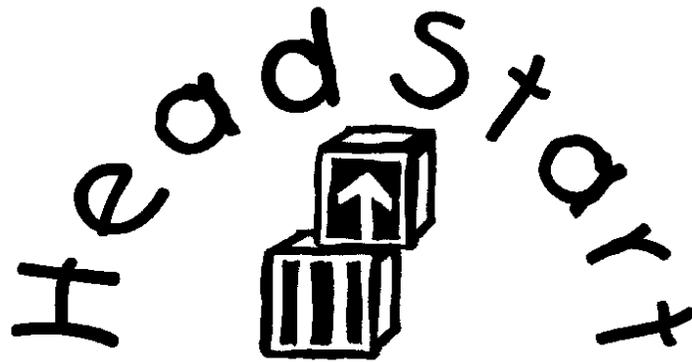


Data Archive on Head Start Program Performance Measures

A Guide for Data Submission



**Data Archive
Head Start Quality Research Consortium's
Performance Measures Center
Westat
1650 Research Boulevard
Rockville, MD 20850**

December 2001

Table of Contents

Acknowledgments..... iii

I. Data Archive on Head Start Program Performance Measures 1

II. Preparing Data for Archiving: Principal Investigators’ Responsibilities..... 2

 Submitting Materials..... 2

 Media 3

 Format 3

 Data Preparation 4

 Data Collection Instruments 5

 Codebooks..... 5

 Data Description 6

 Bibliography 6

 Summary Statistics..... 7

 Responding to Requests for Additional Information..... 7

 Reviewing Draft Materials..... 7

III. Preparing Data for Dissemination: Data Archive of the HSPMC’s Responsibilities.... 8

 Preparing Data Files..... 8

 Preparing a User’s Guide..... 8

 Reviewing the Codebook..... 9

 Making Data Sets Available 9

APPENDIX A. SAS Transport Files 10

APPENDIX B. SPSS Portable Files 11

APPENDIX C. Data Tape documentation..... 12

APPENDIX D. Staff Directory..... 13

APPENDIX E. Data Deposit Form..... 14

ACKNOWLEDGMENTS

Funding for the Data Archive on Head Start Program Performance Measures is provided under the contract (No. 105-96-1912), titled “the Head Start Quality Research Consortium’s Performance Measures Center,” to Westat from the Administration on Children, Youth, and Families (ACYF) of the U.S. Department of Health and Human Services.

We thank Patrick T. Collins and NDACAN at Cornell University for allowing us to adopt text from *Depositing Data with the National Data Archive on Child Abuse and Neglect: A Handbook for Investigators*.

I. DATA ARCHIVE ON HEAD START PROGRAM PERFORMANCE MEASURES

The archiving of data is a cooperative effort between the investigators and the Data Archive of the Head Start Performance Measures Center (HSPMC). The investigators will submit machine-readable data and documentation, and the Data Archive staff of the HSPMC will examine, edit, and archive the submitted data and documentation. In this process, Data Archive staff may identify technical difficulties with the data or the documentation, and will work with the investigators to clarify discrepancies.

All archived data sets and their documentation are available from the Data Archive of the HSPMC, housed at Westat. A listing of the data sets in the Archive and a brief description of each may be obtained by contacting:

Data Archive of the HSPMC
Westat
(Attn: Kwang Kim, TB386)
1650 Research Boulevard
Rockville, MD 20850
(301) 517-4078
(240) 453-2650 (FAX)
E-Mail: kimk1@westat.com

II. PREPARING DATA FOR ARCHIVING: PRINCIPAL INVESTIGATORS' RESPONSIBILITIES

Principal investigators are responsible for:

- (1) Submitting data and supporting material.
- (2) Responding to requests by the Data Archive staff of the HSPMC for additional or clarifying information, if needed.
- (3) Completion of Data Deposit Form (Appendix E).
- (4) Reviewing and correcting draft materials prepared by the Data Archive staff of the HSPMC.

Submitting Materials

Principal investigators must submit the following materials/information to the Data Archive of the HSPMC:

- (1) Data file(s).
- (2) Description of data files.
- (3) Data collection instrument(s).
- (4) References for data collection instruments, including unpublished citations.
- (5) Codebook or data dictionary.
- (6) Explanation of derived (computed) variables.
- (7) Final project report, project summary, or other description of the project.
- (8) Printout of the first and last data records.
- (9) Data Deposit Form (Appendix E).
- (10) Summary statistics (frequency distributions, means, etc.) of all variables (if available).

Data files and documentation must be submitted in **machine-readable format**. A variety of transmittal media and file formats are acceptable, as listed below.

Media

- (1) **Microcomputer Diskettes or CD-ROM.** Microcomputer diskettes or CD-ROM are the preferred media for receiving electronic files. Submit two copies of each diskette or CD-ROM. Data may not be submitted via Internet.
- (2) **Mainframe tapes.** Although mainframe tapes are not as widely used as in the past, investigators can submit data on mainframe tape. The investigators must provide additional documentation as described in Appendix C. The investigator should indicate on the tape or diskette what operating system (VMS, DOS, UNIX, Mac/OS, etc.) was used to write the data file.

Tracks:	Reel Tapes: 9-track only
	Cartridge Tapes: 18-track only
Recording density:	Reel Tapes: 6250 or 1600 BPI
	Cartridge Tapes: 38K only
Labels:	IBM standards label or non-labeled
Code:	EBCDIC or ASCII
Record Format:	fixed or variable
Blocking:	blocked or unblocked
Block Size:	18-32760 for EBCDIC, 18-2048 for ASCII

Format

- (1) **Data files.** Data files can be submitted as SAS transport files, SPSS portable files, or ASCII (DOS) text files. ASCII data files must be accompanied by a listing of the record and column locations of each variable as well as any other information necessary to accurately read the variable (e.g., missing value codes, implied decimal places, and special formats).

Appendices A and B explain what SAS transport and SPSS portable files are and how to create them. Submission of SAS or SPSS files is encouraged because these file formats preserve metadata such as the variable names, variable labels, and value labels that the investigator has already created; however, the submission of data via ASCII (DOS) files is acceptable.

- (2) **Documentation files.** Documentation files can be submitted in either Microsoft Word or WordPerfect (for DOS or Windows formats). Consult with the Data Archive staff if other common file formats (e.g., ClarisWorks, Pagemaker) are used. All documentation must **also** be submitted in hard copy.

- (3) **Qualitative Data.** Investigators collecting qualitative information also may submit their data for archiving. Before submitting the data, text files should be transmitted as unformatted ASCII (DOS) files.

Data Preparation

Investigators should keep in mind the following important requirements:

- (1) **Confidentiality.** Information that could identify an individual (e.g., name, address, or Social Security Number) should not be included on the data file. If identification variables are needed to link multiple files, these should be created variables and not names or numbers that could be traced to individuals.
- (2) **File Linking.** If the data set consists of two or more related files, variables that link the files should be included on each file and described in the documentation. The nature of the relationship between files and the unit of observation for each file should be described clearly in accompanying documentation. Identification numbers should match between files; cases with identification numbers that do not match should be noted in the documentation. Also, each source file must be sorted in ascending order by the case ID number.
- (3) **Rectangular File.** In general, rectangular files with rows corresponding to cases and columns corresponding to variables are preferred over hierarchical files. Complex hierarchical files should, if at all possible, be converted to a series of linkable rectangular files.
- (4) **Variable Label.** When submitting SAS or SPSS files, investigators are encouraged to include a descriptive variable label for each variable in the file.
- (5) **Variable Code.** Each variable should have a set of exhaustive, mutually exclusive codes. These codes should be thoroughly documented in the codebook. Where possible, standard data codes (e.g., FIPS codes) should be used. The use of such codes facilitates linking and comparison of files and results across studies.
- (6) **Variable Format.** Data should be reduced to numeric codes whenever possible. If alphabetic codes are used, they should be identified clearly in the codebook as alphabetic so that users of the data will know to use alphabetic formats to read such data.
- (7) **Missing Data.** Missing data should be coded differently from inapplicable data. As a rule, blanks should not be used for coding variables since some statistical programs do not provide for labeling of blank values.

- (8) **Derived Variable.** Variables created by transforming other variables may be included. The transformations and the transformed variables' values should be thoroughly documented. Such documentation should include the programs used to create the derived variables, if available. Derived variables that the investigator has determined are not of sufficient quality to be used in any analysis should be removed from the file before submittal.
- (9) **Weighting Variable.** If weights are developed that adjust the data, they should be included in the data file. The locations of the weighting variables should be documented in the codebook. Instructions for using the weights should also be described.
- (10) **Data Integrity.** Investigators should check for out-of-range codes, incorrect skip patterns, and codes that are internally inconsistent. Errors identified through these checks should be corrected on the data file.

Data Collection Instruments

Data collection instruments and instructions to interviewers should be submitted. One clean, unused copy of each instrument, including interview schedules, self-administered questionnaires, data collection forms for transcribing information from records, paper tests and scales, screening forms, and call-report forms, should be submitted. If different forms of the instrument were used, each form should be included along with a description of the circumstances in which each was used (study populations, time periods, etc.).

Codebooks

Codebooks must have complete information for every variable in the data file. Codebooks should be submitted in machine-readable form in ASCII, WordPerfect, or Microsoft Word files. An SPSS dictionary file is acceptable, *providing the dictionary describes the variables and their values completely.*

For each variable in the file, the codebook should contain the following information:

- (1) **Variable Name.** An unique, unambiguous name for each item.
- (2) **Variable Label.** A textual description of the item, or a reference to the question, if from a questionnaire. When submitting SAS or SPSS files, this description should be included in the variable label.
- (3) **Record and Column Location.** When submitting ASCII data files, the exact record and column locations should be included for each variable. If applicable, information about implied decimal places and special formats (e.g., date) should also be included.

- (4) **Variable Value and Range.** A list of the valid values for categorical items and valid ranges for continuous items.
- (5) **Missing and Inapplicable Code.** Missing/inapplicable data codes and their meanings.
- (6) **Variable Format.** The mode in which the variable is represented (i.e., numeric, alphanumeric).

Data Description

The data description should include the theoretical or conceptual framework that informs the study, research questions addressed by the study, hypotheses to be tested, and methods used to collect the data. The following is a list of information that should be addressed in the description of methods.

- (1) Unit(s) of analysis.
- (2) Universe from which the study population is drawn.
- (3) Sampling method (e.g., probability, purposive, convenience, etc.) used to select elements of the universe.
- (4) Use of strata or quota (if any).
- (5) Time period and geographic area covered by the data collection.
- (6) Source(s) of data.
- (7) Response rate (the proportion of the sample about which data were obtained).
- (8) Retention rate (in a multiwave study, the proportion of cases successfully followed up).
- (9) Other elements of the study design (i.e., experimental treatments, assignment to groups, timing of follow up data collection, etc.).

Bibliography

A bibliography of research reports, journal articles, and other publications pertaining to the data set should be provided.

Summary Statistics

Investigators are encouraged to submit summary statistics. This information is useful to the archivist as a standard for comparing and verifying the accuracy of data. The form of statistics should be determined by the characteristics of the data; **frequencies** for categorical variables and **summary statistics** (e.g., means, standard deviations) for continuous variables.

Responding to Requests for Additional Information.

The Data Archive staff of the HSPMC will contact the investigator(s) during the archiving process to obtain necessary information not available in the data and documents submitted. The principal investigator should allocate sufficient time to respond to these communications when planning submissions to the Data Archive.

Reviewing Draft Materials

User's guides and codebooks drafted by the Data Archive staff of the HSPMC will be sent to the investigator(s) for review. This provides investigators with the opportunity to clarify, amplify, or correct any information before it is released to the public. Generally, these materials will be sent to investigators within six months of the original receipt of the data. Investigators are asked to complete and return their reviews within three weeks.

III. PREPARING DATA FOR DISSEMINATION: DATA ARCHIVE OF THE HSPMC'S RESPONSIBILITIES

Data Archive staff of the HSPMC are responsible for:

- (1) Preparing the data files in ready-to-use statistical file formats.
- (2) Preparing a user's guide that describes the project and data.
- (3) Reviewing the codebook for completeness and accuracy and augmenting the codebook as necessary.
- (4) Making copies of the archived data sets available to the research community and providing technical support to data users.

Preparing Data Files

Machine-readable data files are inspected for quality, usability, and correspondence with documentation. Checks are made for out-of-range values, missing values, and data inconsistencies. Data files may be restructured to promote usability. However, no changes will be made in the data values unless authorized by the investigator.

Regardless of the format in which the data are obtained, the Data Archive staff of the HSPMC will produce an ASCII, SAS transport, and SPSS portable version of each file submitted.

While preparing the data set for public use, the Data Archive staff of the HSPMC may contact the investigator directly if more information, documentation, or clarification is needed.

Preparing a User's Guide

A user's guide is prepared as part of the data archiving process. The guide provides the potential user with a general overview of the study to help determine the suitability of the data set for the potential user's purpose. The user's guide, data collection instruments, and codebooks are intended to provide enough information about the study to enable the user to work with the data without recourse to the original investigator. A typical user's guide will include as follows:

- I. Project Overview
- II. Purpose of the Study
Sampling/Selection Information
Data Collection
Instruments and Measures
- III. Description of Files
List of Files and Characteristics

Notes Regarding the Data Files

References

References to publications from the data set

References to publications related to the data set

IV. Appendices

Data Collection Instruments

Codebook

Reviewing the Codebook

The codebook is reviewed for completeness and accuracy. If necessary, the codebook is reorganized or augmented to ensure that it links the data collection instruments to the data file; matches the organization of the data file; and provides variable names, variable definitions, codes, and code definitions for each variable.

Making Data Sets Available

Researchers can obtain copies of data sets from the Data Archive of the HSPMC. Data sets are available in three file formats (SAS, SPSS, or ASCII) and free of charge to qualifying individuals or organizations. See *A Guide for Data Access* for more information, or contact the Data Archive Manager of the HSPMC for details (refer to Appendix D for contact information).

APPENDIX A

SAS TRANSPORT FILES

When you create a data set in SAS, it is (by default) saved as a SAS system file. The format of a system file is unique to the computer's operating system and provides the most efficient means of storing your data while you are conducting analyses. In order to submit your data to the Data Archive of the HSPMC, however, you will need to create a SAS transport file. A SAS transport file is a special type of SAS file that can be transported between different types of computers and different versions of the SAS software. While SAS transport files and SAS system files contain much of the same information, they are not the same. Only transport files can be easily exchanged between different types of computers. Transport files contain all of the data for each variable as well as the variable names, variable labels, missing value codes, and other information input by the user. Once your data are saved in SAS, creating a SAS transport file is very simple. Two sample SAS (version 6) programs are included below to help you get started. The file and path specifications will vary depending on your operating system. Consult your operating system manual for details.

To write a SAS transport file from a SAS data set

```
LIBNAME SASDATA 'path specification for SAS data library';
LIBNAME TRANS XPORT 'file specification for SAS transport file';
PROC COPY IN=SASDATA OUT=TRANS;
SELECT member name;
RUN;
```

To read a SAS transport file and create a SAS data set

```
LIBNAME TRANS XPORT 'file specification for SAS transport file';
LIBNAME SASDATA 'path specification for SAS data library';
PROC COPY IN=TRANS OUT=SASDATA;
RUN;
```

In addition to PROC COPY, another SAS procedure called PROC CPORT can also be used to create SAS transport. If you submit a file created by PROC CPORT, please indicate this in the written documentation accompanying the data. Transport files created using PROC CPORT can only be imported using PROC CIMPORT. The sample programs above assume that you are running version 6 of the SAS software. If your copy of SAS predates Release 6, it is fine to prepare a version 5 SAS transport file but, again, please indicate this in the documentation.

Finally, user written value labels created via PROC FORMAT may be submitted along with a SAS transport file. User written formats are an extremely useful form of documentation for SAS data sets and are very much appreciated by secondary users. Due to the idiosyncrasies of different operating systems and different versions of SAS, however, the most reliable and convenient method of transmitting user written formats is to submit a diskette with an ASCII (text) copy of the PROC FORMAT program that created the value labels, along with the FORMAT statement necessary for associating variables in the SAS data set with their corresponding formats.

APPENDIX B: SPSS PORTABLE FILES

When you save your data in SPSS, it is (by default) saved as an SPSS system file. The format of a system file is unique to the computer's operating system and provides the most efficient means of storing your data while you are conducting analyses. In order to submit your data to the Data Archive of the HSPMC, however, you will need to create an SPSS portable file. An SPSS portable file is a special type of SPSS file that can be transported between different types of computers and different versions of the SPSS software. While SPSS portable files and SPSS system files contain much of the same information, they are not the same thing. Only portable files can be easily exchanged between different types of computers. Another important distinction is between portable files and raw data files. Portable files are ASCII files but they are not raw data files. Unlike raw data files they contain all of the data for each variable as well as the variable names, variable labels, value labels, missing value codes, and other metadata input by the user. Once your data are saved in SPSS, creating an SPSS portable file is very simple. If you are using a window-based application you can use "Save as ..." under the file menu and choose "SPSS portable (*.por)" as the file type. Alternatively you can run a simple program to create a portable file. Two sample programs are included below to help you get started. The file specification will vary depending on your operating system. Consult your manual for details.

To write an SPSS portable file from an SPSS system file:

```
GET FILE='file specification for SPSS system file'.  
EXPORT OUTFILE ='file specification for SPSS portable file'.
```

To read an SPSS portable file and create an SPSS system file:

```
IMPORT FILE='file specification for SPSS portable file'.  
SAVE OUTFILE ='file specification for SPSS system file'.
```

APPENDIX C DATA TAPE DOCUMENTATION

If you submit data on tape, please provide the following additional information:

- Name of the program used to write the tape.
- Number of tracks on the tape (9 or 18).
- Type of labels on the tape, if any (e.g., IBM standard).
- Density of the tape (1600, 6250, or 38K cartridge).
- Names of files on the tape (data set names), if labeled.
- Volume serial number of the tape, if labeled.
- Abstract of the contents of each file.
- Record format of each file.
- Logical record length of each file.
- Block size of each file.
- Code of tape (e.g., EBCDIC, ASCII).
- Number of physical records or blocks in each file.
- Information for contacting sender (name, telephone number, address).
- Additional information on non-standard items.
- Parity (odd). (Nine-track tapes must be odd parity.)

A tape volume table of contents listing should be included with the tape. The table should be in an easy-to-read form, rather than a dump of the text of the labels. If possible, the tape table of contents should be produced by a program that also verifies the readability of the tape. If no tape listing program is available, then the tape table of contents can be produced manually, using information from the job which produced the tape or a dump of the tape.

Warning for Standard Label Tapes: If the internal label (i.e., the volume serial number) and external labels do not match, be sure to note this in your documentation! In some operating environments (e.g., IBM/CMS) there is no way to read a standard label tape without knowing its internal label.

**APPENDIX D
STAFF DIRECTORY**

Data Archive staff of the HSPMC can be reached at the following address and fax number:

Data Archive of the HSPMC
Kwang Kim (TB-386)
Westat
1650 Research Boulevard
Rockville, MD 20850

Phone: 301-517-4078
Fax: 240-453-2650
E-mail: hspmc@westat.com

Phone and E-mail for individual staff members:

Nicholas Zill, Project Director
E-mail: zilln1@westat.com
Direct Line: 301-294-4448

Kwang Kim, Data Archive Manager
E-mail: kimk1@westat.com
Direct Line: 301-517-4078

John Brown, Data System Manager
E-mail: brownj1@westat.com
Direct Line: 301-251-4344

APPENDIX E

Data Deposit Form

**Data Archive of the
Head Start Performance Measures Center
Westat
1650 Research Boulevard
Rockville, MD 20850**

Please complete this form to provide the Data Archive of the Head Start Performance Measures Center (HSPMC) with information about substantive and technical characteristics of your data set. It is vital that the information solicited on this form be provided as completely and accurately as possible.

The deposit form grants permission for the Data Archive of the HSPMC to archive and distribute your data pertaining to performance measures of Head Start programs. Please sign the form in the box below after attesting to the three statements below.

- I hereby give permission to the Data Archive of the HSPMC for this data set to be disseminated.
- I have copyright to this work and have the right to make it publicly available through Data Archive system of the HSPMC (if applicable).
- I have taken steps to protect the anonymity of the subjects in this data set where there is an expectation of confidentiality. In the event that problems with confidentiality are discovered, I will work with the Data Archive staff of the HSPMC to resolve them.

Printed name and title: _____
Signature: _____
Date: _____

Below are a number of items for which information is needed. You may attach a separate document in place of filling out these items if that document provides the necessary information. Please contact the Data Archive of the HSPMC (301-517-4078) if you need assistance in completing this form.

1. Descriptive title of data set:

2. Principal investigator(s) and affiliation(s) at time of data collection (for multiple investigators, give proper name order):
 - 2.1 Sponsoring or funding agency (if appropriate) and procurement or grant number:

 - 2.2 Name of the funding agency project officer (if appropriate) and telephone number:

 - 2.3 Person/organization responsible for collecting data:

 - 2.4 Internal study or project number (if appropriate) and the organization that assigned it:

3. If this is a new edition, extract, or special version of the data set, give appropriate details:

4. City and state of production of data collection, organizational name of data producer:

5. Depositor's name, organization, date of deposit:

6. Type of data collection (e.g., survey, aggregate, census/enumeration, experimental, event/transaction, clinical, program source code, machine-readable text, administrative records, etc.):

7. When were the data collected?

8. Time span covered by the data collection (months/days/years--include discrete years and ranges):

9. Geographic area(s) to which data are relevant:

10. Description of data collection
 - 10.1 Purpose and scope:

 - 10.2 Special characteristics or unique features of the collection (if any):

 - 10.3 Major areas of investigation:

 - 10.4 Unit of analysis:

11. Sample design and methodology
 - 11.1 Type of sample:

 - 11.2 Universe:

 - 11.3 Eligibility criteria:

 - 11.4 Response rate:

- 11.5 Method of collection:
- 11.6 Number of records:
12. Sampling description of data collection:
13. Source of data, if derived from another data file or from printed sources (state all relevant sources):
14. Primary publications describing or resulting from the data set:
15. Is restricted information (e.g., respondent's name, address, or telephone number, employer name, Head Start center name, etc.) on the data set? If so, please list.
16. How many distinctly different data files are included in the data set?
17. Can the data files be used separately for analysis? (If so, please explain.)
18. Can the data files be linked? If so, please indicate the variable(s) that will link data files.

- 18.1 Would linkage pose any risk of violation of confidentiality?
19. Is the data set one of a series or will it be updated regularly? If so, state the frequency.
20. The Data Archive of the HSPMC will not accept system files from any proprietary software package if they require a particular operation system.
21. Please indicate whether the data are compressed or uncompressed; if compressed, please check the appropriate format:
- Uncompressed
 - UNIX TAR file
 - UNIX compressed file
 - pkZIP file
 - GZip file
 - Other compressed file (please specify) _____
22. Describe the medium on which your data are being transmitted.
- 22.1 Diskette or CD-ROM
- a. Diskette CD-ROM
 - b. Total number of diskettes or CD-ROM provided: _____
 - c. Format:
 - IBM DOS
 - Macintosh
 - Acrobat
 - Other _____

22.2 Tape

- a. Density (b.p.i.) 1600 6250 38000
- b. Track 9 IBM 3480
- c. Data: Blocked Unblocked
- d. Labels IBM Standard Volume = _____
 ANSI None
- 8mm cartridge tape in standard UNIX format (tar, dd, or cpio)
- 4mm cartridge tape in standard UNIX format (tar, dd, or cpio)

22.3 File Transfer Protocol

- a. Are your files ASCII or binary?
- b. What was the date and time that you transmitted the first file?
- c. Were all transactions apparently complete and correct?

22.4 Other medium (please describe): Note that the Data Archive of the HSPMC can accept data sets on a variety of media. Please describe the medium you have used in detail. Check with us before sending data or documentation on these "other media."

23. Which of the following processing steps were performed on the data?

	File 1	File 2	File 3	File 4
a. Consistency Check				
b. Inclusion of frequencies				
c. Checks for undocumented codes				
d. Missing data codes standardized within collection				

- e. Do the data contain blanks? Yes No
- f. Do the data contains non-numeric codes? Yes No

24. Summarize all documentation submitted with your data set:

- Computer-readable codebook/documentation

Specify the format of this computer-readable documentation:

- ASCII text
- Word processor file (specify software and version)
- Rich Text Format (RTF) file (specify software and version)
- PostScript file (specify software and version)
- Adobe PDF file (specify software and version)
- Other (describe in detail) _____

Medium containing computer-readable documentation:

- Tape
- CD-ROM
- 3-1/2" diskette
- Other (please describe) _____
- Paper copy codebook
- SPSS Data Definition Statements ("Control Cards")
- Mainframe-compatible
- PC-compatible
- SAS Data Definition Statements ("Control Cards")
- Mainframe-compatible
- PC-compatible
- Database dictionary
- OSIRIS dictionary
- Data collection instrument (i.e., questionnaire)
- Accompanying computer programs
- Frequencies, machine-readable
- Frequencies, paper copy
- Other _____