

Survey Data Collection with Survey Solutions

Perugia, Italy

File Formats and Data Export

Sergiy Radyakin

sradyakin@worldbank.org

Development Data Group (DECDG),
The World Bank

May 28 - June 01, 2018



DEVELOPMENT
RESEARCH



THE WORLD BANK



Food and Agriculture
Organization of the
United Nations

- 1 Definitions
- 2 Export file formats
 - Tab-delimited file format
 - SPSS file format
 - Stata file format
- 3 Export by Question Type
 - Overview
 - Details
- 4 Missing values
- 5 Export of data in rosters
- 6 Special export files

Export file formats

As of version 5.0 Survey Solutions supports export in the following formats:

- in tab separated format and a supplementary Stata script (do-file) that can be used to import meta data (variable and value labels).
- Stata format (*.dta)
- SPSS format (*.sav)

If necessary, file conversion utilities can then be used to transfer data to its final destination file format.

Export file formats


 Reports ▾ Interviews Teams and Roles ▾ Survey Setup **Data Export** Help SerglyHQ ▾

Export range:

(version) Survey template

(ver. 1) DemoQuest5 ▾

Status of exported interviews




Any ▾

(ver. 1) DemoQuest5

Pressing Generate button creates a data archive of interviews available on the server at that time. The time of archive creation and the archive size is shown in a table. The latest generated data archive can be downloaded by pressing Download button. Press Generate button again to produce a new archive.

Main Survey Data

Zip archive with main interview data

	LAST GENERATED	FILE SIZE
 STATA format <input type="button" value="GENERATE"/> <input type="button" value="NOT GENERATED"/> 		
 SPSS format <input type="button" value="GENERATE"/> <input type="button" value="NOT GENERATED"/> 		
 Tabular Format <input type="button" value="GENERATE"/> <input type="button" value="NOT GENERATED"/> 		

Binary Data

Archive with binary data (e.g., pictures)

	LAST GENERATED	FILE SIZE
.zip Binary format <input type="button" value="GENERATE"/> <input type="button" value="NOT GENERATED"/> 		

DDI

Data Documentation Initiative XML data

 XML format <input type="button" value="DOWNLOAD"/> 		
--	--	--

Export file formats

Reports ▾ Interviews Teams and Roles ▾ Survey Setup Data Export Help SergiyHQ ▾

Export range:

(version) Survey template

(ver. 1) LFS ▾

Status of exported interviews

Any ▾

Items in export queue: 4

Main Survey Data

Zip archive with main interview data

	LAST GENERATED	FILE SIZE
STATA format <div style="float: right; text-align: right;"> QUEUED <div style="width: 100%; height: 10px; background: linear-gradient(to right, #ccc, #ccc);"></div> × </div>		
SPSS format <div style="float: right; text-align: right;"> QUEUED <div style="width: 100%; height: 10px; background: linear-gradient(to right, #ccc, #ccc);"></div> × </div>		
Tabular Format <div style="float: right; text-align: right;"> QUEUED <div style="width: 100%; height: 10px; background: linear-gradient(to right, #ccc, #ccc);"></div> × </div>		

Binary Data

Archive with binary data (e.g., pictures)

	LAST GENERATED	FILE SIZE
Binary format <div style="float: right; text-align: right;"> QUEUED <div style="width: 100%; height: 10px; background: linear-gradient(to right, #ccc, #ccc);"></div> × </div>		

DDI

Data Documentation Initiative XML data

</>
XML format
DOWNLOAD

Paradata

Metadata on the interview process (events and timing)

	LAST GENERATED	FILE SIZE
Tabular Format <div style="float: right; text-align: right;"> GENERATE NOT GENERATED </div>		

⏪ 📄 ⏩ ⏴ 📄 ⏵ ⏴ 📄 ⏵ ☰ 🔍 ↺ ↻

Tab-delimited format

- Tab-delimited file format is a text format that uses an invisible tab-character (ASCII code 09) to separate data fields.
- Typical file extension is `.tab`, but can also be saved with `.txt` and other extensions.
- First line contains variable names (also delimited with tab characters).
- Dot is used for fractional numbers.
- One of the advantages of the tab-delimited format is that the tab character itself is usually not part of the data being collected (while a comma can be part of the address, company name, or occupation description), which simplifies storage, removes the need for quotes.

Advantages of the tab-delimited file format

- Text-human readable, printable, portable.
- Open, free.
- Supported by various software including statistical packages, database and spreadsheet applications (including open source and proprietary applications):
 - Microsoft Office Excel;
 - Open Office Calc;
 - Gnumeric;
 - Stata, SPSS, SAS, R;
 - Microsoft office Access;
 - Google Drive, etc.
- Can serve as an intermediate format for information exchange.

Disadvantages of the tab-delimited file format

- Stores values (data), but not metadata (variable labels, types, formats, etc).
- Metadata is usually supplied in additional files and in different format, machine readable: XML, DDI; package-oriented scripts: *.sps (for SPSS), *.do (for Stata); or textual description (for human operators).
- When metadata files are not available, tab-delimited files may become fully or partially unusable.

DDI Metadata

- DDI is a standard of presenting information about the survey, questionnaire and data in a machine-readable form;
- Survey Solutions produces a DDI output (XML file) based on the questionnaire used in the survey;
- It is usually used as an input to other systems that need to learn information about the study.

SPSS file format

- Proprietary format, developed for the statistical package SPSS (*Statistical Package for the Social Sciences*), currently developed and marketed by IBM.
- Exists in many different versions implemented in various versions of SPSS (by generation and platform).
- File format specification is not available, other software supporting this format relies mostly on RE-efforts
- Typical file extension: *.sav

Advantages of SPSS file format.

- Contains embedded metadata: variable labels, formatting, comments, etc.
- Original versions of this format didn't support unicode, but recent versions do provide a possibility to work with unicode as well as a large number of other encodings.
- Basic data compression is implemented in SPSS to reduce the size of the .sav data files.
- Modern versions support strings up to 32,767 bytes long (originally up to 8, then later up to 255 bytes).

Disadvantages of SPSS file format.

- Public description of the file format is not available from the authors.
- Limited support by other software because of lack of file format specification.
- Strings are limited to 32,767 bytes.

Stata file format

- Proprietary format, developed for the statistical package Stata currently developed and marketed by StataCorp.
- Exists in different versions, mostly differing by generation of the product.
- Survey Solutions exports the data in Stata 14 format, which permits unicode content.
- File format specification is available, other software supporting this format rely on official documentation.

Stata File Format Specification

<http://www.stata.com/help.cgi?dta>

- Typical file extension: *.dta

Download data

Attention

- At the end of the survey always download and store all the download files before the server is shut down.
- Download and store the data even in formats that you don't currently intend to use.

Export by Question Type

Comprehensive information about different question types, their parameters, export details, etc can be found in the manual.

Document online:

A guide to different question types in the documentation section of the Survey Solutions homepage: <http://worldbank.org/capi>

New question types are added periodically to the program.

Data export details

<u>AB</u>	Text	▼
-----------	------	---

- Answers to text questions are exported as expected: as text.

Data export details

<u>12</u>	Numeric	▼
-----------	---------	---

- Answers to numerical are exported as text in tab-delimited files. Dot is used as a decimal separator.
- In Stata and SPSS files answers to numeric questions are exported as numeric variables.

Date

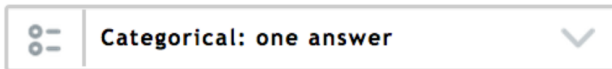
- Variable will contain string representation of the date using the following format:
 - for date:
format: YYYY-MM-DD,
example: 1980/04/19
 - for timestamp:
format: YYYY-MM-DDThh:mm:ss,
example: 2017-04-19T12:34:56
- Application processing the data may need to separate the components from each other (e.g. extract the year separately), this can be done with string manipulation functions.

Geolocation (GPS)



- Multiple variables!
- Original variable name + double underscore + suffix
- Suffixes:
 - Latitude,
 - Longitude,
 - Accuracy,
 - Altitude,
 - Timestamp
- Timestamp format: MM/DD/YYYYTHH:MM:SS

Categorical: single choice



A screenshot of a dropdown menu. On the left, there is an icon consisting of two circles with horizontal lines through them. The text in the center of the dropdown is "Categorical: one answer". On the right side of the dropdown, there is a downward-pointing chevron icon.

- Variable will contain numeric code of the selected option.
- Value labels will be defined for codes stored in such a variable
 - for tab-delimited files: defined in the accompanying do-file;
 - for Stata and SPSS files: stored in the exported files as value labels.

Categorical: multiple choice



- Multiple choice questions: multiple variables will be created in the dataset with indices corresponding to the options' codes. For example x_{101} , x_{102} , x_{103} , and so on.
- For Y/N-mcq: each variable contains a zero, a positive number (selection order), or a missing value
- For non-Y/n-mcq: each variable contains a zero, or a positive number (selection order)

Linked questions

- categorical linked questions contain the codes of the selections, not names

Lists



- Forty variables will be created in the export file for each list question, for example for variable *membername* variables *membername__0*, *membername__1*, *membername__2*, etc will be created.
- Each variable will contain the corresponding item of the list.

Barcodes



- Variable will contain recognized content of the scanned barcode. Typically a string.

Images and audio



- The variable will contain the name of the file and the file itself will be placed in the folder with a unique name, corresponding to the interview.
- Images and audio recordings are part of the binary data (separate download).
- If the question is part of the roster, there may be multiple files created (e.g. portraits of each household member) with indices corresponding to the position in the roster.

Area

- The variable will contain the coordinates of the points selected on the map.

Missing values

- MISSING/BLANK - for values that were skipped due to logic;
- -999,999,999 - for numeric values that were not assigned a value (refusal, don't know, does not apply);
- "##N/A##" - for string values that were not assigned a value (refusal, don't know, does not apply).
- Replace with missing or impute the not specified values before analysis! In Stata files (from v5.26) numeric missing values are exported as .a instead of -999,999,999 and in SPSS files (from v5.26) they are already marked as missing for the SPSS system.

Export of data in rosters

- Survey Solutions exports data as a single zip-archive with multiple files inside.
- Each level of data (such as households, persons, assets, etc) is saved into a separate file.
- Typically each roster in the questionnaire creates a new data level.

Combining the rosters within one data level

Rosters that are triggered by the same source, are automatically combined at data export.

ParentId	ID	demographics				Education				Employment				Health				Investment				Migration			
		v1	v2	v3	v4	e1	e2	e3	e4	p1	p2	p3	p4	h1	h2	h3	h4	i1	i2	i3	i4	m1	m2	m3	m4
1	1																								
1	2																								
1	3																								
2	1																								
2	2																								
2	3																								
3	1																								
3	2																								
3	3																								

Combining data levels

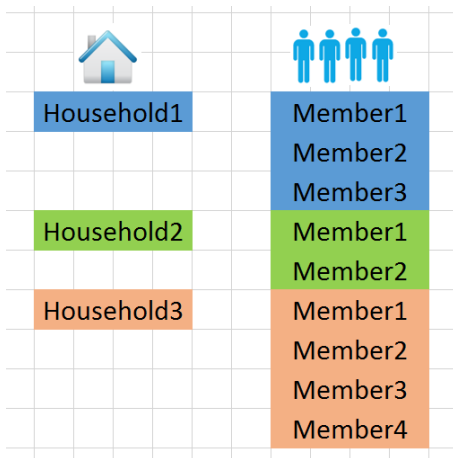
- Often data from different levels is required for analysis, for example urban/rural residence status (characteristic at household level) may be necessary for every person (individual level).
- This process is known as data merging (matching).
- An id (identifier) is required for this process.
- Each interview in Survey Solutions is internally assigned an id variable: *interview__id*
- Each roster item carries an Id of the interview, to which it belongs: *interview__id*
- Our ID is a 32-digit long hexadecimal number (also known as a GUID), for example:

b2937bb2117744f78429e8382ea2be44

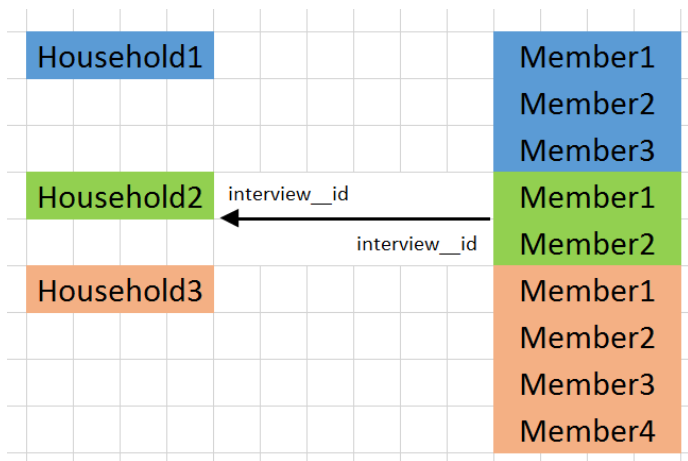
Combining data levels

- To get the urban/rural variable into the individual data one starts with the individual records and combines them with household records with corresponding *interview__id*
- Different statistical packages implement this procedure with different commands, and may require sorting the data by *interview__id*, and/or renaming id variables to be the same in both files.

Combining data levels



Combining data levels



Combining data levels (Stata example)

```
clear
cd "C:\mydata\"
use "DemoQuest5.dta"
sort interview__id
```

```
tempfile tmp
save "'tmp'"
```

```
clear
use "hhmembers.dta"
sort interview__id
```

```
merge interview__id using "'tmp'", nokeep
tabulate _merge
```

Special export files

Survey Solutions exports the following additional files (regardless of the format).

- `export__readme.txt`: contains a catalogue of variables by file.
- `interview__actions`: contains information about the movement of the questionnaire in the system, such as who collected the data, who approved the interview, and when, etc.
- `interview__comments`: contains information on the commentaries entered by any user of the system for questions and for actions.
- `interview__errors`: contains information about any remaining errors in the interviews.

Example of export__readme.txt

Exported from survey solutions Headquarters 18.04.3 (build 21031) on Wednesday, April 18, 2018

DemoQuest5.dta

region, address, visitdate, result, location, memberlist, yearbuilt, material_walls, material_walls_other, material_roof, material_roof_other, watersrc, electricity, reasonNotConnected, consentForPhoto, dwelling_image, assets, coffeemaker, toaster, blender, ricecooker, anyanimals, animals, nadults, oneDecisionMaker, decisionMaker

hhmembers.dta

relhead, gender, age, marstat, attends_school, works, memberPortrait, pubsect, work_desc, work_occupation_code, total_work_hours, total_salary, work_years, commute, com_freq, com_dist, com_time, com_transp_mode

R_inflows.dta

initial_stock, born, bought, received, died, sold, given, lost, food, current_stock

assetsinfo.dta

a_num, a_years, a_pur_price

Interview actions log

Contains the following information:

Field	Purpose	Example
<i>interview__id</i>	Which interview was affected?	5139...f27a2c
<i>Action</i>	Which action was taken?	Completed
<i>Originator</i>	Who performed the action?	JohnSmith
<i>Role</i>	In what capacity?	Interviewer
<i>ResponsibleName</i>	Who is the new person responsible for the interview?	MaryJackson
<i>ResponsibleRole</i>	Role of the new responsible person?	Supervisor
<i>Date</i>	When was the action taken? (day)	09/25/2015
<i>Time</i>	When was the action taken? (time)	18:11:08

NB: date in US format: MM/DD/YYYY

Interview comments log

Contains the following information:

Field	Purpose	Example
<i>Order</i>	Sequential order	1
<i>Originator</i>	Who wrote the commentary?	JohnSmith
<i>Role</i>	In which capacity?	Interviewer
<i>Date</i>	When was the commentary written? (day)	09/10/2015
<i>Time</i>	When was the commentary written? (time)	12:34:56
<i>Variable</i>	Which variable or action the commentary refers to?	hhszise
<i>Roster</i>	Roster name (if variable is within a roster)	members
<i>InterviewId</i>	Which interview was commented?	5139...f27a2c
<i>Id1, etc</i>	Addressing within roster	1
<i>Comment</i>	Actual comment	Refused to answer categorically.

NB: date in US format: MM/DD/YYYY

Interview errors log

Contains the following information:

Field	Purpose	Example
<i>variable</i>	Name of the variable/question name	age
<i>typer</i>	Type of the variable/question	numeric
<i>interview...id</i>	ID of the interview	5139...f27a2c
<i>message_number</i>	Number of the error validation rule	2
<i>message</i>	Text of the error message	Age may not be negative

NB: date in US format: MM/DD/YYYY