Recycling an established framework for data post-harmonization

Lucie Marie

SciencesPo
CENTRE DE DONNÉES SOCIO-POLITIQUES

Recycling an established framework for data post-harmonization

Lucie MARIE
Center for Socio-Political Data (CDSP)

IASSIST 2023, Diversity in Research: Social Justice from Data
June 1, 2023 - Philadelphia

Agenda

- Surveys harmonization perspectives
- Making **the French electoral surveys** data more reusable, comparable, consistent and coherent
- *The True European Voter* project
- How DDI-Lifecycle makes it easier
- Results and next steps

3 types of surveys harmonization

- Input harmonization (prospective)
- Ex-ante out put harmonization (prospective)
- Ex-post output harmonization (retrospective)

2 approaches for ex-post harmonization

- Bottom-Up: begins at the specific and moves to the general
- Top-Down: goes from the general to the specific

How can you compare these 2 variables?

Post-Electoral Survey 2012

 # s7b: Agglomération

Information	[Type= discrete] [Format=numeric] [Range= 1-5] [Missing=*]
Statistics [NW/ W]	[Valid=2782 /-] [Invalid=0 /-]

Value	Label	Cases	Percentage
1	En zone rurale	699	25.1%
2	Dans une ville de 2000 à moins de 20.000 habitants	483	17.4%
3	Dans une ville de 20.000 à moins de 100.000 habitants	382	13.7%
4	Dans une agglomération de 100.000 habitants et plus, en prov	766	27.5%
5	Dans l'agglomération parisienne	452	16.2%

French Electoral Study 2007

 # tu99: Tranche d'unité urbaine

Information	[Type= discrete] [Format=numeric] [Range= 1-10] [Missing=*/666/777/888]
Statistics [NW/ W]	[Valid=2000 /-] [Invalid=0 /-]
Notes	Variable issue du système CATI (Computer Assisted Telephone Interviewing)

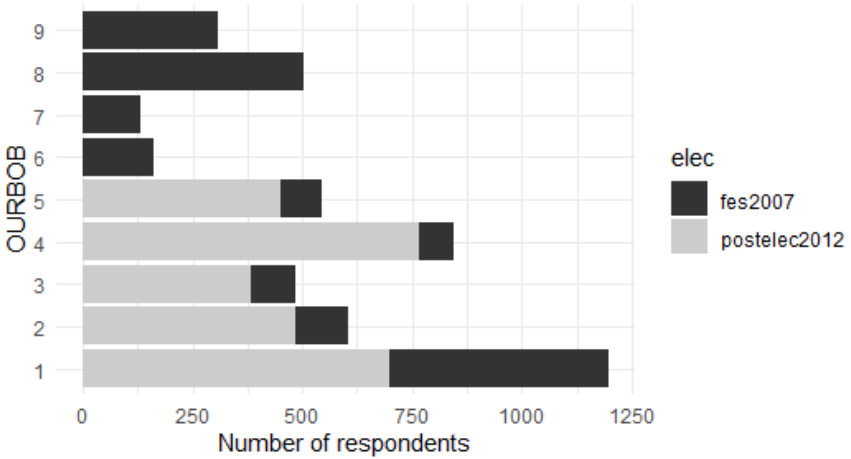
Value	Label	Cases	Percentage
1	Rural	498	24.9%
2	2 à 5000 habitants	123	6.2%
3	5 à 10000 habitants	104	5.2%
4	10 à 20000 habitants	80	4.0%
5	20 à 50000 habitants	92	4.6%
6	50 à 100000 habitants	161	8.0%
7	100 à 200000 habitants	130	6.5%
8	200000 habitants et plus	504	25.2%
9	Agglomération parisienne	308	15.4%

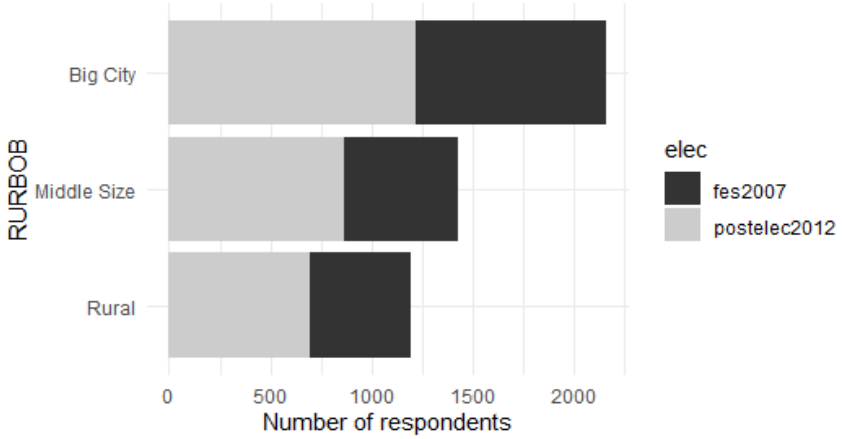
⟨10.5281/zenodo.7405370⟩. ⟨hal-03956459⟩

Making the French electoral surveys data more comparable and reusable

Objective Size of Town

- OURBOB: original variables combined
- RURBOB: standardized recoded variable





⟨10.5281/zenodo.7405370⟩. ⟨hal-03956459⟩

Using an established political science framework

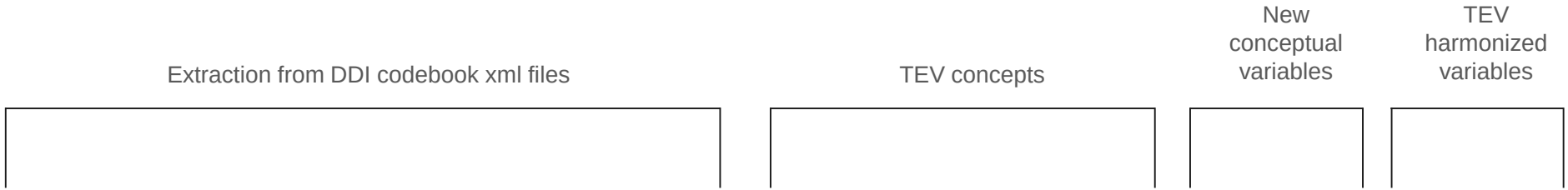
The *True European Voter:* a set of rules for European electoral surveys interoperability

- 23 European countries involved
- common data harmonization guidelines
- 12 concepts and more than 50 sub-concepts
- includes other harmonization initiatives (eg. Manifesto Project, NUTS)

Schmitt, H. (2021). The True European Voter - The True European Voter (1.0.0) [Data set]. GESIS Data Archive. https://doi.org/10.4232/1.13601

⟨10.5281/zenodo.7405370⟩. ⟨hal-03956459⟩

How DDI-Lifecycle makes it easier



title	variable_name	variable_label	concept	sub_concept	conceptual_var_label	dataset_varnam
French Electoral Study 2017	A5	Agglomeration category (9 n	GENERAL BACKGROUND VARIABLES	URBANIZATION	Objective Size of Town	OURBOB
Enquête post-électorale de l'élection pré	s7b	Agglomération	GENERAL BACKGROUND VARIABLES	URBANIZATION	Objective Size of Town	OURBOB
French Electoral Study 2012	in07	Catégorie d'agglomération	GENERAL BACKGROUND VARIABLES	URBANIZATION	Objective Size of Town	OURBOB
French Electoral Study 2007	tu99	Tranche d'unité urbaine	GENERAL BACKGROUND VARIABLES	URBANIZATION	Objective Size of Town	OURBOB
Panel électoral français 2002 Vague 1	agglo	Agglomération	GENERAL BACKGROUND VARIABLES	URBANIZATION	Objective Size of Town	OURBOB
Panel électoral français 2002 Vague 2	agglo	Agglomération	GENERAL BACKGROUND VARIABLES	URBANIZATION	Objective Size of Town	OURBOB
Panel électoral français 2002 Vague 3	agglo	Agglomération	GENERAL BACKGROUND VARIABLES	URBANIZATION	Objective Size of Town	OURBOB
Enquête post-électorale française 1997	agglo	Agglomération	GENERAL BACKGROUND VARIABLES	URBANIZATION	Objective Size of Town	OURBOB
Enquête post-électorale française 1978	t7	Catégorie d'agglomération	GENERAL BACKGROUND VARIABLES	URBANIZATION	Objective Size of Town	OURBOB
Enquête post-électorale française 1962	agglo	Taille de la localité	GENERAL BACKGROUND VARIABLES	URBANIZATION	Objective Size of Town	OURBOB
Panel électoral français 1958	v193	Taille de la localité de réside	GENERAL BACKGROUND VARIABLES	URBANIZATION	Objective Size of Town	OURBOB

⟨10.5281/zenodo.7405370⟩. ⟨hal-03956459⟩

Metadata harmonization: concept identification

Harmonization based on DDI Lifecycle constructs and the TEV conceptual tree

Manual classification of *concepts* and *sub-concepts* according to TEV categories

- pre-defined concepts and a list of all variables make it easy to connect them, especially by text search
- limitation 1: TEV categories are not exhaustive → creation of new concepts/sub-concepts
- limitation 2: TEV categories are generic → creation of conceptual variables

⟨10.5281/zenodo.7405370⟩. ⟨hal-03956459⟩

Outputs

Bilingual metadata base available on data.sciencespo.fr https://doi.org/10.21410/7E4/TSQKNX

- 3,000 variables from 11 selected surveys since 1958
- Approx. 2,500 variables concordance to (at least one) TEV concepts

Harmonized French data currently being integrated into the European database https://doi.org/10.4232/1.13601

- more than 400 variables and over 30,000 individuals
- covering the French elections of the 5th republic

Next steps

- refine the TEV ontology according to the French surveys concepts
- complete curation by building relation and concordance in Colectica
- training research teams to prospective harmonization

Do you have any questions?

lucie.marie2@sciencespo.fr
info.cdsp@sciencespo.fr