French electoral surveys data harmonization based on DDI-L foundational constructs

Lucie Marie, Malo Jan

SciencesPo
CENTRE DE DONNÉES SOCIO-POLITIQUES

French electoral surveys data harmonization based on DDI-L foundational constructs

Lucie MARIE
Center for Socio-Political Data (CDSP)
Malo JAN
Center for European Studies and Comparative Politics (CEE)

EDDI 2022, 14th Annual European User Conference
December 1, 2022 - Paris

Agenda

- Context and objectives
- Why and which ontology?
- Curation and harmonization process
- What outputs on Colectica Portal and Designer?
- Conclusion and next steps

Context

The CDSP dissemination platforms

- Dataverse https://data.sciencespo.fr/
  - Extensive resources (DDI-C)
- Colectica https://explore.cdsp.sciences-po.fr/
  - Selected longitudinal surveys (DDI-L)

UpMet (Upscaling metadata for increasing reuse in the social sciences), WP2

- Question and variable bank in DDI-L with Colectica tools
- Main results:
  - 4 waves of the French Political Barometer publicly accessible on the CDSP Colectica Portal
  - French guide to documenting data in DDI-L with Colectica

Lucie Marie. Guide pour la documentation et la diffusion de données d'enquêtes avec la suite logicielle Colectica. Centre de données socio-politiques (CDSP). 2022. ⟨hal-03845847⟩

UpMet feedbacks

DDI-L documentation:

- Questions, instruments, dataset variables, variable cascade and lineage
- Exploration by study or dataset

Previous challenges:

- Disharmonized datasets and metadata
- Bottom up process of curation
- No conceptual framework

Opportunities:

- Building on the knowledges gained on the DDI-L standard
- Using ontology to drive harmonization

Objectives

Integrating the French historical election studies into the CDSP Colectica Portal

- 11 surveys (from 1958 to 2017)
- Approx. 2500 variables
- More than 30 000 respondents

Enhancing accessibility and comparability

- Harmonized data and metadata
- By making available new functionalities
  - Exploring by concept
  - Multilingual metadata (labels, question texts and categories)

A set of rules for electoral surveys interoperability

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

- 23 european countries involved
- TEV codebook and database
- Common data harmonization guidelines
- Conceptual tree structure (12 concepts and more than 50 sub-concepts)

TABLE OF CONTENTS

Harmonization: metadata extraction

- Metadata extraction from DDI codebook standardised XML files for each survey → ~ 2500 variables

```
−<var ID="V307" name="SD41" files="F1" dcml="0" intrvl="discrete">
   <location width="2" RecSegNo="1"/>
   <labl>Religion [Religion of the respondent]</labl>
  −<qstn>
   −<qstnLit>
       Could you tell me what is your religion, if you have one ?
    </qstnLit>
    <ivuInstr>Show the screen - one answer possible only.</ivuInstr>
```

title	variable_name	variable_label	question_text	category_label
French Electoral Study 2017	A5	Agglomeration category (9 modalities)		2000/5000 inhabitants |
French Electoral Study 2017	O39	Politicians : - the majority of politicians is	Do you strongly agree, somewhat agree, neither agre	Somewhat agree | Neitl
French Electoral Study 2017	PR14	Assessment [Democracy]	On the whole, are you very satisfied, fairly satisfied, r	Satisfied | Not satisfied
French Electoral Study 2017	S3	Age	What is your age ?	
French Electoral Study 2017	SD41	Religion [Religion of the respondent]	Could you tell me what is your religion, if you have or	Protestant | Jewish | M

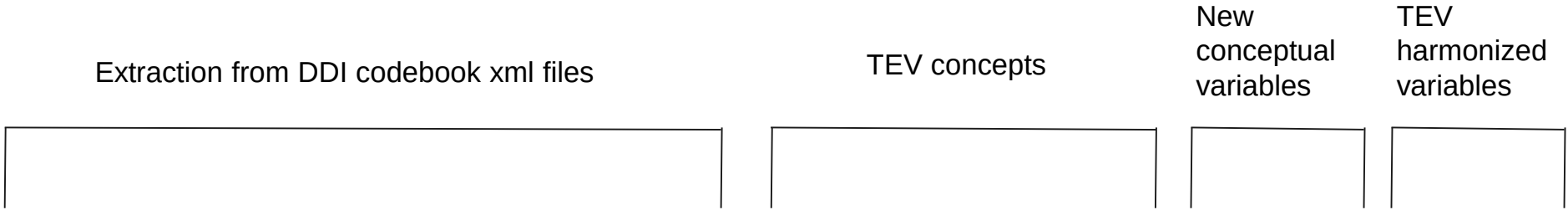
Metadata harmonization: concept identification

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/subconcepts
- Limitation 2: TEV categories are generic → Creation of conceptual variables

Metadata harmonization: concept identification



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

Data harmonization

French Electoral Study 2007

 # 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%

Post-Electoral Survey 2012

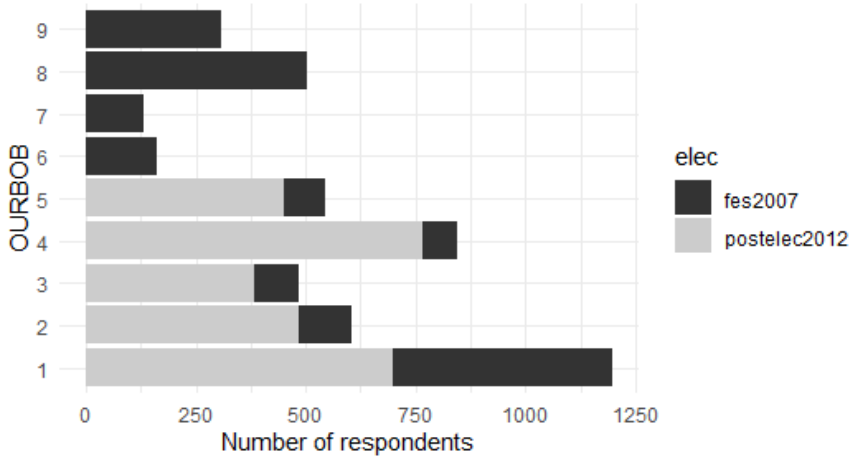
 # tu99: Tranche d'unité urbaine

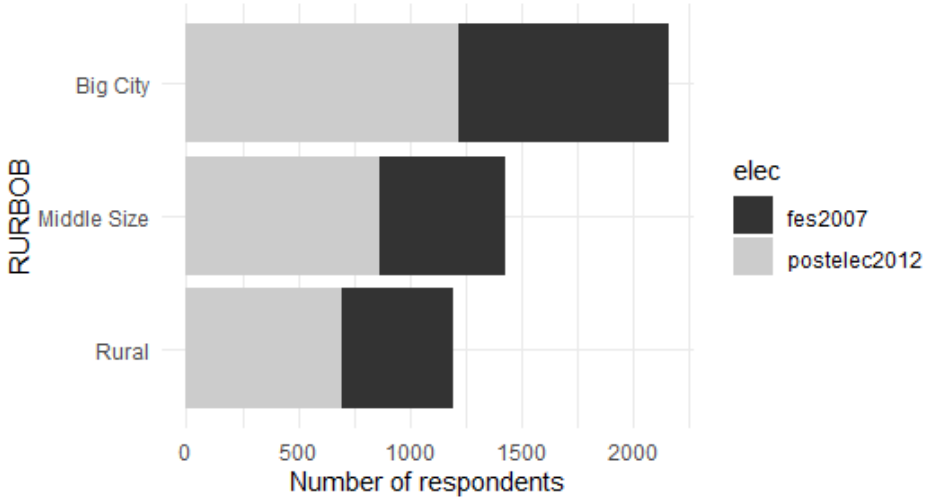
Information	[Type= discrete] [Format=numeric] [Range= 1-10] [Missing=*/666/777/888]
Statistics [NW/ W]	[Valid=2000 /-] [Invalid=0 /-]
Notes	Variable issue du système CATI (Computer Assisted Telephone Interviewing)

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%

Data harmonization

- Ex: Objective Size of Town
- OURBOB: original variables combined
- RURBOB: standardized recoded variable

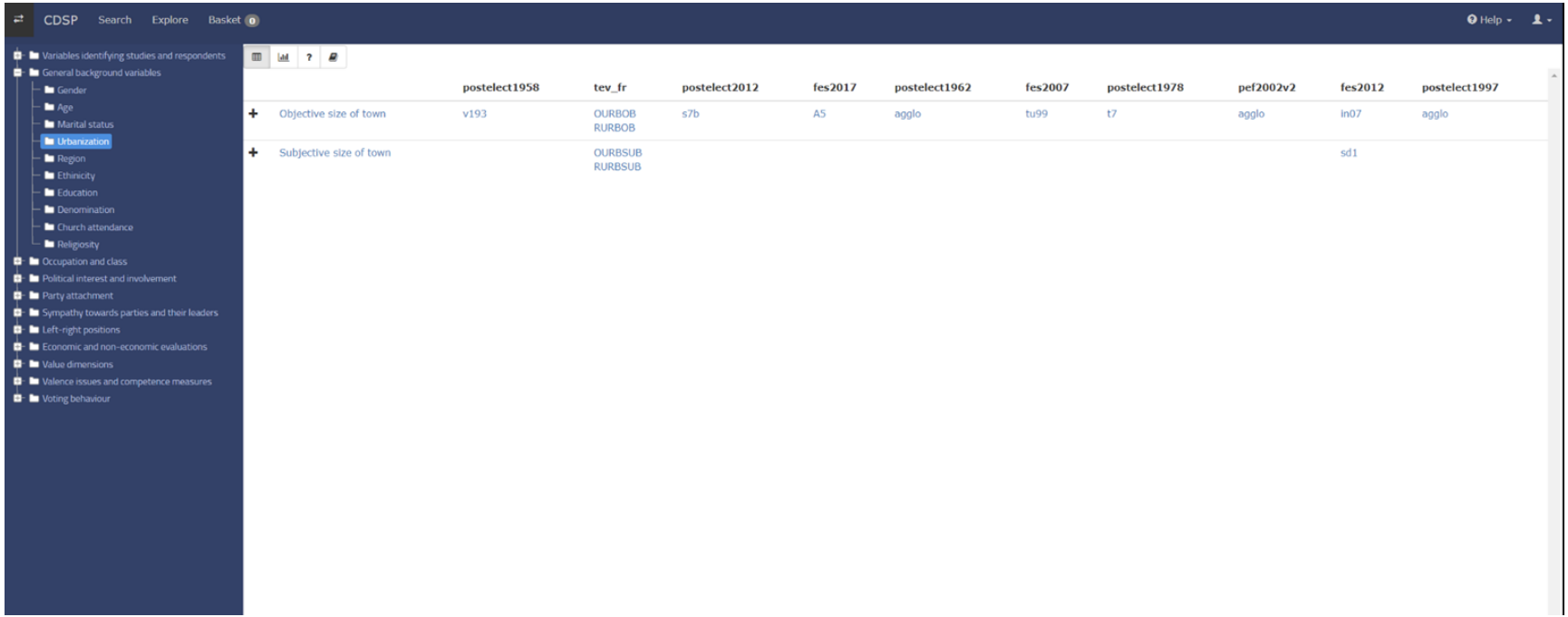




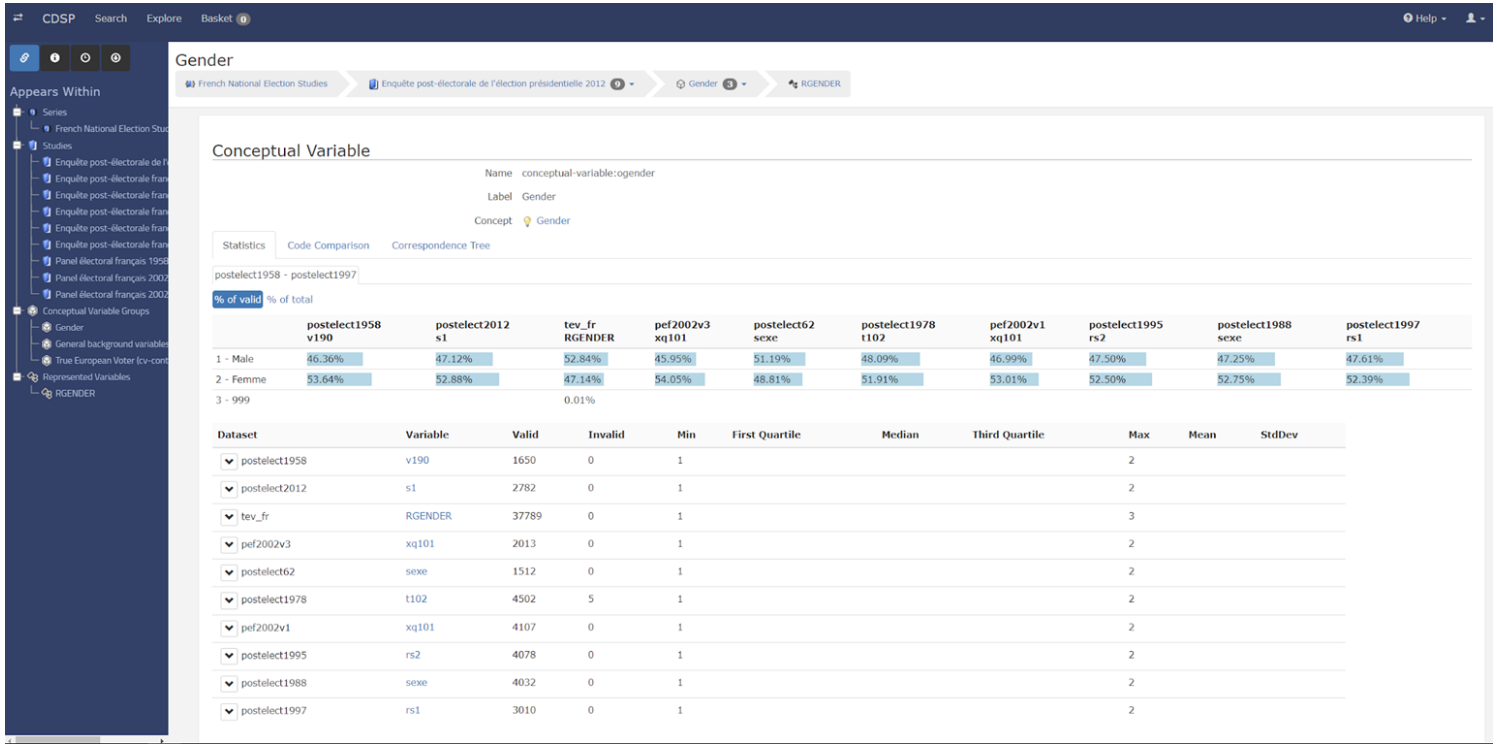
End-users' perspective: explore by concept…

	postelect1958	tev_fr	postelect2012	fes2017	postelect1962	fes2007	postelect1978	postelect1995	postelect1988	pef2002v2	fes2012	postelect1997
+ Gender	v190	RGENDER	s1	S2	sexe	r10	t102	rs2	sexe	xq101	h1	rs1
+ Age	v191		s2	S3	age	rs03	t101	rs1	r1		sd3a_rec	rs2
+ Respondent's family situation			Rs16				t41	rs3	r3	xq102		rs3a2
+ Respondent's married life				SD5		r21					sd5b	
+ Respondent's conjugal status		OMARRIED RMARRIED		SD6		r21b					sd8	
+ Respondent's matrimonial				SD7		r21c					sd9	
+ Objective size of town	v193	OURBOB RURBOB	s7b	A5	agglo	tu99	t7			agglo	in07	agglo
+ Subjective size of town		OURBSUB RURBSUB									sd1	
+ Department	v6a	RREGION OREGION		A4	depart		t6	dep		dep	in06	dep
+ Education level	v189	OEDU1 REDU	Rs8	SD9	q48	rs12	t84	rs6	r8	xq111b	sd10_rec	rs4
+ School leaving age		OEDU2	IRS7					rs7	r9	x111bi		
+ Religion		ODENOM RDENOM	Rs31	SD41		r30	t85	rs21	r28	xq117	sd46	rs12
+ Frequency of religious practice		OCHURCHA RCHURCHA	Rs32	SD40		r31	t86	rs22	r29	xq118	sd44	rs13
+ Sense of religiosity		RRELIG ORELIG				r32					sd45	
+ After the death			Rs33					rs23	q41			

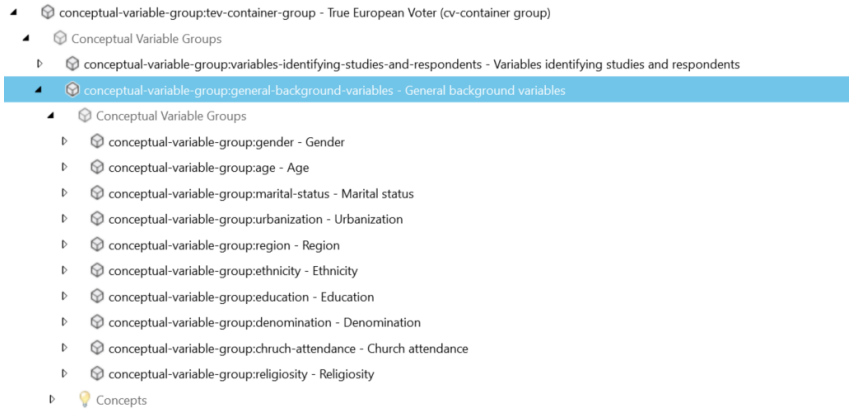
… or by sub-concept



Conceptual variable page



From metadata management perspective



Conclusion

- Relevance of
  - Research team involvement in post-harmonization
  - Conceptual framework in the management of the curation process

- Next steps::
  - Complete curation by building relation and concordance in Colectica
  - Make harmonized and multilingual metadata available on the prod Portal

Do you have any questions?

Merci beaucoup

lucie.marie2@sciencespo.fr
malo.jan@sciencespo.fr