# Rescuing Legacy Data:

## Using Optical Character Recognition Technologies to Make Airline Consumer Data Accessible

Peyton Tvrdy, National Transportation Library, peyton.tvrdy123@gmail.com


## Abstract

Since 1971, The U.S. Department of Transportation (USDOT) has produced Air Travel Consumer Report data tables as physical documents and online as PDFs. These documents contain information and data tables collected by USDOT tabulating grievances from consumer letters and filings against airlines. These data tables are now being extracted and converted into an accessible, tabular format for publication in the Repository and Open Science Access Portal, ROSA P. Using ABBYY FineReader PDF software, this project transforms and rescues PDF-locked data tables into machine-readable formats, ensuring greater accessibility and usability for researchers and the public. This accessibility not only adheres to the FAIR principles but also makes data more accessible for screen readers. Through rescue efforts such as these, other legacy data projects can be executed efficiently by data professionals to provide data accessibility.


## Original PDF Scans

The Air Travel Consumer Report Series, before being published online as PDFs, were originally printed as a physical report. These reports were scanned to PDF but were not made machine readable. These PDFs are image-only PDFs, meaning that modern screen readers cannot read the document and modern features, such as “Find in Document” cannot be used. Additionally, data tables within these PDFs are not in tabular form and accessible as datasets. More recent Air Travel Consumer Reports are machine readable, meaning they are accessible with screen readers; however, they also have their data tables locked within the PDF format. This makes it difficult for the data to be reused.

| CARRIER | Number of Letters | Flights | | | Reservations | | | Baggage | | | | | Fares & Refunds | Customer Treatment | Flight Info | In-Flight Service | Service in General | Discrimination | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Cancld | Delayed | Irregularities | Oversale | Problems | Ticketing | Loss | Damage | Delay | Other | Rates | | | | | | Racial | Pass. |
| AMERICAN | 87 | 8 | 15 | 4 | 8 | 8 | 2 | 3 | 4 | 3 | 6 | 2 | 22 | 6 | 17 | 2 | 0 | 0 | 1 |
| BRANIFF | 46 | 2 | 14 | 3 | 1 | 4 | 2 | 1 | 3 | 3 | 0 | 0 | 6 | 14 | 7 | 2 | 0 | 0 | 0 |
| CONTINENTAL | 14 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 2 | 0 | 1 | 0 | 0 | 0 | 0 |
| DELTA | 39 | 1 | 4 | 1 | 4 | 6 | 2 | 4 | 0 | 2 | 2 | 0 | 7 | 4 | 2 | 0 | 0 | 0 | 0 |
| EASTERN | 73 | 4 | 16 | 2 | 3 | 10 | 2 | 7 | 4 | 4 | 4 | 0 | 14 | 4 | 3 | 3 | 1 | 0 | 0 |
| NATIONAL | 23 | 3 | 6 | 0 | 3 | 1 | 0 | 0 | 1 | 1 | 1 | 0 | 6 | 1 | 4 | 0 | 0 | 0 | 0 |
| NORTHWEST | 23 | 1 | 5 | 2 | 5 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 4 | 3 | 1 | 2 | 0 | 0 | 0 |
| TRANS WORLD | 78 | 9 | 4 | 4 | 2 | 18 | 11 | 7 | 2 | 3 | 1 | 4 | 13 | 6 | 2 | 1 | 0 | 0 | 2 |
| UNITED | 67 | 4 | 10 | 3 | 5 | 7 | 1 | 5 | 2 | 4 | 4 | 0 | 19 | 5 | 3 | 5 | 1 | 1 | 1 |
| WESTERN | 21 | 3 | 1 | 3 | 2 | 3 | 0 | 1 | 0 | 0 | 0 | 0 | 4 | 1 | 1 | 0 | 0 | 0 | 0 |
| PAN AMERICAN | 53 | 3 | 8 | 4 | 1 | 5 | 0 | 8 | 3 | 2 | 3 | 0 | 6 | 1 | 2 | 0 | 3 | 0 | 0 |
| AIR WEST | 20 | 2 | 17 | 0 | 2 | 1 | 0 | 2 | 0 | 1 | 2 | 0 | 0 | 0 | 1 | 2 | 0 | 0 | 0 |

Image 1: Original 1972 Data Table scanned to PDF.

## Using ABBYY FineReader PDF Software

ABBYY FineReader PDF software is a program that uses Optical Character Recognition (OCR) on PDFs to make them machine readable. Running the software first has the program scan and identify characters, which then are verified for accuracy by the user. Characters the machine cannot interpret are then flagged for the user to address.

| CARRIER | Number of Letters | Flights | | | Reservations | | |
|---|---|---|---|---|---|---|---|
| | | Cancl | Delay | Irregulariti | Over | Problems | Ticketing |
| AMERICAN | «T | _8_ | 15 | 4 | 8 | 8 | 2 |
| 0RANIFF | 46 | 2 | 14 | 3 | 1 | 4 | 2 |
| CONTINENT | 14 | 0 | 0 | 1 | 1 | 1 | 0 |
| DELTA | 39 | 1 | 4 | 1 | 4 | ___6 | 2 |

Image 2: Data table after OCR but before user corrections

Once all errors are fixed, the PDF can be exported as a searchable PDF, a structured table such as Excel or CSV format, to PowerPoint or Word formats, and many more.

| CARRIER | Number of Letters | Flights | | |
|---|---|---|---|---|
| | | CancId | Delayed | Irregu-larities |
| AMERICAN | 60 | 7 | 12 | 4 |
| BRAN IFF | 35 | 0 | 6 | 6 |
| CONTINENTAL | 5 | 0 | 1 | 0 |
| DELTA | 55 | 9 | 8 | 3 |
| EASTERN | 104 | 9 | 35 | 6 |
| NATIONAL | 32 | 2 | 16 | 6 |
| NORTHWEST | 25 | 1 | 3 | 2 |
| TRANS WORLD | 63 | 4 | 12 | 2 |
| UNITED | 41 | 3 | 5 | 1 |
| WESTERN | 14 | 3 | 1 | 1 |

Image 3: Data table corrected and exported to XLSX and CSV

By exporting a searchable PDF, a cleaned and formatted Excel file, and a well-structured CSV, you can achieve widespread accessibility for your data. It is accessible in that it has a copy that doesn’t rely on proprietary software, and it is accessible in that both the Excel and the PDF can be optimized for screen readers and other disability software. Adding controlled vocabulary terms and a DOI makes it accessible to search engines and findable in our repository, ROSA P.

## Creating Roadmaps for Large Preservation Projects

For large-scale accessibility initiatives such as this one, organization keeps the project and preservation efforts on track. When setting out to make accessible PDFs, ensure you have a total count of PDFs you will extract. Additionally, time how long it takes to successfully edit, remediate, and export your documents. Timing your efforts will allow you to give a more accurate estimate of hours needed to preserve these documents. This will save you from overcommitting how many you can preserve over a period. Keeping estimates and expectations in check will lead to a balanced workload and more accurate project timelines.

| Title | ROSA P Link | Workroom Accession Number | Date Fixed in Workroom | Done? [Y/N] |
|---|---|---|---|---|
| Air Carrier Traffic Statistics: [2014-06] | https://rosap.ntl.bts.gov/view/dot/6687 | 53067 | 2025-01-14 | Y |
| Air Carrier Traffic Statistics: [2014-05] | https://rosap.ntl.bts.gov/view/dot/6698 | 53078 | 2025-01-14 | Y |
| Air Carrier Traffic Statistics: [2014-04] | https://rosap.ntl.bts.gov/view/dot/6649 | 53029 | 2025-01-13 | Y |
| Air Carrier Traffic Statistics: [2014-03] | https://rosap.ntl.bts.gov/view/dot/6692 | 53072 | 2025-01-14 | Y |
| Air Carrier Traffic Statistics: [2014-02] | https://rosap.ntl.bts.gov/view/dot/6672 | 53052 | 2025-01-14 | Y |
| Air Carrier Traffic Statistics: [2014-01] | https://rosap.ntl.bts.gov/view/dot/6678 | 53058 | 2025-01-14 | Y |
| Air Carrier Traffic Statistics: [2013-12] | https://rosap.ntl.bts.gov/view/dot/6666 | 53046 | 2025-01-06 | Y |
| Air Carrier Traffic Statistics: [2013-11] | https://rosap.ntl.bts.gov/view/dot/6703 | 53083 | 2025-01-06 | Y |
| Air Carrier Traffic Statistics: [2013-09] | https://rosap.ntl.bts.gov/view/dot/6712 | 53092 | 2025-01-06 | Y |
| Air Carrier Traffic Statistics: [2013-08] | https://rosap.ntl.bts.gov/view/dot/6654 | 53034 | 2025-01-06 | Y |
| Air Carrier Traffic Statistics: [2013-07] | https://rosap.ntl.bts.gov/view/dot/6682 | 53062 | 2025-01-06 | Y |
| Air Carrier Traffic Statistics: [2013-06] | https://rosap.ntl.bts.gov/view/dot/6686 | 53066 | 2025-01-06 | Y |
| Air Carrier Traffic Statistics: [2013-05] | https://rosap.ntl.bts.gov/view/dot/6697 | 53077 | 2025-01-06 | Y |
| Air Carrier Traffic Statistics: [2013-04] | https://rosap.ntl.bts.gov/view/dot/6648 | 53028 | 2025-01-06 | Y |
| Air Carrier Traffic Statistics: [2013-03] | https://rosap.ntl.bts.gov/view/dot/6691 | 53071 | 2025-01-06 | Y |
| Air Carrier Traffic Statistics: [2013-02] | https://rosap.ntl.bts.gov/view/dot/6671 | 53051 | 2025-01-06 | Y |
| Air Carrier Traffic Statistics: [2013-01] | https://rosap.ntl.bts.gov/view/dot/6677 | 53057 | 2025-01-06 | Y |
| Air Carrier Traffic Statistics: [2012-12] | https://rosap.ntl.bts.gov/view/dot/6665 | 53045 | 2025-01-03 | Y |

Image 4: Screenshot of the Project Tracking Excel Sheet.

DOI to August 2024: https://doi.org/10.21949/1530675
Access Whole Data Collection via JSON REST API:
https://rosap.ntl.bts.gov/fedora/export/download/collection/dot:38236
Series records are part of larger collections and must be filtered after download. Use the field “mods.related_series” in the JSON to isolate “Air Travel Consumer Report” series entries.
Collection DOI: https://doi.org/10.21949/1504517

## Citation

Tvrdy, Peyton (2025). “Rescuing Legacy Data: Using Optical Character Recognition Technologies to Make Airline Consumer Data Accessible.” National Transportation Library, https://doi.org/10.21949/8k4x-mx08

ORCID iD: https://orcid.org/0000-0002-9720-4725