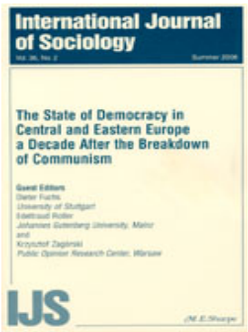

International Journal of Sociology

ISSN: 0020-7659 (Print) 1557-9336 (Online) Journal homepage: http://www.tandfonline.com/loi/mijs20

Working with Data in the Cross-National Survey Harmonization Project: Outline of Programming Decisions

Przemek Powałko & Marta Kołczyńska

To cite this article: Przemek Powałko & Marta Kołczyńska (2016) Working with Data in the Cross-National Survey Harmonization Project: Outline of Programming Decisions, International Journal of Sociology, 46:1, 73-80

To link to this article: https://doi.org/10.1080/00207659.2016.1130433

International Journal of Sociology, 46: 73–80, 2016

ISSN: 0020-7659 print/1557-9336 online
DOI: 10.1080/00207659.2016.1130433

Routledge
Taylor & Francis Group

Working with Data in the Cross-National Survey Harmonization Project: Outline of Programming Decisions

Przemek Powałko

Institute of Philosophy and Sociology, Polish Academy of Sciences

Marta Kołczyńska

Department of Sociology, The Ohio State University; and
Institute of Philosophy and Sociology, Polish Academy of Sciences

In the Harmonization Project we use data from 22 cross-national survey projects that cover a total of 142 countries or territories over a time span of almost 50 years (1966–2013). The large volume of the data, their multilevel structure as well as the large number of source files encouraged us to develop a set of custom tools for extracting, transforming, and loading data into a common database that allows for efficient automation of repeatable and otherwise manual routines. We created an environment based on freeware and open-source software that constitutes an alternative to statistical packages typically used for such purposes in social science research. This platform allows us to store and manage data in a single place, and—what is crucial—enables us to easily manipulate data in order to prepare the harmonized data set for use in substantive analyses. This article presents our motivation for choosing a custom programming and database environment, describes the principles guiding our software choices, and outlines the stages of data processing.

Keywords data harmonization; data management; relational database

The growing wealth of social survey data have the potential to inform a variety of research agendas in the social sciences, but these data are often not comparable or well-documented. Regional coverage of many international surveys, such as the European Social Survey, the Latinobarómetro, and the Asia-Europe Survey, does not facilitate research of truly global issues. The project "Democratic Values and Protest Behavior: Data Harmonization, Measurement Comparability, and Multi-Level Modeling" (hereafter, the Harmonization Project;

Przemek Powałko currently works as a database specialist for the Data Harmonization Project at the Institute of Philosophy and Sociology, Polish Academy of Sciences.

Marta Kołczyńska is a Ph.D. student in the Department of Sociology at The Ohio State University (OSU), and a research assistant at the Institute of Philosophy and Sociology, Polish Academy of Sciences (PAN). She is affiliated with the Cross-National Studies Interdisciplinary Research and Training Program (CONSIRT) of OSU and PAN, and part of the Harmonization Project team. Her research interests include political attitudes, social inequality, and methodology, all in cross-national perspective, as well as area studies in Southeastern Europe.

Address correspondence to Przemek Powałko, Institute of Philosophy and Sociology, Polish Academy of Sciences, Nowy Świat 72, 00-330 Warsaw, Poland. E-mail: ppowalko@ifispan.waw.pl.

http://dataharmonization.org) addresses some of the challenges of cross-national research using survey data by creating comparable measurements of political values, behaviors, and demographics with global coverage and is available online.

In the Harmonization Project we use data from 22 international survey projects that cover a total of 142 countries or territories over a time span of almost 50 years (1966–2013). The large amount of data led us to develop a set of custom tools for extracting, transforming, and loading data into a common database, allowing for efficient automation of repeatable and otherwise manual routines. We created an environment based on freeware and open-source software that constitutes an alternative to statistical packages typically used for such purposes in social science research. This platform allows us to store and manage data in a single place, and, crucially, it enables us to easily manipulate data in order to prepare the harmonized data set for use in substantive analyses.

After a brief description of the Harmonization Project and the data involved, this article presents our motivation for choosing a custom programming and database environment, describes the principles guiding our software choices, and outlines the stages of data processing.

MOTIVATION

The Harmonization Project is motivated by interrelated substantive and methodological goals. Substantively, the project seeks to explore the patterns of protest behavior across a variety of social, political, and economic contexts worldwide. Such analyses require a combination of individual-level survey data with country-level measures with maximal geographic coverage and over an extensive period of time. Hence, the methodological goal of the Harmonization Project is to collect and inspect data from several international survey projects, and merge them into a single data set in a way that would allow for meaningful analyses (Tomescu-Dubrow and Slomczynski 2014).

The survey projects used in our study were selected based on the following criteria: they are (1) noncommercial (mainly academic), (2) freely available for purposes of academic research, (3) designed as cross-national (and preferably multiwave), and (4) sufficiently documented in English (study description, codebook, and/or questionnaire); (5) their samples are intended as representative of the entire adult population of a given country or territory, and (6) they contain questions of substantive interest to the Harmonization Project, that is, items pertaining to political attitudes and protest behavior. For the full list of survey projects, refer to Table 1.

The universe of our research consists of approximately 2.3 million cases originally stored in 81 data files, which we call source files. A single source file may contain data from a single country in one wave up to many countries in many waves, depending on the survey project.[1] The number of data sets, that is, national samples in all survey projects carried out in all waves and in all projects, is 1,726.

Apart from the individual-level survey data, we collected contextual data at different levels. These include survey-level metadata and variables controlling for the methodological variation across surveys, as well as social, political, and economic variables describing either countries or countries in a given year.

These data have a number of features that make them difficult to analyze using single common statistical software such as SPSS (http://www.ibm.com/software/analytics/spss) or Stata

TABLE 1
The List of Investigated Survey Projects

Abbreviation	*Survey project*	*Time span*	*Number of waves*	*Number of source files*	*Number of national surveys**	*Number of cases**
ABS	Asian Barometer	2001–2011	3	3	30	43,691
AFB	Afrobarometer	1999–2009	4	4	66	98,942
AMB	Americas Barometer	2004–2012	5	1	92	151,341
ARB	Arab Barometer	2006–2011	2	2	16	19,684
ASES	Asia-Europe Survey	2000	1	1	18	18,253
CB	Caucasus Barometer	2009–2012	4	4	12	24,621
CDCEE	Consolidation of Democracy in Central and Eastern Europe	1990–2001	2	1	27	28,926
CNEP†	Comparative National Elections Project	2004–2006	1	8	8‡	13,372‡
EB†	Eurobarometer	1983–2012	7	7	152	138,753
EQLS	European Quality of Life Survey	2003–2012	3	1	93	105,527
ESS	European Social Survey	2002–2013	6	2	146	281,496
EVS/WVS	European Values Study/World Values Survey#	1981–2009	9	1	312	423,084
ISJP	International Social Justice Project	1991–1996	2	1	21	25,805
ISSP†	International Social Survey Programme	1985–2013	13	13	363	493,243
LB	Latinobarómetro	1995–2010	15	15	260	294,965
LITS	Life in Transition Survey	2006–2010	2	2	64	67,866
NBB	New Baltic Barometer	1993–2004	6	1	18	21,601
PA2	Political Action II	1979–1981	1	1	3‡	4,057‡
PA8NS	Political Action—An Eight Nation Study	1973–1976	1	1	8	12,588
PPE7 N	Political Participation and Equality in Seven Nations	1966–1971	1	7	7	16,522
VPCPCE	Values and Political Change in Postcommunist Europe	1993	1	5	5‡	4,723‡
		1966–2013	**89**	**81**	**1,721**	**2,289,060**

Source: The table is based on Table 1 in Tomescu-Dubrow and Slomczynski (2014), with appropriate updates. See also Table 1 in Tomescu-Dubrow and Slomczynski (2016).
*Numbers from the source files before any filtering.
†Only selected survey waves.
‡Numbers from the source files after filtering out panel and post-election surveys from CNEP, PA2, and VPCPCE.
#The integrated file.

(http://www.stata.com), which are most frequently used in social science quantitative research. First of all, the sheer volume of the data makes their storage, manipulation, and especially analysis impractical. Second, the large number of source files calls for programmability, which raises the flexibility of implementation of data harmonization, allows for greater control of the data flow through establishing small and well-defined procedural steps, and facilitates repeatability and replicability. Programmable solutions have the additional advantage of providing documentation of each step for quality control of the process. Finally, given the different units of observation (individual respondent, country, country-year, survey, etc.) in our data set, storing them in a single table, as is essentially the case in most statistical packages, would result in enormous redundancy, thus not only slowing any manipulation of the data even more but also leading to great inefficiency of the process. In conclusion, data management in the Harmonization Project required an innovative strategy. In addition to addressing the issues mentioned above, our choice of the solution was guided by the preference for free and/or open-source software, which reduces the costs of the endeavor and extends access to online knowledge bases and user communities.

IMPLEMENTATION

We start with a number of source files downloaded directly from project Web sites or data repositories, extract data and import them to a database, and store in corresponding source tables, one for each source file. We implement the harmonization rules in the form of series of independent database operations, and save results in a single target table, called the master table. Although a database environment provides basic tools for statistical analysis, in order to apply more sophisticated techniques available only in statistical packages, the master table is exported to the master file of the appropriate format. Figure 1 outlines the flow of data in the Harmonization Project. In the following paragraphs we describe the process of importing data from input source file (S_i) into database tables (T_i) and discuss some of the issues pertaining to specific steps of this process.

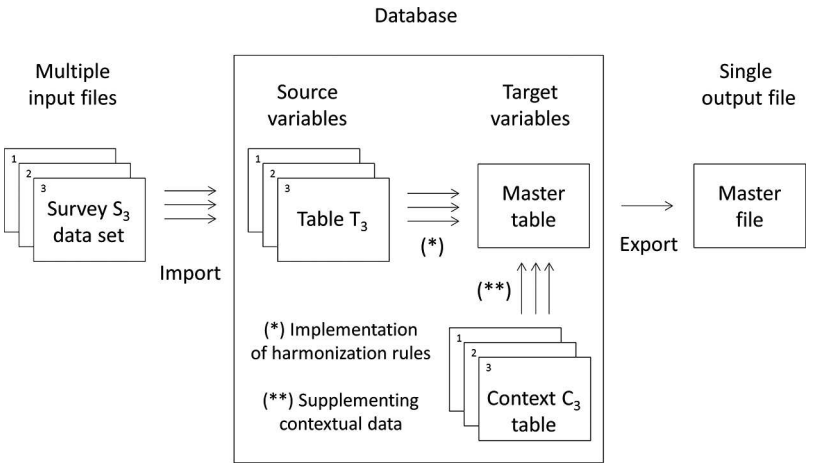

FIGURE 1 Flow of data in the Harmonization Project.

Input source files come in various formats, but for the sake of simplicity, repeatability, and automation of the subsequent steps we decided to work with the SPSS system or portable files, which appear to be the most frequently used. In a few instances, when SPSS files were not available, we converted existing files to the chosen format. To read SPSS data files we use a command-line interface of PSPP (http://www.gnu.org/software/pspp), a free alternative to SPSS.

It is a fact of life that first releases of survey data sets are rarely error-free. Various problems or errors that are identified in the published data sets are highlighted with alerts or corrected with errata or new versions of the data files. We handle these changes in the following way: (1) if a new version of the source file is published, we replace the file in our repository; (2) if an erratum is published in the form of a syntax patch, we apply it to the original data file; and finally, (3) if instructions are given in other ways, usually as explanatory notes, we write them as a corresponding SPSS syntax and then modify the original data file. The data in our repository represent the state of affairs at the end of the first quarter of 2014.[2]

All data processing is scripted and automated in an open-source programming environment called Cygwin (http://www.cygwin.com), which integrates Windows resources and Unix-like capabilities for manipulating data.[3] Cygwin is equipped with a number of desirable features for the automation of repeatable and otherwise manual and time-consuming tasks: command-line interpreters providing interface to the operating system (e.g., Bash), utilities for text processing (e.g., grep, sed), and programming languages for facilitating complex tasks (e.g., AWK, Perl).

Data from the source files are extracted and stored in corresponding tables in a relational database. As indicated earlier, these data include survey data (measured at the level of individual respondents), as well as data on countries (measured on the level of country-years) and survey quality control and methodological variables (measured on the level of national surveys, waves, or projects).

Tables are composed of rows and columns that in a graphical representation resemble matrixes or spreadsheets. Columns correspond to variables and rows correspond to cases, which—depending on the data source—may correspond to individuals, countries, or surveys. In the relational database, tables may be related to each other via explicit or implicit referential integrity mechanisms (keys and joins), which allow for easy, fast, and—at the same time—sophisticated access to related data while reducing its redundancy and merging time. This approach improves scalability of the database, which means that adding more data files to the project's repository—resulting in more tables in the database—scales with the execution time in a predictable, often linear way.

A great advantage of relational databases is that data from all tables are directly available for browsing and querying without the need of having them opened all at once. Executing a database query leads to a series of small data reads from relevant tables so that even if tables are numerous and big, only a fraction of data is read from disk at any time and loaded into internal memory, making the whole process more efficient in terms of time and computing resources than the standard practice of merging and appending all necessary data into a single data file.

After reviewing the most popular databases, we decided on MySQL (http://www.mysql.com), a free and open-source database, because of its ability to handle large numbers of variables that are present in some of the source files we work with. MySQL allows for up to 4,096 columns in a table, which so far has proved to be a sufficient number. Another advantage of MySQL is the existence of a convenient command-line interface whose seamless integration with a Cygwin terminal facilitates the scripting of data processing. Moreover, MySQL has a

variety of built-in storage engines that can be used for different purposes. For storing source data we chose the ARCHIVE engine, which significantly compresses data in tables, therefore making them smaller and making reads from disk faster. This in turn decreases the demand for internal memory and accelerates data processing.

Communication with the relational database is carried out using SQL (Structured Query Language), a high-level declarative language with a simple syntax yet powerful capabilities. For writing queries and browsing data in the database we use HeidiSQL (http://www.heidisql.com), a free SQL editor. For writing programs in scripting languages we use Notepad++ (http://notepad-plus-plus.org), a free text editor. Together with Cygwin, PSPP, and MySQL, all these tools are free and/or open-source software. This best serves our purpose of building a platform for survey data processing without generating unnecessary costs. At the same time, the open-source model attracts both individual users and corporate bodies, creates vigorous user communities boosting online knowledge bases and pushing software developers to deliver new features and solutions to known problems in reaction times much faster than in the world of proprietary software.

DATA PROCESSING

Having described the structure of the programming environment, we turn to the process of loading the survey data into the database. Let us assume that an SPSS source file has been saved to a working directory on our workstation. The following steps are fully automated and executed through a set of Cygwin scripts:

1. Extract data from the SPSS source file and save them to a text file. This is done through a PSPP syntax file. The output is a comma-separated values (CSV) file, a convenient way of storing regularly formatted (e.g., tabular, as in our case) data because of its compactness and readability.
2. Scan and analyze the content of the CSV file and create a corresponding SQL syntax for a source table definition. This is necessary because sometimes SPSS files provide inaccurate information on data types and lengths. We decided that all data in source tables will be stored in columns of the character string type of variable length, that is, what it "looks like" in the source files, without any interpretation. Thus the only parameter of a column's definition is the length of the data, which is evaluated during this step.
3. Connect to the database and execute the SQL syntax file to create an empty table with appropriate parameters for storing the source data.
4. Connect to the database and load the content of the CSV file into the source table.

In addition to the data, an SPSS source file contains information about variable names, formats, acceptable values and missing value codes, as well as variable and value labels. We extract only part of these metadata, that is, code values and labels corresponding to response categories, and save them into a separate text file, which we call a dictionary (DIC) file. Note that dictionary files (and corresponding dictionary tables) are only for reference and as such are not used in the further harmonization process so the following steps are optional:

5. Extract metadata from the SPSS source file and store them in the DIC file.
6. Prepare an SQL syntax file for dictionary table definition according to the values in the DIC file.

7. Connect to the database and execute the SQL syntax file to create an empty dictionary table for storing the metadata.
8. Connect to the database and load the DIC file's content into the dictionary table.

The process of loading the survey data into the database is completed. In the workstation that we use for tests,[4] the processing of a 45MB SPSS source file containing 50,000 cases and 300 variables takes roughly 40 seconds, and the corresponding tables occupy about 2MB. The automation of the whole procedure allows for an easy way of adding new source files and populating the tables. The data are stored in the database and are ready for use.

According to the procedure described above, we have created one data table and one dictionary table for each source file, resulting in a total of 81 data tables and 81 dictionary tables. In addition to survey data, the database contains contextual data from various publicly available sources (country population, gross domestic product, etc.) as well as information describing the survey process (e.g., response rates) and quality indicators (including data-documentation inconsistencies or nonunique cases), all of which are added manually into separate tables, one for each data type.

A final product of the harmonization process is a single data set, which we call the master table, and the corresponding master file exported outside the database to any format read by statistical packages. Harmonization itself, that is, the various operations performed in SQL to transform a number of source variables into target variables with a unified metric across all surveys, and accompanied by a set of case-specific harmonization control variables that capture significant methodological differences between surveys, is outside the scope of this article. Detailed information about the harmonization process as well as all documentation will be published on the project's Web site and announced in the specially established newsletter (http://consirt.osu.edu/newsletter). Suffice to say that the structure of the master file is flexible and depends on the end user's needs and expectations in terms of variables and cases, which can be selected according to specific research objectives.

SUMMARY

In this article we described the structure of the platform for processing data in the Harmonization Project. The essence of this solution lies in the automation of as many procedures as possible with the use of programming languages, efficient use of resources, and exploitation of free and open-source software.

The skills required to develop one's own environment might be perceived as a drawback, but we think that the advantages mentioned so far outweigh the costs. An additional value of storing the data in a relational database is that SQL allows one to perform basic statistical analyses in the database. Although the analytic capabilities of SQL are limited and rather modest compared to dedicated statistical software, there exist statistical and data mining extensions to databases, which are free or reasonably priced for academic purposes.

ACKNOWLEDGMENTS

The authors thank Kazimierz M. Slomczynski, Irina Tomescu-Dubrow, Ilona Wysmułek, Olena Oleksiyenko, and Tadeusz Krauze for insightful comments and support, and the anonymous

reviewer for substantial remarks. An early version of this article was presented by Przemek Powałko at the Conference and Workshop on Survey Data Harmonization (Warsaw, December 18–21, 2013) and the Workshop on Comparability of Survey Data on Political Behavior Following Ex-Post Harmonization of Selected Survey Projects (Columbus, OH, May 8–9, 2014).

FUNDING

This work is part of the project "Democratic Values and Protest Behavior: Data Harmonization, Measurement Comparability, and Multi-Level Modeling in Cross-National Perspective," and is funded by the National Science Centre, Poland, under grant number 2012/06/M/HS6/00322.

NOTES

1. Some survey projects, for example, the European Social Survey, offer the possibility to download single national surveys, whole survey waves, or multiwave cumulative data files. As a rule, for the Harmonization Project we use mainly aggregated files provided by each project.

2. Details about the preparation of the source files can be found in the dataverse at http://dx.doi.org/10.7910/DVN/HPXFA1.

3. UNIX is a family of operating systems that are known for a use of plain text for storing data thus providing a large number of sophisticated tools for text processing.

4. Lenovo Thinkpad T530, Intel Core i5 3210 M CPU 2.5 GHz, 8 GB DDR3, SanDisk SDSSDHP256 G.

REFERENCES

Tomescu-Dubrow, Irina and Kazimierz M. Slomczynski. 2014. "Democratic Values and Protest Behavior: Data Harmonization, Measurement Comparability, and Multi-Level Modeling in Cross-National Perspective." *ASK: Research and Methods* 23(1):103–114.

Tomescu-Dubrow, Irina and Kazimierz M. Slomczynski. 2016. "Harmonization of Cross-National Survey Projects on Political Behavior: Developing the Analytic Framework of Survey Data Recycling." *International Journal of Sociology* 46(1):58–72.