SDR2 MASTER File Overview

1. Introduction

The MASTER file contains harmonized data from 23 international survey projects (Table 1) for 4,402,489 respondents interviewed across 3,329 national surveys that, taken together, span the period 1966-2017 and 156 countries. The MASTER file provides technical and substantive target variables, harmonization controls, and a subset of the survey quality control variables. This document describes the harmonized technical variables and data record quality controls.

Substantive target variables and accompanying harmonization controls are documented in their respective General Variable Reports, GVRs. Here we briefly note that substantive target variables measure respondents' demographics (age, gender, education, civil status, income, household size, place of residence), political behaviors (voting, engaging in demonstrations and boycotts, signing petitions), political attitudes and opinions (interest in politics, trust in government, parliament, the legal system and political parties, and satisfaction with democracy), social capital (membership in political parties, in environmental and religious organizations and trade unions, interpersonal trust, discussing politics) and wellbeing (happiness, life satisfaction and health).

2. Technical variables in the SDR2 MASTER file

2.a Survey projects (T_PROJECT_NAME and T_PROJECT_WAVE)

In the MASTER file, survey projects are identified by the variable T_PROJECT_NAME. The abbreviations we employ are either officially recognized ones or constructed by us from the first letters of the projects' full names.

Most of the survey projects whose data we harmonized in SDR2 are multi-wave (for the details of the survey projects included in the SDR2 MASTER file, see Table 1). Wave numbers are captured by the variable T_PROJECT_WAVE. For example, in the MASTER file we have six World Values Survey waves, which are numbered from 1 to 6, in short WVS/1-6. In the naming of waves, we always follow surveys' conventions. This is why we have ISJP/1991 and ISJP/1996 rather than ISJP/1 and ISJP/2.

In the case of the two Political Action studies, the first, "An Eight Nation Study," conceptualized as a standalone project, was soon followed by a second, yet independent, study called Political Action II. Therefore, we use PA1 and PA2 as the T_PROJECT_NAME entries, where the numbers 1 and 2 are parts of the abbreviations rather than representing waves.

In the case of single-wave surveys, T_PROJECT_WAVE is left as an empty string. Note: T_PROJECT_WAVE is the only variable in the MASTER file which can take empty values; all other variables use missing data schema instead (see SDR2_Missing_Codes_Schema).

Table 1. International survey projects in SDR2

T_PROJECT_NAME	Official project name	Waves
ABS	Asian Barometer Survey	4
AFB	Afrobarometer	6
AMB	AmericasBarometer	7
ARB	Arab Barometer	4
ASES	Asia Europe Survey	1
CB	Caucasus Barometer	7
CDCEE	Consolidation of Democracy in Central and Eastern Europe	2
CNEP	Comparative National Elections Project	2
EB	Eurobarometer	48
EQLS	European Quality of Life Survey	3
ESS	European Social Survey	8
EVS	European Values Study	4
ISJP	International Social Justice Project	2
ISSP	International Social Survey Programme	31
LB	Latinobarómetro	19
LITS	Life in Transition Survey	3
NBB	New Baltic Barometer	.6
NEB	New Europe Barometer	7
PA1	Political Action (An Eight Nation Study)	1
PA2	Political Action II	1
PPE7N	Political Participation and Equality in Seven Nations	1
VPCPCE	Values and Political Change in Postcommunist Europe	1
WVS	World Values Survey	6

2.b Source data files (T_DATAFILE and T_ID)

We downloaded for harmonization 215 source data files that store a total of 174 project-waves. The difference between source data files and project-waves numbers stems from the structure of the source files. While generally a source data file includes all national surveys of a given project-wave, sometimes repositories provide separate files for each national survey of a project-wave, or one single data file for multiple project-waves.

At the time we selected source data for harmonization (January-February 2018), the source files of the international survey projects were freely available online in the public domain. The one exception was the Asian Barometer, which was distributed on demand.

Due to (1) the large number of source documentation and data files of interest to SDR, (2) sources being stored in few repositories, including GESIS, ICPSR, and the UK Data Service, and (3) our intention to check for updated versions, we automated downloads. The SDR team's data science specialist developed a script (`wpc`, for "webpage crawler"), to parse the target web pages and

retrieve the required files. Thanks to this script, the whole process was fully automated given the web link to the online data storage place.

We introduce the variable T_DATAFILE, which in a schematic way indicates the source data file. WVS_1_6 corresponds to WVS/1-6; PA1 and PA2 correspond to their namesakes PA1 and PA2; and ABS_1 corresponds to ABS/1. Note, however, that ISJP_1_2 identifies the integrated 2-wave ISJP file (ISJP/1991 and ISJP/1996, since we avoid putting two years in a file name). Note also that for four project/waves (ABS/4, AMB/2016, CNEP/3-4) the only available level of data aggregation was a country, so we have an additional country term in T_DATAFILE, e.g. ABS_4_TW for ABS/4 Taiwan source data (for countries' naming convention used in SDR 2 see Section "Geographical coverage").

2.c Survey samples

In SDR2, we strive to harmonize ex-post samples that are nationally representative. Thus, in the Master file we include only cross-sectional source samples when the source data provide both cross-sectional and panel samples. On the other hand, if only panel samples are available, we take them. This situation pertains to: (1) Australia in ISSP/1986-2002 (perhaps also other waves, but we cannot tell from the source documentation), where ISSP modules were carried out along with panel modules of National Social Science Survey; (2) Germany in ISSP/1995, 1997, where – similarly – ISSP modules were carried out along with ALLBUS panel studies; and (3) Poland in ISSP/2014, 2015, where the same module of ISSP was carried out on a panel sample (though the available documentation does not provide information about the master study).

If a sample consists of both pre-election and post-election samples, the latter usually being a panel subsample driven from the pre-election sample, we include only the pre-election sample in the MASTER file. Table 2 presents the complete list of samples excluded from the SDR2 MASTER file, due to being supplementary panel or post-election samples.

Table 2. Panel samples excluded from the SDR2 MASTER file (complete list)

Project/waves	**Samples**	**Remarks**
CNEP/3	Mexico	Post-election panel sample
CNEP/4	Dominican Republic	Post-election panel sample
PA2	Germany, Netherlands, USA	Panel samples
VPCPCE	Russian Federation	Post-election panel sample

We had additional considerations when deciding whether to include or not samples in the MASTER file. First, for the United Kingdom, we include source surveys conducted on a sample from Great Britain only (i.e. that lack a sample from North Ireland). Since Northern Ireland covers as little as three percent of the UK population, we decided to take the following source data where the UK surveys lack North Ireland samples: ASES, EB/0.2, ISJP/1991, ISSP/1985-1988, 1992, 1995-1997, 2003-2007, 2009-2015, LITS/2, PA1, WVS/3, 5.

For the same reasons (i.e. the tradeoff between full representativeness and data loss) we include ISSP/1990-1994, 2000 for Israel, although the samples cover only the Jewish population of Israel, but not the Arab population. Note: Due to lack of adequate documentation, we cannot assess the situation in ISSP/1989.

We exclude from the MASTER file so-called "orphaned" samples, i.e. samples that cover only part of a country and where the missing part(s) are too large, e.g. Flanders in Belgium without a corresponding sample for Wallonia. See Table 3 for the complete list of the "orphaned" samples.

Table 3. Complete list of "orphaned" samples excluded from the SDR2 MASTER file

Project/waves	**Sample**	**Remarks**
CDCEE/1	East Germany	Missing West Germany
CDCEE/1	Russia (Krasnoyarsk only)	Missing rest of Russian Federation
ISJP/1996	East Germany	Missing West Germany
ISSP/2002, 2004-2005, 2007-2010	Belgium (Flanders only)	Missing Belgium (Wallonia)
WVS/3 (1998)	Federation of Bosnia and Herzegovina	Missing Republika Srpska

Besides samples enumerated in Table 3, we excluded some other sets for different reasons presented in Table 4.

Table 4. Additional samples excluded from the SDR 2 MASTER file and reasons for doing so

Project/waves	**Samples**	**Remarks**
NEB/5		56 respondents not assigned to any national sample
NEB/6-7	Estonia, Latvia, Lithuania	Samples already included in NBB/5-6
WVS/3	Colombia (sample size: 3029)	One of two samples available in this wave, shorter fieldwork, fewer variables
WVS/3	Mexico (sample size: 854)	One of two samples available in this wave, undocumented

We end these considerations noting that, according to the available source documentation, some samples included in the MASTER file have known limited representativeness of their target populations. Notable examples are presented in Table 5.

We store information about differentiation in population coverage in the source samples in F_SAMPLECOVER_SVY variable in the PLUG_SURVEY file (see SDR2 PLUG_COUNTRY File Overview for its description).

Table 5. Examples of surveys where the available source documentation admits problems with sample's representativeness

Survey	**Excerpts from the source documentation and other observations**
ASES China	[22324-0001-Codebook.pdf] *Survey area: 8 cities: Beijing, Shanghai, Guangzhou, Chongqi, Xian, Nanjing, Dalian, Qingtao*
ASES Indonesia	[22324-0001-Codebook.pdf] *Sampling in Urban areas*; variable "V0333" captures 5 provinces out of 34
LB/1995 Brazil	[F00002256-Latinobarometro_Codebook_Brasil_1995.pdf] *Universe: National Population / Representative of 33% of the population*
LB/1996 Bolivia	[F00002212-Latinobarometro_Codebook_Bolivia_1996.pdf] *Representative of 32% of the population*
LB/1996 Brazil	[F00002214-Latinobarometro_Codebook_Brasil_1996.pdf] *Universe: National Population / Representative of 33% of the population*
LB/1996 Colombia	[F00002216-Latinobarometro_Codebook_Colombia_1996.pdf] *Representative of 25% of the population*
LB/1997 Bolivia	[F00002176-Latinobarometro_Codebook_Bolivia_1997.pdf] *Representative of 32% of the population*
LB/1997 Brazil	[F00002178-Latinobarometro_Codebook_Brazil_1997.pdf] *Universe: National Population / Representative of 31.8% of the population*
LB/1997 Colombia	[F00002180-Latinobarometro_Codebook_Colombia_1997.pdf] *Representative of 25% of the population*
LB/1998 Bolivia	[F00002138-Latinobarometro_Codebook_Bolivia_1998.pdf] *Representative of 35% of the population*
LB/1998 Brazil	[F00002140-Latinobarometro_Codebook_Brasil_1998.pdf] *Universe: National Population / Representative of 11.58% of the urban population*
LB/2000 Brazil	[F00002106-Latinobarometro_Codebook_Brasil_2000.pdf] *Universe: National Population / Representative of 11.58% of the urban population*
PPE7N India	[07768-0006-Codebook.pdf] only four states [out of 29] were sampled in India: *Andhra Pradesh, Kujarat, Uttar Pradesh, and West Bengal*
PPE7N Nigeria	[07768-0006-Codebook.pdf] *In Nigeria, the situation is even more complicated. While we had originally planned a study of the entire nation, severe political tensions began to escalate rapidly as we entered the period of our actual field work.*
PPE7N Yugoslavia	[07768-0006-Codebook.pdf] *the study was conducted in only four of the six republics of the Yugoslavian Federated Nation. The Republics were Croatia, Macedonia, Serbia, and Slovenia.*
WVS/2 China	variable "X048WVS" captures 7 out of 22 provinces, 3 out of 4 municipalities, and 1 out of 5 autonomous regions (Xinjiang)
WVS/4 Indonesia	[F00002500-WV4_Results_Indonesia_2001_v_2015_04_18.pdf] *There were some limitations in the sample*; variable "X048WVS" captures only 5 out of 34 provinces
WVS/5 Indonesia	[F00001324-WVS_Sample_Design_Indonesia_2006.pdf] *This year, we would like carry on WVS Survey from June to November 2006 for 2 000 respondents within 10 provinces*; variable "X048WVS" captures only 10 out of 34 provinces

2.d Geographical coverage (T_COUNTRY_L1U and T_COUNTRY_L2U)

The MASTER file provides two technical variables related to a sample's geographical identification, T_COUNTRY_L1U and T_COUNTRY_L2U, where L1U and L2U stand for level-1 and level-2 units. There are two reasons for distinguishing sample levels. The first are geographic considerations. The second reason is related to data weighting.

T_COUNTRY_L1U identifies the smallest *country-like* sample unit provided in the source data. This usually corresponds to a national sample, e.g. Poland. Sometimes, the sample unit covers (i) only part of a country, e.g. East Germany, (ii) an ethnic group within a country, e.g. Arabs in Israel, (iii) an area larger than a country, e.g. a whole island of Cyprus (which we know is divided into two de facto independent countries, Republic of Cyprus and Turkish Cypriote Community), or (iv) a formerly united country, which at the time of data gathering or/and for the sake of the cross-survey comparability may be treated as a compound of two separate countries, e.g. Czechoslovakia in ISSP/1992 (which in the year of the fieldwork was separating into the Czech Republic and Slovakia).

T_COUNTRY_L2U corresponds to a harmonized, *country-level* sample. Following the above examples, it is Poland, (united) Germany (provided we have both East and West samples covered by L1U), Israel (with both Jew and Arab communities covered by L1U), the Republic of Cyprus and the Turkish Cypriote Community, as well as the Czech Republic and Slovakia, all of these treated as stand-alone samples.

In most instances, T_COUNTRY_L1U and T_COUNTRY_L2U are the same (since L1U and L2U match). Table 6 shows when T_COUNTRY_L1U and T_COUNTRY_L2U differ. Note: In SDR 2, West Germany before the 1990 reunification is treated as Germany.

Table 6. Instances where T_COUNTRY_L1U and T_COUNTRY_L2U differ

L2U samples	**L1U samples**	**Project/waves** (*)
Belgium	Belgium (Brussels), Belgium (Flanders), Belgium (Wallonia)	ISSP/2011-2015
Germany	East Germany, West Germany	CDCEE/2, EB/34.1-87.3, EVS/2-4, ISJP/1991, ISSP/1990-2015, WVS/3-6
United Kingdom	Great Britain, Northern Ireland	ASES, EB/0.2-87.3, EVS/1-4, ISJP/1991, ISSP/1985-2015, LITS/2, PA1, WVS/3-5
Israel	Israel (Arab population), Israel (Jew population)	ISSP/1996-2015
Czech Republic, Slovakia	Czechoslovakia	ISSP/1992
Cyprus, Turkish Cypriote Community	Cyprus (island)	WVS/5-6
Montenegro, Serbia	Yugoslavia	NEB/5 (1998)

(*) For simplicity, we provide only first and last waves within projects. It is possible that some interim waves do not contain sub/supra-country samples that lead to the difference between T_COUNTRY_L1U and T_COUNTRY_L2U.

Once samples in Tables 3 and 4 are excluded, a national survey can be uniquely identified in the SDR 2 MASTER file using a combination of three variables that is conceptually equivalent to *project–year–country*: T_PROJECT_NAME, T_PROJECT_WAVE, and T_COUNTRY_L2U. However, remember that T_COUNTRY_L1U is still needed for identifying the source sample. For the complete list of national surveys in SDR 2 refer to the PLUG_SURVEY file (see SDR2 PLUG_COUNTRY File Overview for its description).

Data weighting is the second reason for distinguishing between T_COUNTRY_L1U and T_COUNTRY_L2U. Since weighting factors created by data producers are usually provided at L1U, and since in SDR 2 we are interested in harmonizing all aspects of source data, we also harmonize weighting factors, recalculating them into L2U factors, captured by the technical variable T_WEIGHT_L2U. The procedure is described in detail in the General Target Variable Report – WEIGHTS.

Both T_COUNTRY_L1U and T_COUNTRY_L2U are typically coded with ISO 3166-1 alpha-2 codes [https://en.wikipedia.org/wiki/ISO_3166-1], e.g. PL for Poland or DE for (united) Germany. In the case of the sub-country division of a sample, we use an up-to-3-letter extension, e.g. DE-E for East Germany or GB-NIR for Northern Ireland. In rare situations, we further extend the ISO 3166-1 standard. Table 7 provides the complete list of all deviations from ISO 3166-1.

Table 7. The complete list of SDR2 sample codes extending the ISO 3166-1 standard

Extension of the ISO 3166-1 code	**L1U/L2U samples**
BE-BRU	Belgium (Brussels)
BE-FLA	Belgium (Flanders)
BE-WAL	Belgium (Wallonia)
CS	Czechoslovakia
CY-ISL	Cyprus (island)
CY-TCC	Turkish Cypriote Community
DE-E	East Germany
DE-W	West Germany
GB-GBN	Great Britain
GB-NIR	Northern Ireland
IL-ARB	Israel (Arab population)
IL-JEW	Israel (Jew population)
KS	Kosovo
YU	Yugoslavia

Since T_COUNTRY_L1U and T_COUNTRY_L2U share the same standard of coding, we provide a single dictionary, PLUG_COUNTRY file, for translating ISO codes to a proper country, sub-country, or specific population names. Use T_COUNTRY variable in PLUG_COUNTRY file for this purpose. The PLUG_COUNTRY file is described in SDR2 PLUG-COUNTRY Documentation.

2.e Time coverage (T_INT_YEAR and T_COUNTRY_YEAR)

Survey projects included in SDR2 span the years 1966-2017. In the MASTER file, we have two variables related to the dimension of time: the year when the respondent was interviewed (T_INT_YEAR) and the year assigned for a national survey as a whole (T_COUNTRY_YEAR). Usually, a single case (respondent) shares the same value on both variables. However, differences appear when the survey's fieldwork was carried out over two or more years, while a specific interview is always pinpointed to a specific year.

In 114 source data files (out of 215), we identified variables related to the interview date that allowed us to straightforwardly extract T_INT_YEAR. There are, however, some situations worth mentioning here:

(1) A source variable takes only one value, generally, a year (we call it a nominal variable). Usually it identifies a wave rather than the year of the fieldwork, which do not necessarily overlap, e.g. the NEB/3 value in data is 1993, while the documented year of the fieldwork is usually 1994.

(2) A variable takes empty, impossible, wrong, or suspicious values, e.g. nulls in AMB/2016 Haiti; 1582 in AMB/2010; 1959 in CB/2009; or 1982, 2000, 2009, 2010, 2011, 2014, 2015, 2016 in LITS/3, while documentation clearly says the fieldwork was carried out in 2015-2016.

(3) We find more than one time-like variable. Usually only one of them is a true interview date, while the other is nominal.

If we found no source variable related to the interview date, or the variable was ambiguous or problematic, we inspected the available survey documentation. Fieldwork dates were documented – in more or less detailed way – for most national surveys.

Only on rare occasions did we take further steps. Usually, the survey year was apparent from the context (e.g. an ISSP module being carried out along with another study, which was, in turn, documented in some external resources). In one case, PPE7N Japan, a detective investigation was needed: first, in the codebook we found a reference to a 1967 paper on the sampling procedure; second, the order of parties mentioned in one of the questionnaire items was identical to the results of the 1967 general elections; third, another question asked about the then prime minister, Eisaku Sato. Although Sato was in office three times in a row, in 1964-1967, 1967-1969 and 1969-1972, since the next general elections after 1967 were held in 1969, we decided that the fieldwork would have been carried out in the period 1967-1969, so we assigned the middle value, 1968 to T_INT_YEAR.

When the fieldwork period spanned more than a year, as indicated either in the data (T_INT_YEAR) or documentation, we adopted the following rules to decide on a single value for a national survey's T_COUNTRY_YEAR:

(1) The year with the larger number of respondents in survey data;

(2) The year with the larger number of days/months in survey documentation;

(3) The year which maximizes consistency between data and documentation.

When we did not find an interview date source variable, we derive T_INT_YEAR from T_COUNTRY_YEAR with the rule (2) taken in consideration.

3. **Data record quality control variables (QR_ID_NONUNIQUE, QR_ID_MISSING, QR_DUPLICATE)**

During source data checks we found that case identifiers (respondent's IDs) are sometimes missing (i.e. not provided) or non-unique. Such cases are flagged by dichotomous quality control variables QR_ID_NONUNIQUE and QR_ID_MISSING, respectively, which code the undesirable situation as 1, otherwise 0. To prevent ambiguity in indicating cases within source data files, we introduce the variable T_ID which uniquely numbers cases (respondents) from 1 to the respective last one within T_DATAFILE.

In the MASTER file we also provide QR_DUPLICATE. This is a dichotomous quality control variable which takes the value 1 when the source record in a source data file is duplicated on all substantive variables, otherwise 0.