NATIONAL
ARCHIVES

Introduction to Using Raw Data

I. Data at the National Archives

Most of the data sets held by the Electronic Records Division of the National Archives and Records Administration consist of raw data and are in a software-independent format. This means:

- Most files will not have a header row nor field delimiters. Instead the files contain fixed-length or fixed-width fields and records. Researchers will have to insert that information, usually as part of the import process, using the technical documentation. The files may also contain coded values.

- Files in a software-independent format do not require a specific software program for use. Researchers can use the files with whatever database, spreadsheet, statistical, word processing, etc., software available to them.

- Most files will not have a file extension, such as .csv or .txt, that today's computers recognize for opening the files in certain applications. Instead, researchers often have to import the data into the appropriate software program.

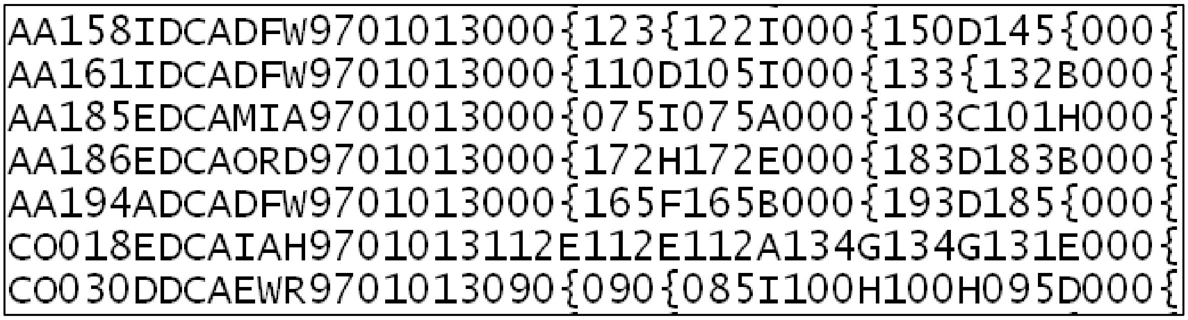
- Researchers need technical documentation, layouts and code lists, to import the data into the appropriate software program and interpret the data.

- Occasionally some additional processing may be required in order to use the data with today's programs.

Federal agencies create, collect, and/or compile data in the course of doing business. Thus the data and documentation held by the Electronic Records Division reflect agencies' different missions, legislative requirements, and computing processes. Formats, coding schemas, and documentation will vary from data set to data set.

In addition, the data in the holdings date as early as the 1940s to the present time. As computer technology has evolved so has the sophistication of the data. How agencies compiled or created the data is a product of the computer capabilities at the time of creation. For example, some data may have been compiled on computer punch cards or maintained by older programming languages. Consequently some data sets may reflect computing conventions standard at the time of creation, but are no longer recognized by most of today's software programs. Some of these computing conventions include zoned decimal or signed numeric formats, binary counters, and packed data.

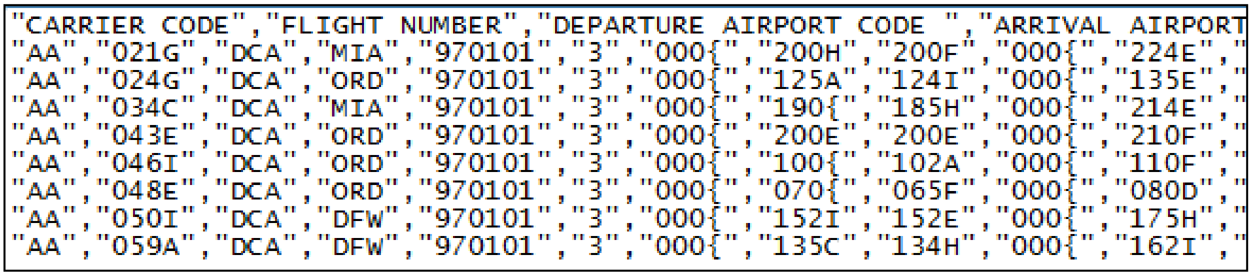
Here is a sample of raw data from the Airline Service Quality Performance File, 1997 (Record Group 398). The data consists of fixed-length fields and each row is a record:



```
AA158IDCADFW9701013000{123{122I000{150D145{000{
AA161IDCADFW9701013000{110D105I000{133{132B000{
AA185EDCAMIA9701013000{075I075A000{103C101H000{
AA186EDCAORD9701013000{172H172E000{183D183B000{
AA194ADCADFW9701013000{165F165B000{193D185{000{
CO018EDCAIAH9701013112E112E112A134G134G131E000{
CO030DDCAEWR9701013090{090{085I100H100H095D000{
```

The data is a string of numbers, letters, and other characters. Nothing in the data itself indicates the individual fields or the meanings of the data. The file does not have a header row. The layouts and code lists in the documentation is needed to interpret the data.

Some data files may include field-delimiters to indicate the individual fields. Here is an example of the same file with field-delimited records:

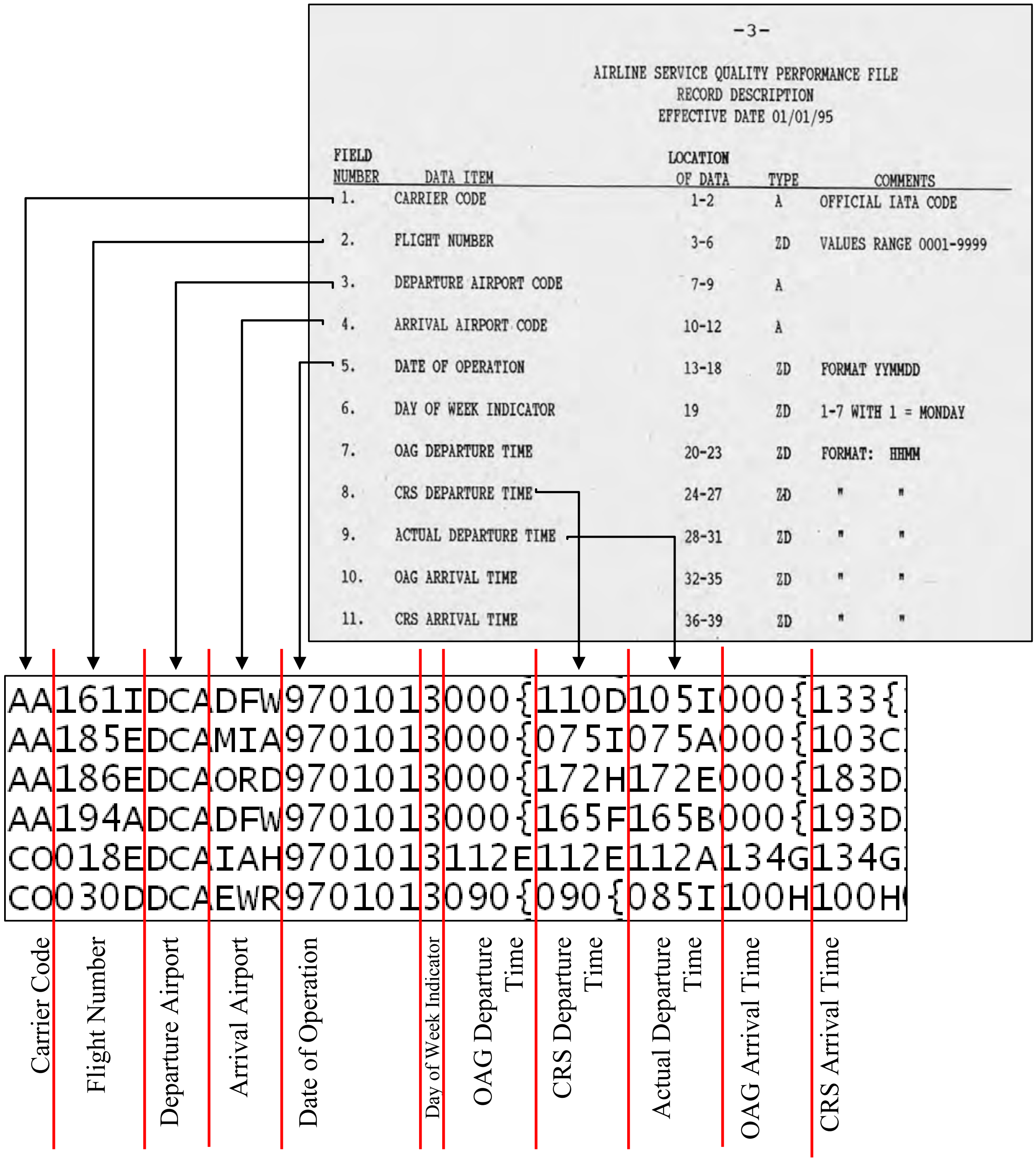


```
"CARRIER CODE","FLIGHT NUMBER","DEPARTURE AIRPORT CODE ","ARRIVAL AIRPORT
"AA","021G","DCA","MIA","970101","3","000{","200H","200F","000{","224E","
"AA","024G","DCA","ORD","970101","3","000{","125A","124I","000{","135E","
"AA","034C","DCA","MIA","970101","3","000{","190{","185H","000{","214E","
"AA","043E","DCA","ORD","970101","3","000{","200E","200E","000{","210F","
"AA","046I","DCA","ORD","970101","3","000{","100{","102A","000{","110F","
"AA","048E","DCA","ORD","970101","3","000{","070{","065F","000{","080D","
"AA","050I","DCA","DFW","970101","3","000{","152I","152E","000{","175H","
"AA","059A","DCA","DFW","970101","3","000{","135C","134H","000{","162I","
```

In this case, the fields are separated by commas (,) and enclosed by quotations ("). The file also includes a header row. Various types of characters may be used as field delimiters, such as pipe (|), carat (^), semicolon (;), or tab.

II. Interpreting raw data, fixed-length fields, using a layout

The layout identifies the fields, how many characters in a field, and the format of the data in the field. For example, the below layout for the Airline Service Quality Performance File indicates that the first field consists of characters 1-2 and contains the airline carrier code. Characters 3-6 contain the flight number and so on. By counting the number of characters, researchers can insert the field delimiters as indicated in the layout.



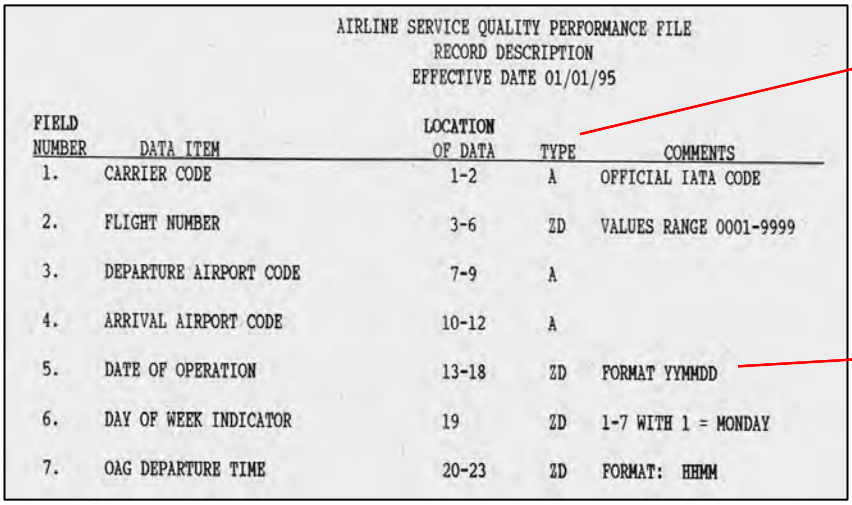
-3-

AIRLINE SERVICE QUALITY PERFORMANCE FILE
RECORD DESCRIPTION
EFFECTIVE DATE 01/01/95

FIELD NUMBER	DATA ITEM	LOCATION OF DATA	TYPE	COMMENTS
1.	CARRIER CODE	1-2	A	OFFICIAL IATA CODE
2.	FLIGHT NUMBER	3-6	ZD	VALUES RANGE 0001-9999
3.	DEPARTURE AIRPORT CODE	7-9	A	
4.	ARRIVAL AIRPORT CODE	10-12	A	
5.	DATE OF OPERATION	13-18	ZD	FORMAT YYMMDD
6.	DAY OF WEEK INDICATOR	19	ZD	1-7 WITH 1 = MONDAY
7.	OAG DEPARTURE TIME	20-23	ZD	FORMAT: HHMM
8.	CRS DEPARTURE TIME	24-27	ZD	" "
9.	ACTUAL DEPARTURE TIME	28-31	ZD	" "
10.	OAG ARRIVAL TIME	32-35	ZD	" "
11.	CRS ARRIVAL TIME	36-39	ZD	" "

```
AA161IDCADFW9701013000{110D105I000{133{
AA185EDCAMIA9701013000{075I075A000{103C
AA186EDCAORD9701013000{172H172E000{183D
AA194ADCADFW9701013000{165F165B000{193D
CO018EDCAIAH9701013112E112E112A134G134G
CO030DDCAEWR9701013090{090{085I100H100H
```

For the Airline Service Quality Performance File, the layout also indicates the type of information in each field and specific comments about the format of the field and coded values.



AIRLINE SERVICE QUALITY PERFORMANCE FILE
RECORD DESCRIPTION
EFFECTIVE DATE 01/01/95

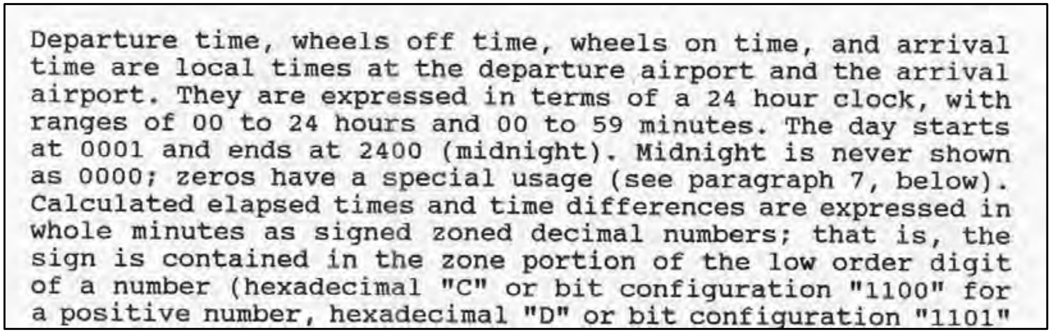
FIELD NUMBER	DATA ITEM	LOCATION OF DATA	TYPE	COMMENTS
1.	CARRIER CODE	1-2	A	OFFICIAL IATA CODE
2.	FLIGHT NUMBER	3-6	ZD	VALUES RANGE 0001-9999
3.	DEPARTURE AIRPORT CODE	7-9	A	
4.	ARRIVAL AIRPORT CODE	10-12	A	
5.	DATE OF OPERATION	13-18	ZD	FORMAT YYMMDD
6.	DAY OF WEEK INDICATOR	19	ZD	1-7 WITH 1 = MONDAY
7.	OAG DEPARTURE TIME	20-23	ZD	FORMAT: HHMM

A = alphabetic
ZD = zoned decimal numeric

Dates are displayed as YYMMDD (e.g. 970101 means January 1, 1997)

The format and types of information captured in a layout can vary. In some cases, the layout may be available electronically as a separate file or NARA may have prepared the layout based on the documentation provided.

Additional user notes in the documentation may provide more details on the collection, content, and format of a field. For the Airline Service Quality Performance File, a user note explains the use of zoned decimal numbers in some of the fields documenting departure and arrival times:

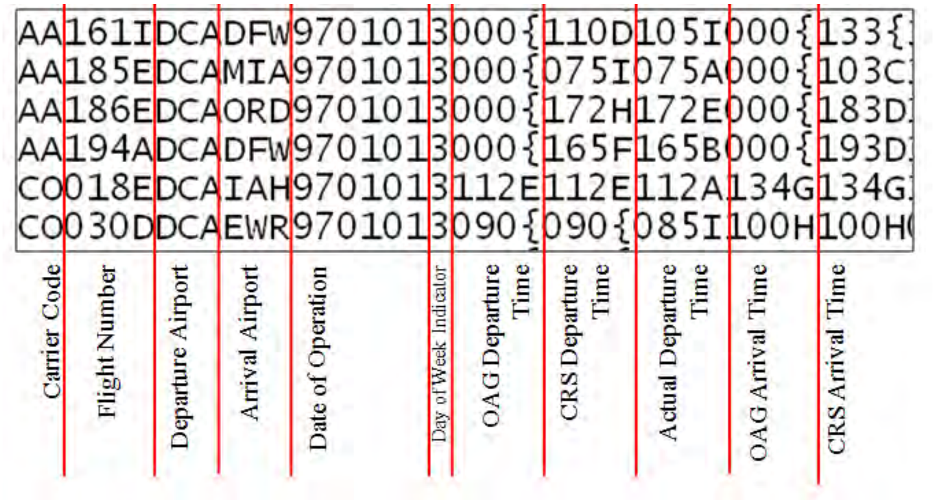


Departure time, wheels off time, wheels on time, and arrival time are local times at the departure airport and the arrival airport. They are expressed in terms of a 24 hour clock, with ranges of 00 to 24 hours and 00 to 59 minutes. The day starts at 0001 and ends at 2400 (midnight). Midnight is never shown as 0000; zeros have a special usage (see paragraph 7, below). Calculated elapsed times and time differences are expressed in whole minutes as signed zoned decimal numbers; that is, the sign is contained in the zone portion of the low order digit of a number (hexadecimal "C" or bit configuration "1100" for a positive number, hexadecimal "D" or bit configuration "1101"

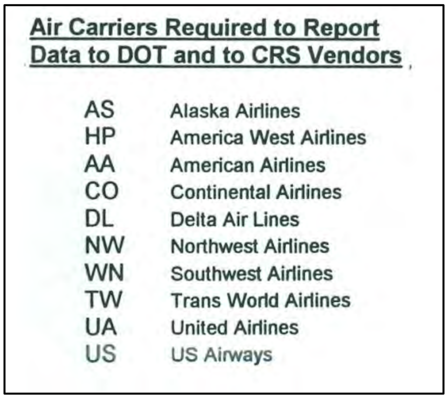
III. Interpreting coded values

Many data sets contain coded values. A coded value is used in place of a lengthier, more detailed value. Code lists in the documentation provide the meanings for the coded values.

In the example of the Airline Service Quality Performance File, coded values are used for carrier code, departure and arrival airport, and day of week indicator fields.

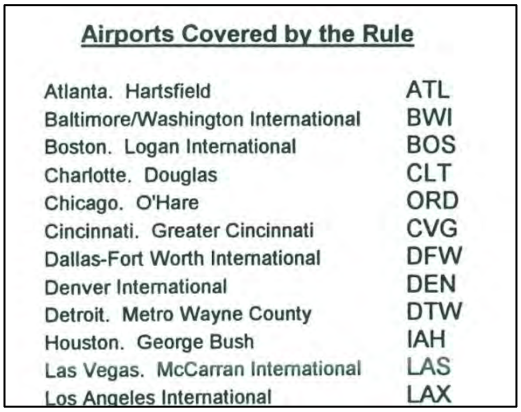


```
AA161IDCADFW9701013000{110D105I000{133{
AA185EDCAMIA9701013000{075I075A000{103C
AA186EDCAORD9701013000{172H172E000{183D
AA194ADCADFW9701013000{165F165B000{193D
CO018EDCAIAH9701013112E112E112A134G134G
CO030DDCAEWR9701013090{090{085I100H100H
```



Air Carriers Required to Report Data to DOT and to CRS Vendors

AS	Alaska Airlines
HP	America West Airlines
AA	American Airlines
CO	Continental Airlines
DL	Delta Air Lines
NW	Northwest Airlines
WN	Southwest Airlines
TW	Trans World Airlines
UA	United Airlines
US	US Airways



Airports Covered by the Rule

Atlanta. Hartsfield	ATL
Baltimore/Washington International	BWI
Boston. Logan International	BOS
Charlotte. Douglas	CLT
Chicago. O'Hare	ORD
Cincinnati. Greater Cincinnati	CVG
Dallas-Fort Worth International	DFW
Denver International	DEN
Detroit. Metro Wayne County	DTW
Houston. George Bush	IAH
Las Vegas. McCarran International	LAS
Los Angeles International	LAX

The code list for Air Carrier shows that the value AA represents American Airlines and the value CO represents Continental Airlines. The Airports code list shows DCA is for Washington Reagan National Airport, DFW is for Dallas-Fort Worth International Airport, and so forth.

The coded values for Day of Week Indicator were described on the layout with 1 equaling Monday so the value 3 would represent Wednesday.

IV. Importing raw data into software programs

Files in a software-independent format do not require a specific software program in order to use them. Researchers can use whatever appropriate software is available to them. Appropriate software may include database, spreadsheet, statistical, and word-processing (text) programs. Researchers often have to import the data into the appropriate software program.

Researchers should check their program's options for importing text or other types of files. The steps to import the files will vary based on the program.

For some common spreadsheet and database programs, users can start the import process by opening the program, select to open a file, change the file type to open to "All Files," and then select the data file.

In general, to import the files researchers will need the documentation in order to specify the length and type of each field. Some common spreadsheet and database programs step users through the process of adding delimiters or specifying field lengths for files containing fixed-length or fixed-width fields. This process may include inserting lines for the field delimiters (similar to the example above) or entering the beginning and ending character location for each field.

If the data file contains field-delimiters, the import process may require identifying the type of delimiter (comma, pipe, semicolon, etc) and if it is enclosed by quotes or not.

The import process may also have the option of identifying the type of data in each field (text, date, number) and the specific format for certain types of data.

Electronic Records Division
May 15, 2015