Step-by-step: Data Harmonisation Process

Anja Neundorf (University of Nottingham)[1]

Rosalind Shorrocks (University of Manchester)

1 September 2017

Part of the ESRC-funded project
"Legacies of Authoritarian Regimes on democratic citizenship"[2]

This document details the documentation and harmonisation process for the creation of the dataset for the Legacy of Authoritarian Regimes on Democratic Citizenship project. In order to create the harmonised dataset four steps need to be followed:

- Download all datasets, which belong to separate repeated, cross-national studies. See section 1 how to access these studies.
- Decide which countries, years, and variables are needed for your research. We provide extended data documentation to make this step easier. See section 2 for details.
- Prepare datasets that belong to each study first by recoding the variables into common variable names and categories and if necessary merge single files for each study to create longitudinal files. See section 3.
- Merge all separate, prepared studies into one big dataset.

[1] Email: anja.neundorf@nottingham.ac.uk

[2] The project was funded by the ESRC Secondary Data Analysis Initiative (Phase 3) between 1 February 2016 and 31 July 2017. Award reference number: ES/N012127/1.

1. Accessing data documentation and datasets

The following 12 datasets were used for this project, accessed from the below links.

World Values Survey (WVS)
World Values Survey 1981-2014 Longitudinal aggregate v.20150418. World Values Survey Association (www.worldvaluessurvey.org). Aggregate File. Producer: JDSystems, Madrid, Spain
http://www.worldvaluessurvey.org/WVSDocumentationWVL.jsp

European Values Study (EVS)
EVS (2011): European Values Study 1981-2008, Longitudinal Data File, GESIS Data Archive, Cologne, Germany, ZA4804 Data File Version 2.0.0 (2011-12-30)
DOI: 10.4232/1.11005
http://zacat.gesis.org/webview/index.jsp?object=http://zacat.gesis.org/obj/fCatalog/Catalog5

European Social Survey (ESS)
Round 1: European Social Survey Round 1 Data (2002). Data file edition 6.5. NSD - Norwegian Centre for Research Data, Norway - Data Archive and distributor of ESS data for ESS ERIC. Round 2: European Social Survey Round 2 Data (2004). Data file eddition 3.5. NSD - Norwegian Centre for Research Data, Norway - Data Archive and distributor of ESS data for ESS ERIC. Round 3: European Social Survey Round 3 Data (2006). Data file edition 3.6. NSD - Norwegian Centre for Research Data, Norway - Data Archive and distributor of ESS data for ESS ERIC. Round 4: European Social Survey Round 4 Data (2008). Data file edition 4.4. NSD - Norwegian Centre for Research Data, Norway - Data Archive and distributor of ESS data for ESS ERIC. Round 5: European Social Survey Round 5 Data (2010). Data file edition 3.3. NSD - Norwegian Centre for Research Data, Norway - Data Archive and distributor of ESS data for ESS ERIC. Round 6: European Social Survey Round 6 Data (2012). Data file edition 2.3. NSD - Norwegian Centre for Research Data, Norway - Data Archive and distributor of ESS data for ESS ERIC. Round 7: European Social Survey Round 7 Data (2014). Data file edition 2.1. NSD - Norwegian Centre for Research Data, Norway - Data Archive and distributor of ESS data for ESS ERIC.
http://www.europeansocialsurvey.org/data/round-index.html

International Social Survey (ISSP)
1985: International Social Survey Programme: Role of Government I - ISSP 1985. GESIS Data Archive, Cologne. ZA1490 Data file Version 1.0.0. 1990: ISSP Research Group (1992): International Social Survey Programme: Role of Government II - ISSP 1990. GESIS Data Archive, Cologne. ZA1950 Data file Version 1.0.0. 1991: ISSP Research Group (1993): International Social Survey Programme: Religion I - ISSP 1991. GESIS Data Archive, Cologne. ZA2150 Data file Version 1.0.0. 1996: ISSP Research Group (1999): International Social Survey Programme: Role of Government III - ISSP 1996. GESIS Data Archive, Cologne. ZA2900 Data file Version 1.0.0. 1998: ISSP Research Group (2000): International Social Survey Programme: Religion II - ISSP 1998. GESIS Data Archive, Cologne. ZA3190 Data file Version 1.0.0. 2000: ISSP Research Group (2003): International Social Survey Programme: Environment II - ISSP 2000. GESIS Data Archive, Cologne. ZA3440 Data file Version 1.0.0. 2001: ISSP Research Group (2003): International Social Survey Programme: Social Relations and Support Systems/Social Networks II - ISSP 2001. GESIS

Data Archive, Cologne ZA3680 Data file Version 1.0.0. 2002: ISSP Research Group (2013): International Social Survey Programme: Family and Changing Gender Roles III - ISSP 2002. GESIS Data Archive, Cologne. ZA3880 Data file Version 1.1.0. 2003: ISSP Research Group (2012): International Social Survey Programme: National Identity II - ISSP 2003. GESIS Data Archive, Cologne. ZA3910 Data file Version 2.1.0. 2004: ISSP Research Group (2012): International Social Survey Programme: Citizenship - ISSP 2004. GESIS Data Archive, Cologne. ZA3950 Data file Version 1.3.0. 2006: ISSP Research Group (2008): International Social Survey Programme: Role of Government IV - ISSP 2006. GESIS Data Archive, Cologne. ZA4700 Data file Version 1.0.0. 2007: ISSP Research Group (2009): International Social Survey Programme: Leisure Time and Sports - ISSP 2007. GESIS Data Archive, Cologne. ZA4850 Data file Version 2.0.0. 2008: ISSP Research Group; Saflianto, Muhammad; Omondi, Paul; Thavajara, Joseph; Wanyama, Evangeline (2013): Religion Around the World Study of the 2008 International Social Survey Programme (ISSP). GESIS Data Archive, Cologne. ZA5690 Data file Version 1.0.1. 2010: ISSP Research Group (2012): International Social Survey Programme: Environment III - ISSP 2010. GESIS Data Archive, Cologne. ZA5500 Data file Version 2.0.0. 2013: ISSP Research Group (2015): International Social Survey Programme: National Identity III - ISSP 2013. GESIS Data Archive, Cologne. ZA5950 Data file Version 2.0.0
DOI: 1985: 10.4232/1. 1990: 10.4232/1.1950. 1991: 10.4232/1.2150. 1996: 10.4232/1.2900. 1998: 10.4232/1.3190. 2000: 10.4232/1.3440. 2001: 10.4232/1.3680. 2002: 10.4232/1.11564. 2003: 10.4232/1.11449. 2004: 10.4232/1.11372. 2006: 10.4232/1.4700. 2007: 10.4232/1.10079. 2008: 10.4232/1.11762. 2010: 10.4232/1.11418. 2013: 10.4232/1.12312
http://zacat.gesis.org/webview/index.jsp

Eurobarometer – Mannheim Trend File (EB)[3]
Schmitt, H., Scholz, E., Leim, I., Moschner, M. (2008): The Mannheim Eurobarometer Trend File 1970-2002 (ed. 2.00). European Commission [Principal investigator]. GESIS Data Archive, Cologne. ZA3521 Data file Version 2.0.1.
DOI: 10.4232/1.10074
http://zacat.gesis.org/webview/

Latinobarometro (LB)
Latinobarometro Corporation: Latinobarometer 1995-2015
http://www.latinobarometro.org/latContents.jsp

Afrobarometer (AfB)
Afrobarometer Data, Cape Verde; Benin; Algeria; Swaziland; Botswana; Burundi; Cameroon; Ghana; Guinea; Cote d'Ivoire; Kenya; Lesotho; Liberia; Madagascar; Malawi; Mali; Mauritius; Morocco; Mozambique; Namibia; Niger; Nigeria; Senegal; Sierra Leone; South Africa; Zimbabwe; Sudan; Togo; Tunisia; Uganda; Egypt; Tanzania; Burkina Faso; Zambia, Rounds 1-5, 1999-2001; 2004; 2005; 2008; 2015
http://www.afrobarometer.org/data/merged-data

[3] We decided to only use the Mannheim Trendfile of the Eurobarometer, which only covers the time 1970 to 2002. The reason is that the European countries that are part of the Eurobarometer were covered by the ESS from 2002 onwards.

Americas Barometer (AB)
The Americas Barometer by the Latin American Public Opinion Project (LAPOP)
http://vanderbilt.edu/lapop/raw-data.php

Asian Barometer (ANB)
Asian Barometer
http://www.asianbarometer.org/data/data-release

Comparative Study of Electoral Systems (CSES)
Modules 1-3: Heiko Giebler, Josephine Lichteblau, Antonia May, Reinhold Melcher, Aiko Wagner & Bernhard Weßels. CSES MODULE 1-3 HARMONIZED TREND FILE [dataset]. May 31, 2016 version. Module 4: The Comparative Study of Electoral Systems (www.cses.org). CSES MODULE 4 THIRD ADVANCE RELEASE [dataset]. June 22, 2016 version.
DOI: Modules 1-3 harmonised trend file: 10.7804/cses.trendfile.2016-05-31. Module 4: 10.7804/cses.module4.2016-06-22
http://www.cses.org/datacenter/download.htm

Central and Eastern Eurobarometer (CEEB)
Central Archive for Empirical Social Research (1997): Central and Eastern Eurobarometer 1990-1997: Trends CEEB1-8. European Commission [Principal investigator]. GESIS Data Archive, Cologne. ZA3648 Data file Version 1.0.0
DOI: 10.4232/1.3648
http://zacat.gesis.org/webview/

Asia Barometer (AsiaB)
The Asia Barometer Project
https://www.asiabarometer.org/en/data

2. Data documentation

The following three spreadsheets contain the key data documentation for these datasets. The spreadsheets cover the following:

- **Waves by Country**: provides detail on coverage for each country. Use this spreadsheet to assess the number of years and which surveys a country appears in.
- **Variables by Dataset Wave:** variable names for each dataset. Use this dataset to find which variables are available in each dataset and for which year.
- **Question Wording:** question wording for each variable in each dataset. Use this dataset to assess similarity and difference in question wording between the datasets for the same variable.

More detail on these spreadsheets is given below

2.1 Waves by Country

This excel spreadsheet documents the years there are surveys available for, for each country. The rows of this spreadsheet show the country, the columns show the years, and the cells show the datasets available for that country-year. This spreadsheet was created first and gives a broad overview of coverage over time for each country. It should be noted that specific variables may appear in some countries and not others, even if they are both in the same survey for the same year.

2.2 Variables by Dataset Wave

This spreadsheet gives the variables contained in each dataset. Each tab refers to a group of variables, and these tabs correspond to the tabs in the **Question Wording** spreadsheet. Our variable names are given in the first column, and these also correspond to the variable names assigned to different questions in the **Question** Wording spreadsheet. Dataset and survey wave are detailed across the first two rows. The variable names from the original dataset are included in the cells.

At the start of the documentation process, we documented all possible variables in each dataset within each of the variable groups. However, as the research process went on, we refined our research questions and noted which questions were asked in many datasets and years, and which were more one-off or had significantly less coverage. This means that the documentation process was discontinued for certain variables, and so some variable groups are documented comprehensively across all datasets whilst others are not. In general, the documentation for the Latinobarometer (LB), Americas Barometer (AB), and World Values Survey (WVS) are comprehensive. For other datasets the absence of a certain variable does not necessarily mean it does not exist in that dataset; it may be that we discontinued its documentation.

Below we give more detail on the level of coverage across datasets for all variable groups.

1. Complete coverage: *Political trust, Political interest, Gender, Party ID.*

All variables listed in these tabs are documented for all datasets. If a variable on these lists existed in a dataset, it was documented here in all cases.

2. Partial to complete coverage: *Democratic development, Attitudes towards democracy, Political efficacy, Voting, IVs*

The variables highlighted in blue in these tabs are documented for all datasets; the details of their coverage are complete.

For the variables not highlighted, the documentation is complete for the LB, WVS, AB, and ANB, incomplete for the ISSP and EVS, and was discontinued before the AfB, ESS, EB, CSES, CEEB, and AsiaB were documented.

3. Partial coverage: *Life satisfaction, Attitudes towards country, Governance, Role of Government, Government powers, Policy attitudes, Political participation, Interpersonal trust, Civic actions,*

Meaning of democracy, Ideology, Attitudes to military rule, Attitudes to judiciary, Rights and freedoms, Social attitudes, National pride, Political knowledge, Attitudes towards politics, Misc.

For all variables in these tables, the documentation is complete for the LB, WVS, AB, and ANB incomplete for the ISSP and EVS, and was discontinued before the AfB, ESS, EB, CSES, CEEB, and AsiaB were documented.

2.3 Question Wording by Dataset

This documents the question wording in each dataset for the various variables, as well as the codes used in the data files. The rows give the variable name we have assigned, the columns give the dataset, and the cells give the question wording. The variable names in the rows correspond to the same variable names in the

NB: This information was taken from the codebooks for each dataset. In the actual datafiles, sometimes the codes used differed in practice from those in the codebooks.

3. Within-dataset harmonisation

The datasets are provided in various forms:

- The WVS, AB, EVS, EB and CEEB provide fully harmonised longitudinal datafiles.
- The ESS provides separate datafiles for each wave but the same variable names and codes are used in each file.
- The CSES provides waves1-3 in a fully harmonised longitudinal datafile; wave 4 is provided separately.
- The ISSP, LB, ANB, AfB, and AsiaB are provided as separate datafiles for each wave, which may have different variable names and codes (for the same question).

The first stage in the harmonisation process was therefore to merge and harmonise the individual waves of the ESS, CSES, ISSP, LB, ANB, AfB, and AsiaB. The spreadsheets **Variables by Dataset Wave** and **Question Wording** were used for this process.

The ESS and CSES waves were merged straightforwardly, since variables and coding were already consistent across the separate waves. The separate waves were appended and saved as one, longitudinal, dataset.

For the ISSP, LB, ANB, AfB, and AsiaB, the following process was followed:

1. The separate waves were appended, retaining only the variables required.
2. Using the **Question Wording** spreadsheet and the codebooks for the datasets, it was determined whether there was a consistent set of categories across all waves within the dataset. For the majority of variables, this was the case. The categories were thus left as they were and given consistent codes and variables names for each wave.

3. Where the categories were inconsistent across waves, a harmonised variable was created which enabled the use of all waves of the dataset where the question was asked. This sometimes involved collapsing multiple categories into one. More details on this can be found in a separate **Data Documentation** document

At the end of this process we had 12 datafiles: a longitudinal dataset including all waves from all 12 datasets. At this point, each dataset had its own, dataset-specific, variable names and codes: names and codes were consistent within a dataset, but not across them. For example, marital status was x007 in the EVS but marital_issp in the ISSP.

3.1 Cross-dataset harmonisation

Each of these 12 longitudinal datafiles was then prepared with the creation of variables with variable names and response codes, which were consistent across all datasets. For example, marital status was then named married in all datafiles.

This is the stage at which both independent and dependent variables were recoded so that they were the same in all datasets. This generally involved collapsing categories down to the lowest number that existed in any dataset. The details of how each variable was recoded to make it comparable across all datasets can be found in the **Data Documentation.**

At the end of this stage of the process, we had 12 longitudinal datafiles (one for each dataset), which had consistent variable names and response codes for all the variables that we included.

We provide a template STATA do-file for step 3 called "`template_within_data_prep.do`". We also provide a STATA do-file to create the country labels.

4. Dataset merge

In the final stage, these 12 longitudinal datafiles were appended into one datafile. This created a dataset with over 3 million respondents. The dataset from which each respondent came from was denoted with the variable *data*, created during stage 3. We provide a template STATA do-file for step 4 called "`template_merge.do`".