INSTYTUT FILOZOFII I SOCJOLOGII
POLSKIEJ AKADEMII NAUK

The Organization and Processing of Data in the HARMONIA Project

Przemek Powałko

Conference and Workshop on Survey Data Harmonization
December 18-21, 2013
Palac Staszica, Nowy Swiat 72, 00-330 Warsaw

Topics

- Introduction
- Database and programming environment
- Data processing
- Database structure
- Master file
- Final questions

Introduction

- What is large?
  - 50 datafiles of total size 1.9 GB
  - 15 surveys
  - 69 surveys × waves
  - 1421 surveys × waves × countries
  - 1.9 million cases
  - 382 variables in a dataset on average (can be 1000+)
  - 1.9 billion elements

  - approx. 600 million elements in the master file

Introduction

- Why a computer programmer in the project?
- Why not SPSS, Stata, R, and the like?
- Why relational databases?
- Why freeware/open-source tools?

Environment

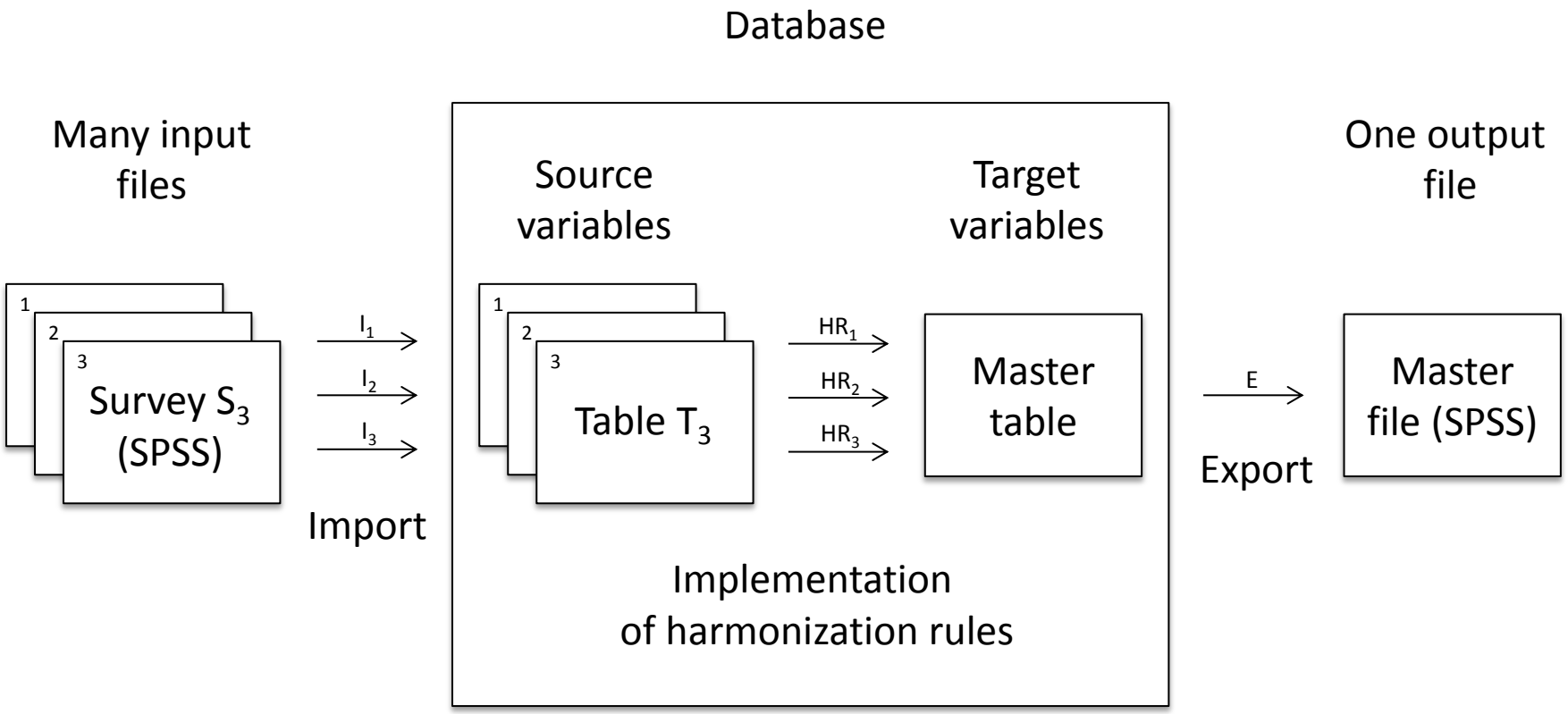
- Operating system
- Programming platform (Cygwin, UNIX-like)
- Languages (Perl, bash, awk, sed, SQL)
- Database (MySQL)
- HeidiSQL
- PSPP

Data processing (algorithm)

- Download SPSS system file (SAV)
- Convert SAV to plain text file (CSV)
- Extract metadata (codebook) from SAV file
- Create corresponding tables in database
- Load CSV data into data table
- Load SAV metadata into dictionary table

```
$ go orig/EB/ZA5612_v1-0-0.sav.gz
-rwxr-xr-x 1 Administratorzy None 7.6M Dec  1 14:57 /home/ppowalko/orig/EB/ZA5612_v1-0-0.sav.gz
-rwxr-xr-x 1 Administratorzy None 33M Dec  1 14:57 /home/ppowalko/orig/EB/ZA5612_v1-0-0.sav
  writing to csv... Done in 47 seconds.
  -rw-r--r-- 1 ppowalko None 57M Dec 16 13:21 /home/ppowalko/orig/EB/ZA5612_v1-0-0.csv
  generating table definition... Done in 19 seconds.
  executing sql instructions... Done.
  preparing csv for import... Done in 6 seconds.
  loading data to ZA5612_v1_0_0... Done in 6 seconds.
  writing to dic... Done.
  Creating /home/ppowalko/orig/EB/ZA5612_v1-0-0.dic.sql... Done.
  executing sql instructions dic_ZA5612_v1_0_0... Done.
```

Data processing (flow chart)



Database structure

- Data tables
  - 1:1 correspondence between data files and data tables
  - Compressing data (and time)
- Dictionary tables
- Country-level data tables

- Indexes
- SQL functions

Master file

- Master table vs master file
- Source, target, and control variables
- Implementation of harmonization rules
  - Playing with SQL

Master file

- Implementation of harmonization rules, an example

```
select
    'EQLS' as t_survey_name,
    wave as t_survey_edition,
    wave as s_survey_year,
    case wave when '1' then 2003 when '2' then 2007 when '3' then 2011 end as t_survey_year,
    Y11_Country as s_country,
    null as t_country_alpha2, -- to be recoded in a next step
    null as t_country_alpha3, -- to be recoded in a next step
    null as t_country_iso3,   -- to be recoded in a next step
    Y11_HH2b as s_age,
    null as s_birth_year,
    cast(Y11_HH2b as signed) as t_age,
    null as t_birth_year,
    Y11_Q28a as s_tr_parli,
    case when Y11_Q28a = '98' then -1 when Y11_Q28a in ('99') or isnull(Y11_Q28a) then -9 else rescale(Y11_Q28a,10,11,1) end as t_tr_parli_1,
    case when Y11_Q28a = '98' then -1 when Y11_Q28a in ('99') or isnull(Y11_Q28a) then -9 else rescale(Y11_Q28a,10,2,1)  end as t_tr_parli_2,
    Y11_HH2a as s_gender,
    if(Y11_HH2a='1',0,1) as t_gender
from EQLS_2003_2012
```

Final questions

- How to automate the whole process?
  - Batch jobs
  - Automatically generated SQL code
- How to verify output?
  - Tests and checks