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-3.0**

Title: Defining Working Databases and Data Entry Forms (Hand Entry)

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.*

Defining Working Databases and Data Entry Forms

1.0 Purpose and Applicability

The purpose of this procedure is to outline a standard approach to naming and defining variables, data types, and data entry forms. This procedure applies to all working databases created for NHEXAS Arizona. It does not apply to data files created specifically for statistical analysis.

2.0 Definitions

2.1 DATA = Classified under this word are the following definitions: DATA, ELECTRONIC; DATA, ENTERED; DATA, PHYSICAL; DATA, RANGE CHECKED; DATA TYPE; DATA, VERIFIED; DATA RECORD.

2.1.1 DATA, ELECTRONIC = Data stored on some type of magnetic medium (eg., floppy disk, hard disk, bernoulli, tape)

2.1.2 DATA, ENTERED = Electronic data entered for the first time into a computer database. Entered data are the product of "data entry."

2.1.3 DATA, PHYSICAL = Datum or data written on a physical data form

2.1.4 DATA, RANGE CHECKED = A data record where the value of each variable in the record was compared against a preestablished valid range. If the value is out of range and valid, the range parameters are re-defined. If the value is out of range and invalid, it is corrected.

2.1.5 DATA RECORD = In the context of this SOP, this is a row of data in a database.

2.1.6 DATA TYPE = the format in which data is stored in a column or variable within a data record. R:BASE data types include (1) date, (2) time, (3) currency, (4) real, (5) double, (6) integer, (7) text, and (8) note (see *R:BASE for DOS: User's Manual* [cited below in 3.3] for an explanation of each data type).

2.1.7 DATA, VERIFIED = Electronic data re-entered into the same table and database into which it was originally entered, and programatically compared against the original entered data. Verified data are the product of "data verification."

2.2 DATABASE, MASTER = Accumulative database generated from validated data processing batches. Newly cleaned batches are appended to the master database. Copies of this database are used in analyses. All corrections made to copies of the master are made to the master database itself. Thus, it is the most complete and accurate database of its kind.

2.3 DATABASE, WORKING = A database earmarked for or in the process of cleaning that contains one or more data processing batches. When

cleaned, this will be appended to the master database with the same name.

2.4 FORM, DATA ENTRY = A computer screen representation of a physical data form. The data on the physical form is entered into the computer database via the data entry form.

2.5 FORM, PHYSICAL (DATA) = Paper or "hard copy" forms or questionnaires.

2.6 KEY VARIABLE(S) = The variable or set of variables that makes a data record unique from other data records in the same table or file.

2.7 KEYPUNCH = The primary area in which data entry and data verification of NHEXAS Arizona field data takes place. It is located in the Respiratory Sciences Center, Room 2329; Arizona Health Sciences Center (AHSC); 1501 N. Campbell Avenue; University of Arizona; Tucson, AZ 85724.

2.8 LAN = abbreviation for **L**ocal **A**rea **N**etwork. This is any physical network technology that operates at high speed over short distances.

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 PARAMETER ROWS = The five data records in an R:BASE table that define the following for all variables contained therein: (1) the upper range limits, (2) the lower range limits, (3) the global codes for refusal, (4) the global codes for non-applicable, and (5) the global codes for missing.

3.0 References

3.1 *Microsoft MS-DOS Operating Systems V5.0 User's Guide and Reference.* Microsoft Corp. 1984.

3.2 *R:BASE for DOS: Building Applications Command Dictionary* (First Edition, Version 2.0). December, 1987. Redmond: Microrim, Inc.

3.3 *R:BASE for DOS: User's Manual* (First Edition, Version 2.0). December, 1987. Redmond: Microrim, Inc.

3.4 *R:BASE for DOS: Utilities Manual* (First Edition, Version 2.0). December, 1987. Redmond: Microrim, Inc.

3.5 *R:BASE for DOS: Command Summary* (First Edition, Version 2.0). December, 1987. Redmond: Microrim, Inc.

3.6 *R:BASE for DOS: Developer's Express Guide* (First Edition, Version 2.0). December, 1987. Redmond: Microrim, Inc.

3.7 *R:BASE for DOS: Learning Guide* (First Edition, Version 2.0). December, 1987. Redmond: Microrim, Inc.

4.0 Discussion

A standard approach to defining the variable names, data types, and data entry forms in working databases leads to greater ease and efficiency in data entry, cleaning, and analysis. Database definition affects both working and master databases because when the Project Data Analyst builds the master database, then she or he retains most of the names and structure initially defined in the working database. If the names and structure are as standard as possible, then memorization and access is easier for all users, from Student Data Assistant to Principle Investigator.

For example, the standard variable name for household identification number is *HHID* and its standard data type is INTEGER; the standard variable name for data processing batch number is *DPBATCH*, and its standard data type is TEXT (two-character). These should be defined consistently across all NHEXAS databases, with few exceptions if any.

A uniform approach to defining the data entry form (i.e., screen) provides a minimum standard that should result in the minimization of inaccurate data entry. The aim is to reproduce the hard copy formatting as much as possible on the data entry form. For example, if physical data are listed vertically on the physical form, then the data fields corresponding to those data should be placed vertically on the data entry form.

5.0 Responsibilities

5.1 The Project Data Coordinator is responsible for the following:

5.1.1 Defining working databases and data entry forms or overseeing their definition by Student Data Assistants;

5.1.2 Consulting with the Project Data Analyst, Principle Investigator, Project Field Coordinator, and/or author of the physical form in order to gain an understanding of its meaning and purpose, and to inquire about any preferences in names, data types, formats, or additional variables not included on the physical form;

5.1.3 Ensuring that this SOP is followed.

5.2 The Student Data Assistant is responsible to the Project Data Coordinator.

5.3 The Project Data Analyst is responsible for providing, if necessary, preferences for names, data types, formats, or additional variables not included on the physical form.

5.4 The On-Site Principle Investigator is responsible for providing, if necessary, preferences for names, data types, formats, or additional variables not included on the physical form.

5.5 The author of the physical form is responsible for explaining the meaning and purpose of the form, especially all data to be electronically entered.

6.0 Materials and Reagents

6.1 Materials

6.1.1 Personal Computer (PC) connected to the LAN
6.1.2 Dot matrix or laser printer attached to a PC or the LAN
6.1.3 R:BASE for DOS version 2.11
6.1.4 R:BASE instruction manuals (see citation in section 3.0)
6.1.5 Paper and writing instrument for designing data entry form and naming variables
6.1.6 Four to five blank physical forms for which the working database and data entry form are being created
6.1.7 Any relevant information from the On-Site Principle Investigator, Project Data Coordinator, Project Data Analyst, Project Field Coordinator, and/or author of the physical form
6.1.8 Previous form(s) or questionnaire(s) containing the same or similar questions, in order to obtain their previous variable name(s) (or variable naming strategy) and data type(s)
6.1.9 "Database Definition: Phase 2" notebook
6.1.10 "Database Definition: Phase 3" notebook
6.1.11 "Database Definition: Phase 4" notebook
6.1.12 "Database Definition: NHEXAS Phase" notebook (to be compiled)
6.1.13 "Database Definition Record" form
6.1.14 *Data.exe* (date = 7/11/90; 5:37 pm; 124,782 bytes): This executable C program works with all databases defined using R:BASE for DOS version 2.11. It is used for data entry, data verification, range checking, and data correction documentation.

6.2 Reagents

None

7.0 Procedure

7.1 Research previous form(s) or questionnaire(s) containing the same or similar questions, in order to obtain their previous variable name(s) and data type(s), and/or the previous variable naming strategy. Do your research using one or more of the following notebooks: "Database Definition: Phase 2" notebook, "Database Definition: Phase 3" notebook, or "Database Definition: Phase 4" notebook. Take adequate notes.

7.2 Name and define all variables to receive data entry. Use standard names and data types wherever applicable.

7.3 Design and define the data entry form. Consult the R:BASE manuals for suggestions. Maximize computer space by designing an efficient database -- the fewer the variables and characters, the less storage space is needed on the LAN.

7.4 Test the data entry form with "dummy" data (or real data if available) from the physical form in question. Attempt to enter and verify the data using *Data.exe*.

7.5 If the data entry and verification functions work, then define and enter the parameter rows.

7.6 Test the range checking function of data.exe. Be sure all functions of Data.exe work before proceeding to the next step.

7.7 Prepare the final draft of the variable names, data types, and design of data entry form to be filed in the "Database Definition: NHEXAS Phase" notebook.

7.7.1 Write all variable names on a blank physical form.

7.7.2 In the upper left corner, write the names of the
(a) database software package used,
(b) database,
(c) tables, if R:BASE,
(d) files and catalog, if dBase,
(e) data entry form,
(f) the research phase (i.e., "NHEXAS Phase"), and
(g) the research stage, if applicable.

7.7.3 On another blank physical form, for each data field, write an X for each character in the field. For date and time fields, include slashes or colons between the X's in the appropriate place(s). For text fields greater than eight characters, enclose the total number of characters in parentheses to the right of the X's. This is done so that Keypunch knows which data to code and enter, as well as how many leading zeros to include, and where.

For example:

```
            HHID: XXXX
 Completion date: XX/XX/XX
      First Name: XXXXXXXXXXXXXXX (15 char. allowed)
```

7.8 Make print-outs of the data entry form and table listing(s).

7.9 File all items generated in steps 7.7.1 through 7.7.3 in the "Database Definition: NHEXAS Phase" notebook according to database name.

7.10 The following is a chart of current standard variable names, their data types, and their meaning. Text data types are followed by a number indicating how many characters are in the data field. If the number of characters in a text data type varies from physical form to physical form, then this is indicated with an *n*.

Variable Name	**Data Type**	**Meaning**
COMMENT	Integer	Comment code indicating status of field sampler, lab result, or question
COM_____	Integer	Prefix for comment codes of multiple field samplers, lab results, or questions on the same physical form

Variable Name	Data Type	Meaning
DATECOMP	Date	Date of questionnaire completion
DATELAB	Date	Date of laboratory analysis
DATESAMP	Date	Date of field sampling
DEDATE	Date	Date of data entry
DPBATCH	Text 2	Data processing batch code
ENDDATE	Date	Stop date of field sampling
ENDTIME	Time	Stop time if field sampling
HHID	Integer	Household identification number
HHIDEXT	Text 2	Code used for tracking HHID
IRN	Integer	Individual Respondent Number
LOCATION	Integer	Location of field sampling
LOC_____	Integer	Prefix for locations of multiple sampling sites on the same physical form
PKTCOLOR	Text 3	Packet color code
QABY	Text 3	Initials of Quality Assurance performer
QCBY	Text 3	Initials of Quality Control performer
QADATE	Date	Date Quality Assurance check performed
QCDATE	Date	Date Quality Control check performed
QXV	Text *n*	Questionnaire version code
STDATE	Date	Start date of field sampling
STTIME	Time	Start time of field sampling
VERIF	Integer	Verification code (used by *Data.exe*)
VISIT	Integer	Number of household visit (sequential)

7.11 Standards and Blanks

None

7.12 Quality Control

The Project Data Coordinator reviews any working databases and data entry forms before they are installed at Keypunch. She or he tests the screens for function, ease of use, and clarity; and approves it by writing initials and current date on the "Working Database Definition Record" form. A second person, other than the Project Data Coordinator, also tests the data entry form. He or she also writes initials and current date on the "Working Database Definition Record" form.

8.0 Records

8.1 Location of Data Processed by this Procedure

N/A

8.2 Record Forms (attached)

8.2.1 Figure 1: Physical form containing variable names and other pertinent information described in section 7.7.2 of this SOP (example)

8.2.2 Figure 2: Physical form containing X's for variable lengths described in section 7.7.3 of this SOP (example)

8.2.3 Figure 3: Print-out of data entry form (example)

8.2.4 Figure 4: Print-out of R:BASE table listing (example)
8.2.5 Figure 5: "Database Definition Record" form

8.3 Location of Forms

The forms generated from this procedure are filed by database name in the "Database Definition: NHEXAS Phase" notebook. This notebook is housed in room 128 of Health Related Professions; 1435 North Fremont Avenue; Tucson, AZ 85719.

Figure 1: Physical form containing variable names and other pertinent information described in section 7.7.2 of this SOP (example)

Database : ENSUPP
Table : ADSUPP1
Form : ADSUPP1
Phase 4/Allergy Panel

PKTCOLOR
HHID: HHIDEXT
HHID label
DATECOMP

ADULT SUPPLEMENT #1

- The following information will be used only to categorize the large population we are studying. This information will not be released in ANY FORM to ANYONE that could identify YOU as an individual. Please circle the best response to each question and answer all questions as completely as possible. If you choose not to respond to a question, please cross out the question (My dog has fleas.). This way we will not think the question was "missed" and will not ask you for a response. After completing this questionnaire, PLEASE make sure that ONLY your first name and identification number are on this form, and SEAL IT IN THE ENVELOPE PROVIDED. This is to assure complete confidentiality.

1. First name only): NAME (PID# PID)

2. What is the approximate combined income of your household BEFORE TAXES? INCOME

a. Less than $11,000
b. $11,000 to $14,999
c. $15,000 to $19,999
d. $20,000 to $24,999
e. $25,000 to $29,999
f. $30,000 to $34,999
g. $35,000 to $39,999
h. $40,000 to $44,999
i. $45,000 to $49,999
j. $50,000 to $54,999
k. $55,000 to $59,999
l. $60,000 to $64,999
m. $65,000 to $69,999
n. $70,000 to $74,999
o. $75,000 to $79,999
p. $80,000 to $84,999
q. $85,000 to $89,999
r. $90,000 to $94,999
s. $95,000 to $99,999
t. $100,000 to $149,999
u. Over $150,000

3. In which religion were you RAISED? RELIGION

a. Catholic
b. Protestant (i.e. Lutheran, Presbyterian, Methodists, Baptist)
c. Fundamentalist Christian
d. Judaism
e. Mormon
f. Islamic
g. None/Atheist
h. Other (please specify): RELIGOTH

4. What is your ethnic group? ETHNIC

a. White
b. Hispanic
c. African American
d. American Indian
e. Other (please specify): ETHNIOTH

THANK YOU FOR YOUR TIME AND SUPPORT!!

For Office Use Only
Comp Check; by: QCBY date: QCDATE
QA Check; by: QABY date: QADATE
Data entry date: DEDATE Batch#: DPBATCH
Qx version/FORM: #2HC allergy panel
QXV

Figure 2: Physical form containing X's for variable lengths described in section 7.7.3 of this SOP (example)

PACKET COLOR: XXX
HHID: XXXX:X
HHID label
COMPLETION DATE: XX/XX/XX

ADULT SUPPLEMENT #1

The following information will be used only to categorize the large population we are studying. This information will not be released in ANY FORM to ANYONE that could identify YOU as an individual. Please circle the best response to each question and answer all questions as completely as possible. If you choose not to respond to a question, please cross out the question (My dog has fleas.). This way we will not think the question was "missed" and will not ask you for a response. After completing this questionnaire, PLEASE make sure that ONLY your first name and identification number are on this form, and SEAL IT IN THE ENVELOPE PROVIDED. This is to assure complete confidentiality.

1. First name (only): XXXXXXXXXXXXXXXXXX (20 char.) (PID# XXX)

2. What is the approximate combined income of your household BEFORE TAXES? X

a. Less than $11,000
b. $11,000 to $14,999
c. $15,000 to $19,999
d. $20,000 to $24,999
e. $25,000 to $29,999
f. $30,000 to $34,999
g. $35,000 to $39,999
h. $40,000 to $44,999
i. $45,000 to $49,999
j. $50,000 to $54,999
k. $55,000 to $59,999
l. $60,000 to $64,999
m. $65,000 to $69,999
n. $70,000 to $74,999
o. $75,000 to $79,999
p. $80,000 to $84,999
q. $85,000 to $89,999
r. $90,000 to $94,999
s. $95,000 to $99,999
t. $100,000 to $149,999
u. Over $150,000

3. In which religion were you RAISED? X

a. Catholic
b. Protestant (i.e. Lutheran, Presbyterian, Methodists, Baptist)
c. Fundamentalist Christian
d. Judaism
e. Mormon
f. Islamic
g. None/Atheist
h. Other (please specify): XXX

4. What is your ethnic group? or ancestry? XXX

a. White
b. Hispanic
c. African American
d. American Indian
e. Other (please specify): XXX

THANK YOU FOR YOUR TIME AND SUPPORT!!

For Office Use Only
Comp Check: by: XXX date: XX/XX/XX
QA Check: by: XXX date: XX/XX/XX
Data entry date: XX/XX/XX Batch#: XX
Qx version/FORM: #2HC allergy panel
XXX

Figure 3: Print-out of data entry form (example)

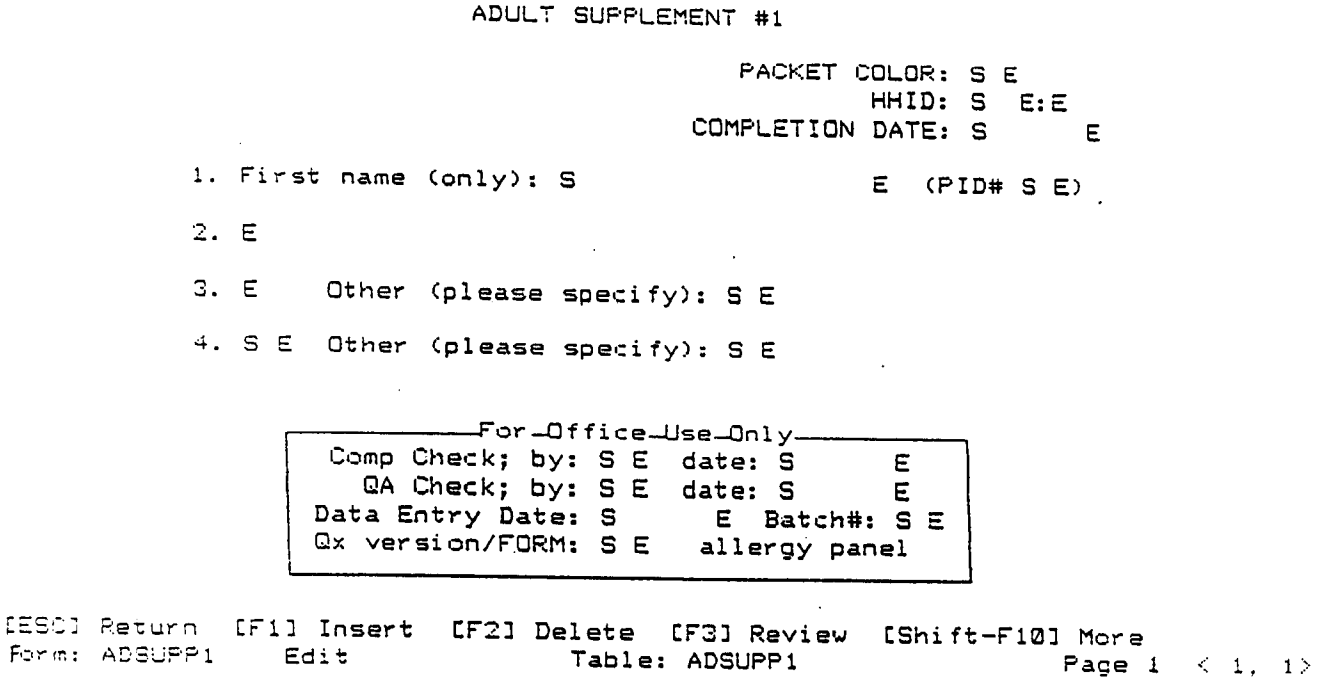

Figure 4: Print-out of R:BASE table listing (example)

Table: ADSUPP1 No lock(s)
Read Password: No
Modify Password: No

Column definitions

#	Name	Type	Length	Key	Expression
1	HHID	INTEGER		yes	
2	HHIDEXT	TEXT	1 characters	yes	
3	DATECOMP	DATE		yes	
4	VERIF	INTEGER			
5	PKTCOLOR	TEXT	3 characters		
6	NAME	TEXT	20 characters		
7	PID	INTEGER		yes	
8	INCOME	TEXT	1 characters		
9	RELIGION	TEXT	1 characters		
10	RELIGOTH	INTEGER			
11	ETHNIC	INTEGER			
12	ETHNIOTH	INTEGER			
13	QCBY	TEXT	3 characters		
14	QCDATE	DATE			
15	QABY	TEXT	3 characters		
16	QADATE	DATE			
17	DEDATE	DATE			
18	DFBATCH	INTEGER			

Column definitions

#	Name	Type	Length	Key	Expression
19	QXV	TEXT	3 characters		

Current number of rows: 5

Figure 5: "Database Definition Record" form

DATABASE DEFINITION RECORD

NHEXAS Arizona

Form ID: UA-D-3.0-1.0

Database Name:_____
Data Entry Form Name:_____
Table Name(s):_____
Physical Form Name:_____
NHEXAS Form ID:_____

Created by:_____ Date:___/___/___
Tested by:_____ Date:___/___/___
Parameter rows defined by:_____ Date:___/___/___
Data Coordinator Approval:_____ Date:___/___/___

[ ] Dictionary Name(s):_____
Created by:_____ Date:___/___/___
Tested by:_____ Date:___/___/___
[ ] Dictionary updated by:_____ Date:___/___/___
Describe update:_____
_____
_____

Database updated by:_____ Date:___/___/___
Describe update:_____
_____
_____

Database Installed
(please check, initial, date, and write sub-directory, if applicable, for each location installed):

[ ] AHSC Site Keypunch:
Installed By:_____ Date Installed:___/___/___
Sub-directory Location:_____
Node Name:_____

[ ] HRP Site:
Installed By:_____ Date Installed:___/___/___
Sub-directory Location:_____
Node Name:_____

[ ] Other:_____
Installed By:_____ Date Installed:___/___/___
Sub-directory Location:_____
Node Name:_____

Items Filed in "Database Definition: NHEXAS Phase" Notebook
(please check each item filed):

[ ] Physical form on which variable names are written
[ ] Physical form on which field lengths are indicated with X's
[ ] Physical form on which ranges are written for each field
[ ] Variable name and type output (generated by database software)
[ ] Data entry screen format: # of pages:_____

Above items filed by:_____ Date filed:___/___/___