



HAL
open science

Strategy to document and disseminate longitudinal surveys: case study of using DDI-Lifecycle and Colectica

Lucie Marie

► To cite this version:

Lucie Marie. Strategy to document and disseminate longitudinal surveys: case study of using DDI-Lifecycle and Colectica. IASSIST 2022 "Data by Design: Building a Sustainable Data Culture" (IASSIST 2022), Jun 2022, Gothenburg, Sweden. <10.5281/zenodo.6646079>. <hal-03956495>

HAL Id: hal-03956495

<https://sciencespo.hal.science/hal-03956495v1>

Submitted on 25 Jan 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons CC BY-NC-ND 4.0 - Attribution - Non-commercial use - No Derivative Works - International License

Strategy to document and disseminate longitudinal surveys: case study of using DDI- Lifecycle and Colectica

[Lucie MARIE](#)

Center for Socio-Political Data (CDSP)

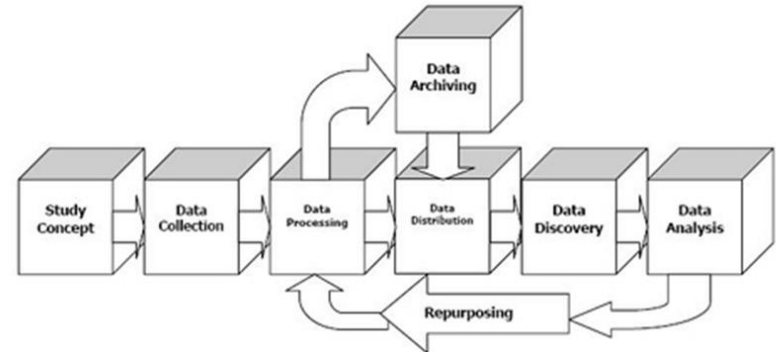
*IASSIST 2022, Data by Design - Data Management and Archiving
Gothenburg - June 10, 2022*

Agenda

- DDI-Lifecycle principles
- Why does DDI-L suit longitudinal surveys data?
- UpMet experimentation
- Conclusion and next steps

What is DDI-Lifecycle?

- Data Documentation Initiative
 - Open standard
 - 3 versions: Codebook, Lifecycle and Cross Domain Integration
- From survey conception to data repurposing
- Reuse and connect information
- Data provenance, lineage and concordance



Source: Thomas, Gregory, & Piazza, 2005

Fika, the Swedish coffee break

- Study concept: Fika consumption seasonality
- Survey instrument: questionnaire to collect data
 - Question: “How often do you have Fika per day?”
 - Never
 - Once per day
 - Twice per day
 - More than twice per day
- Variable to measure concept: Fika consumption frequency (fika_freq)
- A longitudinal survey: same data collected several time among the same sample of individuals
 - fika_freq_jan, fika_freq_feb, fika_freq_mar, etc.



Describe your data and reuse information



Study concept

Fika consumption seasonality



Data capture

*Questionnaire: How often do you have
Fika per day?*



Data processing

Data variable
fika_freq_jan

Data variable
fika_freq_feb

Data variable
fika_freq_mar



DDI-L describe each stage of the data lifecycle

- From study concept to dataset variable level
- More than 500 specifications

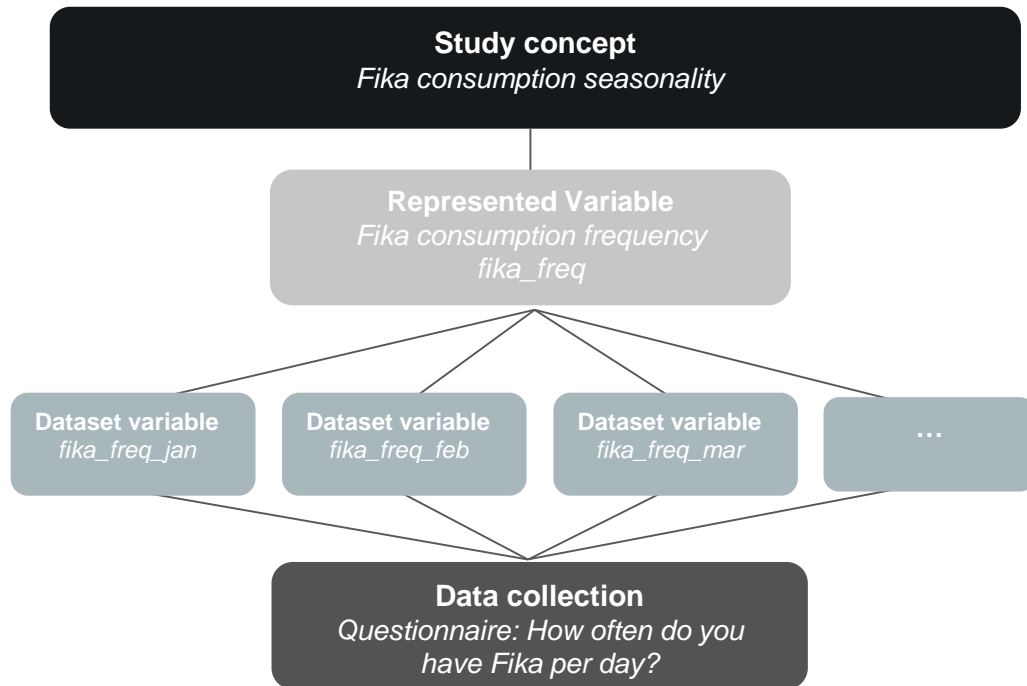
Metadata reusability: unique xml item reusing several times

- Survey information: eg. Author(s), producer(s), mode of collection
- Variable information: eg. data type, codelist, etc.

Advantages

- Increase data homogeneity
- Less time consuming

Describe lineage and concordance



Data concordance

Connect variables to concepts

Data lineage

Provide data sources

Advantages

- Rich documentation
- Interact with information in multiple ways

When should you use DDI-Lifecycle?

Longitudinal or repeated surveys (eg. Comparative program)

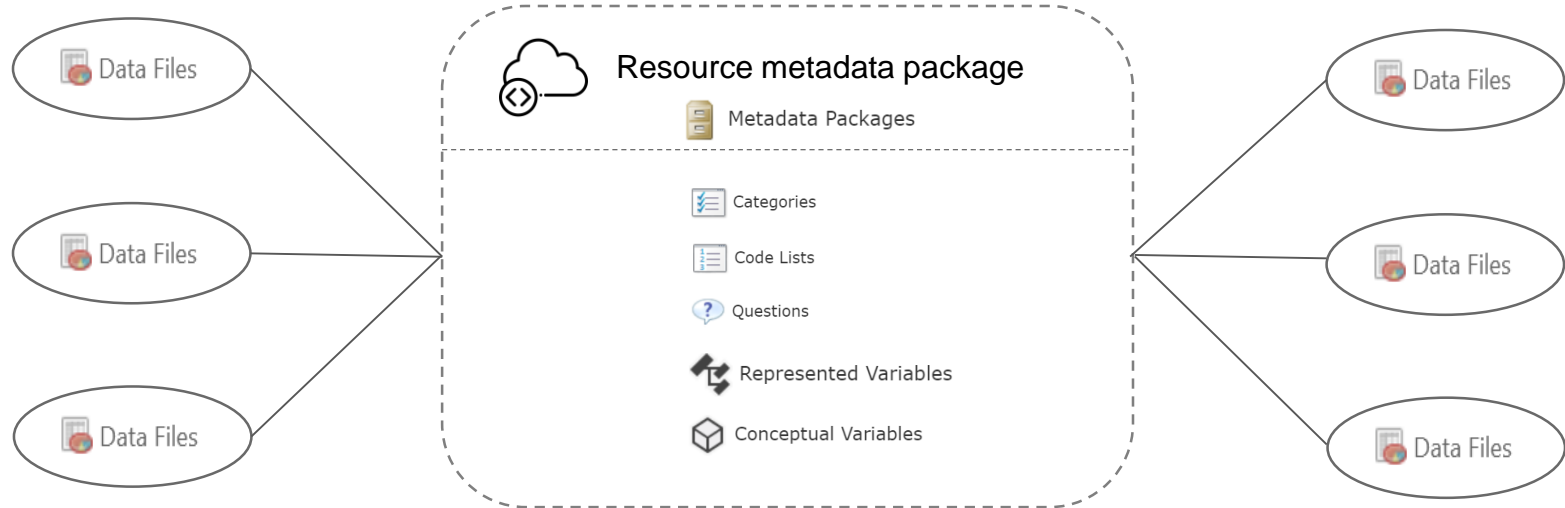
- At an early stage of conception
- Designed by concepts
- Backed by steady project resources

UpMet project


- “Upscaling metadata for increasing reuse in the social sciences”
- WP2: Question and variable bank in DDI-Lifecycle
- Colectica tools (Designer, Repository and Portal)



Set up a post-harmonisation workflow



Main outcomes

- Post harmonization of the French Political Barometer (CEVIPOF)
- Guidelines: how to document surveys in DDI-L with Colectica
- Training research team and hosting the resulting metadata
- Harvested by the  CESSDA Euro Question Bank
- Portal public access: <https://explore.cdsp.sciences-po.fr/>

Home page

The screenshot displays the CDSP Home page. At the top, a dark blue navigation bar contains the CDSP logo, links for Search, Explore, and Basket (with a 0 count), and a Help menu. Below the navigation bar, the page title "Home" is visible. A search bar is located at the top of the main content area. The dashboard features eight data cards arranged in a 2x4 grid, each with a colored circular icon, a count, and a category name. The categories and their counts are: Series (3), Studies (8), Data Files (6), Variables (1,490), Instruments (6), Questions (547), Represented Variables (205), and Conceptual Variables (181).

Category	Count
Series	3
Studies	8
Data Files	6
Variables	1,490
Instruments	6
Questions	547
Represented Variables	205
Conceptual Variables	181

Variable description and representation

The screenshot displays the CDSP (Canadian Data System) interface. On the left is a navigation sidebar with a list of variables: QUEST, VAGUE, DATE_TER, DPT, REG, CCM, AGGLO, GR, REGA, NATIO, INSCR, SEXE, ANNEE, RAGE, AGE, and RS1. The main content area is titled 'Variable Description' and shows details for the variable 'INSCR' from the dataset 'BPF2007-R1 (12 of 236)'. Below the title is a table with columns: Value, Label, Frequency, % of valid, and % of all. The table shows two rows: '1' for 'Homme' (2,704, 47.86%) and '2' for 'Femme' (2,946, 52.14%). Below this is a summary table with columns: Valid, Invalid, Min, and Max. The summary table shows 5650 valid values, 0 invalid values, a minimum of 1, and a maximum of 2. The 'Representation' section shows the variable type as 'Code List', selection style as 'SelectOne', and codes as 'sociodemo.sex000001'. It also lists the values '1' (Homme) and '2' (Femme) with checkboxes. Other metadata includes 'Blank values represent missing values' (True), 'Role' (Input), 'Aggregation Method' (Unspecified), 'Temporal' (False), 'Geographic' (False), and 'Represented Variable' (SEXE000001). The 'Source Questions' section is partially visible at the bottom.

CDSP Search Explore Basket 0 Help

No groups

- + QUEST Identifiant
- + VAGUE Vague d'enquête
- + DATE_TER Date de terrain
- + DPT Département
- + REG Région
- + CCM Catégorie de commune selon espace Urbain / Rural
- + AGGLO Catégorie d'agglomération
- + GR Grande région
- + REGA Région administrative
- + NATIO Nationalité
- + INSCR Inscription sur listes électorales
- + SEXE Sexe
- + ANNEE Année de naissance
- + RAGE Tranche d'âge
- + AGE Age
- + RS1

< INSCR BPF2007-R1 (12 of 236) ANNEE >

Variable Description

Value	Label	Frequency	% of valid	% of all
1	Homme	2,704	47.86%	47.86%
2	Femme	2,946	52.14%	52.14%

Valid	Invalid	Min	Max
5650	0	1	2

Representation

Type Code List
Selection Style SelectOne
Codes sociodemo.sex000001

- 1 Homme
- 2 Femme

Blank values represent missing values True
Role Input
Aggregation Method Unspecified
Temporal False
Geographic False
Represented Variable SEXE000001

Source Questions

Variable provenance and lineage

The screenshot displays the CDSP (Canadian Data System Platform) interface. The top navigation bar includes 'CDSP', 'Search', 'Explore', and 'Basket'. The left sidebar shows a tree view under 'Appears Within' with categories like 'Series', 'Studies', and 'Instruments'. The main content area is titled 'RS1' and shows the variable's lineage through three questionnaires. The 'Question' section details the variable's name, text, type, selection style, and codes. The 'Usage' section lists the questionnaires where the variable is used, with a list of questions below. The question 'RS1' is highlighted with a red border.

CDSP Search Explore Basket

Help

RS1

Baromètre Politique Français 2006-2007 Baromètre Politique Français 2006-2007 Vague 1 Questionnaire - Baromètre Politique Français - Vague 1

Question

Name RS1

Question Text Sexe

Type Code List

Selection Style SelectOne

Codes sociodemo.sex000001

- 1 Homme
- 2 Femme

Usage

Questionnaire - Baromètre Politique Français - Vague 3 Questionnaire - Baromètre Politique Français - Vague 1 Questionnaire - Baromètre Politique Français - Vague 4

Questionnaire - Baromètre Politique Français - Vague 2

Q0b Etes-vous inscrit(e) sur les listes électorales pour pouvoir voter ?

RS1 Sexe

RS2 En quelle année êtes-vous né(e) ?

63 questions after...

Concordance section

Concordance

Statistics

Code Comparison

Correspondence Tree

BPF2007-R4 - BPF2007-R1

% of valid % of total

	BPF2007-R4 RS15	BPF2007-R3 RS15	BPF2007-R2 RS15	BPF2007-R1 RS15
1 - Catholique	64.84%	66.07%	65.63%	64.46%
2 - Protestante	1.58%	1.91%	1.63%	1.95%
3 - Juive	0.50%	0.46%	0.44%	0.48%
4 - Musulmane	2.69%	2.16%	2.07%	2.60%
5 - Bouddhiste	0.40%	0.36%	0.41%	0.42%
6 - Autre religion	0.86%	1.09%	0.74%	1.03%
7 - Sans religion	28.99%	27.81%	28.99%	28.85%

Code comparison and correspondence tree

Concordance

Statistics

Code Comparison

Correspondence Tree

BPF2007-R4 - BPF2007-R1

This representation is used
by 4 variables.

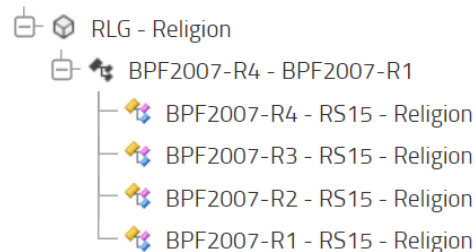
- 1 Catholique
- 2 Protestante
- 3 Juive
- 4 Musulmane
- 5 Bouddhiste
- 6 Autre religion
- 7 Sans religion

Concordance

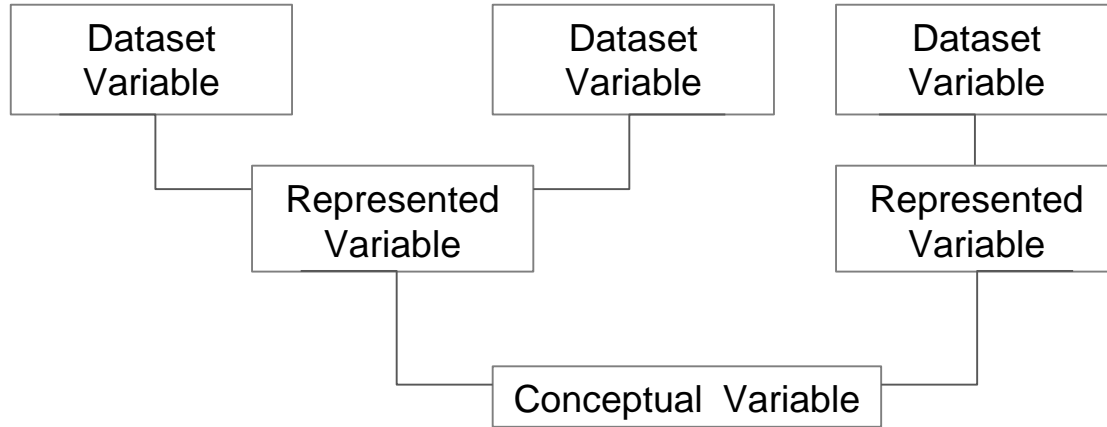
Statistics

Code Comparison

Correspondence Tree

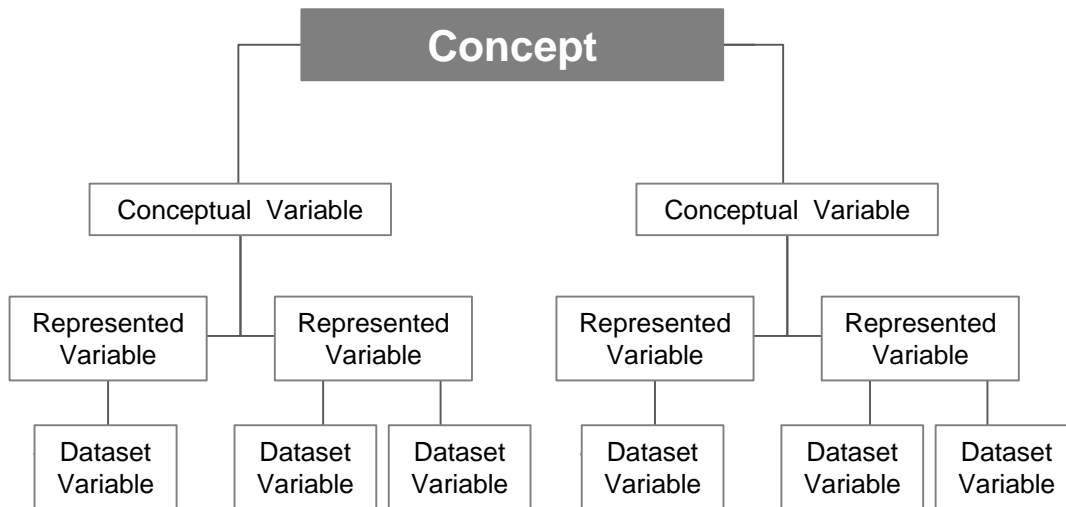


Current concordance architecture

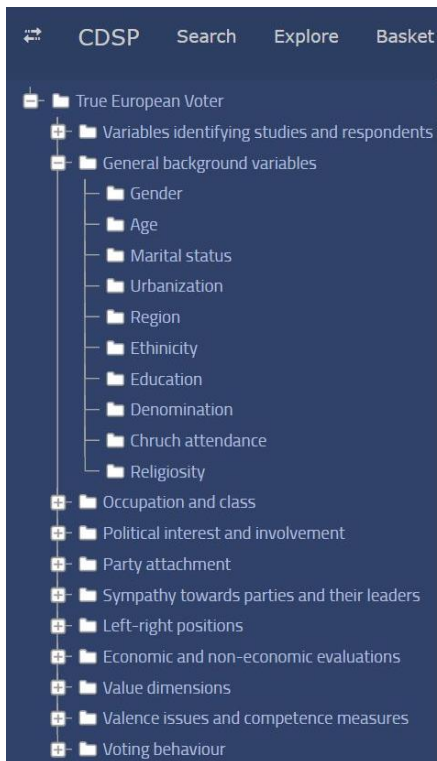


What's next?

- Scale up to **concept system**
- Concept: Units of thought
- Facilitate variable data **discovery** and **exploration**



What's next?



- Reusing existing ontology from: Schmitt, H. (2021). *The True European Voter* (1.0.0) [Data set]. GESIS Data Archive. <https://doi.org/10.4232/1.13601>
- Post harmonization of French historical electoral surveys
- Explore data at variable level with TEV conceptual hierarchy in Colectica


Do you have any questions?

Tack så mycket!

Thank you!

Let's keep in touch

 lucie.marie2@sciencespo.fr

 [@luciemariejsph](https://twitter.com/luciemariejsph)

 [linkedin.com/in/lucie-marie-josephine](https://www.linkedin.com/in/lucie-marie-josephine)