



# Preservation Handbook

## Statistical Data<sup>1</sup>

---

Author	AHDS History
Version	03
Date	15/07/2005
Change History	

---

---

<sup>1</sup> This document is based upon the Process Guides produced by Alasdair Crockett of Depositors Services, UKDA.



## Definition

Most simply, statistical data is comprised by a rectangular matrix of rows and columns (cases and variables) that denote the numeric or alphanumeric values of each datapoint. Most statistical packages also contain what is known as 'internal metadata' describing the information potentially contained in the dataset in addition to the data. Typically this equates to variable names, variable labels, value labels, and missing values. Variable names uniquely identify each variable or column; variable and value labels usually give a textual interpretation of a variable name or value code.

## Description

Statistical data sets typically consist of the following data types:

- Nominal (categorical or coded): At least one variable consists of discrete values, normally integers, with specifically defined meanings. Associated internal metadata: variable name, variable label, value label, missing values.
- Ordinal (rank): Integer values conveying some form of scalar meaning. Associated internal metadata: variable name, variable label, missing values.
- Interval/Ratio (continuous): The data consist of an actual quantity (e.g. degrees in Celsius). Interval data is defined in terms of an absolute referent whereas ratio data is relative. Associated internal metadata: variable name, variable label, missing values.
- Alphanumeric (string): Words rather than numbers. Associated internal metadata: variable name, variable label, missing values.
- Dates/Time: Simple date information recording the year as a two digit number or more detailed date information, which records the month, week, day and so on. Associated internal metadata: variable name, variable label, missing values, date format information.

These different data types govern how data is processed

## Additional Information

- <http://www.icpsr.umich.edu/help/dds.html> [Last checked 12/10/2004]



## Technical Environment

The most commonly used statistical package is SPSS, although STATA and SAS proprietary formats are also popular. Specialist researchers may also make use of more advanced statistical packages such as S-Plus, Matlab, GLIM, NSD-Stat or BUGS. Users of these packages, however, commonly use SPSS or a text editor as their data management software and often produce output in the form of open text files.

SPSS is the de facto standard preservation format. SPSS portable (.por) format is an easily interpretable file structure which consists of tagged ASCII. Moreover, SPSS portable files have been supported by SPSS since the early 1970s and there is a continuing commitment for every version of SPSS (on every operating system) to read and write portable files. Further, they are operating system independent (in terms of the four main SPSS platforms of MS WINDOWS, UNIX, MACOS and MS DOS), so any SPSS user can open an SPSS portable file. These guidelines are dependent on having access to recent version(s) of SPSS or relevant proprietary software. It is not advisable to prepare preservation versions without recourse to these software applications, unless absolutely necessary.

- <http://www.esds.ac.uk/aandp/create/data.asp> [Last checked 12/10/2004]

Statistical data may also be presented in other proprietary formats such as spreadsheets (e.g. MS Excel) or databases (e.g. MS Access). Although some reference is made to these formats later in this document, the reader is advised to consult the relevant preservation manual. In addition, data may also be received in various open formats; typically tab-delimited text, comma-separated text, or fixed-width (undelimited) text data.

## Common Formats

Format	File Extension	Notes
SPSS	*.sav, *.por	*.por preferred preservation format
STATA	*.dta	
SAS (for MS Windows)	*.sd2, *.sd7, *.sas7dbat	
SAS (for Unix)	*.ssd01, *.sas7dbat	
Comma separated	*.csv	
Tab delimited	*.dat, *.tab, *.txt	*.tab preferred preservation format
Fixed width	*.dat, *.txt	
SYLK	*.slk	a Microsoft proprietary format for interchange of data

## Additional Information

### SAS

Useful information on converting old or unknown SAS file types can be found at:

<http://ftp.sas.com/techsup/download/technote/ts271.html> [last checked 16/03/2005]



<http://ftp.sas.com/techsup/download/technote/ts140.html> [last checked 16/03/2005]

<http://www.nber.org/data/sasport.html> [last checked 16/03/2005]



# Ingest Checklist

## Level 1 (Essential)

- Purpose of each rectangular object described
- Content of each rectangular object described
- Content of each variable described
- Content of each case described
- Data type of each variable described
- Coding schemes fully described
- For fixed-width data, especially that with no 'visible' boundaries between variables, information must be provided by the depositor as to what columns belong to which variables.
- Missing data must be defined and described.

## Level 2 (Preferred)

- Calculations checked

## Level 3 (Best Practice)

- Contextual information in user documentation
- Details of how the source(s) have been converted to digital form.
- List of sources, including archival or bibliographic references.
- Copy of original material
- Internal metadata intuitively comprehensible without any documentation



## Preservation

Although proprietary, the SPSS portable (.por) format is an easily interpretable file structure which consists of tagged ASCII. As such, these files can be easily converted into a completely open format without using SPSS software and therefore can be considered a de facto standard preservation format. The other standard preservation format for statistical packages is the more typical tab-delimited text (ASCII or preferably UNICODE character set).

### Significant Characteristics

SPSS portable files contain both the data and internal metadata. It is essential to preserve both of these alongside a number of processing outputs (see below). Typically, internal metadata consists of variable names, variable labels (descriptions of variables), value labels (descriptions of values of variables) and definitions of missing values. The limitation of tab-delimited format is the fact that it can only store the rectangular matrix of datapoints and variable names. Internal metadata cannot be stored as part of the same file. The techniques for preserving this metadata are considered below.

### Technique

#### *Preservation in SPSS portable format*

If the preservation format is SPSS portable, all processing should be carried out in SPSS. It is possible to proceed the other way round but, for the sake of consistency, unless there is a pressing reason otherwise, the ingest format should be converted to SPSS portable at the beginning of processing (with accompanying version control checks, see below) and all subsequent processing checks should then be performed in SPSS format.

#### *Converting to SPSS format*

SPSS contains many import filters, so that data from several common statistical, spreadsheet, database and open formats, can be imported directly into SPSS and saved in SPSS portable format.

#### *Fixed-width data*

For fixed-width (undelimited) data SPSS command ('setup') files can be written to read open formats. The key SPSS command is the DATA LIST command, which reads the data in. The VARIABLE LABELS, VALUE LABELS and MISSING VALUES commands add the internal metadata.

#### *Delimited data*

Delimited data can be directly read into SPSS using 'import wizard' function. Setup files, however, may be required to add any internal metadata. Similarly, common spreadsheet and database data formats can be read directly into SPSS. As is the case with tab delimited preservation formats each database table or spreadsheet worksheet should be archived as a separate data file. Further data transfer is possible via the ODBC (open database connectivity) system, whilst SPSS for UNIX (version 6) offers import filters for the three main heavyweight UNIX databases: Ingres, Oracle and Informix.

#### *Conversion from other statistical packages*

### SAS



All the commonly used types of SAS data file can be directly imported into SPSS (versions 11 and higher) via the import menu, or by using StatTransfer (see also additional information under the "Technical Environment" major heading).

Detailed information can also be found in Alasdair Crockett's *Process Guide 1* held by the UKDA.

## STATA

STATA is the only common format that SPSS will not directly import, and consequently specialist translation software is required. The two main data translation packages used for this purpose are StatTransfer and DBMSCOPY. When using StatTransfer, always convert data files from STATA to SPSS system (.sav), not portable (.por) format, the .por files can be created within SPSS itself.

### *Preservation in tab-delimited format*

If the preservation format is tab-delimited text, processing can be carried out in any package that enables problem-free conversion to tab-delimited text (see also the preservation handbooks concerned with databases and spreadsheets). 'Internal' metadata, however, needs to be stored in an additional series of tab-delimited files or in the documentation. No such standard exists for preserving this information though a standard is emerging from the ongoing development of the DDI (data documentation initiative) using XML (extensible markup language) as the preservation format of metadata storage.

### *Weighted datasets*

All datasets should be processed and archived in an unweighted format. If the dataset is weighted the phrase 'weight on' will appear in the bottom of the SPSS data window in SPSS for Windows or SPSS for UNIX (when running under XWindows). If SPSS is being used in batch mode, or to perform a double check in GUI mode (as the message does not always display if the data are in portable format), the command for finding out whether the dataset is weighted is: show weight. The output should read 'the file is not weighted'

If the output reads 'weighting variable = <variable name>', the data are currently weighted by that variable. Weighting data will not alter the core data matrix (i.e. the actual data values SPSS displays in its data viewer) but will alter any statistical output, such as descriptive statistics. The weight should be removed, using the command: WEIGHT=OFF.

## Validation of Exported Data

The information upon which the checks for consistency in data and metadata conversion are based is generated by the SPSS 'descriptive' and 'data dictionary' commands output. The output from these commands should be archived.

### *Data*

The version control checks should include:

- The number of cases (rows) and variables (columns) is the same in both formats
- The number of decimal places is equal in both formats
- No truncation has occurred in alphanumeric (string) variables
- Date variables do not lose any formatting
- Check that all string variables are inherently meaningful.



More sophisticated validation exercises can be carried out using the output from the SPSS DESCRIPTIVES command. This command forms the basis for checking anomalous values; especially for nominal (categorical) variables, undefined code or one or more errors in the data. For example:

- Check that nominal (categorical) variables have values within the range defined.
- Check interval (and ratio) data for any apparent anomalies – such as ages over 110, incomes less than zero, etc.

Care should be taken with apparent errors in both the above examples. If validation exercises reveal that such errors are the result of data processing activities then processing will have to be repeated. If, however, the apparent error is also present in the ingest data, it is possible that the 'erroneous' value represents a true transcription of the source; such occurrences should be recorded in the study's read file, but the data should not be altered.

In addition to the information provided by the DESCRIPTIVES command, the FREQUENCIES command can be used to reveal the number of cases (observations) in every value category or just variables revealed as problematic by previous validation actions.

The most basic data validation checks should consider a sample of 30 + 10% of the remaining nominal (categorical) / interval (and ratio) variables. If errors are discovered, a more systematic check of nominal (categorical) variables should be made.

### *Internal Metadata*

Internal metadata checks typically involve cross-referring between the documentation provided by the depositor and the internal metadata in the data file. These checks should be performed using the DISPLAY DICTIONARY command output, along with the output from any other commands the processor deems useful (such as 'display labels'). This command generates what SPSS calls the 'data dictionary': a detailed description of the contents of the data file, giving details of each variable in terms of the variable name, variable label, variable type, format, missing values, and value labels.

## **Problems and Issues**

The preservation of statistical data embodies four potentially problematic issues: relational data, the handling of internal metadata in tab-delimited preservation formats, the transfer of internal metadata between different statistical packages and the truncation or alteration of cell contents during transfer.

Issues arising with relationally held data - i.e. explicitly linked datasets - can be resolved by reference to the preservation handbook concerned with complex databases.

Should the tab-delimited format be selected as the preservation media, internal metadata - variable labels, value labels and missing values along with definition/data dictionary - is inevitably lost. All such information should be extracted and saved as part of the dataset documentation and a note made in the study's read file.

All types of internal metadata can cause problems when converting the file into or out of SPSS portable format. Variable names, however, seldom cause problems, although prior to SPSS version 12 their length is limited to eight characters. This can become an issue if longer names as allowed by STATA, for example, are involved. In practice, however, variable names can be stored as an additional row (case) of the dataset (typically the first row).



Variable labels, value labels and missing values cause greater problems. Value labels and missing values cannot be stored as part of a rectangular matrix, since there can be more than one piece of information per variable. Data files have to store this information outside the rectangular matrix that defines the data. SAS provides particular problems in this regard since it generally stores value labels in a completely separate file to the dataset, and this information must be carried across when converting the data into SPSS format. Although SPSS limits are generally generous relative to other statistical packages, STATA value labels have a maximum character limit of 80, exceeding the SPSS limit of 60 characters and therefore can be liable to truncation on conversion. [Similar truncation problems arise if the ingest format supports textual strings longer than the SPSS limit of 255 characters (e.g. MS Access memo fields). In both instances it may be advisable to treat the dataset as one would a database].

Finally, there is no way of automatically transferring missing values (other than null values) from a SAS dataset in SPSS. Likewise all missing values in SPSS will be converted to a single system missing value in STATA, which does not support multiple missing values prior to version 8). Consequently, missing values to be individually specified (e.g. did not apply, answer refused, did not know, etc.) are not preserved when transporting between SAS and SPSS. Rather, they translate as system missing values ('.') in SPSS). To preserve all missing value information, it is necessary to remove the user-defined missing values in SAS and then redefine that value(s) as missing within SPSS. One should always note the presence of missing values in the study's read file and inform users that importing such values into software other than SPSS may produce unforeseen results.