

ONBound-Harmonization User Guide (Stata/SPSS), Version 1.0

Insa Bechert
Antonia May
Markus Quandt
Katharina Werhan

With the collaboration of:

Annette Schnabel, Heinrich-Heine University Düsseldorf
Kathrin Behrens, Heinrich-Heine University Düsseldorf

Student assistants:

Emma Ischinsky
Jaanika Juntson
Isabell Lülfi
Michaela Richter
Sebastian Stecker

2020, GESIS Leibniz Institute for the Social Sciences

The suggested citation for this user guide is:

May, Antonia, Werhan, Katharina, Bechert, Insa and Quandt, Markus (2020): ONBound User Guide.
Retrieved from <http://onbound.gesis.org/wizard> [date of download].

Table of Content

Table of Content.....	2
This Document.....	3
Introduction.....	3
1. Citation(s).....	5
2. ONBound Data.....	7
2.1 Additional Files.....	7
2.2 Basic Structure.....	8
2.3 Scope, Variable Selection and Coverage.....	9
2.4 Included Topics.....	9
2.4.1 Micro-Level Topics.....	9
2.4.2 Macro-Level Topics.....	10
2.5 Included Datasets.....	11
2.5.1 Micro-Level Datasets.....	11
2.5.2 Macro-Level Datasets.....	14
2.6 Detailed Structures.....	14
2.6.1 Variable Names.....	15
2.6.2 Harmonization of Scales (Micro-Level).....	16
2.6.3 Respondent IDs.....	18
2.6.4 Preparation and Expansion of Macro-Level Data.....	19
2.6.5 Harmonization of Country Codes.....	20
2.6.6 Harmonization of Further Categorical Variables.....	22
2.7 Missing Data.....	24
3. ONBound Harmonization in Practice.....	26
3.1 Overview.....	26
3.2 The ONBound Harmonization Wizard.....	26
3.3 Structure (zip.-Folder).....	27
4. How to use ONBound Harmonization Syntax.....	29
4.1 BEFORE using the ONBound Syntax Package (Stata-Users).....	29
4.2 BEFORE using the ONBound Syntax Package (SPSS-Users).....	30
4.3 Special Remarks.....	31
4.3.1 (A) Necessary name changes (Stata and SPSS users).....	31
4.3.2 (B) Unpacking files (Stata and SPSS users).....	31
4.3.3 (C) Unpacking files (Stata-users).....	31
4.4 Saving Storage on your Device.....	32
4.5 USING the ONBound Syntax Package for Harmonization.....	32
4.6 A deeper insight to the ONBound Universe.....	33
4.7 Updating ONBound to newer versions.....	37
4.7.1 Micro-Level Datasets.....	37
4.7.2 Macro-Level Datasets.....	38

This Document

This document provides information on how to work with the data of the ONBound Old and New Boundaries: National Identities and Religion project. It outlines the project's starting point, the data harmonization procedures, the harmonization repository, the ONBound Harmonization Wizard, and most importantly, guides through the steps and procedures needed to compile a customized ONBound dataset. Chapter One of this user guide informs about citations for ONBound data and metadata. Chapter Two gives an overview of the variables available and original datasets that were included. It also introduces our harmonization strategies. Chapter Three introduces the ONBound Harmonization Wizard. Finally, Chapter Four leads users through the application of the ONBound harmonization routines step by step. Before running our ONBound Syntax Package, please read at least chapter four carefully.

Introduction

In times of increasing institutional activities at the supra-national level and growing diversities in e.g. wealth and life chances at the individual level, national and religious identities seem to re-gain salience as markers between "us" and "them". Potential consequences are changes in the patterns of social cohesion and societal solidarity. The ONBound project contributes to existing research in this area regarding two core questions: The first concerns the relationship between religious and national identities at the individual level. How strongly do these two important cognitive frames for the social world overlap, or conversely, create cross-pressures? How consistent are the patterns of overlap or independence across countries, and how stable over time? The second question extends the first one, in that it asks about the effect of institutional characteristics and historical events at the country level on the relationship between religious and national identities. For this, a vast array of existing individual level and contextual data were merged and enriched. The result is a publicly available 'virtual' multi-level database that is designed to facilitate advanced statistical analyses by e.g. multi-level regression. This data infrastructure allows to address not only the above research questions, but also a wide variety of related issues, such as populism, anti-migration attitudes or trust in institutions in a comparative perspective. In terms of methodological research, the ONBound data compilation enables researchers to compare different measurement instruments for the same constructs. Researchers may concentrate on the currently included data sources, but also have the possibility to add more data individually.

Beyond including numerous existing micro-level and macro-level data sources, the ONBound final data embeds data from the ONBound subproject *Religion and Nation in Constitutions Worldwide* which extracts information from national constitutions on our core elements, national identity and religion. The database is most conveniently accessed by the ONBound Harmonization Wizard. The Harmonization Wizard is an online tool that enables users to select variables from the database by topics and scope of research. The Wizard then provides the user with a bespoke set of syntax files that can create a data set with the customized version of the ONBound harmonized data on the user's personal computer.

In total, the ONBound database contains 750 harmonized target variables based upon 1500 source variables from about 280 survey waves stemming from 18 studies for the micro-level variables, as well as 1650 variables from 17 macro-level repositories. Data on the micro-level reach from 1970 to 2018. For the macro-level we included all data points from 1945 to 2019, if possible. The country composition for both levels is entirely based on data availability. In other words, we included all countries in the world for which data were available in our source surveys. All sources were selected based on their conceptual fit. For the micro-level all comparative and publicly available surveys were selected, while on the macro-level coverage and data quality was considered for supplementing and enhancing the micro-level.

We are grateful to Emma Ischinsky, Jaanika Juntson, Isabell Lülfi and Michaela Richter for superb research assistance. With their engagement, accuracy and dedication they made great contributions, especially to improving our coding schemes for countries, religious denominations and parties, but also to translating data and syntax files between SPSS and Stata. We are also very thankful for the highly professional contribution of Matthäus Zloch, who brought our vision of user-friendly harmonization automation to life. For this project, not only did the department *Knowledge Technologies for the Social Sciences* enrich our project, but also the exchange with other colleagues from our department *Data Archive for the Social Sciences* helped during the development. We are especially happy to have shared the process with the HaSpaD project and its staff Sebastian Sterl and Sonja Schulz. We are grateful for the fruitful meetings in which both harmonization projects shared ideas, discussed and enriched each other. Our gratitude also extends to our workshop participants, who pointed towards merits and remaining demerits during the process, added value through their substantial and technical advice and motivated us to continue our path. Special thanks go to our collaborators from the University of Milan, Ferruccio Biolcati Rinaldi and Cristiano Vezzoni, and Mikael Hjerm from Umeå University. All users of pre-versions of our database are thanked for their active participation. Their feedback is much appreciated and helped to improve the project output.

Finally, we are especially grateful to have worked with our collaborators and their colleagues at the Heinrich-Heine University Düsseldorf. Annette Schnabel and Kathrin Behrens accompanied our journey with regular feedback and theoretical expertise. The data they compiled in the ONBound subproject *Religion and Nation in Constitutions Worldwide* can now be used as a stand-alone product but are also integrated in the ONBound database.

The ONBound project has been funded by the German Research Foundation (DFG). This enabled us to compile a large-scale harmonization database that will foster a deeper understanding of the mechanisms of national identity and religion. We hope that the roads we took will inspire future harmonization projects and provide starting points for their journeys.

1. Citation(s)

Below the official citations for all ONBound related data and documentation are listed:

For this *User Guide*:

May, Antonia, Werhan, Katharina, Bechert, Insa and Quandt, Markus (2020): ONBound User Guide. Retrieved from [URL Wizard Website](#) [date of download].

For the *ONBound Micro-Level-Data_Overview (v4.0)*:

ONBound (2020/06/30): 'ONBound Micro-Level-Data_Overview_v4.0'. Retrieved from [URL Wizard Website](#) [date of download].

For the *ONBound Micro-Level-Data_Documentation (v4.0)*:

ONBound (2020/06/30): 'ONBound_Documentation_Micro_v4.0'. Retrieved from [URL Wizard Website](#) [date of download].

For the *ONBound Macro-Level-Data_Documentation (v4.0)*:

ONBound (2020/06/30): 'ONBound_Documentation_Macro_v4.0'. Retrieved from [URL Wizard Website](#) [date of download].

For the *ONBound_APPENDIX I_Country Codes (v4.0)*:

May, Antonia and Werhan, Katharina (2020): ONBound_APPENDIX I_Country Codes_v4.0. Retrieved from [URL Wizard Website](#) [date of download]

For *ONBound_APPENDIX II.a Concept Classification Religions*:

Bechert, Insa and Ischinsky, Emma (2020): ONBound_APPENDIX II.a Concept_Classification Religions. Retrieved from [URL Wizard Website](#) [date of download].

For *ONBound_APPENDIX II.b Coding Scheme Classification Religions*:

Ischinsky, Emma and Bechert, Insa (2020): ONBound_APPENDIX II.b CodingScheme_Classification Religions. Retrieved from URL Wizard Website [date of download].

For *ONBound_APPENDIX III_Subnational Regions*:

Werhan, Katharina, Stecker, Sebastian, and Ischinsky, Emma (2020): ONBound_APPENDIXIII_Subnational Regions. Retrieved from URL Wizard Website [date of download].

For *ONBound_APPENDIX IV_Micro-Level-Data_Weights (v4.0)*:

Werhan, Katharina (2020): ONBound_Micro-Level-Data_Weights_v4.0. Retrieved from URL Wizard Website [date of download].

For *ONBound_APPENDIX V.a_Concept_Party Codes*:

May, Antonia (2020): ONBound_APPENDIX V.a Concept_Party Codes. Retrieved from URL Wizard Website [date of download].

For *ONBound_APPENDIX V.b Coding Scheme Party Codes*:

May, Antonia (2020): ONBound_APPENDIX V.b Coding Scheme_Party Codes. Retrieved from URL Wizard Website [date of download].

The customized ONBound dataset should be cited as follows:

Bechert, Insa, May, Antonia, Quandt, Markus and Werhan, Katharina.2020. ONBound - Old and new boundaries: National Identities and Religion. Customized dataset. GESIS Data Archive, Cologne. <http://onbound.gesis.org/wizard>. [Date of dataset compilation].

Please make sure to also cite the original data sources you downloaded to create the harmonized ONBound data set. The correct citations are available on the single survey programs' websites or on the ONBound website <http://onbound.gesis.org/wizard> where you started the data download, as well as our ONBound_List_of_source_datasets.xlsx.

To avoid copyright infringement with the source survey programs and macro data providers, please, do not distribute your compiled data set to any third parties.

2. ONBound Data

For the ONBound project we first had to identify possible data sources, common variables and potential pitfalls of the harmonization process. This chapter presents our approaches, the steps taken, introduces core information and explains decisions before the next chapter introduces the actual harmonization procedure.

2.1 Additional Files

In addition to the actual harmonization files and this user guide, ONBound provide further documents that can be consulted to address specific questions:

- ONBound_Documentation_Micro_v4.0.xlsx
- ONBound_Documentation_Macro_v4.0.xlsx
- ONBound_Micro-Level-Data_Overview_v4.0.xlsx
- ONBound_List_of_source_datasets.xlsx
- ONBound_Appendix_I_Country_Codes_v4.0.pdf
- ONBound_Appendix_II.a_Concept_Classification Religions.pdf
- ONBound_Appendix_II.b_Coding Scheme_Classification Religions.xlsx
- ONBound_Appendix_III_Subnational Regions.pdf
- ONBound_Appendix_IV_Micro-Level-Data_Weightsv4.0.xlsx
- ONBound_Appendix_V.a_Concept_Party Codes.pdf
- ONBound_Appendix_V.b_Coding Scheme_Party Codes.xlsx

The first two documents, **ONBound_Documentation_Micro (v4.0)** and **ONBound_Documentation_Macro (v4.0)**, serve as the ONBound codebook. They include the information on harmonized target variable names, labels and coding. Additionally, **ONBound_Micro-Level-Data Overview (v4.0)** includes information on the origin of micro-level variables. Thus, these three sources build the core of ONBound documentation. Since they comprise a great number of variables across two research fields and two levels of analysis, for the single variables our documentation cannot be as comprehensive as in the original documentations. For in-depth analyses of specific variables, we therefore recommend to also consult the original documentation sources.

The **ONBound_List_of_source_datasets** lists all included datasets with information on the versions used during the project phase and offers links to their repositories.

Please note: Since our harmonization procedures are developed for the exact version available at the time, in order to use our database, you may not be able to use the latest data versions. Including updated versions of datasets requires advanced coding skills in either Stata or SPSS. How to handle changes like this in the ONBound repository will be explained in the last chapter.

In addition to the previous documents providing detailed information on substantial harmonization, the appendices inform users about our coding strategies of categorical core and merge key variables. They offer attempts of standardized coding schemes. We advise users to carefully consider these documents. Especially Appendix I, listing the country codes, should be consulted for all cross-national analyses. Since, all ONBound datasets are based on a country_year structure, standardization is a technical necessity. To include all micro- and macro-level datasets from 1945 to 2018, we had to introduce a coding scheme that captures unifications, splits, name changes and territorial changes of countries. As a result, we introduced one variable (`ctry_ISO3onb`) that accounts for historical political changes and one variable (`ctry_ISOnumonb`) that follows territorial entities as of today despite political changes. When analyzing data from countries that experienced such changes, country codes and country names differ across time according to these changes. Please consult the `ONBound_Appendix_I_Country_Codes_v4.0`.

Please note: `ctry_ISO3onb` (alphanumeric codes) and `ctry_ISOnumonb` (numeric country codes) follow different coding rules and therefore are not interchangeable. Please refer to `ONBound_Appendix_I_Country_Codes_v4.0`

Appendix II to Appendix V inform about coding strategies of categorical variables (Religious Denominations, Subnational Regions, Parties) and composition of technical variables (Weights).

2.2 Basic Structure

The comprehensive database consists of two main components. The first component comprises all micro-level variables, which are based on observations by respondents nested in their countries and years of the interview. The second component consists of all country- (or macro-) level variables, which are based on observations by countries and years. In order to facilitate merging of macro- to micro-data, some macro-data have been duplicated to fill systematically missing year observations. Please refer to `ONBound_Documentation_Macro` (v4.0) or chapter 2.5.4. Both databases are structured in order to be linked. However, it is of course also possible to generate a purely micro-level or purely macro level dataset to work with.

Please note: Compiling a complete ONBound dataset including all linked sources is **not recommended**. Such a dataset includes 5 million observations on 2400 variables and therefore overloads the capacity of most devices. It is highly recommended to select variables, countries (and possibly timeframes) before downloading the data from the ONBound Harmonization Wizard.

2.3 Scope, Variable Selection and Coverage

The decisions made for the variable selection were preceded by an intense literature review on existing research concerning religious identities, national identities and their interface. The project team carefully identified the independent and dependent variables used for research in this field or interlinked topic areas. Beyond, we constructed further possible research questions within the project's scope and made sure to include the variables necessary for analyzing them.

2.4 Included Topics

For a better usability, we sorted variables into topics. These topics help users to navigate through the variety of variables included on both levels. Moreover, these topics serve as main selection key on the level of individual data when compiling harmonized data files. Users are not selecting individual variables for their target files, but topical blocks of variables, and our syntax-packages perform the dataset compilation at the topic level. In contrast to the micro-level data, macro-level topics are based on the original macro-level sources which have to be selected completely.

2.4.1 Micro-Level Topics

The micro-level section includes various subdimensions of national identity and religion. Additionally, ONBound offers many more topics that are related to those both core elements. Moreover, variables of socio-demographic background of respondents as well as protocol variables are appended for every observation on the micro-level.

Micro-level data offer variables on the following topics:

Background and Protocol Variables

- (6100) Background Variables
- (6100_2) Regions
- (6200) Protocol Variables

National Identity

- (1000) Formal Membership/ Citizenship
- (1100) Origin respondent and family
- (1200) Feeling of belonging/ self-identification
- (1300) Membership Boundaries
- (1400) National pride

Religious Identity

- (2000) Formal Membership: Denominations
- (2100) Practices
- (2200) Belief
- (2300) Religiosity/ Centrality of religion in life
- (2400) Attitudes towards religion
- (2800) Religion and Nation

Interrelation of Groups

- (3000) Identification with different groups of community
- (3100) Tolerance towards different groups of community
- (3200) Discrimination
- (3400) Racism

Legitimacy

- (4100) Civic Attitudes
- (4200) Support for Regime Type Principles
- (4300) Regime Performance
- (4400) Regime Institutions

Miscellaneous

- (5400) Politics
- (5700) Immigration

2.4.2 Macro-Level Topics

For the macro-level data, we included the source datasets as a whole, without prior pruning down. Here, some variables or variable groups may be assigned to more than one topic. Therefore, we had to make decisions for “major topics” in order to assign them. Besides substantial variables, administrative variables were added for every entry.

Macro-level data offer variables on the following topics:

Administrative Variables

- ONBound-Variables
- Harmonized Administrative Variables
- Study Specific Administrative Variables

Policies

- Constitutions: Religions, National Identity, Context
- Policy Settings: Religion, Integration, Immigration

State Indicators

- Economy
- Standard of Living
- Politics & Polity
- Special Topics: Human Rights, Globalization,

Civil Society

- Population: Basic Population Figures, Migration, Education,
- Religion: Religious Relevance
- Diversity: Religious Diversity, Ethnic Diversity

2.5 Included Datasets

The following paragraphs offer an overview on the datasets included. For detailed information, please refer to the documentation (ONBound_Documentation_Macro_v4.0.xlsx) and the master matrix (ONBound_Datamatrix_Micro_v4.0.xlsx).

2.5.1 Micro-Level Datasets

The selection of surveys and variables was based on their thematic fit, a comparative scope and their coverage. We extracted relevant variables for our research area, recoded and rearranged them for completely new perspectives and new insights. Please refer to the ONBound List_of_source_datasets for citations.

These are the datasets included:

- **Afrobarometer (AfroB)**
 - Merged Round 1 Data (12 countries) (1999-2001), Merged Round 2 Data (16 countries) (2004), Merged Round 3 Data (18 countries) (2005), Merged Round 4 Data (20 countries) (2008), Merged Round 5 data (34 countries) (2011-2013), Merged Round 6 data (36 countries) (2016)

- **Americas Barometer 2004-2018 Grand Merge File - FREE (AmericasB_2004_2014)**
(reduced to 2004-2014)
- **Arab Barometer (ArabB)**
 - Arab Barometer Wave I (2006-2009), Arab Barometer Wave II (2010-2011), Arab Barometer Wave III (2012-2014), Arab Barometer Wave III (2016-2017)
- **AsiaBarometer (AsiaB)**
 - AsiaBarometer 2001, AsiaBarometer 2003, AsiaBarometer 2004, AsiaBarometer 2005, AsiaBarometer 2006, AsiaBarometer 2007
- **Asia Europe Survey (ASES): A Multinational Comparative Study in 18 Countries, 2001**
- **Asian Barometer**
 - WAVE 1: 8 Countries in East Asia 2001-2003, WAVE 2: 13 Countries in East Asia, 5 Countries in South Asia 2005-2008, WAVE 3: 13 Countries in East Asia 2010-2012, WAVE 4: 14 Countries in East Asia 2014-2016
- **EU Neighborhood Barometer**
 - EU Neighbourhood Barometer Wave 1, EU Neighbourhood Barometer Wave 2, EU Neighbourhood Barometer Wave 3, EU Neighbourhood Barometer Wave 4, EU Neighbourhood Barometer Wave 5, EU Neighbourhood Barometer Wave 6
- **Eurobarometer**
 - **Applicant Countries Eurobarometer (AECB) / Candidate Countries Eurobarometer (CCEB)**
 - Applicant Countries Eurobarometer 2000, Applicant Countries Eurobarometer 2001.1, Candidate Countries Eurobarometer 2002.1, Candidate Countries Eurobarometer 2002.2, Candidate Countries Eurobarometer 2002.3, Candidate Countries Eurobarometer 2003.2, Candidate Countries Eurobarometer 2003.3, Candidate Countries Eurobarometer 2003.4, Candidate Countries Eurobarometer 2003.5, Candidate Countries Eurobarometer 2004.1
 - **Central and Eastern Eurobarometer 1990-1997: Trends CEEB 1-8 (CEEB)**
 - **Standard and Special Eurobarometer (EB)**
 - Mannheim Eurobarometer Trend File 1970-2002, Eurobarometer 58.1, Eurobarometer 59.1, Eurobarometer 59.2, Eurobarometer 60.1, Eurobarometer 61, Eurobarometer 62.0, Eurobarometer 62.2, Eurobarometer 63.1, Eurobarometer 63.4, Eurobarometer 64.2, Eurobarometer 65.1, Eurobarometer 65.2, Eurobarometer 66.1,

Eurobarometer 66.3, Eurobarometer 67.2, Eurobarometer 68.1,
Eurobarometer 69.1, Eurobarometer 69.2, Eurobarometer 70.1,
Eurobarometer 71.1, Eurobarometer 71.3, Eurobarometer 72.4,
Eurobarometer 73.1, Eurobarometer 73.3, Eurobarometer 73.4,
Eurobarometer 74.2, Eurobarometer 75.3, Eurobarometer 76.3,
Eurobarometer 76.4, Eurobarometer 77.3, Eurobarometer 77.4,
Eurobarometer 78.1, Eurobarometer 78.2, Eurobarometer 79.3,
Eurobarometer 79.5, Eurobarometer 80.1, Eurobarometer 81.2,
Eurobarometer 81.4, Eurobarometer 82.3, Eurobarometer 83.1,
Eurobarometer 83.3, Eurobarometer 83.4, Eurobarometer 84.1,
Eurobarometer 84.3, Eurobarometer 85.2, Eurobarometer 86.1,
Eurobarometer 86.2, Eurobarometer 86.3

- **European Social Survey (ESS) Cumulation Waves 1-8**
- **European Values Study (EVS) & World Values Survey (WVS)**
 - EVS Longitudinal Data File 1981-2016, European Values Study 2017: Integrated Dataset (EVS 2017)
 - WVS_Longitudinal_1981-2016
- **International Social Survey Programme (ISSP)**
 - ISSP 1985/1990/1996/2006 Cumulation “Role of Government”, ISSP 1985/1990/1996/2006 Cumulation Add-On “Role of Government”, ISSP 2016 “Role of Government”, ISSP 1991 “Religion”, ISSP 1998 “Religion”, ISSP 2008 “Religion”, ISSP 1995 “National Identity”, ISSP 2003 “National Identity”, ISSP 2013 “National Identity”, ISSP 2004 “Citizenship”, ISSP 2014 “Citizenship”
- **IntUne - Integrated and United: A quest for Citizenship in an "ever closer Europe"**
 - IntUne 2007, wave 1, IntUne 2009, wave 2
- **Latinobarometer**
 - Latinobarómetro 1995, Latinobarómetro 1996, Latinobarómetro 1997, Latinobarómetro 1998, Latinobarómetro 2000, Latinobarómetro 2001, Latinobarómetro 2002, Latinobarómetro 2003, Latinobarómetro 2004, Latinobarómetro 2005, Latinobarómetro 2006, Latinobarómetro 2007, Latinobarómetro 2008, Latinobarómetro 2009, Latinobarómetro 2010, Latinobarómetro 2011, Latinobarómetro 2012-2013, Latinobarómetro 2015, Latinobarómetro 2016, Latinobarómetro 2017
- **New Europe Barometer I-XI Trend Dataset, 1991-2007 (NHEB)**
- **PewGAP**

- 2002_1, 2005_1, 2005_2, 2006, 2007, 2008, 2009_1, 2009_2, 2010, 2011, 2012, 2013, 2014_1, 2014_2, 2015, 2016

2.5.2 Macro-Level Datasets

Macro-level datasets are included completely. Changes to the macro-level datasets are limited to restructuring, limiting the time-coverage and expanding the entries to systematically missing years, e.g. for datasets that offer data in a 5-year rhythm. These cases than have been duplicated for the missing years. The procedure is explained in chapter 2.6.4. Please refer to the ONBound List_of_source_datasets for citations.

These are the datasets included:

- **KOG Globalization Index (KOF)**
- **Religious Characteristics of States Dataset Project, Demographics (RCS)**
- **Church Attendance and Religious Change, Pooled European Dataset (CARPE)**
- **United Nations: Demographic Statistics Database (Religious Affiliations) (UNSD2)**
- **World Religion Project – National Religion Dataset (WRP)**
- **Religious Diversity Index (RDI)**
- **The Ethnic Power Relations (EPR) Core Dataset 2018 (EPR)**
- **Religion and State Project, Round 3 (RAS)**
- **Religion and Nation in Constitutions Worldwide (RNCw)**
- **Migration Integration Policy Index 2015 (MIPEX)**
- **Indicators of Citizenship Rights for Immigrants (ICRI)**
- **The IMPIC Project: Immigration Policies in Comparison (IMPIC)**
- **World Development Indicators (WDI)**
- **UNHCR’s populations of concern (UNHCR)**
- **Polity IV Annual Time-Series, 1800-2018 (POLITY)**
- **Democratic Electoral Systems around the world, 1946-2016 (DES)**
- **CIRI Human Rights Data Project (CIRI)**

2.6 Detailed Structures

In this section we introduce our main harmonization steps and decisions. This includes consistent labeling and naming of variables across topics and data levels, an elaborated re-scaling that offers various coding schemes per variables included, necessary systematic coding

of variables used for matching across studies and country-level data as well as further coding schemes for categorical variables.

2.6.1 Variable Names

Both components of the ONBound database follow standardized rules when naming and labeling variables.

ONBound variable names contain a great deal of information:

First, variable names on the micro-level are written in lower case and variables on the macro-level are written in upper case with the exception of variables used for matching both levels. Due to case sensitivity of Stata, variable names of those variables are identical for both levels. This applies for *ctry_year*, *ctry*, *ctry_ISOnumonb*, *ctry_ISO3onb*.

Second, the first part of micro-level variable names refers to the main topic the variable was assigned to. The second element includes information on the measured concepts. The third is a number indicating different versions of the concept's measures. Variable names with lowercase letters denote different scaling versions of the variables included. For further information on scale harmonization, please refer to the chapter 2.6.2.

For example, the variable *pol_interest1b* refers to the main topic of politics (*pol*) and the concept of political interest (*interest*). This is the first political interest measurement (*1*) with the scaling version *b*. Micro-level data variable labels roughly follow this strategy. They also contain information of the character of the variable at hand. In the example above the variable is labeled "*Politics: Political interest1: How interested in politics, 5-p. interest*". It contains the topic ("*Politics*"), the measured concept ("*Political interest*"), the measurement version ("*1: How interested in politics*") and scaling information of this version ("*5-p. interest*")

Third, in analogy to the micro-level data, the first element of macro-level variable names refers to the main topic the variable was assigned to. The second element includes information on the sub-dimension of a topic. The middle parts refer to the original variable name and the last part is the acronym used to identify the data source.

For example, the variable *POL_MIG_AVGS_E01_IMPIC* is part of the main topic of *Policies* (POL) and the sub-topic *Migration* (MIG). The original variable name is *AvgS_e01* (AVGS_E01) and stems from the IMPIC dataset (The Immigration Policies in Comparison (IMPIC) Dataset, Bjerre et al. 2016).

Here, the variable labels also contain a reference to the elicitation mode and year coverage of the original dataset. In this example the variable label is “COI: Illegal residence (av) [annually; 1980-2010]” The label refers to the content (“COI: Illegal residence” = Control of immigration: Illegal residence), to the character of the variable (“(av)” = Average). Continuity and time-coverage of the included macro-level dataset are represented in brackets (here: annual data from 1980-2010).

2.6.2 Harmonization of Scales (Micro-Level)

This chapter explains our general harmonization principles and gives an example of concrete harmonization decisions taken. In the beginning of the harmonization processes we thought about general harmonization rules ideally applicable to every single harmonization task. However, we quickly had to realize that given the vast amount and variety of data, there are no ‘catch-all’ rules. So, we had to construct general rules, allowing for individual case decisions.

Our aim was to harmonize as comprehensively as possible, but at the same time maintain as much information as possible. Therefore, we decided on the principle of creating different levels of harmonized variables and keeping the original variable versions in the data set.

Please note: A generalized recode syntax for the harmonization process of a variable can be found in the respective line of the ONBound_Documentation_Micro_v4.0. This chapter explains the general strategies we used to harmonize scales.

The following table shows the example of variable natid_pride1 (Proud to be [COUNTRY NATIONALITY]) which was asked on a 4-, 5- and 7-point scale across different surveys.

ONBound Varname	ONBound Varlabel	Target Coding
natid_pride1	National Identity: Pride1: How proud are you being [COUNTRY NATIONAL], harmonized 4-point pride	(1) Very proud (2) Somewhat proud (3) Not very proud (4) Not proud at all
natid_pride1a	National Identity: Pride1: How proud are you being [COUNTRY NATIONAL], 4-point pride	(1) Very proud (2) Somewhat proud (3) Not very proud (4) Not proud at all
natid_pride1b	National Identity: Pride1: How proud are you being [COUNTRY NATIONAL], 7-point pride	(1) Not at all (2) (3) (4) (5) (6) (7) A Lot
natid_pride1c	National Identity: Pride1: How proud are you being [COUNTRY NATIONAL], 5-point agree	(1) Strongly disagree (2) Disagree (3) Neither agree nor disagree

(4) Agree
(5) Strongly agree

The fully harmonized variable is natid_pride1. For this measure we chose a 4-point scale, because this scale was used most widely and, thus, most of the data was already available on a 4-point scale. Versions a, b and c provide the data in their original scales, but also cumulated across surveys (and adapting the missing values to the ONBound logic). That way, researchers can decide for themselves on which harmonization level they would like to use the data.

Generally, we used the harmonization methods “Semantic Judgement of Response Options”, “Random Split” and “Linear Stretching”. The next example shows how we used the first two methods.

ONBound Varname	ONBound Variable	Target Coding
religid_centr3	Religious Identity: Centrality3: Religiousness respondent self-description, harmonized 4-point scale	(1) Very religious (2) Somewhat religious (3) Somewhat not religious (not asked in ArabB) (4) Not religious at all
religid_centr3a	Religious Identity: Centrality3: Religiousness respondent self-description, 4-point scale	(1) Very much practicing / Very religious (2) Practicing / Moderately religious (3) Not very much practicing / Lightly religious (4) Not practicing / Not religious at all
religid_centr3b	Religious Identity: Centrality3: Religiousness respondent self-description, 7-point scale	(1) Extremely religious (2) Very religious (3) Somewhat religious (4) Neither religious nor non-religious (5) Somewhat non-religious (6) Very non-religious (7) Extremely non-religious
religid_centr3c	Religious Identity: Centrality3: Religiousness respondent self-description, 3-point scale	(1) Religious (2) Somewhat religious (3) Not religious

For the measurement of self-assessed religiosity, most study programs used a 4-point scale, therefore, we decided again against a linear stretching towards a 7 or even 11-point scale. Thus, we harmonized the ISSP-7-point scale into the 4-point target variable (randomly splitting the mid-category) and matched the ArabB categories into the 4-point according to the label wording.

Another good example for our harmonization methods is the trust in institutions item battery, because it has been asked in many studies and waves. Most studies use a 4-point scale, others use 5-point, 7-point, 11-point or 2-point scales. Since we wanted to avoid having too many versions of the variables, we decided to code the different scales into three partly harmonized variable versions and have one completely harmonized variable on a 11-point scale.

ONBound Varname	ONBound Varlabel	Target Coding
-----------------	------------------	---------------

trust_instnat1	Trust in institutions: National Institutions1: National parliament, harmonized 11-point scale	(0) No trust at all ... (10) Complete trust
trust_instnat1a	Trust in institutions: National Institutions1: National parliament, partly harmonized 4-point scale	(1) A great deal / Trust a lot (2) Quite a lot (3) Not very much (4) None at all / Don't trust at all
trust_instnat1b	Trust in institutions: National Institutions1: National parliament, partly harmonized 5-point scale	(1) Complete confidence (2) A great deal of confidence (3) Some confidence (4) Very little confidence (5) No confidence at all
trust_instnat1c	Trust in institutions: National Institutions1: National parliament, dichotomous	(1) Tend to trust (2) Tend not to trust

For the fully harmonized measure (trust_instnat1) all variables were “stretched” into the 11-point scale. This means that response options were stretched to the larger scale range, by assigning the lowest response option to 0 and the highest to 10, while all intermediate options are given equally distanced numbers in between. The dichotomous variables do not ask for extreme categories "No trust at all" or "Complete trust". Instead, the categories are "tend to trust" and "tend not to trust". We therefore decided to place these categories between the extreme points and the center of the scale, i.e. on 2.5 for "tend not to trust" and 7.5 for "tend to trust". Variable version a covers all original 4-point scales and version b combines all scales with a natural mid-point (5-, 7- and 11-point scales), using the semantic judgement method. Finally, version c contains all dichotomous data. As mentioned before, the recode syntax for these variables is available in the respective lines in the ONBound_Documentation_Micro_v4.0.

2.6.3 Respondent IDs

The micro-level data include two respondent IDs: *caseid* and *caseid_onbound*. The first is the original Respondent ID coded into a string variable. This original caseid combined with the onbound_wave allows users to add variables not included in the ONBound datasets to their harmonized files. The second variable is a numeric ID that is unique through the whole ONBound micro-level data comprising about 5.8 million respondents. In order to create this ID, we first sorted the source dataset by ctry_year, onbound_wave and caseid and assigned a consecutive number from 1 to N to each case in the dataset. In a second step we composed the caseid_onbound from the onbound_wave (1-280) and the consecutive number assigned in step 1.

2.6.4 Preparation and Expansion of Macro-Level Data

Matching macro-level data to individual-level data in our setting always means matching by common country-keys on both the individual-level and the macro-level side. For this to be feasible, two modifications of the macro-level resources were necessary. First, we needed to rearrange the macro-datasets to a country_year structure. Sometimes this forced us to create rather large wide-format macro datasets.

Second, when harmonizing macro-level data, we included datasets with different time-coverages and regularity. In order to have a large and consistent structure, we expanded the data in order to fit an annual structure, if possible. This enabled us to increase the chance of matching both parts with the micro-level data being surveyed at specific time-points.

There are four different types of time-coverages:

1. Annual Data

Datasets providing annual data are kept as in the original. Their variable labels are followed by [annually; XXXX-XXXX], while XXXX-XXXX represent the total timespan, e.g. 1960-2017.

2. Regularly repeated data

Some datasets provide data in regular intervals. Their variable labels are followed by “[every Y years; XXXX-XXXX]”, with Y indicating the interval and XXXX-XXXX the total timespan, e.g. every 5 years; 1965-2000]. In this case, we expanded the data to an annual structure. The entries for the year of data collection were repeated for the two years before and two years after the recorded year. Therefore, data from 1970 were copied to 1968, 1969, 1971 and 1972 and so on. While other modes of imputation might have been preferable, this approach makes the fact of ‘data extension’ very visible to users. Variables that have been copied can be identified by the label addition “every X years”.

3. Single Year Data

For purely cross-sectional macro datasets, providing data for only a certain year, we deployed two methods. Depending on whether the cross-sectional observations must be assumed to be invalid for wider time spans or not, we either included the one time point only, or, if possible, we extended the existing data. For example, the *Religion and Nation in Constitutions Worldwide (RNCw)* dataset offered the variable “most recent constitution” (CON_CONTEXT_C2B_RNCw). The entries of 2017 were repeated

accordingly, here: until the last change of the constitution. In the case of USA, the data could be repeated for the years 1992-2017, since the last change took place in 1992. For the first case, with no expansion, the variable labels are followed by [single year data; XXXX], where XXXX represents the time of the data collection. For the second case, where expansion is feasible, the variable labels are followed by [single year data; XXXX-XXXX] representing the maximal timespan, e.g. 1953-2017.

4. Periodic Data

The last type of data collection is mostly found in election-related datasets. Here, the datasets provide e.g. seats won in an election. In those cases, the data was repeated for the country specific legislature periods. Variable labels are followed by [periodic data; XXXX-XXXX]. Again, XXXX-XXXX represents the total timespan of the dataset, even if not all countries offer data for every time point.

2.6.5 Harmonization of Country Codes

For our matching structure that is bound to a country by year structure, we encountered several problems connected to the historical stability of states. In order to overcome these, we introduced a coding scheme for country codes accounting for territorial changes, changes of constitutive people, separations and unifications since 1945. The coding scheme is based on ISO 3166-1 alpha-3 and numeric country codes. In contrast to the strategies deployed by the International Organization for Standardization, we introduced different coding logics which are represented by *ctry_ISO3onb* (ONBound historized alpha numeric codes (based on ISO 3166 ALPHA-3), 'political' coding) which accounts for historical political changes across time, and *ctry_ISOnumonb* (ONBound numeric country codes (based on ISO 3166), 'territorial' coding) which follows geographical territories and national populations of today. For example, Czech Republic was assigned the alphanumeric codes of CSK_CZE before separation from Slovakia and CZE after this separation. The numeric code for this case is "203", regardless of political changes. However, data for CSK (Czechoslovakia as a whole) are coded with a different numeric code "200".

Please note: In this section we introduce the main coding rules. Since countries change surprisingly often, we documented our decisions in detail in the ONBound_Appendix_I_Country_Codes_v4.0. Please refer to this appendix when looking for specific country code decisions.

These are the main coding rules:

1. Country coding roughly follows ISO 3166-1 alpha-3 and numeric country codes. If possible, current ISO3166-1 codes were assigned.
2. For countries that changed their name but not their territory or national population, latest ISO 3166-1 codes were assigned retrospectively, e.g. Congo changed its name from Zaire (ZR, ZAR, 180) to the Democratic Republic of the Congo. All entries are coded COD (CD, 180).
3. Alphanumeric codes (*ctry_ISO3onb*) were assigned in order to identify and flag changes in a country territory or national population, using the following logic:
 - For countries that separated during the time series and got assigned new ISO 3166-1 codes separately, old ISO3166-1 codes were assigned to the formerly unified countries before the separation. The subsequent countries were coded according to their current status. E.g. Czechoslovakia was coded CSK until 1992. From 1993 on, Czech Republic and Slovakia are coded CZE and SVK. However, some datasets include separated data for CZE and SVK between 1945 and 1992. These are assigned the codes CSK_CZE and CSK_SVK for the time period before 1993.
 - Countries that *separated* during the time series and retained official ISO 3166-1 codes during the process are assigned different codes for the former and the subsequent countries. E.g. SUN is assigned to Soviet Union while RUS only contains data for contemporary Russia – contrary to sometimes retrospectively applied codes of successor states to the former entities.
 - Countries that *unified* during the time series and retained official ISO 3166-1 codes during the separation are assigned different codes for the former and the subsequent countries. E.g. DEU_W was assigned for the Federal Republic of Germany before 1990, while DEU implies data for the unified Germany.
4. The string variable *ctry* follows the historized logic of *ctry_ISO3onb* and includes standardized country names.
5. However, numeric codes (*ctry_ISOnumb*) were assigned according to recent borders. Data referring to the Czech Republic before 1993 and after 1993 are assigned the same numeric code (203), since CSK_CZE and CZE would describe the same territory.
 - Territory that officially belongs to another greater entity got assigned the code of the greater entity + 0X, e.g. Germany-West got assigned 26701 (267 indicating Germany) and the German Democratic Republic got assigned 26702.
6. Since countries are mostly unified or separated during a year or over a process of separation that may take more than one year, we apply the codes from the first

observation following the year in which the change took place. Nevertheless, some source data forced us to code countries slightly differently. Especially within transitional phases, codes might not follow this rule. If possible, for the macro-level data, the assignment has been verified based on country codings used in other datasets, e.g. the coding of the Correlates of War datasets. Otherwise population figures have been used in order to identify whether codes should be assigned to former or current political entities.

7. If needed, we introduced special codes for certain regions (e.g. Gaza vs. West Bank, Kurdistan).
8. We did not include any changes after 2017.

2.6.6 Harmonization of Further Categorical Variables

In addition to country codes, we harmonized further categorical variables across all surveys and macro-level resources. These are: subnational regions, religious denominations, and political parties. These various sources were characterized by a substantial heterogeneity of coding schemes for the relevant data, which forbade that we employed a unified systematic standardization approach to all these variables. Rather, we identified internationally applicable code resources and tried to follow the existing coding schemes as closely as possible. If necessary to account for different levels of complexity, we introduced a hierarchical coding scheme that facilitates using data despite of the level of complexity used by original sources. Existing codes have been pooled and assigned to their respective level of complexity and to the respective codes and names. Below, we introduce the different coding schemes shortly. For more information please refer to the respective appendices.

Subnational Regions

In order to harmonize the subnational regions in the micro-level data we first had to get an overview on all existing coding schemes in the source data. The result of this is Appendix III. When developing the target codings, we conducted literature and online searches in order to find the best coding for the respective country. The online source according to which we ultimately developed the target codings can also be found in Appendix III. Whenever possible, we used official coding classifications such as NUTS-2. The variable label includes the level that has been coded (e.g. NUTS-2 or Province). For some countries we had to introduce different versions of the region variable due to the variety of coding schemes in the source data. Here, version a represents the most harmonized classification, while versions b and c classifications are more detailed. In Sweden we coded for example:

bv_region_SWEa: Background variables: Geographical Variables: Region SWE, 3 categories

bv_region_SWEb: Background variables: Geographical Variables: Region SWE, NUTS-2

bv_region_SWEc: Background variables: Geographical Variables: Region SWE, NUTS-3

We coded a source variable into the most detailed version possible. If we had enough information and the categories were clearly assignable, we coded the more detailed versions into the less detailed versions. In the case of Sweden this means that for respondents who have information on NUTS-3 level we also have information on the less detailed levels. For some countries however, the categories of different study-programs could not be harmonized. You can see in Appendix III whether we harmonized the different versions of the variable or not.

Religious Denominations

The ON_Bound coding scheme of religions and denominations does not aim at providing a universally valid classification scheme. It was designed to serve the needs of the data on religious affiliation, which appear in any data set (micro or macro) used by the ON_Bound project. The four-level classification scheme lists all present entries and assigns denominations, subgroups and individual churches to major religious (or unreligious) groups as precisely as possible.

The categories are ordered by their size, i.e. the number of members, of the respective religious groups (source: Appendix E of the Pew Research Center's report "The Changing Global Religious Landscape" (2017)). The four levels unfold the diverse religious affiliations: the first level indicates nine major groups, (1) Christianity, (2) Islam, (3) Those who are not religious, (4) Hinduism, (5) Buddhism, (6) Folk Religions (7) "Other" / "New Religions", (8) Believers without specific religious affiliation and (9) Judaism. On the second level, the largest subgroups of these major groups are recognized, also sorted by membership size (as of 2017 and as far as possible). The third level divides the subgroups further, while the fourth level is the most specific. In many cases it covers individual churches. For more information, please, see ONBound Appendix II a and b on the concept and coding scheme of religious denominations.

Political Parties

Even though, questions on political parties are asked frequently and consistently in surveys, across studies there is little agreement on party codes and party names. The ONBound project includes questions on vote choice (hypothetical prospective and retrospective vote choice), party affiliation (closeness to any party), and party membership. Because party vote and party affiliation can be used as fruitful short-cut to political positions and they can also be viewed as relevant dependent variables, we decided to take on the challenge of harmonizing political

parties. Luckily, researchers previously have already offered well-recognized harmonized coding schemes. Mainly, we follow the coding scheme by ParlGov (Döring and Manow 2019) and Party Facts (Bederke et al.). The former includes all parties winning more than 1,0% of votes or at least two seats in parliaments. The latter furthermore includes party codes for countries not contained in the ParlGov dataset. Both account for historical changes of the parties. To adopt the coding schemes, we created a Stata-ado that refers to various spellings of each party to detect parties within the datasets. The results of this ado are saved in the ONBound_Appendx_V.b_Coding Scheme_Party Codes, which is later used to perform the recodes within the topic file of 5400. Unfortunately, this mechanism only works for labeled data. Numeric codes had to be extracted from codebooks and placed into the ONBound_Appendx_V.b_Coding Scheme_Party Codes. However, this table can be used as repository for party harmonization. It includes all spellings and links our data with ParlGov Party Facts. In that way, our data can be easily linked to further party related information, e.g. party families, ideological orientation and vote shares. Furthermore, the gathered spellings help to link further micro-level datasets on parties with our repository.

2.7 Missing Data

The micro-level data includes eight different missing codes:

- (-1) Variable not available for this country in this survey-wave
- (-2) Country specific variable not applicable for this country
- (-3) Not asked in this survey-wave

- (-4) Not applicable
- (-5) Missing; Various / unspecified
- (-6) Study specific missings / unable to code into scheme

- (-7) Other
- (-8) Don't know, Can't choose
- (-9) No answer, refused

- (-10) No party preference / Did not vote / invalid ballot
- (-11) Independent candidate

Codes (-1) to (-3) account for the different kinds of missing values that occur because variables are not available in all study-waves and in some studies not available for all countries. Code (-1)

means that a variable is included in the study_wave at stake but was not asked in the country. (-2) applies for country-specific variables (e.g. *bv_region*) and has been assigned to all cases not belonging to the relevant country. For example, all countries which are not Austria are coded to -2 in *bv_region_AUT*. Code (-3) indicates that a variable is not available for the whole study_wave.

Codes (-4) to (-9) stem from the original data. (-4) refers to questions that were not asked to certain respondents (irrespective of their country). In most cases this is due to filters in questionnaires and the value label gives further information on the excluded respondents. (-5) is the code for all missing cases that could not be further specified, e.g. system missing cases in some studies. Like code (-4), code (-6) can mean different things in different variables. It accounts for study specific missing codes (e.g. “Do not understand the question” or “None of these”) or indicates that we were not able to code a certain category into the given scheme (e.g. regions, religious denominations). This missing code is always labelled so that it becomes clear, what it means for the current variable. Values (-7) indicates an undefined “Other” category, while (-8) represents the classical “Don’t know” and (-9) a “No answer, refused” category.

For variables related to party choices or vote choices, we had to introduce further missing codes and expand the meaning of others. (-10) indicates answers by respondents without party preferences (*pol_party2*) or respondents who did or would cast an invalid ballot (*pol_vote1* and *pol_vote2*). (-11) refers to answers concerning independent candidates. (-7) includes all other party mentions that were not coded or we could not

3. ONBound Harmonization in Practice

This chapter introduces the technical implementation of our harmonization strategies described in chapter 1 and 2. It introduces the ONBound Harmonization Wizard (<http://onbound.gesis.org/wizard>) and gives an overview on the implementation. The following Chapter 4 explains the technical realization, enables users to apply changes to their ONBound dataset, and to update dataset versions.

3.1 Overview

The ONBound Harmonization Syntax Package has been developed for Stata for Windows (version 14.0 or higher) and SPSS (version 17 or higher). Due to specific syntax commands not being available in older versions of the statistics systems, the syntax cannot be expected to run properly with such older versions. The Syntax Package offers harmonization- and merge-syntax for all micro- and macro-level datasets the ONBound project team decided to include.

Please note: Before running the main.do/main.sps-files users have to perform some steps manually (as described in Chapter 4).

Furthermore, users need to install further packages and software. Stata users who want to include Afro Barometer, Asia Barometer or PewGAP to their dataset need to download and install R (<https://cran.r-project.org/bin/windows/base/>). For SPSS users the Essentials for Python package that comes with SPSS must be installed (further information can be found here: <https://www.spss-tutorials.com/python-for-spss-installing-and-testing/>).

3.2 The ONBound Harmonization Wizard

The ONBound Harmonization Wizard is an online tool that enables users to select source datasets and the appropriate syntax elements by two dimensions and across the micro- and the macro-level: first, blocks of variables nested in topics for the micro-level data, topics clustered in datasets on the macro-level, and a selection of countries and time frames for their research interests. In an intermediate step, the wizard shows users the availability of their selection within our data. Based on the selections users make, the wizard creates a customized ONBound Syntax Package – both for SPSS and Stata users. The packages can be downloaded as .zip-folders. In addition to this download, the ONBound Harmonization Wizard creates a table of links to the original datasets based on the selection. These original datasets contain variables/topics, country and year-points according to the researcher's choice.

To use the ONBound Harmonization Wizard, users need to download the .zip-folder *and* follow the links to the original datasets as indicated on the website, (possibly) register and agree to the terms of use, download the original datasets and save them in the folder “2_source” of the unzipped ONBound-folder (please also see descriptions in chapter 3.3 and instructions at chapter 4). After applying minor necessary changes in the respective main syntax file (see chapter 4.3), the ONBound Syntax Package is ready to use. After running the main files, a customized dataset with all variables, datasets, countries and years selected on the webpage is compiled on the users’ computer and ready to be analyzed.

If users wish to include further data from the ONBound Repository beyond those initially selected, the ONBound Harmonization Wizard is the first address to do so. By simply selecting a new combination of variables/topics, datasets, countries and years is created and can be downloaded. With the additional source datasets also downloaded into the previous folder structure, executing of the new, extended syntax package will recreate the extended merged file.

3.3 Structure (zip.-Folder)

All necessary documents are included in the .zip-file that can be downloaded from our website (<http://onbound.gesis.org/wizard>). This .zip-Folder contains four main folders and one main.do/main.sps-file. It is available for Stata and SPSS.

These are the main folders and our main-files:

- 1_documentation*
- 2_source*
- 3_harmonization*
- 4_target*
- main.do (do-file for Stata-Users)
- main.sps (Syntaxfile for SPSS-Users)

1_documentation contains all additional information documents as described in chapter 2.1:

- ONBound_Documentation_Micro_v4.0.xlsx
- ONBound_Documentation_Macro_v4.0.xlsx
- ONBound_Micro-Level-Data_Overview_v4.0.xlsx
- ONBound_List_of_source_datasets.xlsx
- ONBound_List_of_source_datasets_customized.xlsx
- ONBound_Appendix_I_Country_Codes_v4.0.pdf

- ONBound_Appendix_II.a_Concept_Classification Religions.pdf
- ONBound_Appendix_II.b_Coding Scheme_Classification Religions.xlsx
- ONBound_Appendix_III_Subnational Regions.pdf
- ONBound_Appendix_IV_Micro-Level-Data_Weightsv4.0.xlsx
- ONBound_Appendix_V.a_Concept_Party Codes.pdf
- ONBound_Appendix_V.b_Coding Scheme_Party Codes.xlsx

2_sources is an empty folder. Here, users save all downloaded original source datasets.

3_harmonization contains all syntax files compiled and provided by ONBound. Here, all files necessary to perform the harmonization are stored.

do_files (contains all dofiles related to harmonizational work on single datasets)

dofiles_macro

dofiles_micro

help_files (additional files for ONBound harmonization)

adofiles

labels

sav_to_dta_with_R (for Stata-Users only)

translationuni_files (for Stata-Users only)

... and additional help files

prepared_datasets starts as an empty folder. After running the main.do/main.sps this folder contains all fragments and single datasets before the final merge/append routine is initiated.

4_target also starts as an empty folder. After running the main.do/main.sps, this folder contains the customized final dataset, based on the selection made by users using the ONBound Harmonization Wizard. Additionally, it contains single datasets for micro-level data and macro-level data separately.

4. How to use ONBound Harmonization Syntax

Chapter 4 guides our users through the necessary steps to use the ONBound Package. We advise users to carefully follow our instructions and apply the changes suggested here.

4.1 BEFORE using the ONBound Syntax Package (Stata-Users)

Some steps are required before using the ONBound Package.

- I. Unzip the folder downloaded from <http://onbound.gesis.org/wizard> and place the unzipped folder in the desired folder on your computer.
- II. Download all datasets you intend to harmonize from the ONBound_List_of_source_datasets.xlsx or as indicated by the ONBound Harmonization Wizard Website.
- III. Please save the downloaded files in the empty folder “[...] /2_sources”.

Please make sure to use original filenames as provided by the original data providers.

- IV. Apply necessary changes to the original filenames if necessary. The cases and special steps that need to be taken are described under 4.3.

There are two occasions that require the application of changes:

- A) Some data-files need name changes due to date-tags.
- B) Some datasets are only available in the .sav-dataset format.

- V. Alter the folder path in the “main.do” file:
 - a. Open the file “main.do”
 - b. In the 56th line of code, modify the path “...\ONBound” to match the folder path where you placed the unzipped folder
This line has a comment, “/***CHANGE THIS PATH TO MATCH THE LOCATION OF THE ONBOUND-FOLDER ON YOUR MACHINE! ***/”
It should look like this now:

```
cd “[path to the ONBound-folder]”
```
- VI. Save the changes.

The ONBound Syntax Package for Stata is now ready to use.

4.2 BEFORE using the ONBound Syntax Package (SPSS-Users)

Some steps are required before using the ONBound Package.

- I. Unzip the folder downloaded from <http://onbound.gesis.org/wizard> and place the unzipped folder in the desired folder on your computer.
- II. Download all datasets you intend to harmonize from the ONBound_List_of_source_datasets.xlsx or as indicated by the ONBound Harmonization Wizard Website.
- III. Please save the downloaded files in the empty folder "[...]/2_sources".

Please make sure to use original filenames as provided by the original data providers.

- IV. Apply necessary changes to the original data-file names if necessary. The cases and special steps that need to be taken are described under 4.3.

There are two occasions that require the application of changes:

- A) Some data files need name changes due to date-tags.
- B) Some data sets come in .rar-format and need need to be unpacked

- V. Alter the folder path in the "main.do" file:
 - a. Open the file "main.sps"
 - b. In the 55th line of code, modify the path ".../ONBound" to match the folder path where you placed the unzipped folder
It should look like this now:

```
spss.Submit("[path to the ONBound-folder]".''')
```
 - c. In the 57th line of code, modify the path ".../ONBound" to match the folder path where you placed the unzipped folder
It should look like this now:

```
wd = ".../ONBound/"
```

Please note: The path in line 53 should not end with "/" whereas the path in line 55 has to end with "/"!

- VI. Save the changes.

The ONBound Syntax Package for SPSS is now ready to use.

4.3 Special Remarks

The following changes are only needed for few datasets. These datasets are: UN Demographics Database, Asian Barometer, PewGAP, Asiabarometer and Afrobarometer. If none of those mentioned above are included in your customized ONBound dataset, you can skip this chapter.

4.3.1 (A) Necessary name changes (Stata and SPSS users)

Special Steps before running the main.do and main.sps (renaming of datasets)

Refers to: **UN Demographics Database** (Religious Affiliation, macro-level)

In case you decided to include any of the datasets provided by the UN Demographics Database / United Nations Statistics Division you must:

1. unzip the ZIP-folder you downloaded.
2. save the .csv-file into the “[...]/2_source” folder
3. rename the .csv-file as follows:

Religious Affiliation: UNdata_Export_Religion.csv

4.3.2 (B) Unpacking files (Stata and SPSS users)

Special Steps before running the main.do and main.sps (unpacking files)

Refers to: **Asian Barometer**

In case you decided to include waves of the Asian Barometer, please unpack the .rar-folders and save the .sav-files to the folder “2_source”.

4.3.3 (C) Unpacking files (Stata-users)

Special Steps before running the main.do (datasets only available in .sav)

Refers to: **PewGAP, Asiabarometer** and **Afrobarometer**-datasets (micro-level)

These datasets are only available in .sav-format. We included an R-script to our harmonization routine which converts the respective .sav-files to .csv-files. To run the script, you need to take these further steps:

1. R needs to be installed on your device. Please download and install R on your device. You can download it for free here: <https://cran.r-project.org/>.
2. Alter the path in the “main.do” folder:

a. Open the file “main.do”

b. In the 57th line of code, modify the path `global Rterm_path `"...\\R.exe"'` to match the folder path where you stored R

This line has a comment: `"/***CHANGE THIS PATH TO MATCH THE LOCATION OF R ON YOUR MACHINE - IF NECESSARY! (refer to USERS GUIDE)***/"`

It should look like this now:

```
global Rterm_path `[path to program-folder]/Program Files/R/R-4.0.0/bin/R.exe`
```

3. Please erase the * in line 82-84

4.4 Saving Storage on your Device

During the process of harmonization, several datasets and other files are going to be created. To not occupy too much space, at the end of the `main.do/.sps` all unnecessary files are deleted. This section erases all temporary files within the macro-level dataset preparation (1), temporary Unicode translation files (2), all fragments created of topic sections of micro-level datasets (3), and all prepared fragments including recodes (4). If you are not satisfied with our definition of “necessary files”, you may want to comment out the respective lines of code using * (Stata) or # (SPSS using Python) for every line or use `/*` at the beginning and `*/` at the end of the code you do not want to execute (Stata only).

Furthermore, we included code that deletes all prepared macro-level datasets (5) and compiled micro-level topic datasets (6). These lines of code are commented out. If you wish to delete these datasets to save more space on your device, erase * (Stata) or # (SPSS) in the respective lines of codes. Numbers 1-6 mark the parts responding to these procedures in the `main.do/.sps`.

4.5 USING the ONBound Syntax Package for Harmonization

After you applied the changes as described in 4.1 or 4.2 and 4.3, and possibly applied additional changes described in 4.4 you can run the `main.do` – or `main.sps` respectively. This main file creates the harmonized dataset with your set presets applied on our website <http://onbound.gesis.org/wizard>. Just press “run”.

The final customized dataset will be saved in the “4_target” folder.

Please note: Since ONBound intends to harmonize a vast array of datasets into a single dataset, the actual process might take you some minutes – or rather hours or even a whole night. Depending on your selection, this might even produce a dataset too large to handle. Please

reconsider your selection made at the ONBound Harmonization Wizard and narrow down the variable selection.

4.6 A deeper insight to the ONBound Universe

With the initial download of the ONBound .zip-folder you gain access to the whole ONBound universe. This implies the customized main file, but also a complete documentation and further resources to harmonize other resources or to modify your personal ONBound harmonized dataset. This section offers tips and tricks as well as a deeper insight into the structure. Please be advised that the ONBound universe entails black holes if coders are not careful.

In this chapter, we are giving a glimpse at the internal ONBound structure. The following exemplifies steps necessary within the network of syntax files communicating with each other. This is the simplest modification possible. If you wish to modify the syntax further, and add more variables, topics or original datasets, we advise to carefully retrace the files and their commands and how they relate to each other. Chapter 2 explained the folder structure; this chapter digs deeper into the ONBound code. A first insight in the function of the network is shown here.

In ONBound, the main.do/main.sps is the centerpiece of the ONBound Syntax Package. But the actual work is done within the syntax-files. Together with special task files and documents, they are hidden in the folder "3_harmonization". Let's recall the main structure for a moment.

Level 1 of ONBound:

- 1_documentation*
- 2_source*
- 3_harmonization*
- 4_target*
- main.do (do-file for Stata-Users)
- main.sps (Syntaxfile for SPSS-Users)

Below this level, within the chapter 3_harmonization the work is done:

Level 2, 3, 4 (below 3_harmonization)

Level 2	Level 3	Level4
<i>do_files</i>		
	<i>dofiles_macro</i>	
		dofile_d_DES.do.do /.sps
		dofile_d_RNCw.do /.sps
		<u>dofile_d `datasets_macro'.do /.sps</u>

dofiles_micro

dofile_d_1000.do /.sps
dofile_d_3400.do /.sps
dofile_d `selectedtopics'.do /.sps
dofile_d ACEB_2000.do /.sps
dofile_d ISSP_2013.do /.sps
dofile_d `datasets_micro'.do /.sps

help_files

adofiles (contains files for country code harmonization via labels.xlsx)
labels (contains all files with label-information)
sav_to_dta_with_R (contains special files for R-routine)
tranlationuni_files (contains Unicode-translation for Stata datasets)
[other specific .do /.sps-files for special topics / special needs]

prepared_datasets

prepared_files_macro

(contains all produced macro-level datasets before final merges/append)

prepared_files_micro

([contains all produced micro-level datasets before final merges/append)

anchor.dta (STATA only)

The following table guides through the important syntax files and how they relate to each other:

Locals	Code	Examples
Main.do/main.sps		
GENERAL SECTION		
<ul style="list-style-type: none">• Sets working paths• Runs other general preparations (e.g. R-routine, ados etc.)		
* FILTER SECTION *		
<ul style="list-style-type: none">• defines all filters (as written by ONBound Harmonization Wizard)		
dataset_micro	Stata: local dataset_micro dataset1 dataset2 dataset3 SPSS datasets_micro = [„dataset1“,“dataset2“,“dataset3“]	ISSP_2013 EB_84_1_2015 LatinoB_2010
dataset_macro	Stata: local dataset_macro dataset1 dataset2 dataset3 SPSS: datasets_macro = [„dataset1“,“dataset2“,“dataset3“]	RNCw WDI IMPIC
selectedtopics	Stata: local selectedtopics topic1 topic2 topic3 SPSS: selectedtopics = [„topic1“,“topic2“,“topic3“]	1000 1200 5700
selectedcountries	Stata: local selectedcountries country1 country2 country3	DEU

	SPSS: selectedcountries = [„country1“,“country2“,“country3“]	FRA DNK
obltopics	Stata: local obltopics 6100 6200 SPSS: obltopics =[“6100”,“6200”]	Notes: refers to protocol and background variables
selectedyears	local selectedyears year1 year2 year3 selectedyears = [„year1“,“year2“,“year3“]	1945-2019

*** PREPARATION SECTION ***

- creates empty folders on your device within the ONBound structure
- runs syntax files by **dataset_micro**
[folder: 3_harmonization/do_files/dofiles_micro/dofile_d_dataset_micro.do/.sps]

dofile_dataset_`dataset_micro'.do/.sps

- checks if dataset has been downloaded by users

if yes:

- runs general preparations and defines anchor variables (version, download, onbound_wave, study_wave, caseid, etc.)
- partitioning of dataset to fit topics
- saves fragmented datasets with obligatory variables, administrative variables and substantial variables sorted by topics

[saved: “**help_`selectedtopic’_`dataset_micro’.dta/.sav**”

in folder: 3_harmonization/prepared_datasets/prepared_files_micro/fragments]

e.g.: help_1300_ISSP_2013.dta/.sav

- runs syntax files by **selectedtopics**
[folder: 3_harmonization/do_files/dofiles_micro/dofile_d_dataset_micro.do/.sps]

dofile_d_`selectedtopics'.do/.sps

- defines variables of this topic
- BY dataset_micro:
- checks if help_`selectedtopics’_`dataset_micro’ exists

if yes:

- creates all variables that belong to the topic
- replaces target variables with original variables
- performs recodes
- saves topic-fragment dataset with obligatory variables, administrative variables and substantial variables, sorted by topics and recoded to target variable codes

[saved: “**dataset_`selectedtopic’_`dataset_micro’.dta/.sav**”

in folder: 3_harmonization/prepared_datasets/prepared_files_micro/`selectedtopic’]

e.g.: “dataset_1300_ISSP_2013.dta/.sav”

- appends all **datasets_micro**
- runs syntax files to label the appended data
[folder: 3_harmonization/help_files/labels/LABEL_`selectedtopic’.do/.sps]

LABEL_`selectedtopic’.do/.sps

- labels target dataset

LABEL_oblvars.do/.sps

- labels obligatory variables in target dataset

- applies filter: **selectedcountries** and **selectedyears** [drops all data not wanted]

[saved: “**dataset_`selectedtopic`_all.dta/.sav**”

in folder: 3_harmonization/prepared_datasets/prepared_files_micro/all]

e.g.: “dataset_1300_all.dta/.sav”

- runs syntax files by **obltopics**

[folder: 3_harmonization/do_files/dofiles_micro/dofile_d_`dataset_micro`.do/.sps]

dofile_d_6100.do/.sps

dofile_d_6200.do/.sps

- analog to dofile_d_`selectedtopics`.do/.sps

- runs syntax files by **dataset_macro**

[folder: 3_harmonization/do_files/dofiles_macro/dofile_d_`dataset_macro`.do/.sps]

dofile_d_`dataset_macro`.do/.sps

- checks if dataset has been downloaded by users

if yes:

- (Stata only) translates dataset to Unicode
[runs syntax file 3_harmonization/help_files/dofile_translateuni.do]
- (Stata only) runs automatic country code recodings using the
countriesonbound_macro.ado
- (SPSS only) harmonization of country codes
- generates anchor variable (ctry_year)
- performs recodes, expansions and reshaping
- applies filter: **selectedcountries** and **selectedyears** [drops all data not wanted]

[saved: “**dataset_`selectedtopic`_all.dta/.sav**”

in folder: 3_harmonization/prepared_datasets/prepared_files_macro]

e.g.: “dataset_RNCw.dta/.sav”

* MERGE / APPEND SECTION *

- merges all micro-level datasets, respectively micro-level topic datasets
[“dataset_`selectedtopic`_all.dta/.sav”
in folder: 3_harmonization/prepared_datasets/prepared_files_micro/all]
- saves micro-level dataset “dataset_merged_micro.dta/.sav”
["3_harmonization/prepared_datasets/prepared_files_micro"]
- merges all macro-level datasets
[“dataset_`selectedtopic`_all.dta/.sav”
in folder: 3_harmonization/prepared_datasets/prepared_files_macro]
- saves macro-level dataset “dataset_merged_macro.dta/.sav”
["3_harmonization/prepared_datasets/prepared_files_macro"]
- merges final micro-level and macro-level dataset
[dataset_merged_micro.dta/.sav + dataset_merged_macro.dta/.sav]

* CLEANING WITHIN FINAL DATASET*

- orders variables for the final dataset

* SAFE *

- saves final dataset
[folder: “4_target”]

* CLEANING OF REPOSITORY *

- (1) deletes temporary files of macro-level datasets
- (2) deletes temporary files of Unicode translation (macro-level, Stata only)
- (3) deletes temporary help-files consisting of fragments of original micro-level datasets
- (4) deletes temporary prepared help-files consisting of fragments of original micro-level datasets
- (5) deletes single prepared macro-level datasets (optional)
- (6) deletes single prepared topic files on the micro-level (optional)

4.7 Updating ONBound to newer versions

In general, if the original datasets that were used when we created the ONBound universe are updated in their source repositories, users should employ the possibility to request these earlier versions of these datasets. The `ONBound_List_of_source_datasets.xlsx` indicates the versions used by ONBound. Alternatively, users can try to update datasets to the newer versions within the ONBound Syntax Package. To update the ONBound Syntax Package requires only few, but vigilant coding routines.

To update datasets versions, two things are important:

1. Did the name of the new version change when downloading the updated dataset?
2. What did internally change in the new versions of the datasets, i.e. regarding variable names and contents? Deviant codings might corrupt the harmonization processes and make the harmonized variables useless.

While both questions are important to both levels of datasets, they require different procedures for micro- and macro-level data. This section therefore offers two different approaches.

4.7.1 Micro-Level Datasets

1. Save the updated version in the “2_source”-folder
2. Open the dataset file of the updated dataset:
`dofile_dataset_`micro_datasets'.do/.sps`
 [Folder: “3_harmonization/do_files/dofiles_micro/]
3. Update the name of the updated and downloaded version:

a. Stata:

- Check for existence of file:
`capture confirm file `2_source/[...]'`
 (Data format is not relevant here. If the original data comes in a zip-folder, the name of the zip-folder needs to be in this line)
- *if* the original dataset is downloaded with a zip-folder:
`unzipfile `2_source/[...]'.zip"`

- **Load dataset**
`insheet using "2_source/[...] .xlsx"`
OR `import excel "2_source/[...] .xlsx" , firstrow clear`
OR `use "2_source/[...] .dta" , clear`
OR `import delimited "2_source/[...] .csv" , clear`

b. SPSS

- **Load dataset**
`file=wd+"2_source/BBBB.csv"`
 - *if* the original dataset is downloaded with a zip-folder:
`file=wd+"2_source/[...] .zip"`
insert the zip-file name
`file2=wd+"2_source/[...] .sav"`
insert the dataset-file name (if different)
4. Change variable names according to version updates by original datasets, if necessary. Check which topic files are affected by the changes of original variables.
 5. Save changes
 6. Open the topic file were variables have been changed:
`dofile_d_`selectedtopics'.do/.sps`
[Folder: "3_harmonization/do_files/dofiles_micro/"]
 7. Find the section referring to the original dataset that changed
 8. Change variable recodes according the updated made
 9. Save changes
 10. Re-run main.do/main.sps (If you wish to save a new version of the final dataset, change the dataset name in the very last line of code within the main.do/.sps)

4.7.2 Macro-Level Datasets

1. Save the updated version in the "2_source"-folder
2. Open the dataset file of the updated dataset:
`dofile_d_`macro_datasets'.do/.sps`
[Folder: "3_harmonization/do_files/dofiles_macro/"]
3. Update the name of the updated and downloaded version:
 - a. Stata:
 - Check for existence of file:
`capture confirm file "2_source/[...]"`
(Data format is not relevant here. If the original data comes in a zip-folder, the name of the zip-folder needs to be in this line)
 - *if* the original dataset is downloaded with a zip-folder:
`unzipfile "2_source/[...] .zip"`
 - **Load dataset**
`insheet using "2_source/[...] .xlsx"`

```
OR import excel "2_source/[...] .xlsx" , firstrow clear  
OR use "2_source/[...] .dta" , clear  
OR import delimited "2_source/[...] .csv" , clear
```

b. SPSS

- Load dataset

```
file=wd+"2_source/BBBB.csv"
```

- *if* the original dataset is downloaded with a zip-folder:

```
file=wd+"2_source/[...] .zip"
```

insert the zip-file name

```
file2=wd+"2_source/[...] .sav"
```

insert the dataset-file name (if different)

4. Change the do-/syntax-file according to the changes indicated by the original datasets throughout the files
5. Save changes
6. Re-run main.do/main.sps (If you wish to save a new version of the final dataset, change the dataset name in the very last line of code within the main.do/.sps)

Have fun with the data, don't forget to cite us and please send any questions and errata you might have or find within the ONBound universe.