



**ined**  
INSTITUT  
NATIONAL  
D'ÉTUDES  
DÉMOGRA  
PHIQUES

---

# Data documentation in the new round of the Generations and Gender Survey

---

Arianna Caporali (INED, [arianna.caporali@ined.fr](mailto:arianna.caporali@ined.fr))

Julie Lenoir (INED)



*2022 CSDI Workshop  
April 4-5, 2022, Condorcet Campus, Paris, France*

# Metadata in cross-national longitudinal surveys

- Key to document data quality, issues of cross-national comparability, and data harmonization processes (Mohler et al. 2010)
- Important to describe variations across waves
- Should be in compliance with international standards :
  - ✓ DDI-codebook (DDI-C or DDI 2) designed for simple surveys
  - ✓ DDI-Lifecycle (DDI-L or DDI 3)
    - Designed to document surveys across their entire life; particularly suited for longitudinal surveys
    - Contains features (i.e., represented and conceptual variables) for explicit comparisons between metadata and items of different datasets (DDI 2015; Hansen et al., 2011; Kramer et al., 2011; Mohler et al. 2010)

1. Main characteristics of the Generations and Gender Surveys (GGS)
2. Data documentation in the first round of GGS (GGS-I)
  - Main limits and needs for improvements
3. Data documentation in the second round of GGS (GGS-II)
4. Next developments

# Main characteristics of the GGS (1)

- It is a comparative survey in Europe and beyond launched by UN in 2000 and run by a consortium of research institutes (e.g., NIDI, INED)
- It studies demographic and socio-economic challenges (e.g., low fertility, population ageing) (Gauthier et al. 2018)
- It is designed as a longitudinal panel survey (3 waves; 3 years interval) on 17-79 year olds

# Main characteristics of the GGS (2)

- GGS-I (20 countries)
  - different levels of compliance to the standard questionnaire and methodology
  - *ex-post* data harmonization
- GGS-II since 2020
  - centralization of fieldwork operations at NIDI
  - implementation of the software *TMT* and *Blaise*
  - *ex-ante* data harmonization

- In compliance with DDI-C
- Fieldwork metadata collected from national teams
- Description of country-specific deviations from the GGS standard questionnaire and methodology
- Prepared and disseminated online through the software Nesstar

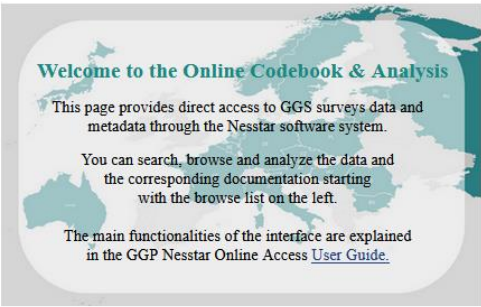
# GGG-I - Online data browsing with Nesstar

## Online Codebook & Analysis

User Guide

- GGS Pooled datasets
- GGS country data files
- GGS Variable Availability
- GGP 2020 - Wave 1 - Version 1.0

EN ?



**Welcome to the Online Codebook & Analysis**

This page provides direct access to GGS surveys data and metadata through the Nesstar software system.

You can search, browse and analyze the data and the corresponding documentation starting with the browse list on the left.

The main functionalities of the interface are explained in the GGP Nesstar Online Access [User Guide](#).

Dataset: Generations and Gender Survey Wave 2 - Consolidated

Variable **bactstat: Activity Status Respondent (W2)**

DESCRIPTIVE TEXT  
Consolidated variable.

Country-specific additional values:  
 - DEU: "1401-maternity/parental leave", "1402-employed or self-employed".  
 - FRA: "1501-in apprenticeship or paid training", "1502-employed or self-employed".  
 - GEO: "1301-on maternity leave, parental leave or child-care leave".  
 - POL: "2601-maternity leave".

Values	Categories	N	
1	student, in school, in vocational training	2211	2.8%
2	employed	35017	44.0%
3	self-employed	5923	7.4%
4	helping family member in a family business or farm	566	0.7%
5	unemployed	5384	6.8%
6	retired	21183	26.6%
7	in military or social service	18	0.0%
8	homemaker	4172	5.2%
9	maternity leave	370	0.5%
10	parental leave, care leave	750	0.9%
11	ill or disabled for a long time or permanently	3066	3.8%
12	other status	856	1.1%
1301	on maternity leave, parental leave or child-care leave	9	0.0%
1401	maternity/parental leave	37	0.0%
1402	employed or self-employed	0	0.0%
1501	in apprenticeship or paid training	16	0.0%
1502	employed or self-employed	0	0.0%
2601	maternity leave	80	0.1%
96	unknown	6517	

### SUMMARY STATISTICS

Valid cases 79658  
 Missing cases 6517  
 This variable is numeric

### NOTES

Copied from variable b801. Labels Wave 2 | = Wave 1

Dataset: Generations and Gender Survey Wave 2 - Consolidated

Activity Status Respondent (W2): Categories | Country (W2): Categories | Type: Column percentage

Country (W2)	Bulgaria	Russia	Georgia	Germany	Total
Activity Status Respondent (W2)					
student, in school, in vocational training	2.2	12.7	0.8	3.5	4.8
employed	45.6	46.0	25.4	46.5	39.9
self-employed	5.3	2.7	16.0	6.1	7.8
helping family member in a family business or farm	0.2	0.3	4.1	0.4	1.4
unemployed	11.4	2.8	19.5	5.5	10.7
retired	30.5	27.2	18.1	27.7	25.7
in military or social service	0.0	0.0	0.1	0.1	0.1
homemaker	1.1	6.2	13.2	7.7	6.8
maternity leave	0.4	0.2	0.0	0.0	0.2
parental leave, care leave	1.2	1.4	0.0	0.0	0.8
ill or disabled for a long time or permanently	1.9	0.4	2.6	0.6	1.5
other status	0.3	0.2	0.0	0.6	0.2
on maternity leave, parental leave or child-care leave	0.0	0.0	0.1	0.0	0.0
maternity/parental leave	0.0	0.0	0.0	1.1	0.1
Total	100.0	100.0	100.0	100.0	100.0
N=	9,282.8	7,723.6	8,291.2	3,226.4	28,524.0

Filter is on  
 Weight is on

- Only partially automatized
- Country-specific information entered “by hand”
- Comparisons of variables across countries and waves only through Nesstar functionalities
- Uncertainty around the future of Nesstar software



# Test of three software

---

1. Dataverse (<https://dataverse.org/>)
2. NADA Microdata Cataloging Tool (<https://nada.ihsn.org/>)
3. Colectica Designer & Portal (<https://www.colectica.com/>)

# GGP-II - Implementation of Colectica and DDI-Lifecycle

- Compatible with Blaise software
  - possibility to use the Blaise output files to document the variables
- Compliant with DDI-L standard
  - possibility to document comparability across countries and waves
- Used by other longitudinal surveys (e.g., UK longitudinal studies)



# GGG-II – Variable documentation

---

Datasets in Stata  
and SPSS format

Core questionnaire  
in Blaise format

National  
questionnaire in  
Blaise format (in  
national language)

# GGG-II – Variable documentation

Datasets in Stata  
and SPSS format

Core questionnaire  
in Blaise format

National  
questionnaire in  
Blaise format (in  
national language)



# GGG-II – Variable documentation

Datasets in Stata  
and SPSS format

Core questionnaire  
in Blaise format

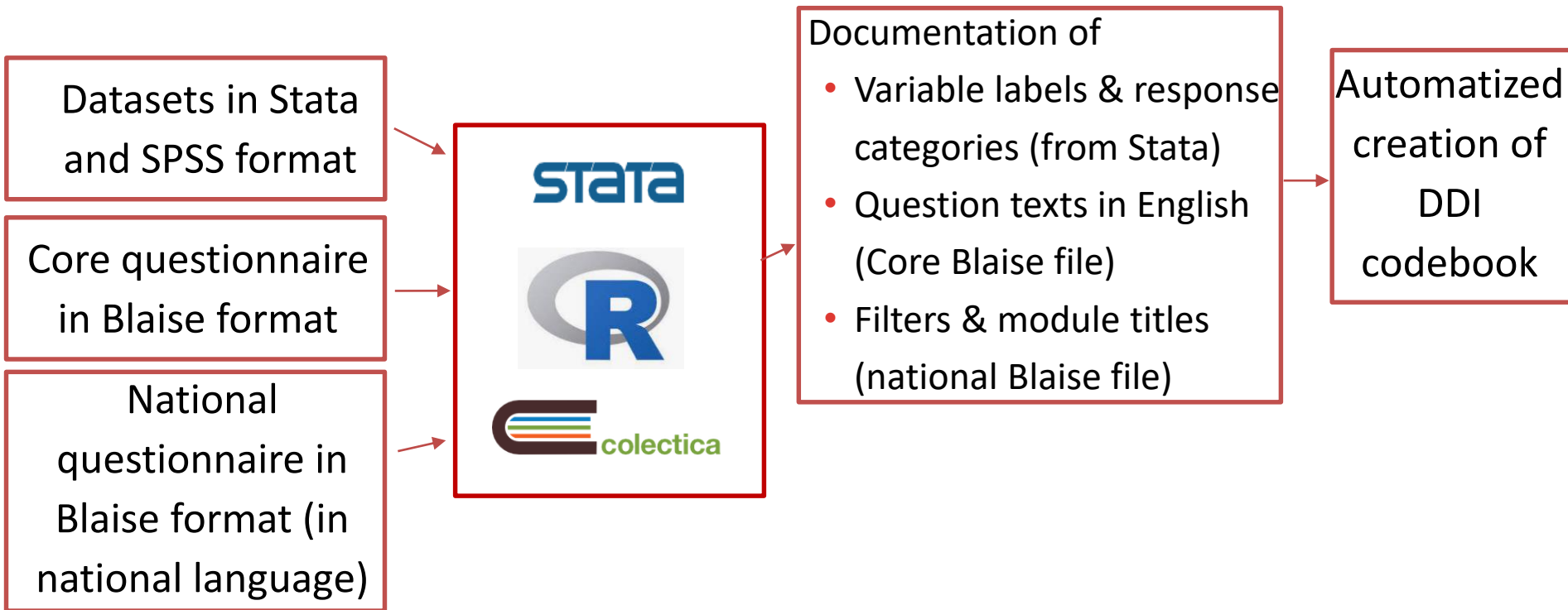
National  
questionnaire in  
Blaise format (in  
national language)



Documentation of

- Variable labels & response categories (from Stata)
- Question texts in English (Core Blaise file)
- Filters & module titles (national Blaise file)

# GGG-II – Variable documentation



# GGG-II – Variable documentation



Metadata Input sheet

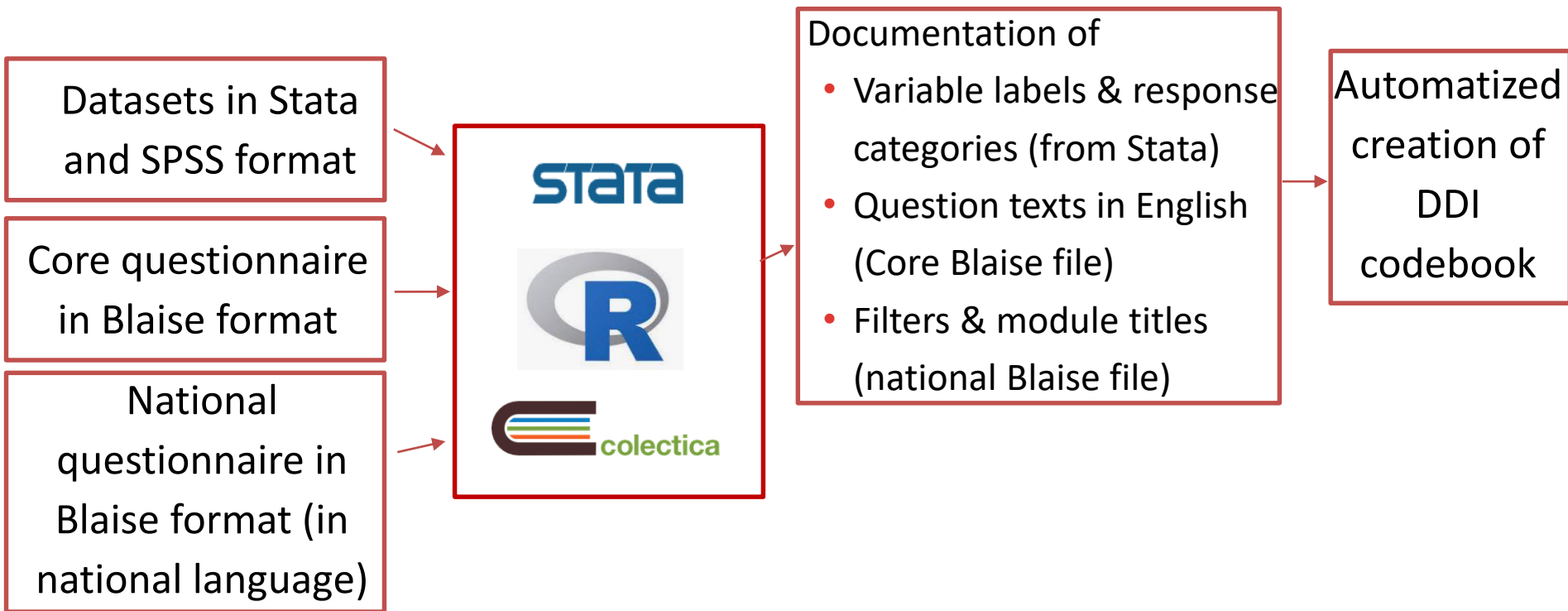
1	A	B	C	D	E	F	G	H	I	J
	Name	Label	Description	Type	Measurement	Question Text	Codes	Missing Value Codes	Topic	custom-Condition
2	country	Country of Residence	Country of Resident Code				20, Norway   a, Dor a, Don't know   b, Refusal   c, No		INTRODUCTION	
3	weight	Individual weight	Country specific ad	Numeric					INTRODUCTION	
4	respid	Respondent ID	NOR: variable gene	Numeric					INTRODUCTION	
5	mode	Mode of survey	Variable generated Code				1, face to face   2, v a, Don't know   b, Refusal   c, No		INTRODUCTION	
6	instrument	Questionnaire version	NOR: variable gene	Numeric					INTRODUCTION	
7	intdatem	Month of Interview	Variable generated Code				1, January   2, Febr a, Don't know   b, Refusal   c, No		INTRODUCTION	
8	intdatay	Year of Interview	Variable generated	Numeric			a, Don't know   b, a, Don't know   b, Refusal   c, No		INTRODUCTION	
9	dem01	Gender	Code			INTRODUCTION TO 1, Male   2, Female a, Don't know   b, Refusal   c, No			DEMOGRAPHICS	
10	dem02m	Date of Birth - M	Code			When were you bo 1, January   2, Febr a, Don't know   b, Refusal   c, No			DEMOGRAPHICS	
11	dem02y	Date of Birth - Y	Numeric			When were you bo a, Don't know   b, a, Don't know   b, Refusal   c, No			DEMOGRAPHICS	
12	dem03	Born in country	Code			Were you born in t 1, Yes   2, No   a, D a, Don't know   b, Refusal   c, No			DEMOGRAPHICS	
13	dem04a	Place of birth	Only country specifi	Code		In which region wa 2001, Agder   2002, a, Don't know   b, Refusal   c, No			DEMOGRAPHICS	DEM03 - yes
14	dem04b	Country of Birth	Numeric			In which country were you born? Please type in (the first letters of) the co			DEMOGRAPHICS	
15	dem04bis	Country of Birth (ISO C00)	Generated variable	Numeric		In which country w a, Don't know   b, a, Don't know   b, Refusal   c, No			DEMOGRAPHICS	
16	dem05m	Date of Immigration - M	Code			When did you first 1, January   2, Febr a, Don't know   b, Refusal   c, No			DEMOGRAPHICS	DEM03 - no
17	dem05y	Date of Immigration - Y	Numeric			When did you first a, Don't know   b, a, Don't know   b, Refusal   c, No			DEMOGRAPHICS	DEM03 - no
18	dem06	Activity Status	Country specific ad	Code		Which of the items 1, In education or t a, Don't know   b, Refusal   c, No			DEMOGRAPHICS	
19	dem07	Education Level (ISCED 20 Only country specifi	Code			What is the highest 0, ISCED0 Baneskol a, Don't know   b, Refusal   c, No			DEMOGRAPHICS	Educat
20	dem08m	Date Education Reached - M	Code			When did you reac 1, January   2, Febr a, Don't know   b, Refusal   c, No			DEMOGRAPHICS	Educat NOT(DEM07 = ISCED0)
21	dem08y	Date Education Reached - Y	Numeric			When did you reac a, Don't know   b, a, Don't know   b, Refusal   c, No			DEMOGRAPHICS	Educat NOT(DEM07 = ISCED0)
22	dem09	Number of Rooms	Numeric			//Accommodation   a, Don't know   b, a, Don't know   b, Refusal   c, No			DEMOGRAPHICS	Accom



GGG Stata file

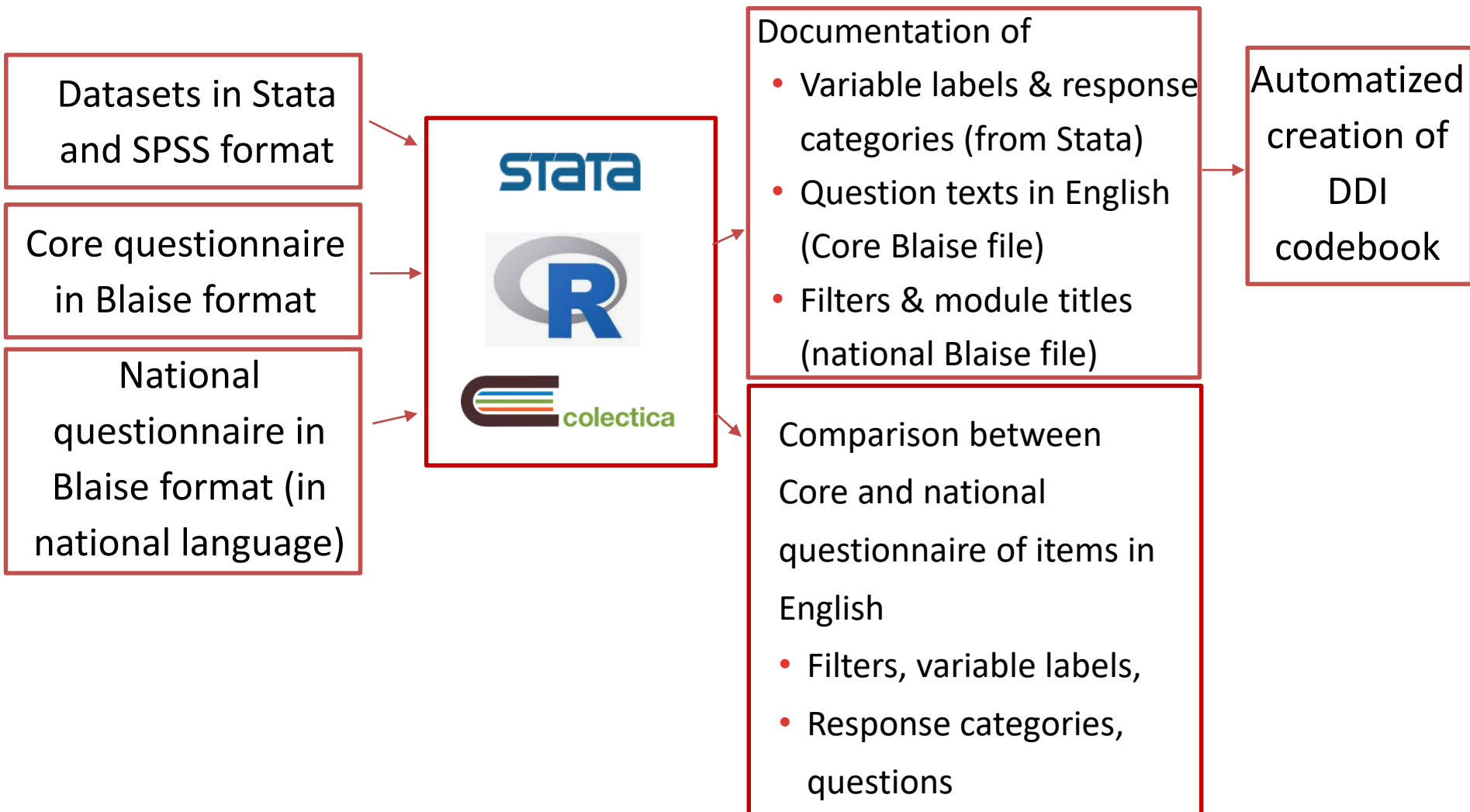


# GGG-II – Variable documentation

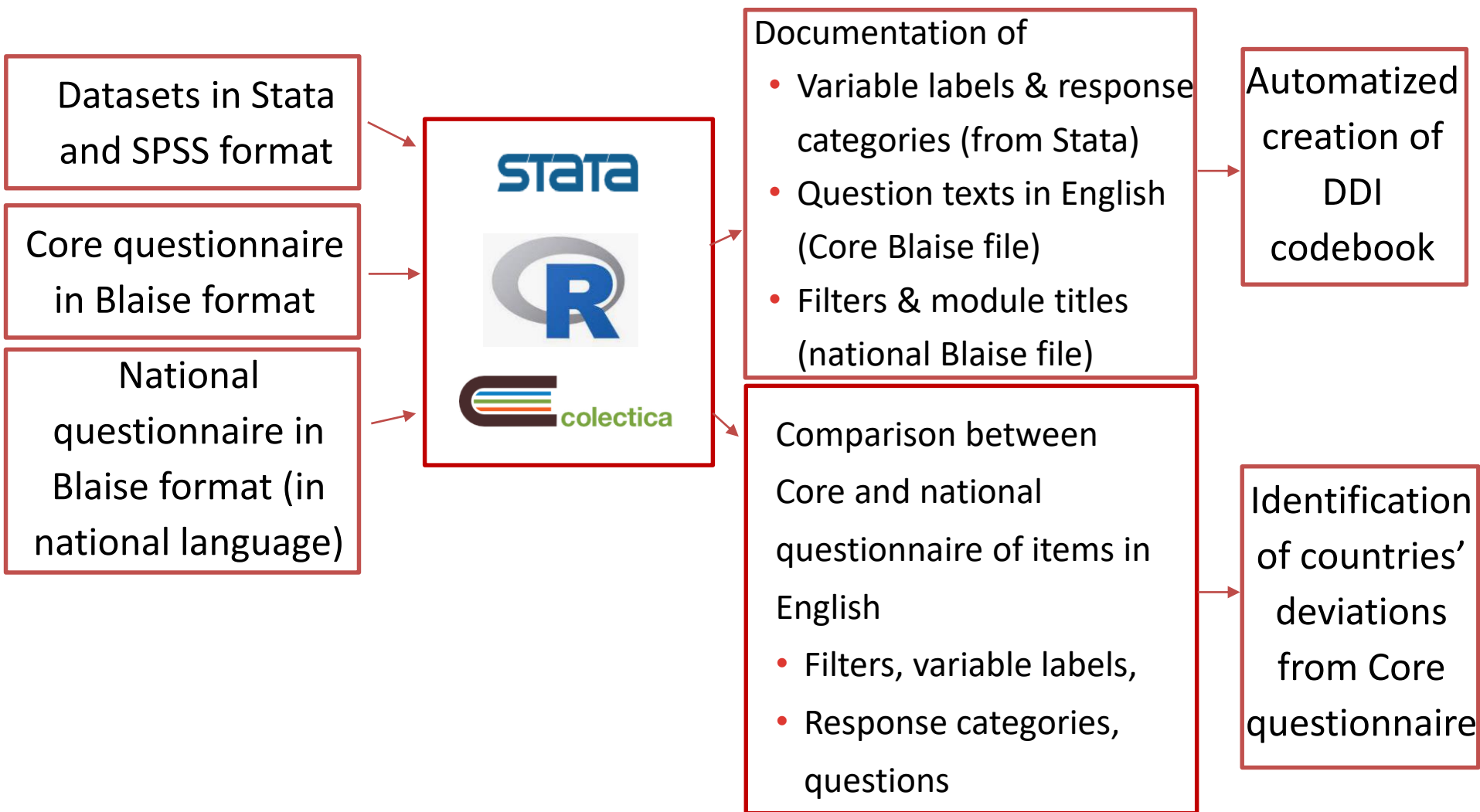




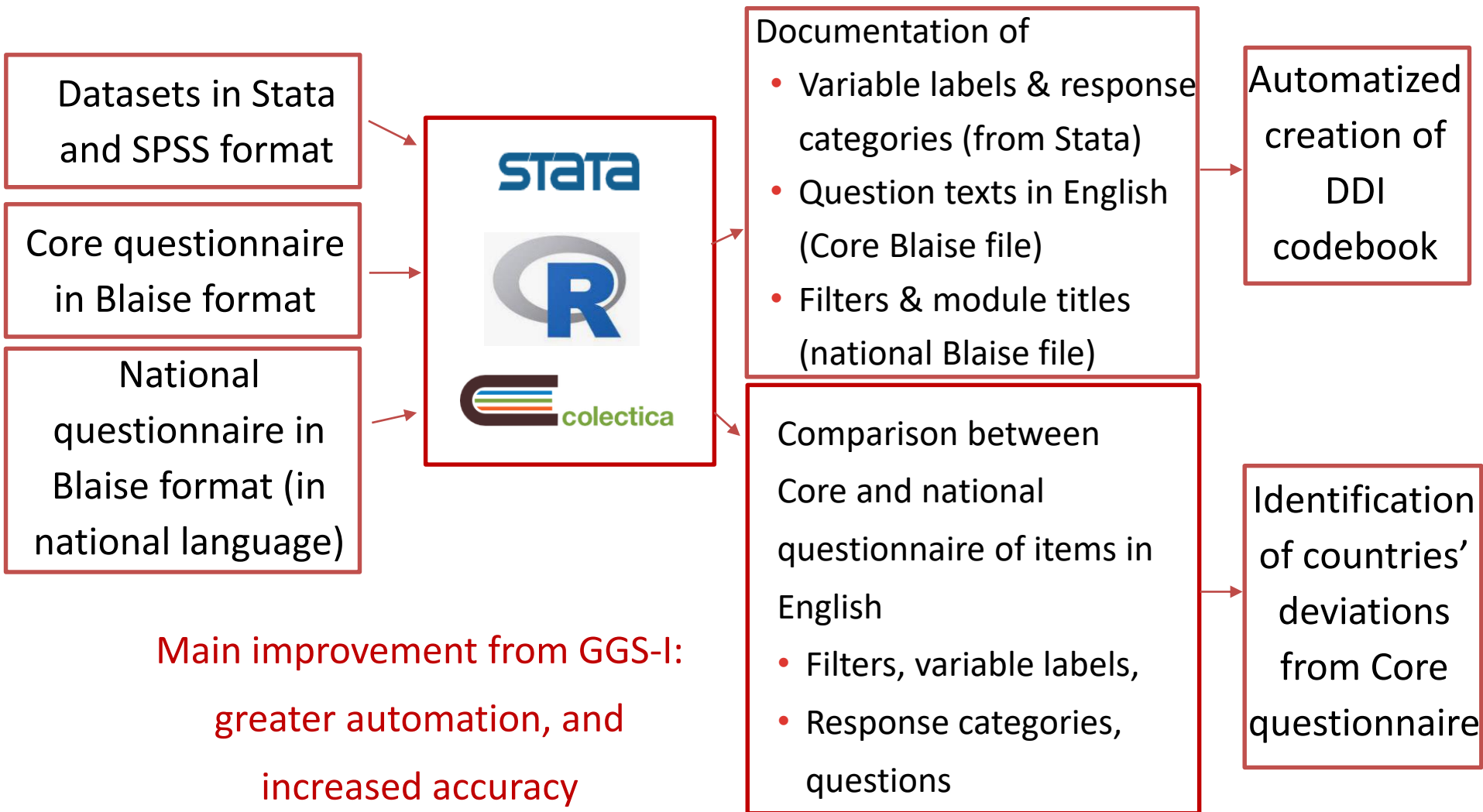
# GGG-II – Variable documentation



# GGG-II – Variable documentation



# GGG-II – Variable documentation

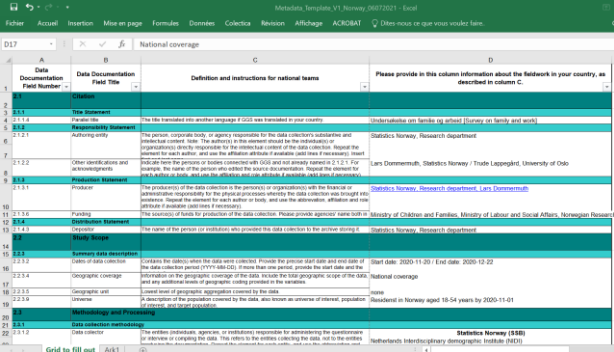


# GGG-II – Survey documentation

- Information on fieldworks provided by the national teams in a template in Excel format,

e.g. :

- ✓ Sampling, mode of data collection, etc.
- ✓ References to related material
- ✓ Data processing and weights



Data Documentation Field Number	Data Documentation Field Title	Definition and instructions for national teams	Please provide in this column information about the fieldwork in your country, as described in column C.
1	1.1.1	1.1.1.1	
2	1.1.2	1.1.2.1	
3	1.1.3	1.1.3.1	
4	1.1.4	1.1.4.1	
5	1.1.5	1.1.5.1	
6	1.1.6	1.1.6.1	
7	1.1.7	1.1.7.1	
8	1.1.8	1.1.8.1	
9	1.1.9	1.1.9.1	
10	1.1.10	1.1.10.1	
11	1.1.11	1.1.11.1	
12	1.1.12	1.1.12.1	
13	1.1.13	1.1.13.1	
14	1.1.14	1.1.14.1	
15	1.1.15	1.1.15.1	
16	1.1.16	1.1.16.1	
17	1.1.17	1.1.17.1	
18	1.1.18	1.1.18.1	
19	1.1.19	1.1.19.1	
20	1.1.20	1.1.20.1	
21	1.1.21	1.1.21.1	
22	1.1.22	1.1.22.1	



# GGP Colectica Portal

<https://ggp.colectica.org/>



The screenshot shows the home page of the GGP Colectica Portal. The navigation bar at the top includes 'GGP', 'Search', 'Explore', and 'Baskets 0'. On the right, there are links for 'Help' and 'Feedback'. The main content area is titled 'Home' and features a 'Welcome' section with a message about the portal's purpose and a search bar. Below the search bar, there are six statistics cards: 2 Projects, 5 Series, 25 Studies, 39 Data Files, 45,514 Questions, and 57,003 Variables. At the bottom, there is a banner for 'The Generations and Gender Survey - GGS Round 1 (2004-2011)'. Logos for GGP, DDI, and Colectica are also visible.

GGP Search Explore Baskets 0 Help Feedback

## Home

Welcome

Welcome to the Generations and Gender Programme (GGP) Colectica Portal [Beta Version](#).

Through this portal, you can search and browse Generations and Gender Survey (GGS) data and the corresponding documentation.

The Portal uses [Data Documentation Initiative metadata standards](#) and is powered by [Colectica software](#). Click [here](#) or on the Help tab on the right top of this page to get the main functionalities of the interface.

Click [here](#) to download the GGS microdata files.

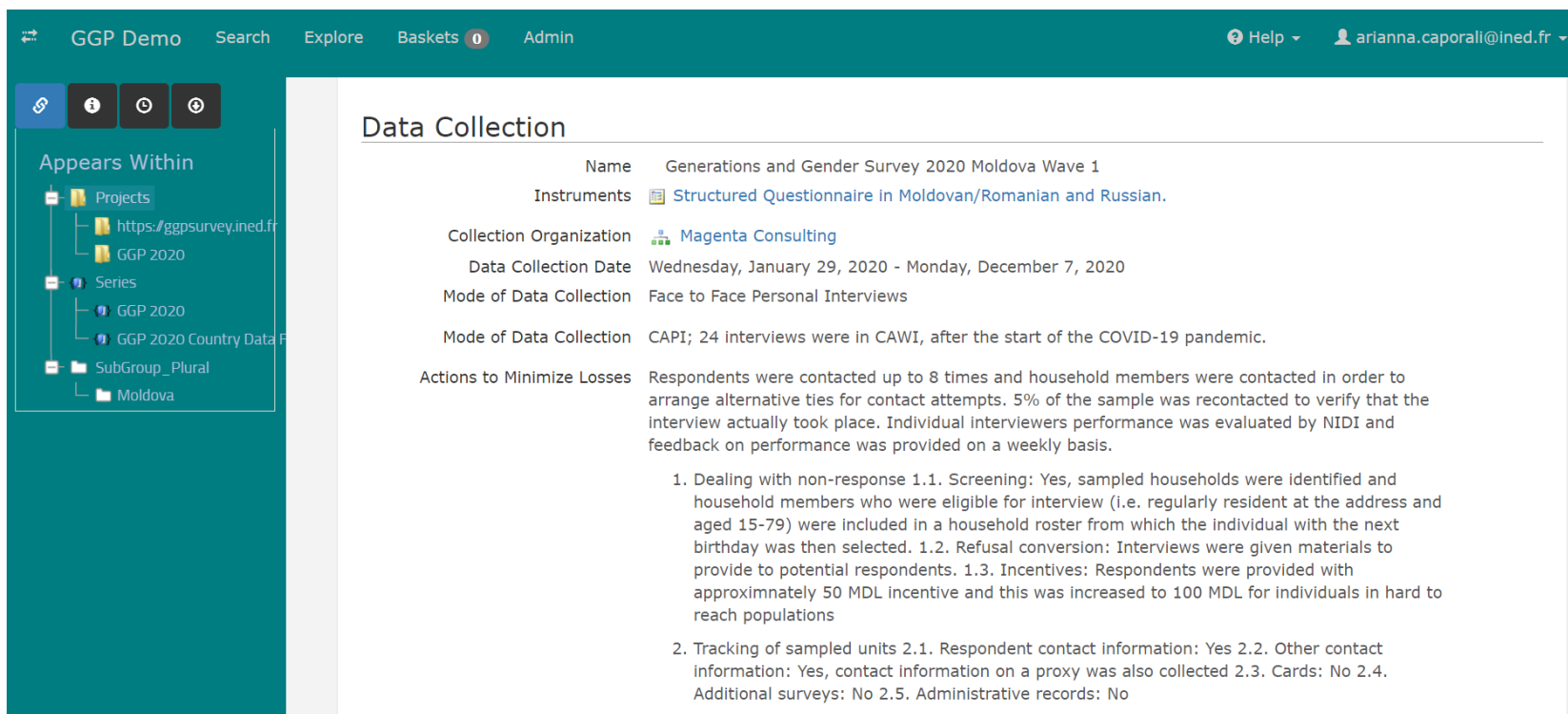
Generation and Gender Programme

Search

- 2 Projects
- 5 Series
- 25 Studies
- 39 Data Files
- 45,514 Questions
- 57,003 Variables

The Generations and Gender Survey - GGS Round 1 (2004-2011)

# Example of online survey metadata



The screenshot displays the 'Data Collection' metadata page in the GGP Demo application. The interface includes a top navigation bar with 'GGP Demo', 'Search', 'Explore', 'Baskets', and 'Admin'. A user profile 'arianna.caporali@ined.fr' is visible in the top right. On the left, a sidebar shows a tree view under 'Appears Within' with categories: Projects (https://ggpsurvey.ined.fr, GGP 2020), Series (GGP 2020, GGP 2020 Country Data F), and SubGroup\_Plural (Moldova). The main content area lists the following metadata:

Name	Generations and Gender Survey 2020 Moldova Wave 1
Instruments	Structured Questionnaire in Moldovan/Romanian and Russian.
Collection Organization	Magenta Consulting
Data Collection Date	Wednesday, January 29, 2020 - Monday, December 7, 2020
Mode of Data Collection	Face to Face Personal Interviews
Mode of Data Collection	CAPI; 24 interviews were in CAWI, after the start of the COVID-19 pandemic.
Actions to Minimize Losses	<p>Respondents were contacted up to 8 times and household members were contacted in order to arrange alternative ties for contact attempts. 5% of the sample was recontacted to verify that the interview actually took place. Individual interviewers performance was evaluated by NIDI and feedback on performance was provided on a weekly basis.</p> <ol style="list-style-type: none"><li>1. Dealing with non-response 1.1. Screening: Yes, sampled households were identified and household members who were eligible for interview (i.e. regularly resident at the address and aged 15-79) were included in a household roster from which the individual with the next birthday was then selected. 1.2. Refusal conversion: Interviews were given materials to provide to potential respondents. 1.3. Incentives: Respondents were provided with approximately 50 MDL incentive and this was increased to 100 MDL for individuals in hard to reach populations</li><li>2. Tracking of sampled units 2.1. Respondent contact information: Yes 2.2. Other contact information: Yes, contact information on a proxy was also collected 2.3. Cards: No 2.4. Additional surveys: No 2.5. Administrative records: No</li></ol>

# Example of online variable metadata



< fer13

GGP2020\_WAVE1\_MDA\_V\_0.1. (1477 of 1791)

## Variable Description

Name fer14  
 Label Intention to have a child in next 3 years  
 Description The response category "Currently expecting a child" was not fielded in MDA.  
 Condition NOT(FER05 = defno) age < 50  
 Dataset [GGP2020\\_WAVE1\\_MDA\\_V\\_0\\_1](#)

fer14

### Summary Statistics for 'fer14'

Value	Valid	Label	Frequency	% of Valid	% of All				
.	✘		5,994		59.65%				
.a	✘	Don't know	93		0.93%				
.b	✘	Refusal	18		0.18%				
.c	✘	Not Applicable	0		0.00%				
1	✓	Definitely not	1,176	29.83%	11.70%				
2	✓	Probably not	659	16.71%	6.56%				
3	✓	Unsure	722	18.31%	7.19%				
4	✓	Probably yes	798	20.24%	7.94%				
5	✓	Definitely yes	588	14.91%	5.85%				
<b>Total Responses</b>	<b>Valid</b>	<b>Invalid</b>	<b>Min</b>	<b>Max</b>	<b>Mean</b>	<b>StdDev</b>	<b>Variance</b>	<b>Skewness</b>	<b>Kurtosis</b>
10048	3943	6105	1	5	2.737002282526	1.44581407085888	2.09037832749352	0.168648704167343	-1.35364908522416

# Example of online comparison between two variables (ongoing work)

## Concordance

GGP 2020 Country Data File		
	Norway	Moldova
	Generations and Gender Survey 2020 Norway Wave 1	Generations and Gender Survey 2020 Moldova Wave 1
	GGP2020_WAVE1_NOR_V_0_1	GGP2020_WAVE1_MDA_V_0.1.
	fer14	fer14
Don't know	.a	.a
Refusal	.b	.b
Not Applicable	.c	.c
Not at all	.f	
Does not work or homemaker	.g	
Incomplete survey	.h	
Definitely not	1	1
Probably not	2	2
Unsure	3	3
Probably yes	4	4
Definitely yes	5	5
Pregnant	6	



# Next developments

---

- Finishing up the migration of GGS-I DDI files
- Implementation of DDI-L fields to document variables across survey
- Harvest of GGP documentation by other data catalogues (e.g. CESSDA catalogue)

# References

- DDI (2015). Data Documentation Initiative (DDI) Technical Specification. Part II: User Guide. Version 3.2, 2014-02-15. © 2015 DDI Allianc. Available at: [https://ddialliance.org/Specification/DDI-Lifecycle/3.2/XMLSchema/HighLevelDocumentation/DDI\\_Part\\_II\\_UserGuide.pdf](https://ddialliance.org/Specification/DDI-Lifecycle/3.2/XMLSchema/HighLevelDocumentation/DDI_Part_II_UserGuide.pdf)
- Emery T. and Caporali C. (in prep). The Added Value of Cross-National Studies: Compliance and Usage in the GGP.
- Fokkema T., Kveder A., Hiekel N., Emery T., and Liefbroer A. C. (2016). Generations and Gender Programme Wave 1 data collection: An overview and assessment of sampling and fieldwork methods, weighting procedures, and cross-sectional representativeness. *Demographic Research* 34(18): 499-524. DOI: 10.4054/DemRes.2016.34.18.
- GAUTHIER, Anne H.; CABAÇO, Susana Laia Farinha; EMERY, Tom. Generations and Gender Survey study profile. *Longitudinal and Life Course Studies*, [S.l.], v. 9, n. 4, p. 456-465, oct. 2018. ISSN 1757-9597. Available at: <<https://www.llcsjournal.org/index.php/llcs/article/view/500>>. Date accessed: 04 apr. 2022. doi:<http://dx.doi.org/10.14301/llcs.v9i4.500>.
- Hansen S.E., Iverston J., Jensen U., Orten H., Vompras J. (2011). Enabling Longitudinal Data Comparison Using DDI, *DDI Working Paper Series – Longitudinal Best Practice*, 2.
- Kramer, S., Banks, R., Chang, V., Sieber, I., Vardigan, M., et Zenk-Möltgen, W. (2011). Presenting longitudinal studies to end users effectively using DDI metadata. *DDI Working Paper Series – Longitudinal Best Practice*, 4. DOI: 10.3886/DDILongitudinal04
- Mohler, P.P., Hansen, S.E., Pennell, B.E., Thomas, W., Wackerow, J., and Hubbard, F. (2010). A survey process quality perspective on documentation. In: Harkness, J.A., Braun, M., Edwards, B., Johnson, T.P., Lyberg, L.E., Mohler, P.P., Pennell, B.-E., and Smith, T.W. (eds.). *Survey Methods in Multinational, Multiregional, and Multicultural Contexts*. Hoboken, N.J.: Wiley & Sons: 299–314. DOI: 10.1002/9780470609927.ch15.

# Thank you!

Arianna Caporali

[arianna.caporali@ined.fr](mailto:arianna.caporali@ined.fr)

