Harmonised data for comparative research

Webinar

15 June 2020, Webinar

Dr Jen Buckley, University of Manchester

Professor Irina Tomescu-Dubrow, Institute of Philosophy and Sociology, Polish Academy of Sciences

Dr Kristi Winters , GESIS - Leibniz-Institute for the Social Sciences

cessda.eu

@CESSDA_Data

DOI: 10.5281/zenodo.3894146

Outline

- Sources of harmonised data for comparative research
  - Jen Buckley
- Ex-post harmonization in the Survey Data Recycling, SDR, Project
  - Irina Tomescu-Dubrow
- Harmonization and cross-national comparative research
  - Kristi Winters
- Questions and Answers

Ways to get comparable data

Ex-ante
[before data collection]

Ex-post
[after data collection]

Input

Output

Common concepts and procedures

Common outputs (populations/specifications)

Select/correct information from different sources

All these methods have been used to create cross-national databases available to use

Major cross-national projects

- Many cross-national survey projects
  - including major collaborative endeavours with infrastructure around data collection, processing and dissemination

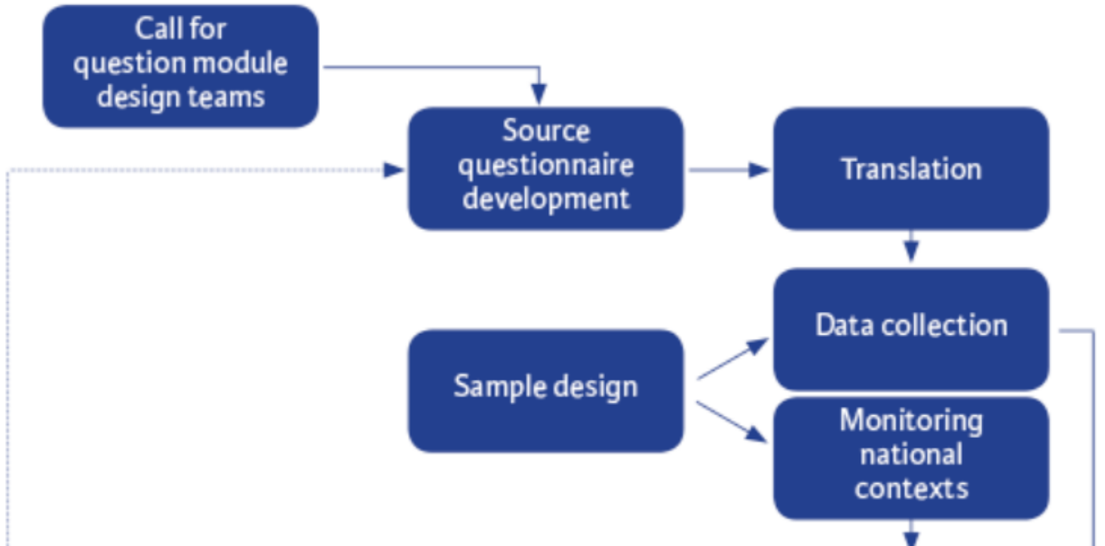
European Social Survey (ESS)

- europeansocialsurvey.org
- Attitudes, beliefs and behaviour patterns
- 2002 – ongoing
- 9 rounds of data, 36 countries
- Core and rotating modules designed by expert teams
- Survey managed by National Co-ordinators
- ESS multilevel contains contextual data for regions and countries

Methodology Overview

A key aim of the ESS has always been to implement high quality standards in its methodology and to improve standards in the field of cross-national surveys more generally. Measuring attitudes cross-nationally has challenges that go beyond those of surveys conducted in a single country or language. The methods used on the ESS are outlined in the Survey specifications, which all countries are required to adhere to when conducting the ESS in their country.

Additional information about the ESS methodology and methodological research can be found in the sub-sections on the left-hand side menu, also accessible from the flowchart below.



Documentation

- ESS Round 10 Survey Specification for participating countries
- ESS Round 9 Manual

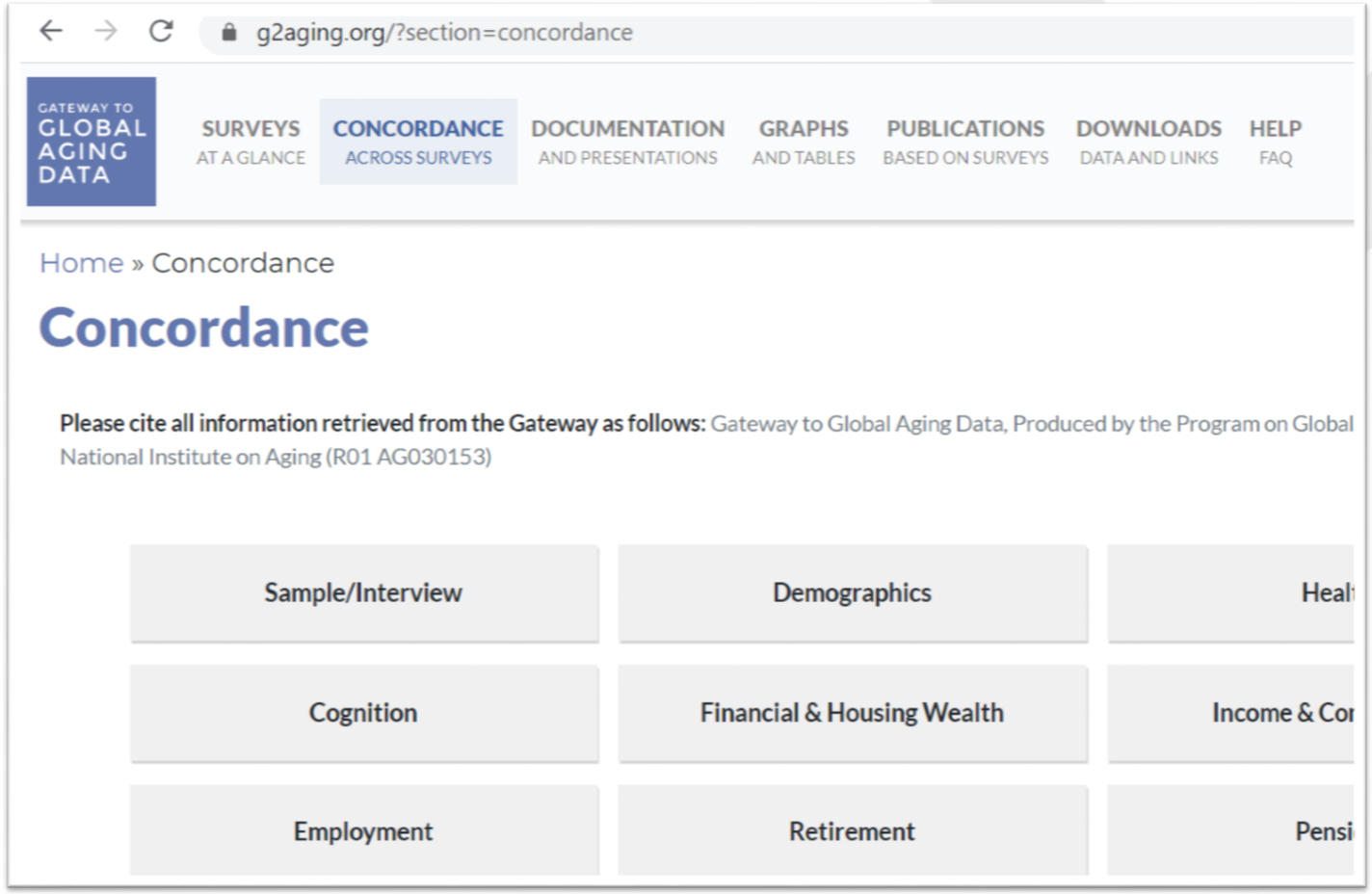
Major cross-national projects (2)

- SHARE – Survey of Health, Ageing and Retirement in Europe
- share-project.org
- Longitudinal study involving more than 123,000 individuals aged 50+
- 27 European countries and Israel
- Data on health, socio-economic status and social and family networks
- 7 waves between 2004 and 2017

- Part of a family of ageing studies
  - HRS - U.S. Health and Retirement Study
  - ELSA - English Longitudinal study of aging
  - Tilda - the The Irish Longitudinal Study of Ageing
- Comparable topics and questions

Gateway to Global Aging

- G2aging.org
- Tool for navigating the HRS studies
- Harmonised datasets
  - Via original studies or
  - Can be created with a Stata programme



Collaborative programs for common modules of survey questions

International social survey programme (ISSP)

w.issp.org

- Collaboratively developed thematic modules including:
  - Work, religion, role of government, inequality, citizenship, health, family, citizenship
- Fielded in national survey such as Germany ALLBUS
- Access via Gesis: https://www.gesis.org/issp/modules/issp-modules-by-year/

Comparative study of electoral systems (CSES)

cses.org

Collaboration between election study teams across the world

Five thematic modules since 1996

Integrated dataset available via the CSES website

Harmonisation across studies

- Use of international classifications
  - Education - International Standard Classification of Education (ISCED)
  - Occupation - International Standard Classification of Occupations 08 (ISCO-08)

seriss SYNERGIES FOR EUROPE'S RESEARCH INFRASTRUCTURES IN THE SOCIAL SCIENCES

Synergies for European Research Infrastructures for the Social Sciences (SERISS)

- Collaboration between major research infrastructures including CESSDA, SHARE and ESS
- Developing and sharing coding modules (Surveycodings.org)
  - Education
  - Occupation
  - Industry

Data discovery resources from CESSDA and national data archives (1)

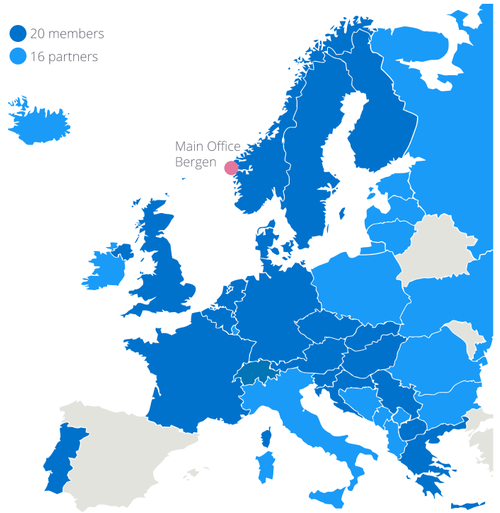
[CESSDA web directory on international surveys]:

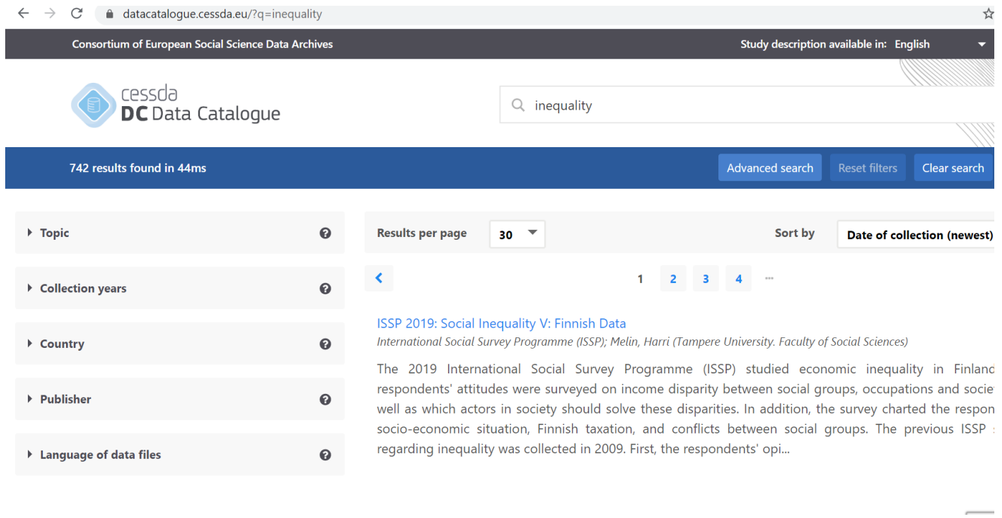
- Lists continuous international survey research programmes with accessible primary data
- Includes descriptions and links to the data resources

Also [GESIS also have a chronological Overview of comparative studies (last updated 2015)]

Data discovery resources from CESSDA and national data archives (2)

- Data catalogues of national data archives
  - Details of national data archives are on the CESSDA website cessda.eu/About/Consortium
- CESSDA Data Catalogue: datacatalogue.cessda.eu
- In development: European Question Bank (EQB)





Eurostat

Statistical office of the European
Access to many harmonised microdata databases:
https://ec.europa.eu/eurostat/web/microdata

- European Community Household Panel
- European Union Labour Force Survey
- Community Innovation Survey
- European Union Statistics on Income and Living Conditions
- Structure of Earnings Survey
- Adult Education Survey
- European Road Freight Transport Survey
- European Health Interview Survey
- Continuing Vocational Training Survey
- Community Statistics on information Society
- Micro-Moments Dataset
- Harmonised European Time Use Survey
- Household Budget Survey

Household panel studies & Cross-national Equivalent File (CNEF)

- Many household panel studies examining comparable topics:

Understanding Society
Household Income and Labour Dynamics in Australia (HILDA)
Korea Labor and Income Panel Study (KLIPS)
Panel Study of Income Dynamics (PSID)
Russia Longitudinal Monitoring Survey (RLMS-HSE)
Swiss Household Panel (SHP),
Canadian Survey of Labour and Income Dynamics (SLID)
German Socio-Economic Panel (SOEP)

Cross-national Equivalent File (CNEF), 1970-2017
(http://cnef.ehe.osu.edu/),

- Dataset with simplified version of the panels for cross-national analysis
- Guidelines for formulating equivalent variables across countries
- Topics: Demographics, Employment and Income, Health and Psychological well-being

Access: Apply to the data provider in each country for access, then access to the CNEF via: https://cnef.ehe.osu.edu/apply-for-cnef-data-beta/.

Other major harmonisation projects

LIS – Luxembourg Income Study

- Income and wealth
- Access via remote service (LISSY)
- https://www.lisdatacenter.org/

Time-use studies

- Harmonized European Time Use Study (HETUS) - Eurostat
- Multinational Time Use Studyhttps://www.timeuse.org/mtus

International Stratification and Mobility File (ISMF)

- social mobility patterns over time and throughout the world
- Inclusive – 250 surveys, 56 nations
- Files to create dataset and comparable measures on the project website

Survey Data Recycling (SDR)

- multi-country multi-year database with a thematic focus is on political engagement, trust and social capital
- https://www.asc.ohio-state.edu/dataharmonization/

Reference:
Dubrow, Joshua K. and Irina Tomescu-Dubrow. 2015. "**The rise of cross-national survey data harmonization in the social sciences: emergence of an interdisciplinary methodological field**." *Quality and Quantity* DOI 10.1007/s11135-015-0215-z

Ex-post harmonization in the Survey Data Recycling, SDR, Project

Irina Tomescu-Dubrow, IFiS PAN and CONSIRT

tomescu.1@osu.edu

CONSIRT.osu.edu

“What Is”

International survey projects measure *political attitudes & behaviors, social capital,* and *wellbeing*

Asian Barometer
Afrobarometer
Americas Barometer
Arab Barometer
Asia Europe Survey
Caucasus Barometer
Consolidation of Democracy
Comparative National Elections Project
Eurobarometer
European Quality of Life Survey
European Social Survey
European Values Study
International Social Justice Project
International Social Survey Programme
Latinobarometro
Life in Transition Survey
New Baltic Barometer
Political Action II
Political Action - An Eight Nation Study
Political Participation and Equality
Values and Political Change
World Values Survey
New Europe Barometer

“What Should be”

Multi-country multi-years data infrastructure for comparative research on:

Democracy & Political Participation

Social Capital & Political Participation

Social Capital & Wellbeing

Integrated SDR Database v.2.0

Nr. survey projects	23
Nr. project waves	174
Nr. national surveys	3,485
Nr. respondents	4,400,000
Nr. countries/territories	169
Time span	1966-2017
Nr. source data files	214

How to get from “Is” to “Should be”?

Ex-post survey data harmonization methodology that accounts for biases and errors stemming from:

(a) deviations from standards of documenting and preparing survey data suggested in the specialized literature (e.g. Biemer and Lyberg 2003)

(b) inter-survey differences in properties of items measuring the same concept

(c) harmonization procedures

Extended scope of ex-post survey data harmonization methodology

a) Classic definition (e.g. Günther 2003; Minkel 2004; Ehling et al. 2006; Granda, Wolf & Hadorn 2010)

> Methods to select survey measures of the same concept from sources not *a priori* designed as comparative, transform them to achieve/increase the comparability of answers from respondents interviewed in different populations, and create a new, integrated, dataset that can be analyzed as a single data source.

+

b) Methods to define and measure inter-survey methodological variability stemming from source survey quality and ex-post harmonization

SDR database: Target variables

- *Technical variables* (e.g. project name, interview year, survey year, country, weights)

- *Substantive variables measuring respondents' characteristics* (ca. 24 target variables)

(a) socio-demographics (e.g. age, gender, education, place of residence)

(b) reported behaviors (e.g. participation in demonstrations, membership in organizations)

(c) attitudes and opinions (e.g. institutional trust, trust in people, life satisfaction)

asc.ohio-state.edu/dataharmonization/wp-content/uploads/2020/01/Warsaw-2019-12-18-session-2.pdf

+ missing codes schema

dataharmonization.files.wordpress.com/2020/01/powac582ko-data-flow-and-database-structure-in-sdr.pdf

SDR database: Methodological variables (I)

Harmonization controls

Preserve features of the source items and of harmonization that could introduce methodological variability:

Ex:
- diffr. in meaning of the source variables (re time, space, scope, attributes)
- diffr. in formal properties of scale measures (scale length, direction, polarity)
- diffr. number of source variables used to construct a target variable
- some values of a target variable are derived (e.g. age derived from birth year)

Target variable specific

SDR database: Methodological variables (II)

Source data quality controls

Capture variability in source data quality, along three dimensions:

(i) survey documentation (codebook, questionnaires, technical reports, etc.)

inadequate info. in documentation reduces the data's fitness for intended use

(ii) data records in computer files

errors can lead to distortion of empirical results

(iii) consistency documentation – data records (for subset of variables)

processing errors can affect the overall usability of the survey

Practical considerations: harmonization workflow

1. Outline theoretical model(s) and concepts
2. Criteria for data selection -> Time zero: data and documentation downloaded
3. Source variables availability check & preliminary (inclusive) selection
4. (Re)Define target variable based on source variables; final source variable list
5. Creating harmonization control variables
6. Cross-walk coding and additional source variables check
7. Control variables for source data quality
8. Final revisions of harmonization and its documentation; general variable report

Transparency and resource sharing

SDR Database v. 1.1 & documentation available via Harvard Dataverse, (asc.ohio-state.edu/dataharmonization/data)

- General target variable report (Word doc)
- Detailed variable report (Excel)
- Cross-walk table (macro-enabled Excel)
- Syntax (notepad++)

for each SDR target variable

SDR 2.0 Cotton file

Provides overview of 88118 variable names, values, and labels available in the original (source) data files that we retrieved automatically for harmonization purposes in the SDR Project (asc.ohio-state.edu/dataharmonization/data/sdr-2-0-cotton-file)

Harmonization Newsletter (asc.ohio-state.edu/dataharmonization/newsletter)

Acknowledgements

Work presented here reflects the collaboration of the SDR Team of faculty and graduate students from The Ohio State University and the Polish Academy of Sciences (asc.ohio-state.edu/dataharmonization/about/project-team/).

The *Survey Data Recycling: New Analytic Framework, Integrated Database, and Tools for Cross-national Social, Behavioral and Economic Research*, SDR Project, is funded by the National Science Foundation, USA
(NSF SMA-1738502, Sept. 2017- Sept 2021).

Harmonization and cross-national comparative research

A brief overview

Dr. Kristi Winters, GESIS
Mr. Janek Bruker

CESSDA Webinar
15 June, 2020

cessda.eu

@CESSDA_Data

Why Harmonize?

- Necessary for cross-national comparative research
- Contributes to basic (empiric) knowledge
- Allows to conclude impact of policies on societies and human behaviour
- Enables to compare impacts across nations, cultures or over time

Harmonization strategies

- To decide on strategy of data harmonization it is easier to follow a **variable-specific approach** than a survey oriented one
- European Social Survey (ESS) includes some *ex-ante output* and some *input* harmonized variables = *flexible-prospective* harmonization approach
- Some variables follow a strictly standardized pattern while others depend on country-specific conditions

Harmonization Processing

- Decide on your strategy as soon as possible to include data harmonization into the survey lifecycle when applicable
- Planning is everything!
- Create a data (harmonization) processing plan

Data Harmonization Processing Plan

- Follow a systematic design
- Include experts and methodology groups
  - Include country-specific experts if necessary
- Gather feedback on the analytic usefulness of the data
- Implement (external) quality controls
- Use software facilitating variable comparison and harmonization planning

Harmonization Processing

Rules for harmonization
(see Hoffmeyer-Zlotnik, 2008: 11f.; Hoffmeyer-Zlotnik & Wolf, 2003: 405)

- Agree on common definition of each variable that is measured
- Common definition refers to comparable elements of participating countries
- Analysis of national concepts and structures by country specialist
- Identify areas of common ground underlying national concepts and structures

Harmonization Processing

- Decide on valid indicator respecting variable of interest and national manifestations
- Choose harmonization strategy
- Test survey/classification instrument with regard to empirical structures in country and correspondence to logic of joint definition
- Ensure that average respondents from different countries understand survey instrument in intended manner and are able to answer question

Harmonization Processing

Referring to list of Hurst, B. C. & Patrick, D. I. (1998), Harkness, J. et al. (2003) name the following steps for cross-cultural questionnaire design:

(1) Review of literature and existing instruments
(2) Establishment of a conceptual framework
(3) Elicitation of items
(4) Evaluation of cross-cultural equivalence

Harmonization Processing

(5) Development and refinement of draft questionnaire
(6) Evaluation of psychometric properties
(7) Evaluation of responsiveness
(8) Preparation of users' manual and, if relevant, a scoring scheme
(9) Submission to a supervisory council (or other signing-off procedure) and distribution

DOCUMENTATION

Transparency in research

- "Transparency is the cornerstone of social science. Academic discourse rests on the obligation of scholars to reveal to their colleagues the data, theory, and methodology on which their conclusions rest. Unless other scholars can examine evidence, parse the analysis, and understand the processes by which evidence and theories were chosen, why should they trust—and thus expend the time and effort to scrutinize, critique, debate, or extend—existing research?" (Moravcsik, 2007)
- Given the importance of transparency to precise replication and comparable statistical results, the lack of scientific standards for documenting harmonization can be a serious obstacle.

Moravcsik, A. (2007). Transparency: The Revolution in Qualitative Research. *PS: Political Science & Politics*, 47 (1): 48–53. http://dx.doi.org/10.1017/S1049096513001789

Freese, J. (2007). Replication standards for quantitative social science: Why not sociology. *Sociological Methods Research*, 36: 153–172. https://journals.sagepub.com/doi/pdf/10.1177/0049124107306659

Why care about documentation?

Crucial for…

- **Evaluating** data (improvement of *ex-ante* harmonization)
- **Replicating/testing** study (improvement of *ex-ante*, right application in *ex-post*)
- **Re-using** data = secondary analysis (right application in *ex-post*)

Why care about documentation?

Pressure to document increasingly driven by

- more and better **quality** data available to harmonize;
- increasingly sophisticated statistical **methods** used to analyse said data;
- and specific information required to **replicate** variables, thus advancing scientific theories.

→ Not only is replication key to evaluating the quality of a piece of research, but it allows new research paths to be developed and explored

Metadata

- Documenting basically means collecting metadata which is „**data about your data**“
- Metadata are descriptors that facilitate cataloging data and data discovery
- Can be stored in a data repository as well and be transformed into machine-readable metadata

CESSDA Data Management Expert Guide. Documentation and Metadata, https://www.cessda.eu/Training/Training-Resources/Library/Data-Management-Expert-Guide/2.-Organise-Document/Documentation-and-metadata

Metadata

- Machine-readable metadata support your documentation by explaining the purpose, origin, time, location, creator(s), terms of use, and access conditions of research data
- DDI (Data Documentation Initiative) (DDI Alliance, 2017b) is an international standard for describing the data produced by surveys and other observational methods in the social, behavioural, economic, and health sciences
- Expressed in XML, the DDI metadata specification supports the entire research data lifecycle

CESSDA Data Management Expert Guide. Documentation and Metadata, https://www.cessda.eu/Training/Training-Resources/Library/Data-Management-Expert-Guide/2.-Organise-Document/Documentation-and-metadata

Code alone is insufficient

- Including the syntax or do file with an article does not solve the problem:
  - there are no conventions that outline how the syntax code should be published (i.e. SPSS syntax or Stata do files);
  - code alone may not contain the information necessary to replicate others' work (i.e. if questions were answered based on prior routing); and
  - providing code may not include how certain coding decisions were made.

How and what?

- Ideally, **from beginning** of survey life cycle and in most detailed and comprehensible way for all steps
- Often most efforts are put on data collection, but researchers and their team should also be **aware** of the importance of documenting
- Consider documentation as integral part of study design, i.e. should follow **disclosure requirements** and not just refer to what is available

How and what?

Specially important in harmonization:

- *Cultural* context as well as *linguistic* meanings need to be strictly respected when mapping (meta)data
- Loss of information through merging items can create bias in harmonized data

How and what?

- A perfect documentation also includes:
  - Access to the original data (if possible)
  - Original questionnaire wording
  - Information about survey collection process
  - Assistance for check-backs or re-transformation plans of users
  - Restricted-use data agreements to allow access under controlled conditions (deposit in a Secure Data Center, original data might be in a SDC)

How and what?

- A complete data management plan checklist also including an orientation for metadata is provided by CESSDA: https://www.cessda.eu/content/download/4302/48656/file/TTT_DO_DMPExpertGuide_v1.2.pdf

EVS: Harmonization process

- Harmonization including input and ex-ante output elements
- Careful translation process on basis of Master questionnaire

https://europeanvaluesstudy.eu/methodology-data-documentation/survey-2017/methodology/

EVS: Ex-ante harmonized

Religious denomination: Harmonization Mapping

Denmark

Target: v52	Source: v52_DK
2 Protestant	1 People's Church
5 Muslim	2 Muslim
7 Buddhist	3 Buddhist
6 Hindu	4 Hindu
4 Jew	5 Mosaic
9 Other	6 Asatro
9 Other	7 Other
77 not applicable	77 not applicable
88 don't know (spontaneous)	88 don't know (spontaneous)
99 no answer (spontaneous)	99 no answer (spontaneous)

Serbia

Target: v52	Source: v52_RS
8 Orthodox	1 Serbian Orthodox
1 Roman catholic	2 Roman Catholic
5 Muslim	3 Islam
4 Jew	4 Jews
2 Protestant	5 Protestant
2 Protestant	6 Evangelic
2 Protestant	7 Slovakian evangelic
2 Protestant	8 Adventist
9 Other	9 Jehovah's witnesses
2 Protestant	10 Baptist
9 Other	30 Other, please specify (Write in)
77 not applicable	77 not applicable
88 don't know (spontaneous)	88 don't know (spontaneous)
99 no answer (spontaneous)	99 no answer (spontaneous)

https://www.gesis.org/angebot/daten-analysieren/internationale-umfragen/european-values-study/5th-wave-2017

And now…?

All documentation work only becomes meaningful if you…

- Guarantee as much **transparency** as possible to creators and users of the data
- Publish and store your data and documentation in a **FAIR** way

Using Persistent Identifiers (PID)

- A PID will ensure users find and can access data and documentation (see FAIR principles)
- The identifier enables long-term preservation always referring the user to the original source
- Prevents link rot (hyperlinks to unavailable web objects)
- DOI (digital object identifier) is a well-known persistent identifier in academia https://www.doi.org/

Archives

The harmonized data together with the metadata can be stored in data archives (e.g. GESIS supporting ISSP, NDS supporting ESS) providing a link to the documentation files

- Chapter 6 of CESSDA Data Management Expert Guide offers detailed information on archiving and publishing
- Including a list of CESSDA data archives

  https://www.cessda.eu/Training/Training-Resources/Library/Data-Management-Expert-Guide/6.-Archive-Publish/Publishing-with-CESSDA-archives

Benefits of (CESSDA) Data Archives

- Comprehensibility
- Visibility
- Findability
- Reusability
- Longevity
- Overall quality

Read more: https://www.cessda.eu/Training/Training-Resources/Library/Data-Management-Expert-Guide/6.-Archive-Publish/Publishing-with-CESSDA-archives

Literature

Burkhauser, R. V.; D. R. Lillard (2005). The Contribution and Potential of Data Harmonization for Cross-National Comparative Research. *DIW Discussion Papers*, 486. http://hdl.handle.net/10419/18337

Doiron, D., P. Raina, V. Ferretti, F. L'Heureux & I. Fortier (2012). Facilitating collaborative research: Implementing a platform supporting data harmonization and pooling. In *Norsk Epidemiologi*, 21 (2): 221-224.

Ehling, M. (2003). Harmonising data in official statistics. In Hoffmeyer-Zlotnik and Wolf (eds.), *Advances in cross-national comparison* (pp. 17-31). New York, Kluwer Academic/Plenum Publishers.

Fortier, I., D. Doiron, P. Burton & P. Raina (2011). Invited Commentary: Consolidating Data Harmonization—How to Obtain Quality and Applicability? In *American Journal of Epidemiology*, 174 (3): DOI: 10.1093/aje/kwr194

Granda, P., E. Blasczyk (2016). Data Harmonization. *Cross-Cultural Survey Guidelines*. Online retrievable at: https://ccsg.isr.umich.edu/index.php/chapters/data-harmonization-chapter

Granda, P., C. Wolf, R. Hadorn (2010). Harmonizing Survey Data. In Harkness et al. (eds.), Survey Methods in Multinational, Multiregional, and Multicultural Contexts (p. 315). New York, John Wiley & Sons, Inc.

Harkness, J. et al. (2003): Questionnaire design in comparative research. In Harkness, Van de Vijer, Mohler (Eds.): *Cross-cultural survey methods*, pp. 19-34. New Jersey: John Wiley & Sons, Inc.

Harkness, J. (2003): Questionnaire Translation. Harkness, Van de Vijer, Mohler (Eds.): *Cross-cultural survey methods*, pp. 19-34. New Jersey: John Wiley & Sons, Inc.

Hoffmeyer-Zlotnik, J. H. & Warner, U. (2007). How to Survey Education for Cross-National Comparisons: The Hoffmeyer-Zlotnik/Warner-Matrix of Education. In *Metodološki zvezki*, 4(2): 117-148.

Hoffmeyer-Zlotnik, J. H., & Warner, U. (2011). Measuring Occupation and Labour Status in Cross-National Comparative Surveys. GESIS – Leibniz Institute for the Social Sciences.

Mohler, P. & R. Uhler (2003): Documenting Comparative Surveys for Secondary Analysis. In Harkness, Van de Vijer, Mohler (Eds.): *Cross-cultural survey methods*, pp. 19-34. New Jersey: John Wiley & Sons, Inc.

Riley, J. (2017). Understanding Metadata. What is Metadata, and what is it for? Baltimore: National Information Standards Organization (NISO).

Winters, K., S. Netscher (2016). Proposed Standards for Variable Harmonization Documentation and Referencing: A Case Study Using QuickCharmStats 1.1. *PLOS ONE*, 11(2): e0147795. https://doi.org/10.1371/journal.pone.0147795

Questions

cessda.eu

@CESSDA_Data