U.S. Department of Transportation

Office of the Secretary of Transportation

Curators to the Rescue: New Strategies for Making Legacy Data Accessible to the Public

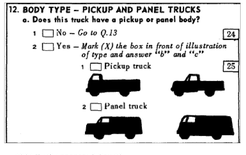
Transportation Research Board 103rd Annual Meeting
Washington, D.C., January 7-11, 2024
Poster: P24-20518

Abstract

As the Bureau of Transportation Statistics (BTS) in 2019 began planning for the 2021 return of the **Vehicle Inventory and Use Survey (VIUS)**, the National Transportation Library (NTL) Data Services team was tasked with locating and sharing the historic, digital data tables and reports from the previous VIUS surveys. The NTL Data Services team was able to locate and provide digital files of the legacy VIUS data beginning with 1977. However, the 1963, 1967, and 1972 **Truck Inventory and Use Surveys (TIUS)** needed to be addressed as well. Unfortunately for the Data Services team, those data tables were trapped in the PDF scans of the original, 50-year-old print documents. How was the NTL Data Services team able to liberate this data and make it reusable for transportation statisticians, researchers, and the public? The NTL had to create new workflows and strategies for rescuing legacy datasets and reports.

Legacy data, which are older datasets that are trapped in non-machine-readable formats, have not been accessible or easily usable by researchers for decades. The NTL Data Services team is working to make these tables trapped behind pdf-scans accessible using **ABBYY FineReader PDF software**. ABBYY FineReader uses optical character recognition to create a machine-readable text layer embedded in the PDF, making each report and table searchable and editable, where the OCR text needs to be corrected. Additionally, the program allows for the export of these data tables into tabular formats. Using these new techniques, legacy Truck Inventory and Use Surveys have become available to the public and researchers in non-print form for the first time in decades, providing an opportunity for a complete longitudinal analysis of the legacy TIUS/VIUS data just as the 2021 VIUS data is being released!

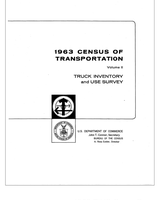
The NTL Data Services team efforts can be replicated by other research programs wishing to liberate useful data from PDF scans. This poster will highlight NTL activities around: Historical TIUS data tables and reports rescue efforts; Incorporating new technologies, such as ABBYY FineReader PDF, into data curation workflows; Increasing accessibility for the entirety of the Truck Inventory and Use Survey/Vehicle Inventory and Use Survey series from 1963 to today; and describe innovative data rescue workflows that can be implemented at other institutions.


12. BODY TYPE – PICKUP AND PANEL TRUCKS
a. Does this truck have a pickup or panel body?
1 ☐ No – *Go to Q.13* 24
2 ☐ Yes – *Mark (X) the box in front of illustration of type and answer "Q" and "a."*
1 ☐ Pickup truck 25
2 ☐ Panel truck

1967 TIUS Appendix B Box 12

Truck Inventory and Use Survey and Vehicle Inventory and Use Survey

The Truck Inventory and Use Survey (TIUS), later known as the Vehicle Inventory and Use Survey (VIUS), is one of the surveys included in the Census of Transportation program from the Census Bureau . Its primary purpose is to collect and publish data on the physical and operational characteristics of the Nation's truck resources. The TIUS survey was completed in the years 1963, 1967, 1972, 1977, 1982, 1987, and 1992. In 1997, the survey name was changed to VIUS, and the survey was completed under the new name in 1997, 2002, and, most recently, in 2021 when it was revised. The goal of these surveys was to collect transportation data not collected by other transportation surveys. These surveys were released initially in advanced reports and then later released as a bound volume of each of the 50 States, the District of Columbia, the 9 geographic divisions of the US, and the whole United States for each survey year. When the Census made this publication available online as PDFs, they divided the full report into individual State summaries. To each of these summaries Census added the same cover, front matter, and appendices. Readers may notice that the page numbers of the summaries are not sequential, as these page numbers refer to the original print edition.


1963 CENSUS OF TRANSPORTATION
Volume 2
TRUCK INVENTORY and USE SURVEY

1963 TIUS Cover Page

Authors

Peyton C Tvrdy
https://orcid.org/0000-0002-9720-4725
Data Management and Data Curation Fellow, National Transportation Library
peyton.tvrdy.ctr@dot.gov

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

Using ABBYY FineReader for Optical Character Recognition of Surveys and their Data Tables

The ABBYY FineReader PDF software was an essential part of making the 1963, 1967, and 1972 TIUS reports and data tables accessible, machine-readable, and reusable. This software was used to take these rough PDF scans of the original reports and fix issues such as skewed pages, cut-off text, and image-only text layers that was not able to be recognized by a machine, copied and pasted, or read by accessibility software. With the changes made with ABBYY FineReader PDF, these previously inaccessible reports are now able to be used and compared across years.

The process of using optical character recognition exponentially sped up the process of fixing these reports and their data. The software does an excellent job of scanning the characters and, overall, has very high accuracy. However, due to artifacts in the scans, skews in the text, the type-font used, and other issues, the software is not 100% independent and still required human review and action. When the software is unsure of a line of text or a word is misspelled, it highlights that passage for human review. The reviewer can then fix the text and move onto the next inaccuracy.

motortrucks during 1963.

1963 TIUS Introduction

The software not only works with paragraphs of text but can also recognize more complex structures such as images and tables. When recognizing tables, ABBYY FineReader PDF can distinguish column headers and row text from the cells of data. By straightening the pages before recognition, this allows the software to better recognize rows of cells, even if the pages are originally skewed and there are no lines between rows and columns, as was the case with the TIUS surveys. In the next section, you will see an image of how the software views and recognizes table structures.

Recognizing Tables and Editing Mistakes

As the TIUS reports from 1963, 1967, and 1972 are all legacy reports set with mechanical type, bound into books, and roughly scanned to image-only PDFs, the ABBYY FineReader PDF software does not recognize complex textual structures, such as tables, with 100% accuracy . Some common issues include not being able to distinguish rows apart, inability to read certain numbers such as 1 and 4 correctly due to the font, misreading what is text and what is a part of the table, inability to put row lines between rows and not in the middle of rows, misreading "-" or null values as row lines or not recognizing them as values at all, and simply misreading large sections of text. Examples of these errors, from the 1967 Wisconsin TIUS, are illustrated below.

Misreading Numbers (1):

Original table from print version

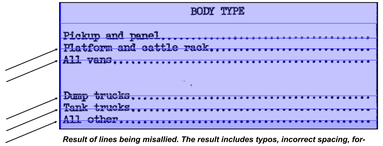
Combinations		
3-axle	4-axle	5-axle
100.0	100.0	100.0

ABBYY Scanned table from showing misread numerals

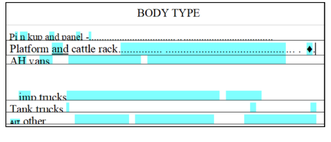
Combinations		
3-axle	4-axle	5-axle
700.0	700.0	700.0

Misaligned Rows and Misrepresented Letters :

Table in ABBYY showing lines cutting through words


BODY TYPE
Pickup and panel
Platform and cattle rack
All vans
Dump trucks
Tank trucks
All other

Result of lines being misallied. The result includes typos, incorrect spacing, formatting and text


BODY TYPE
Pi n kup and panel
Platform and cattle rack
AH vans
imn trucks
Tank trucks
ar other

Missing or Misrepresented Null Values

Original table from the print version

6.4	-
3.8	-
3.4	-
6.4	2.7

"-" or null values are not identified and left blank

6.4	
3.8	
3.4	—
6.4	2.7

Missing Row Lines Between Rows

Table in ABBYY missing lines between rows (ex: no row line between "Wholesale and retail trade" and "Utilities and. services"

Construction	11.1
Manufacturing	3.1
Wholesale and retail trade	13.0
Utilities and services	8.1
For hire	5.3

Result of no line between "Wholesale and retail trade" and "Utilities and services" is incorrect table structure

Construction	11.1
Manufacturing	3.1
Wholesale and retail trade	13.0
Utilities and services	8.1
For hire	5.3

Complexities with Appendices

While the tables had many complications that required many hours of human effort to correct, the appendices for each year were an even harder problem to tackle. The appendices included a blank copy of the survey given out to participants. These surveys, because their non-uniform structure, were recognized as pictures, text, and table structures due to the confusing layout of boxes and their contents. The best way to preserve the structure of the form and its information was through recognizing the entire page as one large table and adding rows, columns, and merging cells. This process was extremely time-consuming, but the user is now rewarded with an accurately structured form with correct text and box placements. This extra effort in correcting the survey leads to more accurate machine-readable text for the user.

Recognizing form as text:

Appendix recognized as a body of text in ABBYY

13. TYPE AND SIZE OF BODY (other than pickup or panel)
Mark (X) ONE box to describe the type of body of the truck or combination. If the power unit is a truck-tractor, report body type of the combination most frequently used with the power unit.
28
Body type
01 ☐ Multi-stop or walk-in
02 ☐ Platform, stake, grain, flatbed or other platform type (with or without dumping device) including low bed and depressed center
03 ☐ Cattle rack (hogs, calves, and other livestock)
04 ☐ Insulated non-refrigerated van
05 ☐ Insulated refrigerated van
06 ☐ Furniture van
07 ☐ Open top van
08 ☐ All other enclosed vans
09 ☐ Beverage

Box 13 of the 1963 Appendix C

Results of recognizing the Appendix as text (no lines, incorrect boxes and structure, etc.)

13. TYPE AND SIZE OF BODY (other than pickup or panel)
Mark (X) ONE box to describe the type of body of the truck or combination. If the power unit is a truck-tractor, report body type of the combination most frequently used with the power unit.
Body type
01 ☐ Multi-stop or walk-in
02 ☐ Platform, stake, grain, flatbed or other platform type (with or without dumping device) including low bed and depressed center
03 ☐ Cattle rack (hogs, calves,and other livestock)
04 ☐ Insulated non-refrigerated van
05 ☐ Insulated refrigerated van
06 ☐ Furniture van
07 ☐ Open top van
08 ☐ All other enclosed vans
09 ☐ Beverage

Recognizing form as a table:

Appendix recognized as a table in ABBYY

TRUCKS		
Type	Number	
	Owned	Leased
Standard panel, sedan delivery, compact van, station wagon, pick-up, multi-stop, walk-in	11	21
Platform, stake, grain, open top van or cattle rack	12	22
Closed top non-refrigerated or furniture van	13	23
Refrigerated van	14	24
Tank	15	25
Dump	16	26
Other trucks	17	27

Image from the 1963 Appendix C form Box 20 of the National Appendix.

Results of recognizing the Appendix as a table (clean lines, clear number, retaining form structure, etc.)

TRUCKS		
Type	Number	
	Owned	Leased
Standard panel, sedan delivery, compact van, station wagon, pick-up, multi-stop, walk-in	13	21
Platform, stake, grain, open top van or cattle rack	12	22
Closed top non-refrigerated or furniture van	13	23
Refrigerated van	14	24
Tank	15	25
Dump	16	26
Other trucks	17	27

Exporting and Cleaning Data Tables:

The reports were not the only part that was cleaned and made accessible. The data tables were also exported as Microsoft Excel sheets and created as additional resources to the PDF. While it is useful that the text and tables are now machine readable in the PDF report, the tables are not accessible in the way that modern data is upon publication. The data tables in each report were exported into Microsoft Excel and cleaned and standardized so that they may be used be researchers as data. While ABBYY FineReader helps verify the text, structure, and numbers, it is not a perfect exporter for data. Some of the changes that needed to be made include font, text size, removing boldface from text, cell borders, formatting numbers as number values and not text, adjusting formatting and spacing of notes and footnotes, and making sure all parts of the table were usable and readable. See below for an initial export of a table vs. the cleaned up final product.

Before

	A	B	C	D
1	Item	1963	1967	1972
2	*Total trucks*	100.0	100.0	100.0
3	MAJOR USE			
4	Agriculture	5.5	3.9	2.3
5	Forestry and lumbering	—	—	—
6	Mining	—	—	—
7	Construction	14.7	14.9	14.4
8	Manufacturing	2.9	—	1.1
9	Wholesale and retail trade	11.3	6.3	6.2
10	For hire	6.1	3.7	3.0
11	Personal transportation	48.7	58.6	59.4
12	Utilities and services	7.4	6.5	9.2
13	All other	3.4	6.1	4.6
14	BODY TYPE			
15	Pickup, panel, multistop, or walk-in	75.4	74.0	80.5
16	Platform and cattlerack	11.3	10.0	8.3
17	Vans	3.6	3.0	4.1
18	Utility truck	—	1.3	1.4
19	Pole or logging	—	—	—
20	Dump truck	5.3	3.8	2.2

After

	A	B	C	D
1	Item	1963	1967	1972
2	*Total trucks*	100.0	100.0	100.0
3	MAJOR USE			
4	Agriculture	5.5	3.9	2.3
5	Forestry and lumbering	—	—	—
6	Mining	—	—	—
7	Construction	14.7	14.9	14.4
8	Manufacturing	2.9	—	1.1
9	Wholesale and retail trade	11.3	6.3	6.2
10	For hire	6.1	3.7	3.0
11	Personal transportation	48.7	58.6	59.4
12	Utilities and services	7.4	6.5	9.2
13	All other	3.4	6.1	4.6
14	BODY TYPE			
15	Pickup, panel, multistop, or walk-in	75.4	74.0	80.5
16	Platform and cattlerack	11.3	10.0	8.3
17	Vans	3.6	3.0	4.1
18	Utility truck	—	1.3	1.4
19	Pole or logging	—	—	—
20	Dump truck	5.3	3.8	2.2

This image is from 1972's Alaska Report: Table 1 Comparative Summary 1963, 1967, and 1972

Conclusion

Using ABBYY FineReader PDF software and a good deal of human curatorial effort, data and text once trapped in PDFs can be cleaned, extracted, and shared via machine-readable formats to power longitudinal research and new discoveries. While not a perfect solution to rescuing legacy data, the time saved by using ABBYY FineReader allows data curators and stewards to make these documents and data tables accessible efficiently. ABBYY FineReader helps reduce inefficiencies resulting from a variety of causes, such as inaccessible, legacy documents and information; previous workflows that were originally print documents or poorly made digital documents, and optical character recognition and accessibility tasks that were originally extremely time consuming and difficult across many pieces of software. The software has an intense learning curve, with many aspects of the scanning process having imperfections. Despite the challenges outlined in this poster, ABBYY FineReader PDF software is a game changer for anyone who works with historical documentation or has image-only PDFs. Now all who wish to view the PDFs and data, including those using accessibility software, can obtain and use these reports as effortlessly as modern documents. With proper investment and training of staff, libraries across the globe can make inaccessible and legacy documents accessible on an unprecedented scale. This project is only the start of an intense effort by the National Transportation Library to make legacy reports and data accessible to all, leading the way for other institutions to do the same.

FAQ and Links

FAQ Page:
https://www.bts.gov/surveys/vius/faqs
VIUS Home Page:
https://www.bts.gov/vius
Complete 2021 Data Tables:
https://www.census.gov/programs-surveys/vius/data/tables.html
2021 Methodology Document:
https://www.census.gov/programs-surveys/vius/technical-documentation/methodology.html
1963 TIUS Datasets:
https://rosap.ntl.bts.gov/view/dot/72625
1967 TIUS Datasets:
https://rosap.ntl.bts.gov/view/dot/72627
1972 TIUS Datasets:
https://rosap.ntl.bts.gov/view/dot/72628

VIUS FAQ

VIUS Series

Citation

Tvrdy, Peyton C., Long, Jesse A., Christiansen, Leighton L., "Curators to the Rescue: New Strategies for Making Legacy Data Accessible to the Public." Transportation Research Board 103rd Annual Meeting. Washington, D.C., USA. https://doi.org/10.21949/1529899.