National Human Exposure Assessment Survey (NHEXAS)

Arizona Study

Quality Systems and Implementation Plan for Human Exposure Assessment

The University of Arizona
Tucson, Arizona 85721

Cooperative Agreement CR 821560

Standard Operating Procedure **SOP-UA-D-4.0**

Title: The Generation and Operation of Data Dictionaries

Source: The University of Arizona

U.S. Environmental Protection Agency
Office of Research and Development
Human Exposure & Atmospheric Sciences Division
Human Exposure Research Branch

Notice: *The U.S. Environmental Protection Agency (EPA), through its Office of Research and Development (ORD), partially funded and collaborated in the research described here. This protocol is part of the Quality Systems Implementation Plan (QSIP) that was reviewed by the EPA and approved for use in this demonstration/scoping study. Mention of trade names or commercial products does not constitute endorsement or recommendation by EPA for use.*

The Generation of Data Dictionaries

1.0 Purpose and Applicability

The purpose of this procedure is to provide a standard method for the writing of data dictionaries. This procedure applies to the dictionaries generated by the NHEXAS study, the Border Study, and other Health and Environment projects.

2.0 Definitions

2.1 ASCII DATA FILE = An ASCII file which holds data to be checked by a specific dictionary. This data will be stored with each variable positioned at a predefined location.

2.2 BORDER STUDY = An alias for "Total Human Exposure Arizona: A Comparison of the Border Communities and the State" conducted in Arizona by the University of Arizona/Battelle/Illinois Institute of Technology consortium.

2.3 DEBUGGING = The act of inspecting a process and fixing any errors that may arise.

2.4 DICTIONARY = A program that evaluates an ASCII data file for a specific structure and performs basic range and logic checking functions.

2.5 DICTIONARY FORMAT = Format that determines dictionary syntactic correctness.

2.6 FORM TABLE = An AWK program associated to a specific form that will take a downloaded data file for that form and create an ASCII data file.

2.7 HEALTH AND ENVIRONMENT PROJECTS (or H&E) = An umbrella title for all projects funded to M.D. Lebowitz and/or M.K. O'Rourke (or their designees) which examine purported or real relationships among environmental factors and any aspect of human health.

2.8 MASS DATA MASSAGE PROCESS (or MDM) = The data processing program used by NHEXAS Arizona, the Border Study, and other Health and Environment projects.

2.9 NHEXAS Arizona: Acronym for **N**ational **H**uman **EX**posure **A**ssessment **S**urvey, a research project conducted in Arizona by the University of Arizona/Battelle/Illinois Institute of Technology consortia.

2.10 RUNNING A DICTIONARY = The act of applying a dictionary to an ASCII data file.

2.11 SCHEMA = A file which lists variables, their types, and their lengths.

3.0 References

"SPSS for UNIX: Operations Guide"

4.0 Discussion

This SOP describes the procedures for dictionary generation. Dictionaries serve two purposes. First, dictionaries provide a complete description of the databases to which they are associated. This description includes each variable name in the database, the position of each variable, and the type of each variable. Also, the values each variable may take are given, and logic checks that may be used to determine the appropriateness of a value are also listed. Note that dictionaries are dynamic objects, and are subject to frequent modifications and updates.

The second purpose is to serve as a template for the MDM to use for data checking purposes. This function is invisible to the user, and is fully described in SOP# UA-D-44.X.

5.0 Responsibilities

The Project Data Manager is responsible for the creation, accuracy, and thoroughness of each data dictionary. This responsibility may be delegated.

6.0 Materials and Equipment

6.1 Hardware: It is necessary to work on the LAN. This requires use of a workstation in the H&E network.

6.2 Software: The following software is required in the construction and operation of each dictionary.

a. Teleform V5 Standard
b. *convert*, UNIX binary
c. *generate*, UNIX binary
d. *gentab*, UNIX binary

7.0 Procedure

7.1 Preliminaries: Obtain schema, form table, and forming dictionary skeleton

1. Run Teleform.
2. Open the form for which dictionary is to be written.

3. Save report to a file. The download directory should be /rsc51/luisf/NhexasDict/reports/dos. The filename (without extension) should reflect the form, and shall be referred to as the *base* name.
4. Log into the UNIX environment
5. Change directory to /rsc51/luisf/NhexasDict/reports
6. Run the *convert* program, which has usage:
   convert <base>
7. Change directory to /rsc51/luisf/NhexasDict
8. Run the *generate* program, which has usage:
   generate <base>
9. Run the *gentab* program, which has usage:
   gentab <base>
10. The dictionary skeleton, schema, and form table will be found in /rsc51/luisf/NhexasDict/dicts/<base>

7.2 Dictionary Command Format: Once a dictionary skeleton is obtained, it may be necessary to add logic and range checking. The following describes the available dictionary commands.

1. Database Identification Record: This record identifies the database and form, and provides its location, as well as the number of records the physical format requires. One and only one record of this type may exist. Logically, this should be the first record of the dictionary:
   a. Field 1: D
   b. Field 2: Form or Questionnaire name
   c. Field 3: Location of database (pathname in UNIX format)
   d. Field 4: Number of tables in this database.
2. Comment Record: This record provides a documentation facility to provide comments at any point for any reason within the dictionary.
   a. Field 1: *
   b. Field 2: Any text desired.
3. Variable Specification Record: This record names the data field, and specifies the record number and starting position on the record, as well as the format under which the variable should be read as though it existed on a fixed-format ASCII record.
   a. Field 1: Q
   b. Field 2: Variable name (One letter followed by up to 7 alphanumeric characters).
   c. Field 3: Record number.
   d. Field 4: Initial column position on record.
   e. Field 5: Format: May be one of the following
      A<n>: ASCII field of length n
      I<n>: Integer field of length n

F<w>.<d>: Floating point field, of width w and d places after the decimal

4. Text of Question Record: This question provides the full, word-for-word text of the question as on the questionnaire or form.
    a. Field 1: T
    b. Field 2: Text of question
5. Variable Label Record: This is the variable label that will be associated with the variable.
    a. Field 1: V
    b. Field 2: Variable label
6. Value Label Record: These specify any value labels that are to be associated with certain values of the variables. If this record is included, these values as well as any declared missing values and ranges will be the only allowable values the variable may have.
    a. Field 1: Y
    b. Field 2: Value (no quotes)
    c. Field 3: Value label (no quotes)
7. Logical Validation Record: These records are actual expressions in the target language that must be met for the data to be considered valid.
    a. Field 1: L
    b. Field 2: X
    c. Field 3: Logical Expression (ref. "SPSS for UNIX: Operations Guide")
8. Logical Expression Continuation: Continue a logical expression initiated by preceding logical validation record. This allows for logical statements to be defined over multiple lines.
    a. Field 1: +
    b. Field 2:
        C: Continue Logical Expression
        F: Finish Logical Expression
    c. Field 3: Logical Expression (ref. "SPSS for UNIX: Operations Guide")
9. Range of Value Record: This record describes the full range of values allowable for the variable. It may be in addition to the values declared on the label records.
    a. Field 1: R
    b. Field 2: (
    c. Field 3: Minimum value
    d. Field 4: ,
    e. Field 5: Maximum value
    f. Field 6:)
10. Exceptional Case Record: This record would not be part of the original definition of the database, but rather may be added to the dictionary at a later date. Its purpose is to document exceptional circumstances within the data that may need to be considered in understanding the values of certain cases.
    a. Field 1: X

b. Field 2: Case ID
c. Field 3: Description of circumstances

11. Keyword Record: This record lists any number of keywords that can be used in locating variables to be used for analysis.
    a. Field 1: I
    b. Field 2: Keyword(s)
12. Define Key Variable(s): Establish parameters on which all proceeding variables are dependent on (until a new set of key variables are defined). If no keys are defined, the default key variables are "HHID" and "PID".
    a. Field 1: K
    b. Field 2: Variable list (comma delimited)
13. Print Related Variables: Print variables mentioned with report of error (after key variables and before current variable). Thus, the error report format would be: "----> <key vars>, <related vars>, <current var>"
    a. Field 1: P
    b. Field 2: Variable list (comma delimited)

7.3 File Placement

1. Place the form table file in:
   /rsc53/NHEXAZdata/download/<base>
2. Place the dictionary file in:
   /rsc53/NHEXAZdata/inproces/<base>

8.0 Records

Each dictionary has, in its header, the date and author of last modification. Also in the header is the date of last verification, and the name of the verifier.

Dictionaries built for each form are found as an Appendix in each coding protocol as described below:

SOP #	SOP Title	Dictionary
UA-D-17.X	Cleaning: Descriptive Qx	Appendix B
UA-D-18.X	Cleaning: Baseline Qx	Appendix B
UA-D-20.X	Cleaning: Time Diary & Activity Qx	Appendix B
UA-D-21.X	Cleaning: Follow Up Qx	Appendix B
UA-D-36.X	Cleaning: Technician Walk-Through Qx	Appendix B
UA-D-38.X	Cleaning: Field Data	
	Floor Dust Sampling	Appendix B
	Household Summary Sampling	Appendix D
	Personal Air Sampling	Appendix F
	PID Sampling Sheet	Appendix H

SOP #	SOP Title	Dictionary
	PM Sampling Data Sheet	Appendix J
	Sentinel Data Sheet	Appendix L
	Soil Sampling	Appendix N
	Surface Sampling	Appendix P
UA-D-45.X	Cleaning: Diet Diary Qx	Appendix B
UA-D-46.X	Cleaning: Food Diary Followup Qx	Appendix B
UA-D-47.X	Cleaning: Questionnaire Feedback	Appendix B