers are values, not variable names

Table 3-A A Summary of Transportation Revenue and Expenditure

Share

Thousands of Current Dollars

Excel | CSV

	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007
Revenue Allocated for Transportation, Total	**146,955,080**	**152,399,220**	**158,747,257**	**172,201,688**	**198,756,772**	**193,734,554**	**206,444,335**	**217,081,964**	**228,283,845**	**232,441,810**	**247,557,200**	**265,361,007**	**239,487,152**
Highway	94,249,469	101,630,138	104,517,620	111,277,807	127,563,761	129,555,925	128,209,694	134,686,098	135,837,966	143,038,823	152,064,167	163,933,766	135,719,615
Transit	22,894,971	23,014,200	23,851,278	24,547,037	32,044,491	29,908,966	34,306,593	34,422,484	36,153,416	38,269,294	39,366,507	42,068,496	41,827,118
Railroads	1,080,000	1,010,000	1,137,000	1,080,000	435,000	759,000	735,000	1,308,000	1,219,000	1,513,000	1,451,000	2,299,000	1,506,000
Air	20,961,769	19,066,511	21,508,074	27,547,567	30,584,123	25,431,972	34,303,744	37,185,557	35,373,368	37,575,887	41,501,558	42,793,993	45,804,803
Water	7,281,871	7,227,371	7,318,285	7,421,277	7,802,396	7,680,691	8,490,305	7,898,825	9,544,095	10,879,806	11,988,968	12,845,752	13,272,615
Pipeline	57,000	65,000	61,000	63,000	64,000	77,000	74,000	100,000	118,000	116,000	123,000	132,000	136,000
General Support	430,000	386,000	354,000	265,000	263,000	321,000	325,000	1,481,000	10,038,000	1,049,000	1,062,000	1,288,000	1,221,000
Expenditure, Total	**143,255,637**	**149,133,221**	**155,953,603**	**163,543,521**	**182,318,129**	**186,374,184**	**211,179,952**	**223,807,795**	**238,092,447**	**237,636,148**	**243,085,735**	**257,226,420**	**221,707,245**
Highway	90,075,441	94,745,874	98,397,988	103,987,699	112,258,778	119,910,746	127,103,943	133,672,268	138,614,597	140,073,077	147,324,680	157,613,045	114,252,805
Transit	25,459,587	26,113,438	27,858,489	28,989,888	39,169,559	34,827,547	38,988,965	41,603,545	41,482,031	44,636,365	41,899,585	44,096,646	48,750,370
Railroads	1,048,800	1,027,900	1,164,100	1,100,200	452,900	778,300	753,300	1,324,300	1,242,000	1,533,000	1,472,000	1,548,000	1,528,000
Air	19,250,284	19,769,728	20,694,088	21,732,265	22,066,270	22,525,014	32,838,620	37,025,402	34,184,510	39,175,571	40,857,654	41,872,912	43,790,938
Water	6,622,987	6,726,762	7,108,336	7,097,762	7,700,768	7,633,757	10,620,395	8,037,742	11,774,611	10,904,772	10,307,111	10,888,262	12,068,846
Pipeline	24,000	34,000	33,000	36,000	38,000	46,000	37,000	48,000	65,000	73,000	82,000	91,000	89,000

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	Transportation_Mode	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	Revenue_Expenditure	
2	Highway	########	########	########	########	########	########	########	########	########	########	########	########	########	Revenue	
3	Transit	########	########	########	########	########	########	########	########	########	########	########	########	########	Revenue	
4	Railroads	1,080,000	1,010,000	1,137,000	1,080,000	435,000	759,000	735,000	1,308,000	1,219,000	1,513,000	1,451,000	2,299,000	1,506,000	Revenue	
5	Air	########	########	########	########	########	########	########	########	########	########	########	########	########	Revenue	
6	Water	7,281,871	7,227,371	7,318,285	7,421,277	7,802,396	7,680,691	8,490,305	7,898,825	9,544,095	########	########	########	########	Revenue	
7	Pipeline	57,000	65,000	61,000	63,000	64,000	77,000	74,000	100,000	118,000	116,000	123,000	132,000	136,000	Revenue	
8	General Support	430,000	386,000	354,000	265,000	263,000	321,000	325,000	1,481,000	########	1,049,000	1,062,000	1,288,000	1,221,000	Revenue	
9	Highway	########	########	########	########	########	########	########	########	########	########	########	########	########	Expenditure	
10	Transit	########	########	########	########	########	########	########	########	########	########	########	########	########	Expenditure	
11	Railroads	1,048,800	1,027,900	1,164,100	1,100,200	452,900	778,300	753,300	1,324,300	1,242,000	1,533,000	1,472,000	1,548,000	1,528,000	Expenditure	
12	Air	########	########	########	########	########	########	########	########	########	########	########	########	########	Expenditure	
13	Water	6,622,987	6,726,762	7,108,336	7,097,762	7,700,768	7,633,757	########	8,037,742	########	########	########	########	########	Expenditure	
14	Pipeline	24,000	34,000	33,000	36,000	38,000	46,000	37,000	48,000	65,000	73,000	82,000	91,000	89,000	Expenditure	
15	General Support	774,539	715,518	697,602	599,707	631,854	652,820	837,729	2,096,538	########	1,240,363	1,142,705	1,116,555	1,227,286	Expenditure	
16																

__init__.py | BTS_GTFS_2007_table_03a_intermediate.csv | BTS_GTFS_2007_table_03a.csv | GTFS_Tidy.py

```
Transportation_Mode,1995,1996,1997,1998,1999,2000,2001,2002,2003,2004,2005,2006,2007,Revenue_Expenditure
Highway,"94,249,469","101,630,138","104,517,620","111,277,807","127,563,761","129,555,925","128,209,694","134,686,098","135,837,966","143,
Transit,"22,894,971","23,014,200","23,851,278","24,547,037","32,044,491","29,908,966","34,306,593","34,422,484","36,153,416","38,269,294",
Railroads,"1,080,000","1,010,000","1,137,000","1,080,000","435,000","759,000","735,000","1,308,000","1,219,000","1,513,000","1,451,000","2
Air ,"20,961,769","19,066,511","21,508,074","27,547,567","30,584,123","25,431,972","34,303,744","37,185,557","35,373,368","37,575,887","41
Water,"7,281,871","7,227,371","7,318,285","7,421,277","7,802,396","7,680,691","8,490,305","7,898,825","9,544,095","10,879,806","11,988,968
Pipeline,"57,000","65,000","61,000","63,000","64,000","77,000","74,000","100,000","118,000","116,000","123,000","132,000","136,000",Revenu
General Support,"430,000","386,000","354,000","265,000","263,000","321,000","325,000","1,481,000","10,038,000","1,049,000","1,062,000","1,
Highway,"90,075,441","94,745,874","98,397,988","103,987,699","112,258,778","119,910,746","127,103,943","133,672,268","138,614,597","140,07
Transit,"25,459,587","26,113,438","27,858,489","28,989,888","39,169,559","34,827,547","38,988,965","41,603,545","41,482,031","44,636,365",
Railroads,"1,048,800","1,027,900","1,164,100","1,100,200","452,900","778,300","753,300","1,324,300","1,242,000","1,533,000","1,472,000","1
Air ,"19,250,284","19,769,728","20,694,088","21,732,265","22,066,270","22,525,014","32,838,620","37,025,402","34,184,510","39,175,571","40
Water,"6,622,987","6,726,762","7,108,336","7,097,762","7,700,768","7,633,757","10,620,395","8,037,742","11,774,611","10,904,772","10,307,1
Pipeline,"24,000","34,000","33,000","36,000","38,000","46,000","37,000","48,000","65,000","73,000","82,000","91,000","89,000",Expenditure
General Support,"774,539","715,518","697,602","599,707","631,854","652,820","837,729","2,096,538","10,729,699","1,240,363","1,142,705","1,
```

```
Run: __init__
['Table 3-A   A Summary of Transportation Revenue and Expenditure', '', '', '', '', '', '', '', '', '', '', '', '', '']

Process finished with exit code 0
```

```
Run: __init__
['Transportation_Mode', '1995', '1996', '1997', '1998', '1999', '2000', '2001', '2002', '2003', '2004', '2005', '2006', '2007', 'Revenue_Expenditure']

Process finished with exit code 0
4: Run   Terminal   Python Console
```

Results

Run: table3a

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 182 entries, 0 to 181
Data columns (total 4 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   Transportation_Mode  182 non-null    object
 1   Revenue_Expenditure  182 non-null    object
 2   Year                 182 non-null    object
 3   USD                  182 non-null    object
dtypes: object(4)
memory usage: 5.8+ KB
None
   Transportation_Mode Revenue_Expenditure  Year        USD
51            Railroads         Expenditure  1998  1,100,200
9             Railroads         Expenditure  1995  1,048,800
4                 Water             Revenue  1995  7,281,871
39                Water         Expenditure  1997  7,108,336
23            Railroads         Expenditure  1996  1,027,900

Process finished with exit code 0
```

	A	B	C	D
1	Transportation_Mode	Revenue_Expenditur	Year	USD
2	Highway	Revenue	1995	94,249,469
3	Transit	Revenue	1995	22,894,971
4	Railroads	Revenue	1995	1,080,000
5	Air	Revenue	1995	20,961,769
6	Water	Revenue	1995	7,281,871
7	Pipeline	Revenue	1995	57,000
8	General Support	Revenue	1995	430,000
9	Highway	Expenditure	1995	90,075,441
10	Transit	Expenditure	1995	25,459,587
11	Railroads	Expenditure	1995	1,048,800
12	Air	Expenditure	1995	19,250,284
13	Water	Expenditure	1995	6,622,987
14	Pipeline	Expenditure	1995	24,000
15	General Support	Expenditure	1995	774,539
16	Highway	Revenue	1996	101,630,138
17	Transit	Revenue	1996	23,014,200
18	Railroads	Revenue	1996	1,010,000
19	Air	Revenue	1996	19,066,511
20	Water	Revenue	1996	7,227,371
21	Pipeline	Revenue	1996	65,000
22	General Support	Revenue	1996	386,000
23	Highway	Expenditure	1996	94,745,874
24	Transit	Expenditure	1996	26,113,438
25	Railroads	Expenditure	1996	1,027,900
26	Air	Expenditure	1996	19,769,728
27	Water	Expenditure	1996	6,726,762
28	Pipeline	Expenditure	1996	34,000
29	General Support	Expenditure	1996	715,518
30	Highway	Revenue	1997	104,517,620
31	Transit	Revenue	1997	23,851,278
32	Railroads	Revenue	1997	1,137,000
33	Air	Revenue	1997	21,508,074
34	Water	Revenue	1997	7,318,285
35	Pipeline	Revenue	1997	61,000
36	General Support	Revenue	1997	354,000
37	Highway	Expenditure	1997	98,397,988
38	Transit	Expenditure	1997	27,858,489

BTS_GTFS_2007_Table_3-A_A_Summa

Conclusion

- I suggest Federal dataset reports to incorporate a data cleaning process standard
- Structuring datasets to facilitate analysis is an essential component for increasing re-usability
- Tidy datasets are easy to
  - Manipulate
  - Model
  - Visualize
  - Have a specific structure
    - column = variable
    - row = observation
    - table = observational unit type

References

- Wickham H., (2014). Tidy Data *The Journal of Statistical Software*, vol. 59(10). Retrieved December 9, 2020, from https://vita.had.co.nz/papers/tidy-data.html
- *Daniel chen: Cleaning and tidying data in pandas | pydata dc 2018—Youtube*. (n.d.). Retrieved December 10, 2020, from https://www.youtube.com/watch?v=iYie42M1ZyU
- *Tidy data in python · jean-nicholas hould*. (n.d.). Retrieved December 10, 2020, from https://www.jeannicholashould.com/tidy-data-in-python.html

- Bertram Ludäscher
  - Professor and Director, Center for Informatics Research in Science and Scholarship
- Leighton L Christiansen
  - Librarian/Data Curator, National Transportation Library, U.S Department of Transportation.
- Hoa Luong
  - Assistant Director Research Data Curation, Research Data Service at University of Illinois at Urbana-Champaign
- Ashley Hetrick
  - Data Analyst at the University of Illinois at Urbana-Champaign

Awknoledgements