Data Flow and Database Structure in the SDR Project

Przemek Powałko, IFiS PAN
ppowalko@ifispan.waw.pl

Building Multi-Source Databases for Comparative Analyses, Conference & Workshop
Warsaw , December 16-20, 2019

Outline

- SDR facts
- Software environment
- Data flow
  - Input files
  - Relational database
  - Harmonization process
  - Output files
- Master Box structure
  - Master file
  - Plug files
- Missing codes schema

SDR facts

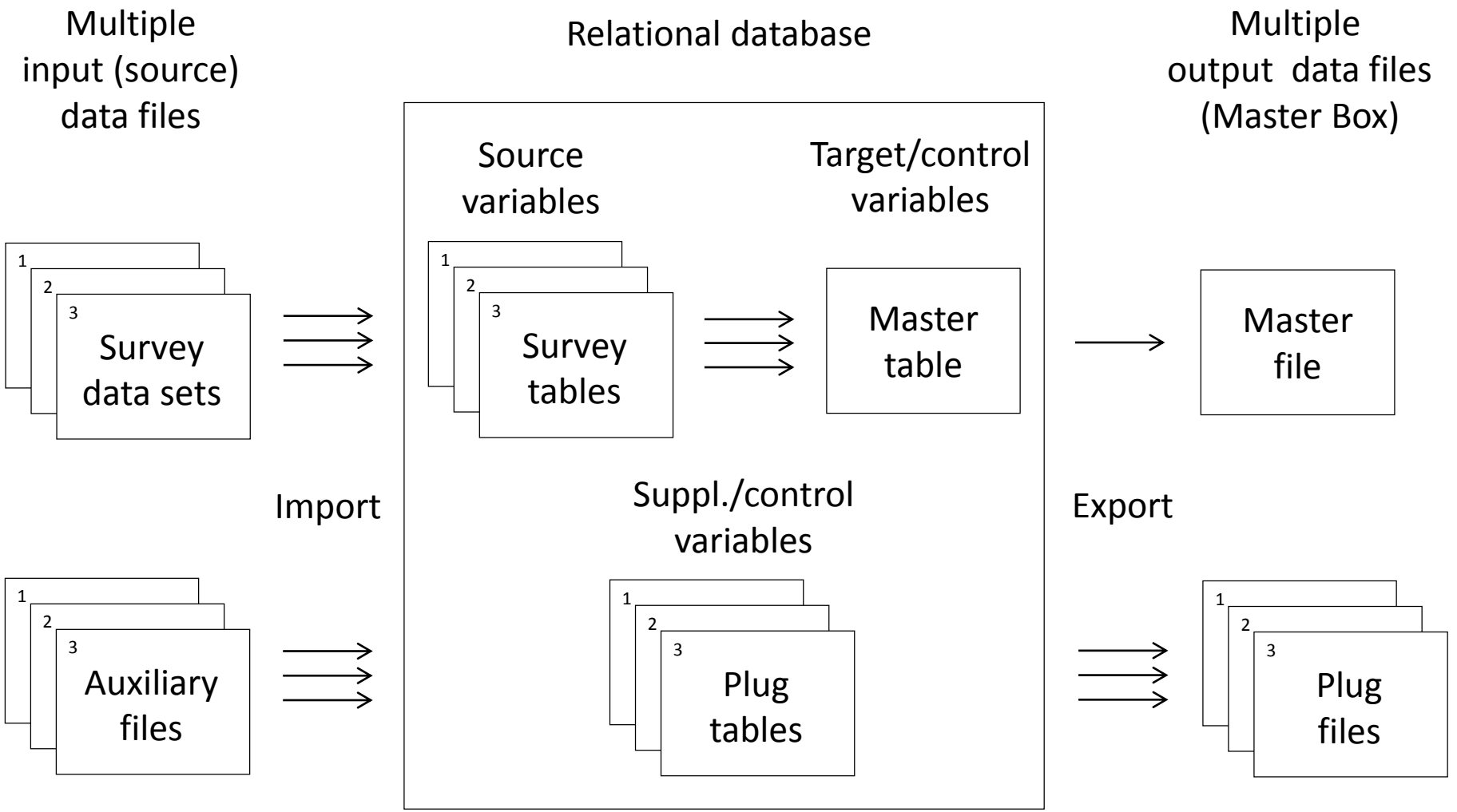
	SDR 1	**SDR 2**
# survey projects	22	23
# waves	89	174
# samples	1721	3485
# cases	2.3×10^6	4.4×10^6
Time span	1966-2013	1966-2017
# countries/territories	142	169
# source data files	81	215
Source data raw size	2.3 GB	3.8 GB
# source variables used	1395	?

Software environment

- Criteria
  - No-extra-cost solution: free & open-source software
  - Automation: batch processing – repeated actions without user interaction
  - Minimizing memory usage

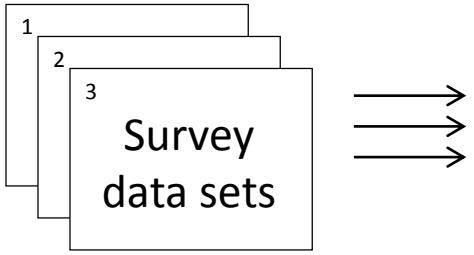
- Software used
  - Operating system: Windows
  - Programming platform: Cygwin → Linux [Fedora] (on virtual machine)
  - Statistical packages: PSPP, R?
  - Spreadsheet software: Excel with Visual Basic for Applications
  - Languages: Perl, Unix shell scripting, awk, sed, SQL
  - Database system: MySQL → MariaDB
  - SQL editor: HeidiSQL
  - Text editor: Notepad++

Data flow

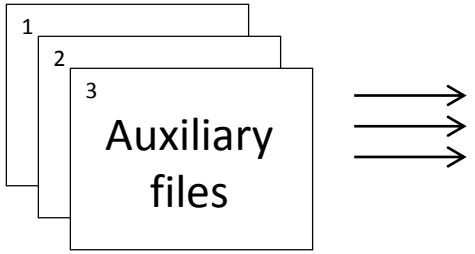

Multiple
input (source)
data files
Relational database
Multiple
output data files
(Master Box)
Source
variables
Target/control
variables
1
2
3
Survey
data sets
1
2
3
Survey
tables
Master
table
Master
file
Import
Suppl./control
variables
Export
1
2
3
Auxiliary
files
1
2
3
Plug
tables
1
2
3
Plug
files

Input files

Multiple input (source) data files

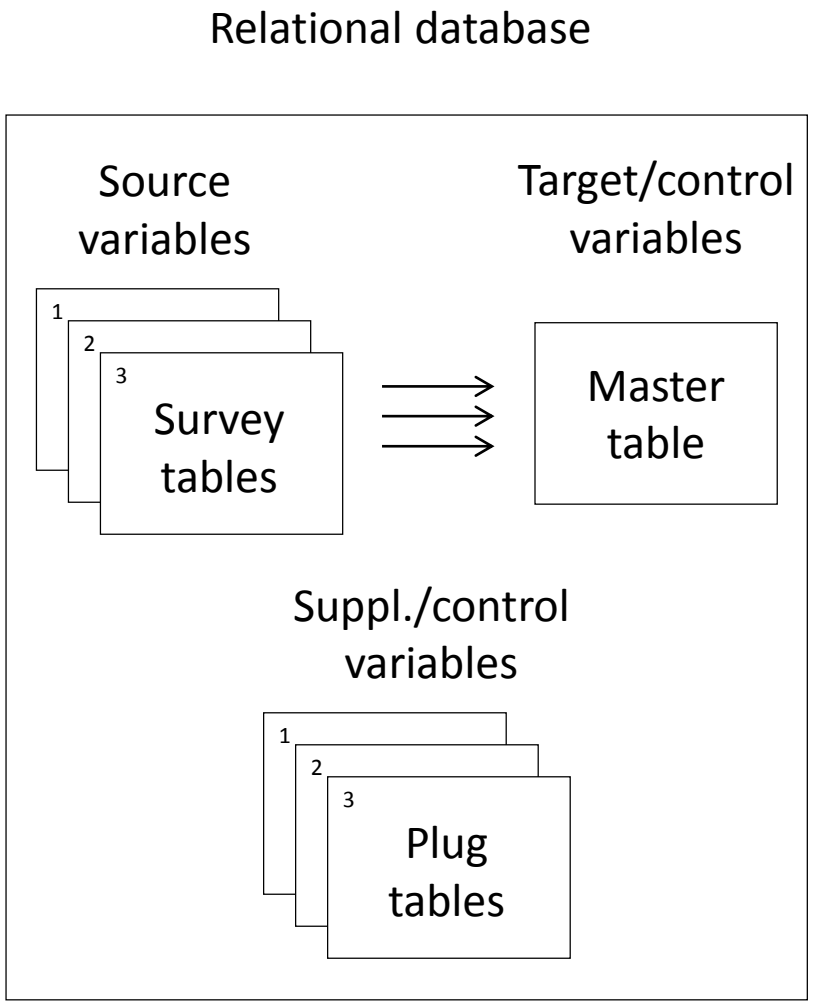


Import



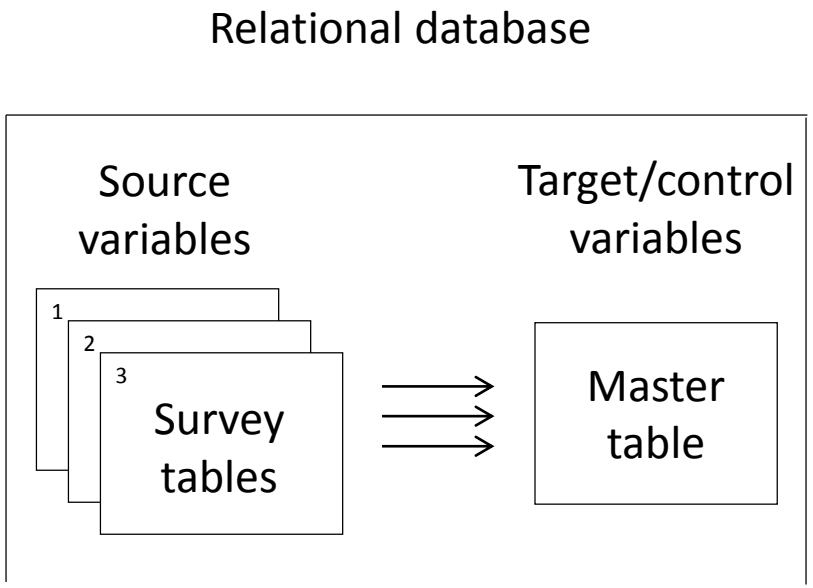
- Source data
  - Survey (micro) data: SPSS / STATA files → SAV → CSV
  - Auxiliary/contextual (macro) data: XLS → CSV
- Time zero
  - SDR 1: turn of 2013/2014
  - SDR 2: January 2018
- Patching survey data
- Automation
  - Source data/documentation download
  - Data extraction (PSPP)
  - Data transformation (shell scripts)
  - Data load – import to database (shell scripts & SQL)

Relational database



- Survey tables (micro data)
  - 1:1 full data sets
  - source variables
- Master table (micro data)
  - selection of cases
  - selection of source variables
  - target variables
  - control variables – source variable context/characteristics; quality indicators
- Plug tables (macro data)
  - supplementary contextual data on country, country/year, survey and wave levels
  - quality control variables

Data transformation

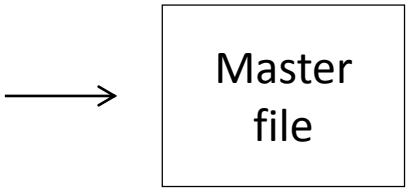


- Data crosswalk – source-to-target data mapping: CWT files (Excel)
- Implementation of harmonization rules (SQL)
  - In-file harmonization (SDR 2) → modularity, flexibility; errors become systematic and are easy to find
  - Source data transformations
    - map (translate)
    - edit (cut, trim)
    - rescale (linear, distributional)
    - calculate
    - merge
- Automation
  - Generation of CWT files
  - Generation the harmonization syntax (CWT)
  - Data description/browsing utilities (*show* tools)

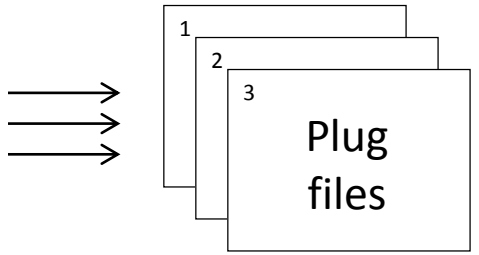
Output files

- Master Box
  - Master file: only target and control variables available for end-users
  - Plug files
- Merge syntax
  - The size of the merged master file may pose a problem
- Automation
  - Export to CSV (SQL)
  - Creation of SPSS and STATA files (PSPP/STATA, R?)
  - Creation of CSV files and do-files (Linux)
- SDR Master Box 1.0 dataset available on Harvard Dataverse
- SDR Master Box 2.0 dataset – in progress

Multiple output data files (Master Box)



Export



Master Box structure: Master file

- Master file contains survey (micro) data on individual (respondent) level

- Variables are of different categories
  - Source variables (not to be published)
  - Target variables, e.g. `T_TR_PARLI`)
  - Control variables
    - Context informing controls, e.g. `C_TR_PARLI_SRC_ASCEND`
    - Quality controls, e.g. `QR_DUPLICATE`, `QC_PROCESSING_ERROR` (SDR 2)

- Key variables (to merge with plug files)
  - SDR 1: `T_SURVEY_NAME`, `T_SURVEY_EDITION`, `T_COUNTRY_L1U`, `T_COUNTRY_SET`
  - SDR 2: `T_PROJECT_NAME`, `T_PROJECT_WAVE`, `T_COUNTRY_L1U`, `T_COUNTRY_L2U`

Master Box structure: Master file

- Variables are of different types
  - Technical variables, e.g. ~~S_CASE_ID~~, T_SURVEY_NAME , C_WEIGHT_L1U_TYPE
  - Quality variables, e.g. QR_DUPLICATE
  - Substantive variables, e.g. ~~S_TR_PARLI~~, T_TR_PARLI_11, C_TR_PARLI_SRC_SCALE_LENGTH
- Variables are grouped in concepts/indicators related to theoretical constructs
  - Concept, e.g. EDUCATION
    - Indicators: EDU (education level), SCHOOL_YRS (schooling years)
      - Source variables: ~~S_EDU~~ , ~~S_SCHOOL_YRS~~
      - Target variables: T_EDU, T_SCHOOL_YRS
      - Control variables, e.g. C_EDU_INCOMPLETE, C_SCHOOL_YRS_STILL_SCHOOL

Master Box structure: concepts/indicators

					# source variables per concept/indicator																	
data set name	**# all source variables in the file**	**# concepts realised**	**# distinct source variables used**	**% distinct source variables used**	CASE_ID	COUNTRY	INTERVIEW_DATE	WEIGHT	GENDER	AGE	BIRTH_YEAR	EDU	SCHOOL_YRS	RURALURB	METRO	TR_PARLI	TR_LEG	TR_PARTY	TR_GOV	TR_PERSONAL	PR_DEMONST	PR_PETITION	INT_POLIT
	31426	**19**	**1395**	**4.44**	79	61	24	94	81	81	18	105	61	137	396	53	48	37	40	45	70	31	59
ABS_1	217	16	24	11.06	1	1	1	8	1	2		1	1	1		1	1	1	1	1	1		1
ABS_2	265	18	27	10.19	1	1	1	10	1	1		1	1	2	1	1	1	1	1	1	1	1	1
ABS_3	274	19	17	6.20	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
AFB_1	144	10	10	6.94	1	1		1	1	1		1		1			1			1	1		
AFB_2	330	12	12	3.64	1	1	1	1	1	1		1		1		1	1				1		1
AFB_3	302	13	13	4.30	1	1	1	1	1	1		1		1		1	1			1	1		1
AFB_4	294	12	12	4.08	1	1	1	1	1	1		1		1		1	1				1		1
AMB_1_5	494	17	17	3.44	1	1	1	1	1	1		1	1	1	1	1	1	1	1	1	2		1
ARB_1	181	12	13	7.18	1	1			1	2		1				1	1	1		1	1	1	1
ARB_2	468	15	17	3.63	1	1		1	1	1		2		2		1	1	1	1	1	1	1	1
ASES	571	16	48	8.41	1	1	1		1	1		18	1	17	17	1	1	1	1		1	1	1
CB_2009	438	15	15	3.42	1	1	1		1	1	1	1	1	1	1	1	1		1	1			3

…

Master Box 1.0 structure: plug files

- PLUG_COUNTRY
  - Country ISO codes, geographical regions
  - Non-standard ISO codes: BA-FBH, BA-RSR, BE-BRU*, BE-FLA, BE-WAL, CY-ISL*, CY-TCC, DE-E, DE-W, GB-GBN, GB-NIR, IL-ARB, IL-JEW, KS, RU-KRA [* in SDR 2]
  - Key: T_COUNTRY_L1U
- PLUG_COUNTRY_YEAR
  - GDP per capita, Freedom House Index, population size, Worldwide Government Indicators, EIU Democracy Index, Gini Index
  - Keys: T_COUNTRY_L1U, T_COUNTRY_YEAR
- PLUG_SURVEY
  - Response rate, information about translation methods, pretesting, fieldwork control, sampling procedure, weights
  - Keys: T_SURVEY_NAME, T_SURVEY_EDITION, T_COUNTRY_L1U, T_COUNTRY_SET
- PLUG_WAVE
  - Quality control indicators : illegitimate variable values, misleading variable values, contradictory variable values, variable values discrepancy, lack of variable value labels
  - Keys: T_SURVEY_NAME, T_SURVEY_EDITION

Missing codes schema

- SDR 1 Master file
  - `-1/.a` don't know
  - `-2/.b` not applicable
  - `-3/.c` <reserved>
  - `-4/.d` value not acceptable
  - `-5/.e` variable not identified
  - `-6/.f` insuff. info for source option
  - `-7/.g` insuff. info for all src options
  - `-8/.h` question not asked in survey
  - `-9/.i` (regular) missing value

  `<null>/.` – used in plug files

- SDR 2 Master file
  - `-1/.a` question not asked in survey [cntry]
  - `-2/.b` insufficient information [cntry]
  - `-3/.c` don't know [resp]
  - `-4/.d` no answer [resp]
  - `-5/.e` refusal [resp]
  - `-6/.f` don't understand question [resp]
  - `-7/.g` combination of dk, na, ref, du [resp]
  - `-8/.h` not applicable [resp]
  - `-9/.i` other missing [resp]
  - `-10/.j` procesing error [resp/SDR decision]
  - `-11/.k` combination of different missing codes (in multiplets) [resp/SDR decision]

  `<null>/.` – to be used in plug files