Chapter 4: Documenting Survey Data Files

File documentation accompanying grantee welfare outcomes survey data files should help researchers use these data and prevent grantees from having to bear the burden of supporting and interpreting data files to researchers in the future.

The Work Group recommends that grantees use a **standardized approach** to producing survey 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 Survey Data File Documentation

Survey data file documentation should include the following five elements:

1. A **narrative** that provides essential information about the survey process and other supporting information on the survey data included in the file.

2. **Technical and programming information** required to use the files.

Tip 4.1 shows an outline for survey data file documentation.

3. A **data dictionary** in the form of an annotated survey questionnaire that describes variables included in the file.

4. **A record layout** that describes where data are located in a file.

5. Other **information that might be helpful to researchers when using the data** (e.g., survey data report, glossary of state-specific terms and acronyms).

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.

Similar to administrative data files, grantees may produce more than one file that include data from different surveys (see Chapter 1 for recommendations for when this is appropriate). Where grantees have engaged in **distinct survey efforts**, such as those with different population groups, survey questionnaires, or survey methodology, a **separate set of documentation should be created for each file**.

To the extent that the **surveys are similar in scope and approach**, grantees may produce a **single set of file documentation** that includes the narrative, technical and programming information, and other information useful to researchers. However, each file should have a separate data dictionary and a record layout.

✓ **Tip 4.1**: Recommended Outline for Survey Data File Documentation

I. Narrative

A. Survey Overview
B. Survey Design and Administration
C. Study Population Definition
D. Survey Sample Approach
E. Survey Response Rate and Calculation

II. Technical and Programming Information

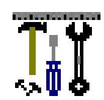
A. Survey Data File Guide
B. Survey Data File Information

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. Survey Data File Documentation Narrative

The survey data file documentation should open with a **narrative** that provides a survey overview, description of the survey methodology, detailed information about survey sampling, survey response rate calculations, and instructions for using the survey data.

A. Survey Overview

The survey overview should **summarize unique aspects of grantees' research and the role surveys played** in this research. Recommended items for the overview include:

- A very brief discussion of the **major research questions** addressed by the survey data analysis, including key topics (e.g., household income and food insecurity) covered by the survey instrument.

- A description of the data analysis **techniques** used in grantee survey data analysis.

- A list of **reports** in which findings from these data sources have been reported. To the extent possible, copies of these reports should be included as appendices to the survey data file documentation.

Grantees summarizing multiple survey efforts in one set of documentation **should include a separate overview for each survey effort.**

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

B. Survey Design and Administration

The survey data file documentation should describe the **survey methodology** applied to the study and other related survey administration outcomes, including:

- The survey **questionnaire development** process (e.g., pre-testing and questions that may be benchmarked against other surveys).

- The **survey mode** or modes (e.g., CATI telephone, CAPI in-person).

- **Languages** other than English into which the questionnaire was translated.

- The length of the survey **fielding period**.

- Survey **management** (e.g., who conducted the survey).

- Average survey **length** (e.g., length in minutes for telephone or in-person surveys).

- Other **relevant information** about the survey process (e.g., respondent location strategies, other problems encountered and their solutions, known data limitations, etc.).

Grantees who are documenting multiple survey efforts **should describe the above items for each survey effort.**

Figure 4.1 shows an example of a summary table.

The Work Group also recommends that grantees summarize this information in a **summary table.** This table should summarize, not supplant information contained in other sections of the survey data file documentation. One table should be completed for each survey effort.

A summary table template is included in the Grantee Toolkit.

An example table is provided in Figure 4.1, and a template for producing this table is included in the Grantee Toolkit.

Figure 4.1: Recommended Summary Table for Survey Documentation

One summary table should be completed for each survey data file.

Survey Description	• A brief description of the survey effort (e.g., Survey of Welfare Leavers from Fourth Quarter 1996: A 24-Month Follow-up Survey). • Survey data file name (e.g., survey1.asc).
Description of the Study Population	• What were the exit month(s)/quarter(s) and year(s) used to define the population? • How was the selection event defined (e.g., exit month)? • Who has been included and excluded from the study population?
Survey Sample Design	• What was the survey sample frame (e.g., all leavers or certain subgroups)? How many cases were in the sample frame? • What was the survey sample design (e.g., stratified sample, cluster sample, oversample for certain subgroups)? • How many cases were sampled? And, if a specialized design was used, how many cases within certain groups were sampled?
Number of Completed Interviews	• Number of completed interviews as calculated for the numerator of the survey response rate.
Survey Response Rate	• Survey response rate. • Description of the response rate calculation methodology.
Survey Approach	• What survey modes were used to interview respondents? • Percentage of completed interviews using each survey mode.
Survey Fielding Period	• When was the survey fielded (e.g., between what dates)?
Survey Questionnaire Length	• What was the average survey administration time (e.g., 45 minutes)?
Known Data Limitations	• What limitations should researchers consider when using these data (e.g., bad survey questions, low response rate, item non-response)?

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 the sample frame).

- **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."

D. Survey Sample Approach

Grantees should provide a description of the sampling approach that includes the following information.

- A description of the **survey sample frame**, including a description of the population groups that are and are not included in the frame (e.g., how child-only cases addressed).

- The **number of cases** in the sample frame.

- A description of the **survey sample design** (e.g., what the sample was designed to accomplish in terms of representation and the approach (es) used).

- The **number of cases sampled** and, if applicable, how these cases are distributed (e.g., across subgroups or strata).

- If applicable, **a description of all survey data weighting** approaches used and **the corresponding sample weights**.

Grantees should provide a separate sample approach description for each survey effort.

E. Survey Response Rate and Calculation

The level of survey response and the response rate calculation methodology should also be noted in the survey data file documentation. Specifically, the documentation should include:

- A table listing the **final survey dispositions** for all sampled cases. This should include a notation of any sampled cases that are determined to be ineligible (e.g., deceased individual) and excluded from the denominator when estimating the response rate.

- A description of the survey's **response rate**, or range of response if multiple rates have been calculated.

- A description of **how the response rate was calculated** and, if applicable, the source for the response rate formula (e.g., AAPOR or CASRO guidelines).

- A description of the **techniques used to correct for possible nonresponse bias** (e.g., multiple imputation or weighting).

2. Technical and Programming Information

The second section of the survey 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. Survey Data File Guide

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

- File name.
- File type (e.g., ASCII).
- 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).
- List of files to which these data may be linked.
- File description.

Figure 4.2 shows an example file guide.

To assist grantees with reporting this information, a template for the **file guide table** has been developed. **A separate table should be completed for each survey data file produced.** An example of this table is provided in Figure 4.2, and the template is included in the Grantee Toolkit.

The Grantee Toolkit includes a file guide template.

Figure 4.2: Example Survey Data File Guide

A separate file guide should be produced for each survey data file.

File Name	• What is the file's name and extension (e.g., survey1.asc)?
File Type	• What is the file's type (e.g., ASCII)?
Number of Observations	• What is the number of observations or records included in the file?
Record Length	• What is the file's maximum record length (e.g., 1,200 columns)?
Record Format	• What is the record format (e.g., fixed field)?
Record Identification Variable/Sort Key	• What is the name of the record identification variable or sort key (e.g., "masterid")?
Unit of Analysis	• What is the file's unit of analysis (e.g., casehead)?
Relating Files	• What are the names of the 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	• How can the file be generally described (e.g., Survey of Welfare Leavers from the Fourth Quarter 1996: A 24-Month Follow-up Study)?

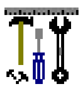
B. Survey Data File Information

In addition to the file guide, grantees should provide brief **instructions** about how the files may be used. These instructions should include:

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

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 grantees using this format do not have to construct an independent description.

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.



The Grantee Toolkit offers standardized descriptions of the recommended file format, data dictionary, record layout format, variable naming conventions, and unique record identifier construction.

Grantees who follow recommendations in these areas may use the standardized descriptions instead of writing their own.

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

Where grantees have elected to produce more than one survey data file, the documentation should describe **how the multiple files were constructed** (e.g., different survey instruments or survey cohorts).

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

The Work Group has also **recommended a format for producing data dictionaries and record layouts** (see Sections 3 and 4 in 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 clearly describe the alternative formats and how researchers may interpret them.

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

The Work Group recommends that **survey data variable labels correspond with question numbering or other labels present in the survey questionnaire.** A brief description of this 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 for survey data, their file documentation should clearly describe their approach.

- **A description of how unique record identifiers were constructed.**

Each record or case included in both survey and administrative 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

A data dictionary should be included in the survey file documentation's third section. To minimize the burden of developing survey data dictionaries, the Work Group recommends that an **annotated questionnaire** be used as the basis for survey files' data dictionaries. In presenting a questionnaire as a data dictionary, grantees should note the following items:

- The **variable name** associated with specific questionnaire responses.

- **Skip pattern instructions.**

- **Allowable values and formats** for specific questions.

- **Data limitations** for specific questions (e.g., perceived reliability issues with questions as written).

It is also important that the file documentation provide additional detail about certain aspects of the survey data. For example, the documentation should describe the following:

- How variables have been constructed for analysis.

- Where survey values have been imputed.

- Where supplementary survey administration variables have been included in the file.

Figure 4.3 shows an example of a survey data dictionary.

The Grantee Toolkit provides a template for producing this data dictionary.

A separate data dictionary should be produced for data elements that are not listed in the survey questionnaire (e.g., constructed variables). This dictionary should follow the format outlined for administrative data variables (see Chapter 3, Section 3). An example of how this format may be adapted for survey data is illustrated in Figure 4.3 and a template for producing this data dictionary is provided in the Grantee Toolkit (see Chapter 6).

Figure 4.3: Recommended Format for Documenting Survey Variables in a Data Dictionary

Separate table should be constructed for each file variable that is not incorporated into the survey's questionnaire.

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

Variable Name:

- The variable name as it appears in the file layout.

Description:

- A variable definition that describes the data element.
- 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.

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 field.
- A list of the allowable values and, if applicable, their codes. For example:

  0 = No
  1 = Yes

Included in Survey Data Report:

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

4. Record Layout

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 survey data file.

- An item number.
- The variable's beginning and ending position within the file.
- The variable's length as measured in columns.
- The variable's data type (i.e., character or numeric).
- The variable name.

Figure 4.4, below, shows an example record layout format.

To help grantees construct a record layout for their data, the Work Group developed a recommended format and template. An example of this format is provided in Figure 4.4 and a template is provided in the Grantee Toolkit (see Chapter 6).

A record layout template is included in the Grantee Toolkit.

Figure 4.4: Recommended Record Layout Format

One row should be completed for each variable included in the survey 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** to use welfare outcomes survey data files. Grantees are encouraged to **append the following items** to their survey data file documentation.

- Copies of welfare outcomes grant **survey data final reports**.

- A statistical program **input statement** (e.g., SAS) or tool that may be used to read an 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).