
Chapter 3: Documenting Administrative Data Files

File documentation should accompany all welfare outcomes administrative data files. This documentation should help researchers use the data and prevent grantees from having to bear the burden of supporting and interpreting data files to researchers in the future.

The Work Group concurred that, although it is essential for grantees to carefully document their welfare outcomes data files, producing file documentation can also be a complicated and burdensome task. The Group's goal was to help grantees **produce file documentation that strikes a balance between meeting researcher needs and minimizing grantee burden.**

Preparing documentation for grantee administrative data files presents many challenges, including:

- The administrative sources from which data were extracted may be poorly documented.
- A large number of administrative data elements may have been used in grantee research and must be reported in the data files.
- Grantees may construct one or more administrative data files and each file requires some level of documentation.

In preparing this guide, the Group recognized these challenges and attempted to address these concerns in its guidance and recommendations.



Chapter 6 contains an overview of documentation templates, and instructions for their use.

The Work Group recommends that grantees use a **standardized approach** to producing administrative data file documentation. This will improve researcher understanding and reduce the amount of work required by grantees to produce these documents. To assist grantees with producing file documentation using a standardized approach, the Work Group developed a set of **documentation templates**. These templates are provided in the Grantee Toolkit (see Chapter 6).

Outline for Administrative Data File Documentation

Administrative data file documentation should include the following five elements:

1. A **narrative** that provides essential background and summary information about the files' data and how the study's selection event was defined.
2. **Technical and programming information** required to use the files.
3. **Data dictionaries** that describe the variables included in the files.
4. **Record layouts** that describe where data are located in the file.
5. Other **information that might be helpful to researchers** using the data (e.g., administrative data reports, glossaries of state-specific terms and acronyms).



Tip 3.1 outlines the recommended presentation for the five administrative data file documentation elements.

A recommended outline for presenting these five elements is presented in Tip 3.1 and the following sections of this chapter provide additional detail about this outline.

Grantees who produce **more than one administrative data file** (see Chapter 1 on recommendations for when this is appropriate) may combine the narrative, technical and programming information, and other information useful to researchers, into a single document. However, **separate data dictionaries and record layouts** will be required for each file.

**✓ Tip 3.1: Recommended Outline
for Administrative Data File Documentation**

I. Narrative

- A. Project Overview
- B. Summary of Reported Data
- C. Study Population Description

II. Technical and Programming Information

- A. Administrative Data File Guide
- B. Administrative Data File Information

III. Data Dictionary

IV. Record Layout

V. Other Items for Researchers



The template described in Chapter 6 and included in the Grantee Toolkit provides grantees with resources to produce documentation that follows this outline.

1. Administrative Data File Documentation Narrative

The administrative data file documentation should open with a **narrative** that provides a project overview, summary of the data included in the files, and a description of the study population.

A. Project Overview

The project overview should summarize how administrative data were used in grantees' welfare outcomes research. This summary may include the following elements:

- A very brief discussion of the **major research questions** addressed by the administrative data analysis.
- A description of the **data analysis techniques** used in grantee administrative data analysis.
- A list of **reports** in which findings from these data sources have been reported and how copies of these reports may be obtained. To the extent possible, copies of these reports should be appended to the file documentation.

The project overview should summarize information contained in the administrative data reports; however, grantees do not need to reiterate project background information at the same level of detail.

B. Summary of Reported Data

Grantees have used a wide range of administrative data drawn from numerous sources. The narrative section of the file documentation should clearly **identify what types of data were used in grantee analysis and from where these data were drawn**, including:

- **The agency, program or other sources that collect these data.**
- **How these data were collected** by an agency, program or other source (e.g., caseworker, electronic filing by employers).
- **The unit of analysis** (e.g., casehead, assistance unit).

- The **time period covered** by data used in welfare outcomes analyses.
- The **population groups** that are and are not covered by these data.
- The **number of cases** from a particular data source that were used in welfare outcomes analyses.
- The **limitations** researchers should consider when using certain types of data.



Figure 3.1 shows an example of a descriptive information matrix for a particular data source.

Given the number of data sources and the potential burden for creating a written description of each data source, the Work Group recommends that a **simple descriptive information table** be constructed for each type of data. An example of this table is provided in Figure 3.1, and a template for producing this table is included in the Grantee Toolkit (see Chapter 6).



A template for the descriptive information matrix is provided in the Grantee Toolkit (see Chapter 6).

Where grantees have chosen to conduct analyses with a random sample of cases from their administrative data files, or are not reporting on a full universe of cases for a given study cohort, the narrative should include a description of the sampling or selection procedures used and, if applicable, the sample weights.

C. Study Population Definition

Grantees should describe **how their study populations or research cohorts were defined**. For example, in the case of TANF leavers studies, this description should note the following:

- **The exit month(s)/quarter(s) and year(s) used to define the population** (e.g., the time period represented by the sample frame such as leavers from the fourth quarter of 1996).
- **An explanation of how the conceptual definition of the study population was operationalized** (e.g., a notation as to the administrative file codes were used to select cases or an explanation about how files were merged).
- **The definition of “exit month” that was applied to the study**. For example, some grantees chose to define the “exit month” as the month in which an individual left assistance; other grantees chose to define “exit” as the first month or quarter in which an individual did not receive a benefit payment.

Additionally, many grantees chose to apply a “two month” criterion when determining whether an individual had left assistance. For example, for an individual to be identified as a leaver, he or she would have to be off assistance for the two months after the “exit” month.

- **A description of who was included or excluded from the study population** (e.g., child-only cases and male caseheads). This description should provide enough detail so researchers can understand what types of cases are and are not included in the study population.

In cases where grantees have chosen to study **multiple cohorts**, the documentation should also include a description of the relationship between the selection event and cohort assignment. For example: “Cohort 1 includes cases that exited in the fourth quarter of 1997; Cohort 2 includes cases that exited in the first quarter of 1998; etc.”

When multiple study cohorts are reported in one administrative data file, **a flag variable to note which cases were included in specific study cohorts should be used**. The file documentation narrative and data dictionary should describe how these flag variables have been applied, and how researchers who wish to replicate cohort-specific results in administrative data reports may interpret them.

Figure 3.1: Sample Descriptive Information Table for Administrative Data

A descriptive information table should be completed for each data type reported in the administrative data files.

Data Type	<ul style="list-style-type: none"> What type of data are reported (e.g., Unemployment Insurance, Temporary Assistance for Needy Families, food stamps)?
Acronym/State Reference	<ul style="list-style-type: none"> What is the program's acronym and/or the colloquial reference used by program personnel or state authorities (e.g., UI, TANF)?
Source Agency & Program	<ul style="list-style-type: none"> What agency or program collected the data and/or was responsible for the data (e.g., State XX Department of Economic Security)?
Data Source	<ul style="list-style-type: none"> How were the data collected (e.g., caseworkers, electronic reporting from employers)? How often were data elements updated within the system?
Time Period Covered in File	<ul style="list-style-type: none"> What time period is covered by the data included in the file (e.g., October through December 1999)?
Unit of Analysis	<ul style="list-style-type: none"> What is the unit of analysis (e.g., casehead, assistance unit, and child)?
Description of Population Included in File	<ul style="list-style-type: none"> Who is and is not included in the reported administrative data (e.g., leavers, no child-only cases)? Are there multiple cohorts? If so, who is included and what time periods are covered (this should also be noted above)? What is the cohort size?
Total Number of Records/Cases Reported in File	<ul style="list-style-type: none"> How many cases and/or records are reported in the data file(s)?
Matching	<ul style="list-style-type: none"> How were records from this administrative data source matched, or linked, across administrative data systems (e.g., one-for-one matches, probabilistic matching)? On what field were data matched (e.g., case I.D., Social Security number)? For how many cases did a match not exist or turn up missing (e.g., number of cases excluded from analysis)? How were missing cases handled?
File Location(s)	<ul style="list-style-type: none"> If multiple welfare outcomes data files are constructed, in which file or files does this data source appear? If data are in multiple files, which portions of the data are in which file?
Data Limitations	<ul style="list-style-type: none"> What limitations should researchers consider when using the data (e.g., data are primarily collected at program entry and is not updated at program exit; reliability issues with certain variables)?

2. Technical and Programming Information

The second section of the administrative data file documentation provides the **technical and programming information** researchers need to understand the files. This information should include a summary “file guide” for each file and other information about how the files may be used.

A. Administrative Data File Guide

For each administrative data file produced, the documentation should include a summary “**file guide**” that includes the following information:

- File name.
- File type (e.g., ASCII).
- Types of data included in the file (e.g., TANF, UI).
- Number of observations or records contained within the file.
- Maximum record length.
- Record format (e.g., fixed, comma delimited).
- Name of the record identification variable, or sort key.
- Unit of analysis (e.g., casehead, child).
- List of files to which these data may be linked.
- File description.



Figure 3.2 shows an example file guide matrix.



A template of the file guide matrix is included in the Grantee Toolkit.

A separate file guide should be completed for each administrative data file produced.

To assist grantees with reporting this information, a template for the **file guide table** was developed by the Work Group. An example of this table is provided in Figure 3.2 and a template for producing this table is included in the Grantee Toolkit.

Figure 3.2: Example Administrative Data File Guide

A file guide should be produced for each administrative data file.

File Name	<ul style="list-style-type: none">• What is the file's name and extension (e.g., admin1.asc)?
File Type	<ul style="list-style-type: none">• What is the file's type (e.g., ASCII)?
Data Included in File	<ul style="list-style-type: none">• What data are included in the file (e.g., TANF, UI, and food stamps)?
Number of Observations	<ul style="list-style-type: none">• What is the number of observations or records included in the file?
Record Length	<ul style="list-style-type: none">• What is the file's maximum record length (e.g., 1,200 columns)?
Record Format	<ul style="list-style-type: none">• What is the record format (e.g., fixed field)?
Record Identification Variable/Sort Key	<ul style="list-style-type: none">• What is the name of the record identification variable or sort key (e.g., "masterid")?
Unit of Analysis	<ul style="list-style-type: none">• What is the file's unit of analysis (e.g., casehead, and child)?
Relating Files	<ul style="list-style-type: none">• What are the names of the other files to which data in this file may be linked?• What is the name of the variable that may be used when linking files?• Are there instances where files may be linked, but researchers should not expect a "one-to-one" match between data in the two files?
File Description	<ul style="list-style-type: none">• How can the file be generally described (e.g., Administrative data for study cohort 1 that includes all TANF leavers cases during the fourth quarter 1996)?

B. Administrative Data File Information

The file documentation should also include the following brief **instructions** about how the files may be used.

- **A statement describing the files' format.**



A description of the ASCII fixed-field format is provided in the Grantee Toolkit.

As discussed in Chapter 1, the Work Group recommends that grantees produce their welfare outcomes data files using an **ASCII fixed-field format**. A brief statement on the ASCII fixed-field format is provided in the documentation template so that grantees using this format do not have to construct an independent description.



Grantees can use this statement instead of having to construct their own descriptions.

If grantees decide not to follow the Work Group's recommendation and produce files using a format other than ASCII, their file documentation should clearly describe the file's format and how it may be accessed and used by researchers.

- **A brief discussion about the relationship among multiple administrative data files.**

Where grantees have elected to produce more than one administrative data file, the documentation should describe **how the multiple files were constructed** (e.g., separate administrative data study cohorts, different time periods).

- **A description of the files' data dictionary and record layout formats.**



Descriptions of recommended formats for data dictionaries and record layouts are provided in the Grantee Toolkit.

The Work Group has also recommended a format for producing data dictionaries and record layouts (see Sections 3 and 4 of this chapter). A brief description of the recommended formats is provided in the documentation template so that grantees following these recommendations do not have to construct an independent description.

If grantees decide not to follow the Work Group's recommendations for the dictionary and layout formats, their file documentation should describe alternative formats and how researchers may interpret them.



A description of recommended variable-naming conventions is included in the Grantee Toolkit.

- **An overview of the variable naming conventions used in the files.**

The Work Group **recommended an approach for constructing variable names** (see Chapter 1) and provided names for common and some auxiliary data elements (see Chapter 2). A brief description of the Group's recommended approach is provided in the documentation template so that grantees do not have to construct an independent description.

If grantees elect to use alternative methods for constructing variable names, their file documentation should describe their approach.

- **An overview of how unique record identifiers were constructed.**



A description of the recommended approach to record identifier construction is included in the Grantee Toolkit.

Each record or case included in both administrative and survey data files should be assigned a **unique record identifier** that may be used to link, or relate, data for particular cases across files. The Work Group **recommended an approach** for creating these unique identifiers that includes two parts: the state FIPS code and a randomly assigned record number (see Chapter 1). A brief description of this recommended approach is provided in the documentation template so that grantees using this approach do not have to create an independent description.

If grantees elect to use another method, their documentation should provide a detailed description as to how these unique identifiers were created and whether they may be used to link data across files.

3. Data Dictionary

Grantees should construct data dictionaries for each administrative data file that include the following information for every variable in a particular data file:

- Variable name.
- Description.
- If applicable, time period covered.
- Known problems or limitations associated with a particular variable.
- Allowable values and formats.
- Whether the data element was included in the administrative report.



Figure 3.3 shows an example format for the data dictionary.



A data dictionary template is included in the Grantee Toolkit.

The Work Group developed a recommended format and template to assist grantees with constructing data dictionaries. An example of this format is provided in Figure 3.3, and a template for producing this matrix is included in the Grantee Toolkit (see Chapter 6).

Figure 3.3: Recommended Format for Documenting Administrative Variables in a Data Dictionary

Separate tables should be constructed for each variable in data set.

Data Dictionary for {Grantee Name}'s {Administrative Data File Name}: Page XX

Variable Name:

- The variable name as it appears in the file layout.
- In cases where a data element is recurring, a list of variable names may be provided here and only one variable description needs to be created.

Description:

- A variable definition that describes the data element and from what source agency or program it was drawn. The source should be listed in the documentation's narrative section.
- If a variable was constructed, grantees should describe what operations were used to produce the data included in the file. This description may take the form of annotated statistical program code (e.g., SAS) used to construct the variable. However, the annotation should be sufficient for an independent researcher to understand and replicate the variable's construction.

Time Period Covered:

- The time period covered by a variable (e.g., month prior to exit, second month following exit, first quarter following exit).

Data Limitations:

- The known problems or limitations that researchers should consider when using this variable in their analysis.

Allowable Values and Formats:

- A description of the data listed in the file field (e.g., dollar values for wages, whether a person received a benefit).
- A list of the allowable values and, if applicable, their codes. For example:

0 = No
1 = Yes

Included in Administrative Data Report:

- Grantees should note whether a particular variable was included in their administrative data report.

4. Record Layout



An example record layout format for data is shown in Figure 3.4, below.



A template for producing the record layout format matrix is included in the Grantee Toolkit.

The fourth section of the documentation should include a **record layout** that specifies where variables are located in a file. A separate record layout that includes the following information should be constructed for each administrative data file.

- An item number.
- The variable's beginning and ending position within the file.
- Variable length.
- Data type (i.e., character or numeric).
- Variable name.

To help grantees construct record layouts for their data, the Work Group developed a recommended format and template. An example of this format is presented in Figure 3.4 and a template for producing this matrix is provided in the Grantee Toolkit (see Chapter 6).

Figure 3.4: Recommended Record Layout Format

One row should be completed for each variable included in a data file.

Item Number	Beginning Position	Ending Position	Item Length	Data Type	Variable Name
Sequential ordering for file variables	Column location where data begins for specific variable	Column location where data ends for specific variable	The number of columns allotted for a specific variable	C = character N = numeric	The assigned variable name

5. Other Items for Researchers

Finally, the Work Group identified a list of items that would **help researchers** use welfare outcomes administrative data files. Grantees are encouraged to **append the following items** to their administrative data file documentation.

- A copy of welfare outcomes **administrative data final reports**.
- Printout of SAS or other statistical program **input statement** or tool that may be used to read ASCII data file.
- **State and county FIPS codes**, if they are used in the data file.
- **Glossary of terms** that provides researchers with an overview of state-specific programs, terminology, and acronyms.
- **Processed cross tabulations for one or more variables** that could be used by researchers interested in replicating results.
- **Statistical program output** that lists file variables with their frequency and means (e.g., similar to output generated by a “**proc univariate**” procedure in SAS).