Guidelines for Depositing Quantitative Data at NACJD for National Institute of Justice Sponsored Research

To facilitate secondary use, it is important to fully document and describe variables in the context of the data as well as in the documentation. This document provides guidance and a checklist to assist in preparation for archiving. Accordingly, NACJD staff review deposits for completeness and accuracy and send a report to NIJ. Any outstanding issues will need to be resolved before the submission can be finalized and assigned to a data processor. Please note the following:

**Confidentiality** — All fields directly identifying research subjects must be removed prior to deposit or transfer to NACJD (unless explicitly permitted and provided for through an IRB-approved process for informed consent). Examples of direct identifiers include names, addresses, and linkable numbers such as social security numbers, driver license number, prisoner identification number, etc. Data submitted with direct identifiers will be rejected and a resubmission of a repaired dataset will be required. Researchers should also carefully consider indirect or secondary identifiers and assess their analytic importance. Such variables make unique cases visible and may need to be removed before deposit. For example, a ZIP code variable may not be problematic on its own, but when combined with race and income variables, it may allow identification of unique individuals (e.g., the extremely wealthy).

**Deposit Checklist and Guidelines**

The following checklist details the items we are looking for with your submission:

- **Data** — Submit the final, sanitized data.
- **Format** — SPSS system files with embedded variable elements are preferred but other formats are acceptable as well (e.g., SAS, Stata).
- **Secondary analysis of existing data** — For projects that only involve analysis of existing data already publicly available where the product of the project is the analysis alone, data does not need to be deposited. Researchers are encouraged to deposit the code that produced the analysis and indicate that the project is a secondary analysis of existing data on the deposit form.
- **Analysis variables** — Submit the final, cleaned data file(s) used to produce analysis and tables in the final report including transformed, computed, or recoded variables that were constructed or derived from variables collected using the data collection instrument(s).
- **Project documents** — Submit copies of the final, approved Data Archiving Strategy, IRB approval, Privacy Certificate and Consent forms.
- **Syntax** — Describe the source of any transformed variables and the method for deriving them by depositing the recode syntax statements and/or providing a more extensive explanation in the documentation.

... checklist continued on reverse

**Guide to Archiving**

The Guide to Social Science Data Preparation and Archiving (http://tinyurl.com/d7ufjd6) is provided as a resource to assist you in the preparation of your data for use by your own project team and the research community. Additional guidance for addressing confidentiality can be found in Phase 5: Preparing Data for Sharing of The Guide.
✓ **Record counts** — Reconcile univariate statistics on each variable. Verify that the case count in the data file(s) match findings published in the final report. Secondary data analysts often confirm that they are reading the data properly by comparing the documentation with univariate statistics they produce from the data. Data producers are best positioned to reconcile record counts, out-of-range codes, skip patterns, and univariate distributions before deposit.

✓ **Variable labels** — Variable labels are extremely important and should provide information on the item or question number in the original data collection instrument (unless the item number is part of the variable name), a clear indication of the variable’s content, and an indication of whether the variable is constructed from other items.

✓ **Value labels/codes** — Assign a set of exhaustive and precisely defined codes to each variable and use the same codes across variables recording the same type of responses (e.g., 0 No, 1 Yes). Each interview response should fit into one and only one category. Provide each code with descriptive labels in the data file to aid proper documentation and understanding of the data content.

✓ **Missing data** — Assign separate missing codes for refusals, non-response, not applicable and other types of missing data for all variables in a data file, or across the entire collection if the study consists of multiple data files. Use numeric codes that are above the maximum valid value for the variable (e.g., 97, 98, and 99) because special missing characters are often lost when converting between software packages.

✓ **Congruence between data and data collection instrumentation** — Verify that the data and the instrument used to collect the data (where applicable) align.

✓ **Character variables** — Remove character (string) variables or comment/note fields that were not a part of the analysis. If character variables were coded to create new variables, they should be differentiated from those that were not used at all. If the data contain many character variables, please review NACJD’s *Deposit Guide for Qualitative Data*.

✓ **Documentation** — Provide all documentation needed for others to sufficiently understand the data. NACJD often distributes the original project documentation with the data. Documentation can include:

- Project report that describes the data collection as a whole including design, study rationale, descriptions of weighting, imputation, recodes
- Questionnaires or other data collection instruments
- Information on instances where variables were reverse coded for analysis purposes
- Codebook(s) that provides details about the variables, values, and formats
- Information about known data anomalies and/or caveats for working with the data
- Guidance for what questions the data can and cannot answer
- Glossary of frequently used terms (including any terms created/defined by project staff)
- Technical documentation that includes additional information about instruments used, decisions or changes made during the data collection process, information for merging files, guidance for selecting and using weights. Examples can be helpful and often secondary data users struggle with making decisions about which weight to select when conducting secondary analysis properly, even on a basic level.
- User guide that provides guidance for using the data appropriately which can include information to cover the points listed above