



# Data Transparency with DDI and Colectica

Introduction and case studies

# About Me

- Colectica
  - ▣ 2006 – present
- DDI Technical Committee, Invited Expert
  - ▣ 2009 - present
- Wisconsin Longitudinal Study
  - ▣ 2004 – 2006

# About Colectica

- Software for standards-based metadata management
  - ▣ Concepts and classifications
  - ▣ Survey design and specification
  - ▣ Data documentation
  - ▣ Data lifecycle, methods, and quality
- Metadata Repository and Portal

# Overview

- Introduction
  - Discoverability, Transparency, and Reproducibility
  - Data Documentation Initiative (DDI) Standard
- Lineage throughout the Data Lifecycle
  - Data Sources, Comparability, and History
- Colectica and DDI Case Studies
  - Official Statistics and Research Surveys
- Recommendations

# Discoverability

- Good standards exist to describe datasets
  - ▣ [schema.org](http://schema.org)
  - ▣ DCAT
  - ▣ Plenty more

# Discoverability

---

- Good places to look
  - ▣ Google
  - ▣ ICPSR
  - ▣ Data.gov



Once we discover a dataset: then what?

# Data

\*M3P1-annotated.sav [DataSet1] — PSPPIRE Data Editor

File Edit View Data Transform Analyze Utilities Windows Help

Open... Save Go To Variable... Go To Case... Find... Insert Cases Insert Variable Split File... Weight Cases... Value Labels

Case	M2ID	M2FAMNUM	SAMPLMAJ	C1STATUS	C1PRAGE	C1PBYEAR	C1PRSE	C1PIDATE_MO	C1PIDATE_YR	C1PIDATE_YR
1	10001	110498	2	4	69	1943	1	7	2013	2013
2	10002	100001	1	1	78	1935	1	6	2013	2013
3	10011	110475	2	4	61	1952	2	6	2013	2013
4	10015	120805	3	4	63	1950	2	11	2013	2013
5	10019	100009	1	4	60	1952	1	6	2013	2013
6	10020	100010	1	1	65	1948	2	3	2014	2014
7	10024	100013	1	4	60	1953	1	8	2013	2013
8	10030	120243	3	4	66	1947	2	7	2013	2013
9	10036	120944	3	4	64	1949	1	12	2013	2013
10	10037	110065	2	4	51	1962	1	6	2013	2013
11	10038	120049	3	4	66	1946	2	7	2013	2013
12	10040	100018	1	4	58	1955	1	6	2013	2013
13	10046	120728	3	4	53	1960	2	10	2013	2013
14	10047	100022	1	4	54	1958	2	7	2013	2013



# Metadata

Variable	Name	Type	Width	Decimals	Label	Value Labels	Missing Values	Column
1	M2ID	Numeric ...	5	0	MIDUS 2 ID number	None ...	None ...	10
2	M2FAMNUM	Numeric ...	6	0	MIDUS 2 Family number	None ...	None ...	10
3	SAMPLMAJ	Numeric ...	8	0	Major sample identification	{1, MAIN RDD}...	None ...	10
4	C1STATUS	Numeric ...	1	0	Completion status of M3 re	{1, COMPLETED M3 CATI ONL\	None ...	10
5	C1PRAGE	Numeric ...	2	0	Respondent's age	None ...	None ...	11
6	C1PBYEAR	Numeric ...	4	0	Respondent's year of birth	None ...	None ...	8
7	C1PRSEX	Numeric ...	1	0	Respondent's sex	{1, MALE}...	None ...	6
8	C1PIDATE_MO	Numeric ...	8	0	Interview date - Month	None ...	None ...	13
9	C1PIDATE_YR	Numeric ...	8	0	Interview date - Year	{9997, DON'T KNOW}...	None ...	13
10	C1PAA1	Numeric ...	1	0	Recession began with spec	{1, YES}...	7, 8 ...	5

# Metadata

4	C1STATUS	Numeric	...	1	0	Completion status of M3 re	{1, COMPLET
5	C1PRAGE	Numeric	...	2	0	Respondent's age	None
6	C1PBYEAR	Numeric	...	4	0	Respondent's year of birth	None

# Statistical tools have limited metadata

---

- ▣ Data types
- ▣ Variable labels
- ▣ Value labels

# No metadata

Open...		Save		Go To Variable...		Insert Variable		Split File...		Weight Cases...		Value Labels	
Variable	Name	Type	Width	Decimals	Label	Value Labels	Missing Values	Col					
1	NEWID	String ...	8			None ...	None ...	8					
2	DIRACC	String ...	1			None ...	None ...	1					
3	DIRACC_	String ...	1			None ...	None ...	1					
4	AGE_REF	Numeric ...	8	0		None ...	None ...	8					
5	AGE_REF_	String ...	1			None ...	None ...	1					
6	AGE2	Numeric ...	8	0		None ...	None ...	8					
7	AGE2_	String ...	1			None ...	None ...	1					
8	AS_COMP1	Numeric ...	8	0		None ...	None ...	8					
9	AS_C_MP1	String ...	1			None ...	None ...	1					
10	AS_COMP2	Numeric ...	8	0		None ...	None ...	8					
11	AS_C_MP2	String ...	1			None ...	None ...	1					
12	AS_COMP3	Numeric ...	8	0		None ...	None ...	8					
13	AS_C_MP3	String ...	1			None ...	None ...	1					
14	AS_COMP4	Numeric ...	8	0		None ...	None ...	8					

# The metadata problem



- ▣ Metadata is needed to understand data
- ▣ Statistical tools have limited metadata capabilities
- ▣ No ability to record information about variable lineage
- ▣ Nothing about methods or quality



Discoverability is not Transparency

# Transparency

- AAPOR Transparency Initiative calls for
  - ▣ Funding
  - ▣ Question Wording
  - ▣ Population, sampling, weighting information
  - ▣ Modes of collection
  - ▣ Contact information

# Reproducibility

---

- In addition to data:
  - ▣ Methods
  - ▣ Full data lineage
  - ▣ Data transformation code



# Data Documentation Initiative



- Since 1995
- Open standard for describing data
  - ▣ Focus on social, behavioral, and economic sciences
  - ▣ XML
- Users
  - ▣ National Statistical Institutes
  - ▣ University Research Groups
  - ▣ Data Archives
  - ▣ Other Data Producers and Publishers

# DDI Content

Data

Classifications

Quality

Data collection

Research Lifecycle

Foundational

Data Processing

# Data

## **Dataset**

A data file, database, or other source of data

## **Data Layout**

Describes the layout of a data file

## **Variable**

A column in a dataset

## **Variable Statistics**

Summary statistics for a single variable

## **NCube**

Aggregate data

## **Represented Variable**

Describes the common information of one or more harmonized variables

## **Conceptual Variable**

Describes the common information of one or more harmonized variables

# Classifications



**Classification Family**

**Classification Series**

**Statistical Classification**

**Classification Level**

**Classification Item**

**Classification Correspondence Table**

**Classification Index**

# Quality



## **Quality Statement**

A set of statements with information about how a study was conducted

## **Quality Standard**

Describes all information that a *quality statement* should record

# Data Collection

## **Instrument**

A survey or other data capture instrument

## **Question**

A question that can appear in a survey instrument

## **Question Grid**

A question grid that can appear in a survey instrument

## **Question Block**

A question block that can appear in a survey instrument

## **Statement**

A statement that can appear in a survey instrument

## **Instruction**

Information for an interviewer or respondent

## **Computation**

Source code that performs calculations, validation, or other actions

## **Sequence**

A set of items in an instrument, used for grouping, paging, or other organization

## **Data Collection**

Describes the processes and methods used to collect data

# Research Lifecycle



## **Study**

A single research project

## **Series**

A repeated set of studies

## **Archive**

Information about how a study is archived for long term preservation

# Foundational



## **Concept**

An abstract idea or general notion

## **Category**

A class of people or things

## **Code List**

A list of categories, each with an assigned value

## **Universe**

A population being studied

## **Organization**

An institution, company, or other group



# Data Processing



## **Processing Event**

Information about who performed data processing and how it was performed

## **General Instruction**

Any sort of data processing instructions

## **Generation Instruction**

Data processing that creates new variables or datasets

# DDI Items

- All items are identified
- Items can be registered in a repository
  - ▣ ISO 11179
- Full audit trail of changes to information



# Lineage throughout the Data Lifecycle

- Methods Documentation
- Quality Reports
- Variable Lineage
- Structured Questionnaires



# Methods

# Methods in DDI

---

- Data Collection
  - ▣ Timing
  - ▣ Weighting
  - ▣ Sampling

# Methods in DDI

---

- Coverage
  - ▣ Time
  - ▣ Geography
  - ▣ Topics

# Methods in DDI



- Collection Events
  - ▣ Organizations
  - ▣ Data Sources
  - ▣ Modes
  - ▣ Actions to minimize losses



# Quality Reports



# Quality Reporting

- Quality Reports for Eurostat
- Concepts, methods, quality assessments
- European standards
  - ▣ SIMS, ESMS, ESQRS
- Supported in DDI through the `QualityStatement` element

# Variable Lineage in DDI

Source questions

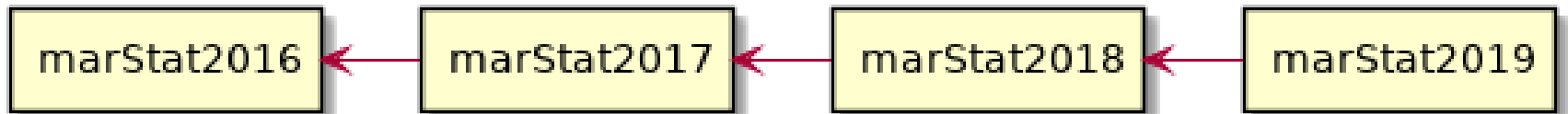
Source variables

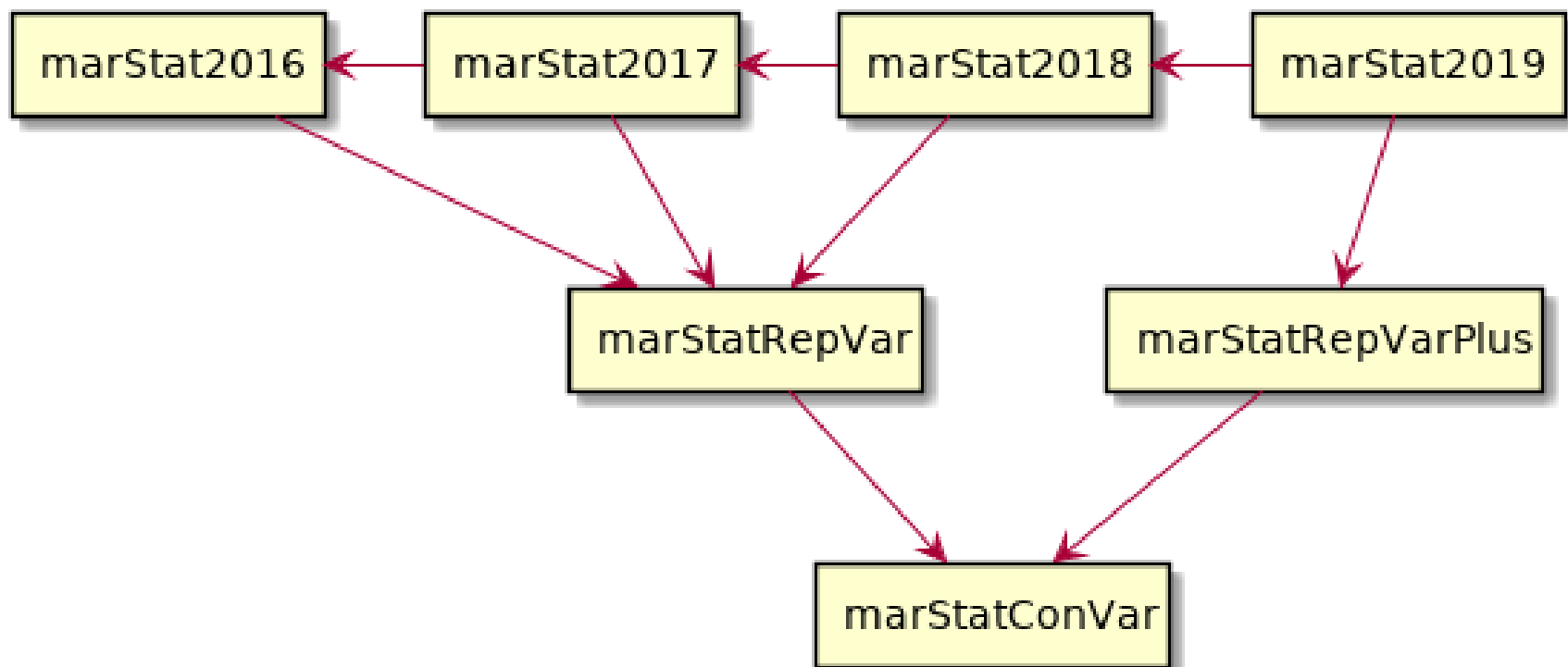
Variables measured across time, in different studies

Versioning

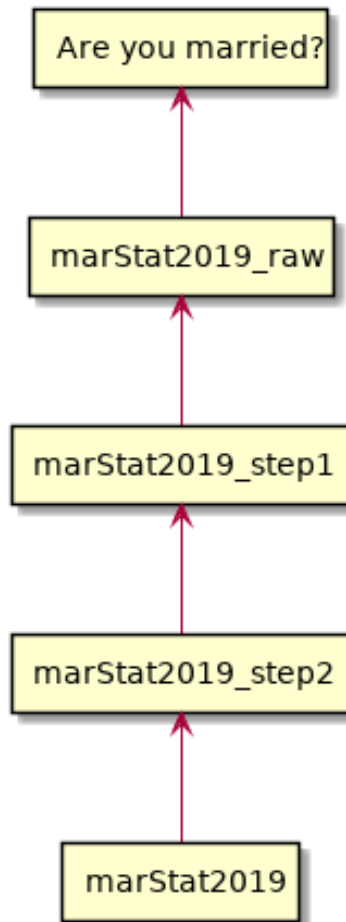
BasedOn

# Variable Concordance

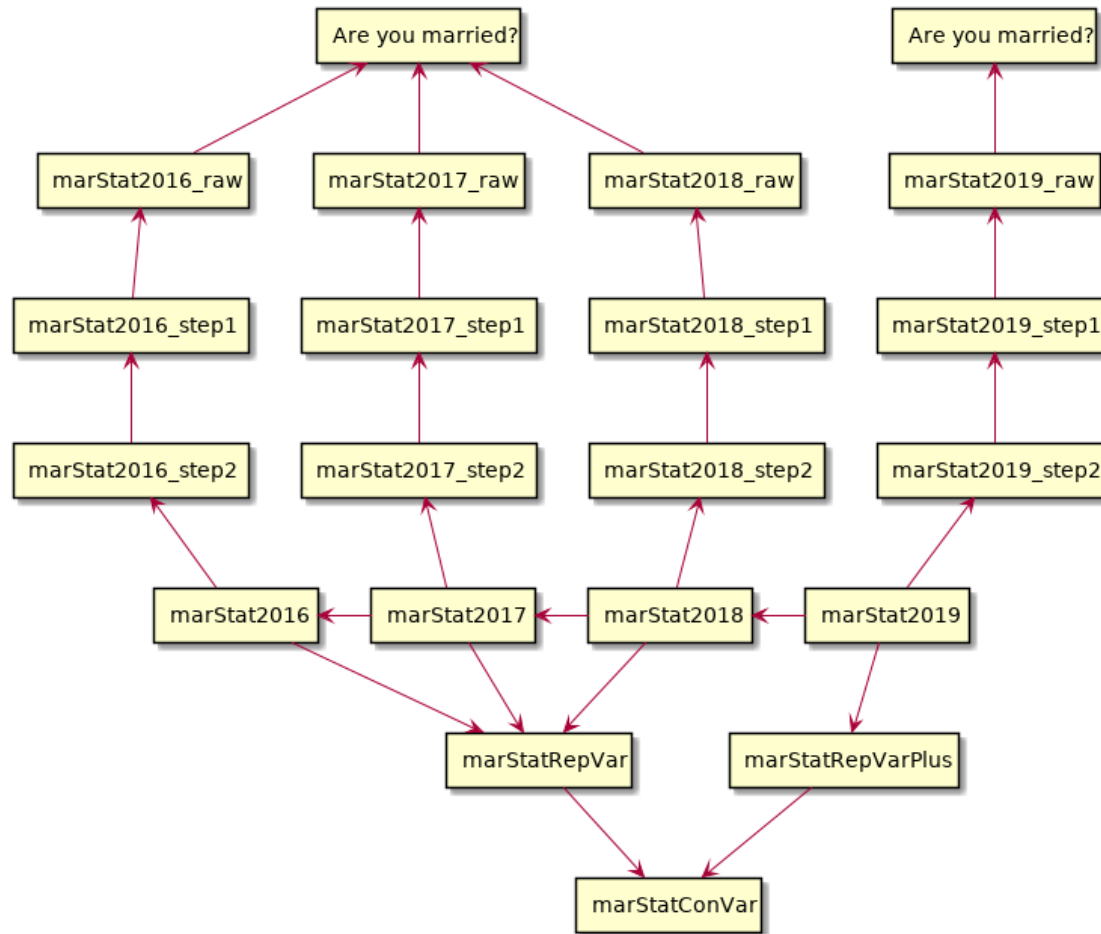


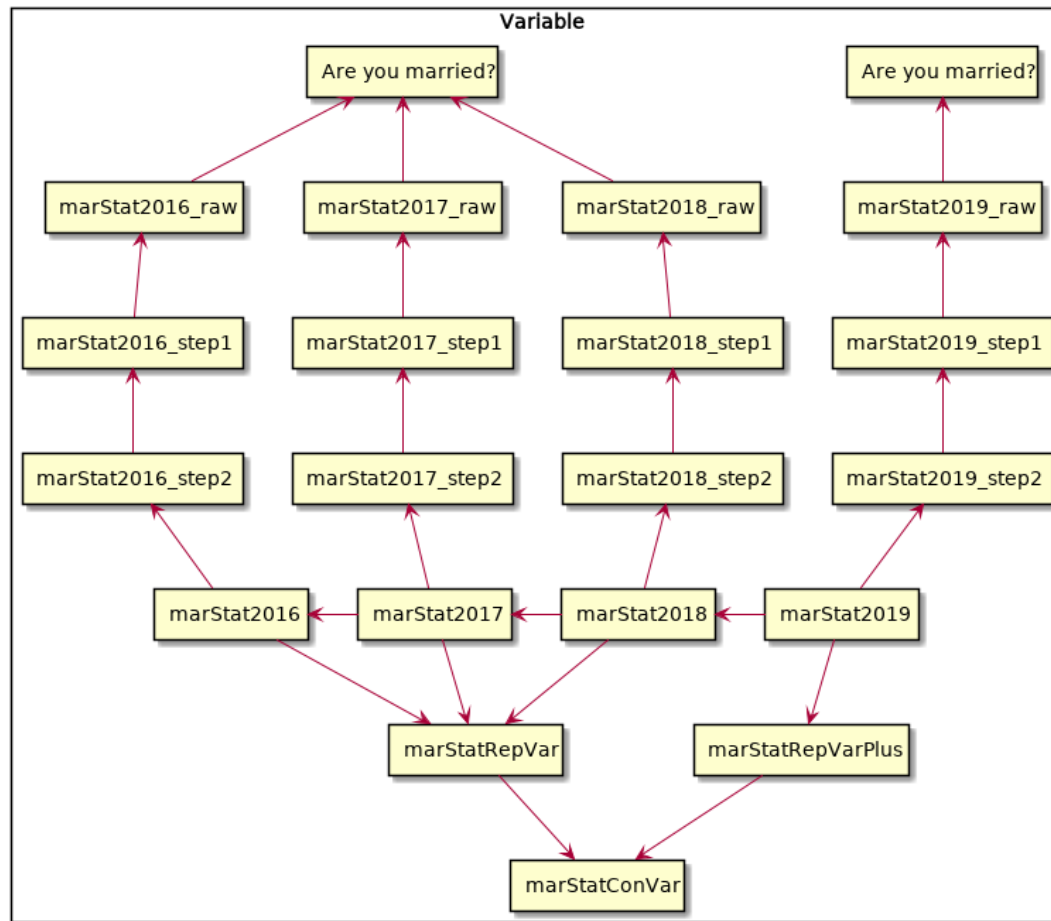


# Variable Sources



# Combined Lineage





# Show the lineage in different ways

---

- ▣ Reports and discovery portals can make use of this lineage





# Documenting Data Transforms

# C2Metadata Overview

- ▣ Continuous Capture of Metadata
- ▣ George Alter, PI
- ▣ Funding from National Science Foundation
  - Data Infrastructure Building Blocks (DIBBs)

ICPSR

colectica

NSD

NORWEGIAN CENTRE  
FOR RESEARCH DATA

ANES  
American National Election Studies

mtna

NORC<sup>75</sup>  
at the UNIVERSITY of CHICAGO

# C2Metadata Overview

- ▣ Extract transforms by parsing statistical source code
  - SPSS
  - Stata
  - SAS
  - R (tidyverse)
  - Python (pandas)
- ▣ Record information about the transforms in a structured way
- ▣ Update DDI metadata with information about the transforms

# C2Metadata Tools

---

- ▣ In active development
- ▣ Desktop, command line, and web tools
- ▣ Developer libraries
- ▣ [c2metadata.org](https://c2metadata.org) and [gitlab.com/c2metadata](https://gitlab.com/c2metadata)

# Structured Data Transform Language

---

- SDTL
- Machine readable descriptions of data transforms
- JSON, XML, RDF representations



# Structured Questionnaires

# Questionnaire Documentation in DDI



- Questions
- Sequences
- Logic
- Full survey instruments

# Questionnaire Documentation in DDI



- Questions can be sources for variables
- Questions can be re-used across many surveys



# Case Studies: Official Statistics



# Statistics New Zealand

## □ DataInfo+



**DataInfo+**  
Explore our metadata

### Find information about our data

DataInfo+ gives you one place to search and browse for information about our statistical activities and data.

#### Find our series



##### Series

Search or browse a list of series we produce. Find descriptions of series, including their frequency and collection methods.

#### Find out more about what we measure and how we produce our statistics



##### Concepts

Search or browse for information about the concepts explored in our studies, including the statistical terms we use.



##### Populations

Search or browse a list of populations we study.



##### Variables

Search or browse a list of variables used in our studies, including their descriptions.



##### Classifications

Search a list of classifications we use.



##### Codes

Search for classification codes by keyword or browse for codes by classification.



##### Questionnaires and forms

Search a list of the instruments used in our studies.

- [-] Consumers Price Index
  - [+] Category Sets (0)
  - [+] Concept Sets (2)
  - [-] Data Collections (3)
    - CPI Data Collection 2011
    - [+] CPI Data Collection 2014
    - CPI Data Collection 2017**
  - [+] Instruments (0)
  - [-] Population Sets (1)
    - Consumers Price Index Population
  - [-] Variable Sets (1)
    - [+] Consumers Price Index published variables



## CPI Data Collection 2017



Data Dictionary

### Data Collection Methodology

#### Methodology

##### Field collection

Statistics NZ price collectors gather prices directly from retail outlets.

##### Sample size

We collected about 100,000 prices from about 2,800 retail outlets and 2,300 other businesses and landlords.

##### General information

##### Imputation

Due to unavailability at the time of price collection, on average we impute 1–2 percent of prices (not including seasonal items such as winter clothing) each quarter. We often do this by carrying forward the previous quarter's price. Other imputation we do is to apply the movements of similar categories of items.

##### Review of the CPI

















Reviews of the CPI are undertaken every three years. We implemented the latest review when the December 2017 CPI was published. The review involved reselecting the basket of representative goods and services, updating the new national expenditure weights, and updating regional expenditure weights.

Consumers price index review: 2017 has more information.

##### Impact of GST rise on the CPI

Results **1** to **25** of **103** for (**0.094** seconds)

Sort By: Alphabetical - Item Type - Metadata Rank - Version Date

Item	Description	Metadata Rank
 2001 Post-enumeration Survey		0
 2006 Post-enumeration Survey		0
 2013 Post-enumeration Survey		0
 2018 Census of Population and Dwellings		0
 Abortion Statistics		0
 Accommodation Survey (2013 to current)		0
 Agriculture Production Surveys and Censuses		0
 Alcohol Available for Consumption		0
 Annual Balance Sheets 2007-17		0
 Annual Enterprise Survey		0
 Balance of Payments and International Investment Position Statistics		0
 Births		0
 Building Consents Issued		0
 Business Demography Statistics		0
 Business Frame		0
 Business Operations Survey		0

# Statistics Denmark

- Eurostat Quality Reporting
  - ▣ 25+ people enter quality information
  - ▣ Review and approvals
  - ▣ Export from DDI Lifecycle to SIMS metadata structure
  - ▣ Deliver to Eurostat
  
- Classifications

# CLASSIFICATIONS

**Population and elections**



**Living conditions**



**Education and knowledge**



**Culture and National Church**



**Labour, income and wealth**



Statistics Denmark's Classification of Occupations (DISCO-08)  
DISCO in wage statistics

**Prices and consumption**



European Classification of Individual Consumption according to Purpose (ECOICOP) - Statistics Denmark  
European Classification of Individual Consumption according to Purpose (ECOICOP) - Eurostat

**National accounts and government finances**



Classification of the functions of government (COFOG)  
Classification by sector in the statistical business register (ESR)  
Classification by sector in the European system of accounts (ESA2010)  
Social protection expenditure (ESSPROS)

## Koder og kategorier

ÅBN HIERAKIET

DOWNLOAD ▾

- + 1: Ledelsesarbejde
- + 2: Arbejde, der forudsætter viden på mellem- eller højt niveau inden for pågældende område
- + 3: Arbejde, der forudsætter viden på mellemniveau
- + 4: Almindeligt kontor- og kundeservicearbejde
- 5: Service- og salgsarbejde
  - 51: Servicearbejde
    - 511: Service- og kontrolarbejde under transport og rejser
      - 5111: Servicearbejde af passagerer i forbindelse med rejser
        - 511110: Passagerbetjening under rejser
        - 511120: Passagerbetjening i lufthavne og havne terminaler
      - + 5112: Kontrolarbejde under rejser
      - + 5113: Turist- og rejselederarbejde
    - + 512: Kokkearbejde
    - + 513: Tjenere og bartendere
    - + 514: Frisørarbejde og kosmetologarbejde samt beslægtede funktioner
    - + 515: Inspektørarbejde inden for rengøring, husholdning og ejendomme
    - + 516: Andet servicearbejde
  - + 52: Salgsarbejde (ekskl. agentarbejde)
  - + 53: Omsorgsarbejde
  - + 54: Rednings- og overvågningsarbejde

# INSEE (France)

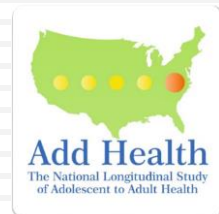
- Survey instrument specification using INSEE tooling, using DDI Lifecycle
- Study level information using Colectica
- Connect information in metadata repository, via DDI Lifecycle



# Central Statistics Office (Ireland)

- Questionnaire specification
  - ▣ Use Blaise to collect data
  - ▣ Designing LFS using DDI Lifecycle
  - ▣ Will be able to connect data back to specified questions

# Case Studies: Research Studies



# Finnish National Election Study (FNES)

## □ Portal with rich variable details

### Welcome

Welcome to the FNESdata longitudinal survey metadata portal. The portal contains the metadata of four surveys in the Finnish National Elections Study series. Studies range from 2003 to 2015 and they are listed below. You can browse individual datasets, search specific variables on the search page and easily explore the datasets to find and investigate new, interesting viewpoints to existing data. By collecting your findings into the basket, you can save custom codebooks from hand-picked variables.



### FSD1260 Finnish National Election Study 2003

The survey consists of two parts which were collected after the 2003 parliamentary elections in Finland with the help of face-to-face interviews and a supplementary, self-administered questionnaire. Swedish-speaking population is over-represented in the data. The interview data is Finland's contribution to the international Comparative Study of Electoral Systems program (CSES). Variables q19-q22, q51-q65\_11, and q67-q80 are national election study variables, variables q23\_1-q50 and q66 are CSES variables, variables beginning with 'p' are self-administered questionnaire variables, and the rest are background variables.

**Keywords**  
Internet; election campaigns; elections; expectations of future; parliamentary elections; political attitudes; political participation; political party affiliation; political party preference; voting

[Browse Data](#)

[Codebook](#)

[Documentation](#)

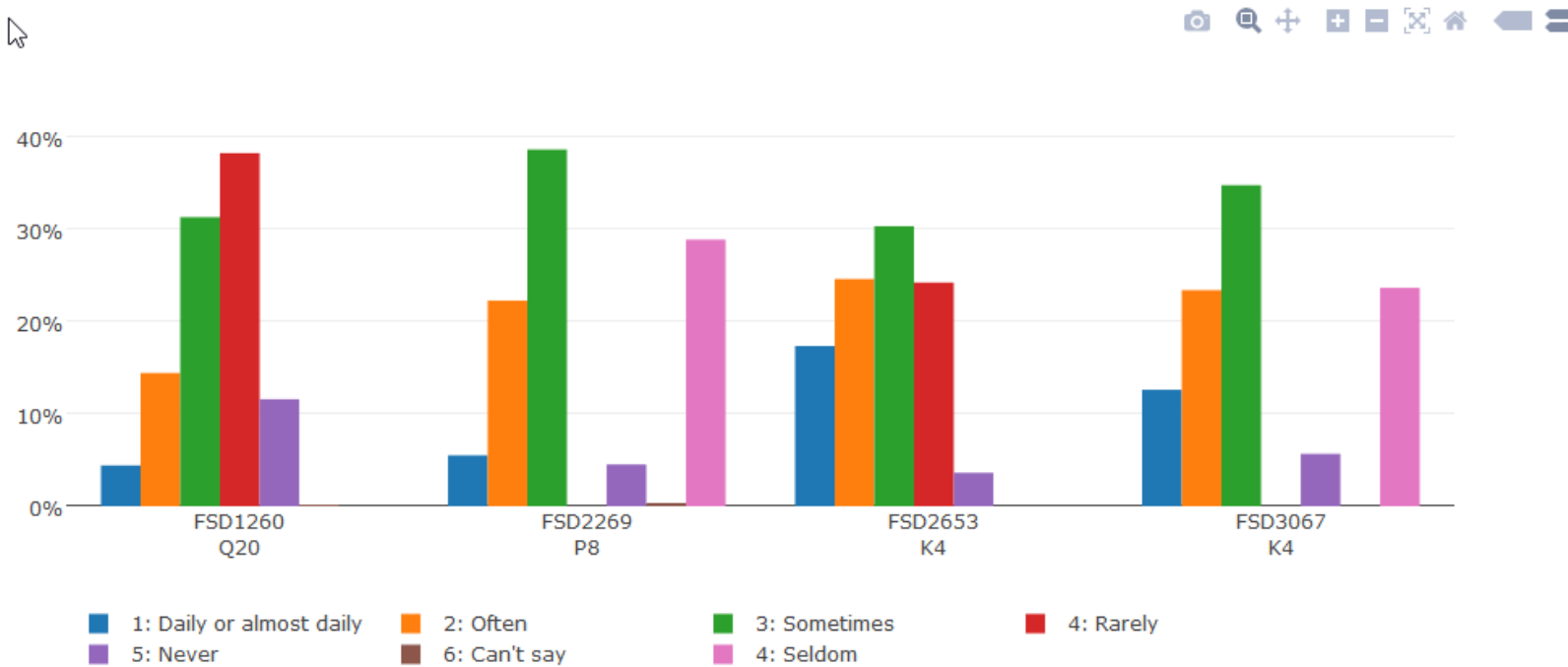
[Data from Aila](#)

### FSD2269 Finnish National Election Study 2007

		FSD1260	FSD2269	FSD2653	FSD3067
+	[fnes10] How often do you discuss politics with others?	Q20	P8	K4	K4
+	[fnes11] How much attention did you pay to media coverage of the parliamentary elections: Television debates and party leader interviews	Q21_1	K2_1	K2_1	K2_1
+	[fnes18] How much attention did you pay to media coverage of the parliamentary elections: Candidate selectors on the Internet	Q21_8	K2_9	K2_11	K2_11
+	[fnes19] How much attention did you pay to media coverage of the parliamentary elections: Television advertisements	Q21_9	K2_10	K2_6	K2_6
+	[fnes20] How much attention did you pay to media coverage of the parliamentary elections: Newspaper advertisements	Q21_10	K2_11	K2_7	K2_7
+	[fnes61] Sometimes politics seems so complicated that I can't really understand what is going on	Q65_1	K17_7	K15_7	K16_7

[Details...](#)

### Wording in Question



Dataset	Variable	Valid	Invalid	Min	First Quartile	Median	Third Quartile	Max	Mean	StdDev
FSD1260	Q20	1270		1				6		
FSD2269	P8	1016		1				6		
FSD2653	K4	1298		1				5		
FSD3067	K4	1587		1				5		

# CLOSER

- Metadata portal
- 10 British cohort studies
- 80,000 variables
- 30,000 questions
- 200+ survey instruments

## Search

Item type

All

Studies

All

Life Stages

All

Search query

Search

Search



10

Studies



93

Sweeps



246

Data Files



29,231

Questions

## Welcome

CLOSER Discovery is an online resource that enables researchers to [search](#) the data from eight leading UK longitudinal studies. We need your [feedback](#) to help us shape this resource to best meet the needs of its users.

[Read more](#) about CLOSER Discovery or take a look at the [FAQs](#) or [How-to guides](#) to get started.

Our studies:

- Avon Longitudinal Study of Parents and Children
- 1970 British Cohort Study
- Hertfordshire Cohort Study
- Millennium Cohort Study
- 1958 National Child Development Study
- MRC National Survey of Health and Development
- Southampton Women's Survey
- Understanding Society



79,412

Variables



214

Questionnaires

Variable

Details

Sources

Lineage

Name

A0043B

Label

SMOKING DURING PREGNANCY

Dataset

UKDA-SN-2666-2

Value	Label	Frequency	
-3	NS or Nk	87	
1	Non Smoker	7179	
2	Stopped Pre-Preg	2031	
3	Stopped Dur-Preg	814	
4	Ctl Smokers 1 - 4	1154	
5	Ctl Smokers 5 - 14	3615	
6	Ctl Smokers >= 15	2316	
Valid		Invalid	MinMax
17196		0	-36



[Variable](#)[Details](#)[Sources](#)[Lineage](#)

## Source Questions

### 11 a

Birth Questionnaire

32 questions before...

 **10 b**

What type of antenatal preparation did the mother receive during this p



 **11 a**

Does the mother smoke now?



 **11 b**

did she ever smoke?



126 questions after...

[View the complete instrument](#)

[Birth Questionnaire](#)

? 23 ii

Was the mother booked for delivery at the place where her confinement occurred?



If YES,



? 23 ii(a)

give the date of booking



? 24 i

Did labour start



- |   |               |
|---|---------------|
| 1 | Spontaneously |
| 2 | Induced       |
| 0 | Not known     |

If INDUCED,



? 24 i(a)

what methods were used?



# US Studies

- Midlife in the United States (MIDUS)
- National Health and Aging Trends (NHATS)
- National Social Life, Health, and Aging Project (NSHAP)
- National Archive of Computerized Data on Aging (NACDA)

# Recommendations

- Use or mandate standards that support lineage
- Fund research projects for new specifications and tools

# Use Standards for Data Lineage



- DDI data documentation standard
- AAPOR Transparency Initiative

# Funding Research

- Specifications and Tooling
- DDI originally funded by Health Canada
  - ▣ Now by its members
- SDTL funded by NSF DIBBs

# Recap



- Transparency requires rich information, not just searchability
  - ▣ Methods
  - ▣ Data Lineage






# Thank You

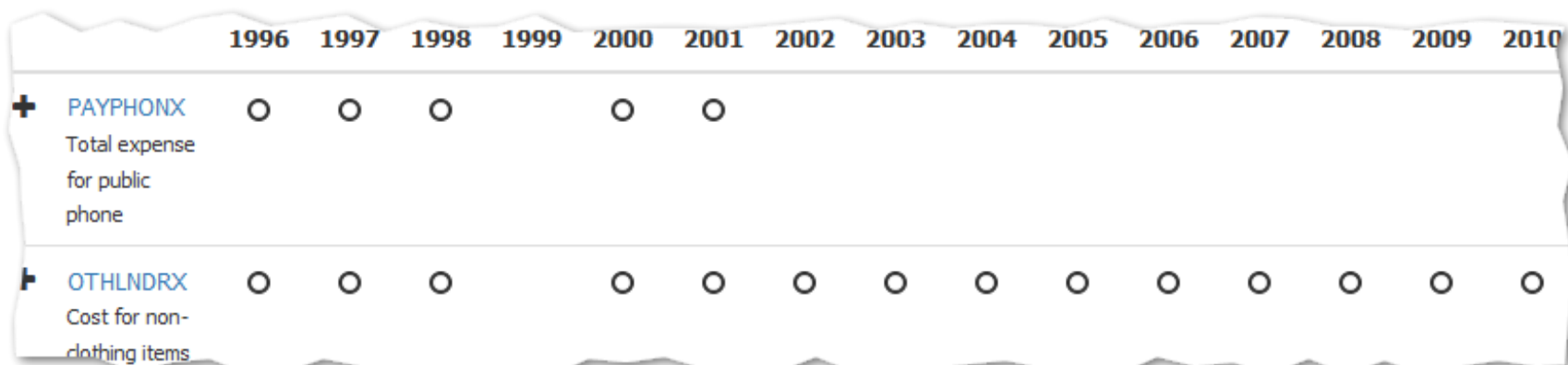
Jeremy Iverson

[jeremy@colectica.com](mailto:jeremy@colectica.com)



# Variables across time in Portal

	<b>M1P1</b>	<b>M2P1</b>	<b>M3P1</b>
 Marital status	A1PB17	B1PB19	C1PB19
 # of Times married	A1PB19	B1PB20	C1PB20
 Month of marriage	A1PB18MO	B1PB21M	C1PB21M
 Year of marriage	A1PB18YR	B1PB21Y	C1PB21Y
 How 1st marriage ended	A1PB20	B1PB22	C1PB22



## Concordance

Statistics

Code Comparison

Correspondence Tree

% of valid % of total

	M1P1 A1PB17	M2P1 B1PB19	M3P1 C1PB19
MARRIED	65.69 %	70.71 %	67.20 %
SEPERATED	2.83 %		
SEPARATED		1.63 %	1.49 %
DIVORCED	13.54 %	12.89 %	13.25 %
WIDOWED	5.00 %	7.04 %	11.00 %
NEVER MARRIED	12.94 %	7.73 %	7.05 %
DON'T KNOW			
REFUSED			
INAPP			

# Sources in Portal

## Lineage



**M3P1 - C1PB19**

Marital status currently



**C1PB19**

Are you married, separated, divorced, widowed, or never married?

# Question text in a data dictionary

 C1PB19

**Label**

Marital status currently

**Question Text**

Are you married, separated, divorced, widowed, or never married?

**Forward Skip**

IF [C1PB19](#) = NEVER MARRIED, DK, OR REFUSED, GO TO [C1PB30](#).

Value	Label	Frequency	%
1	MARRIED	2,211	67.1%
2	SEPARATED	49	1.5%
3	DIVORCED	436	13.2%
4	WIDOWED	362	11.0%
5	NEVER MARRIED	232	7.0%
7	DON'T KNOW	2	0.1%
8	REFUSED	2	0.1%

# Version history in Portal

The screenshot displays the MIDUS Portal interface. The top navigation bar includes links for Search, Explore, Basket (27), and Admin. A left sidebar contains a 'History' section with a list of revisions. The main content area shows a breadcrumb trail for 'Marital status currently' and a 'Variable Description' section for the variable C1PB17B5.

**History**

- Revision 20  
1/17/2019 7:57:22 PM  
jeremy
- Revision 19  
6/14/2018 4:57:15 PM  
jeremy
- Revision 15  
2/21/2017 2:48:13 PM  
jeremy
- Revision 14  
2/13/2017 8:07:08 PM  
jeremy

Recalculate summary statistics

**Breadcrumb:** MIDUS 3 Project 1 > M3P1 > M3P1 VariableGroup 3

**Variable Description**

Name	C1PB19
Label	Marital status currently

# Recording variable lineage

- ▣ Tools let users specify this information manually
- ▣ DDI
  - References
  - Versions
  - Variable cascade

# DDI Lifecycle is pretty good at variable lineage

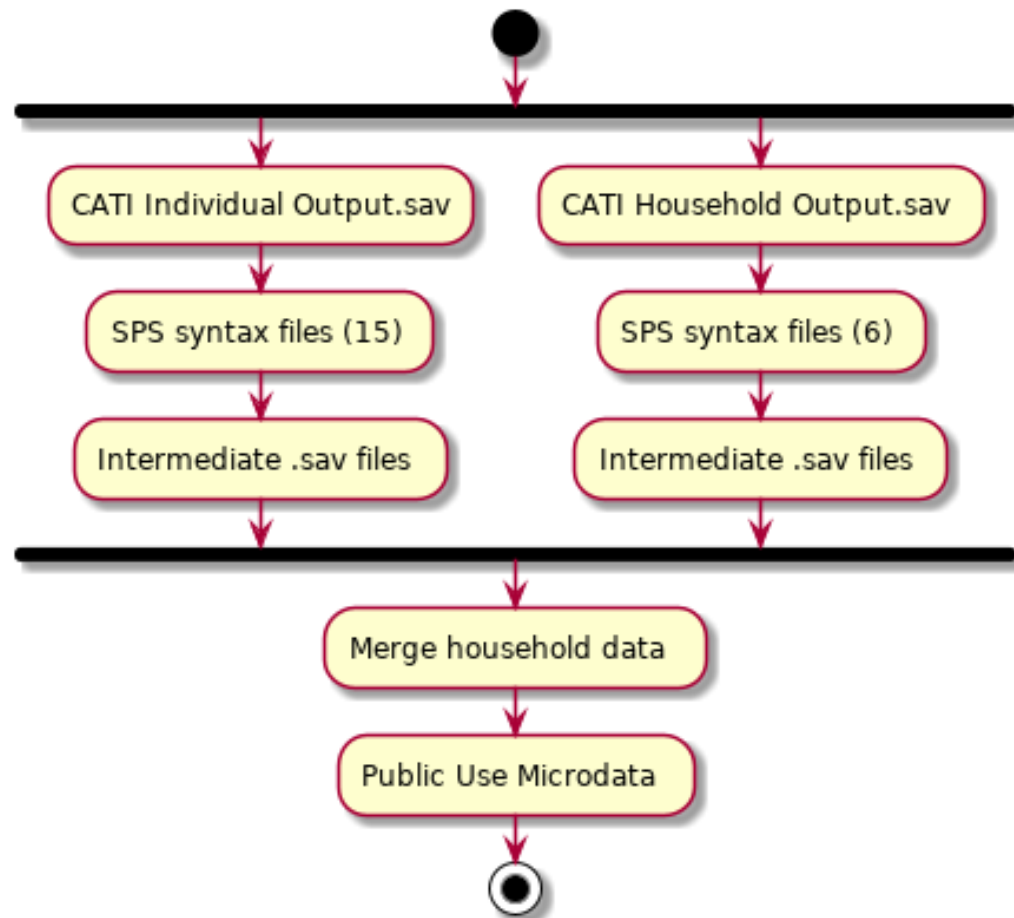
- ▣ Many useful ways to record the information
- ▣ Many useful ways to display the information
- ▣ But: it is time intensive



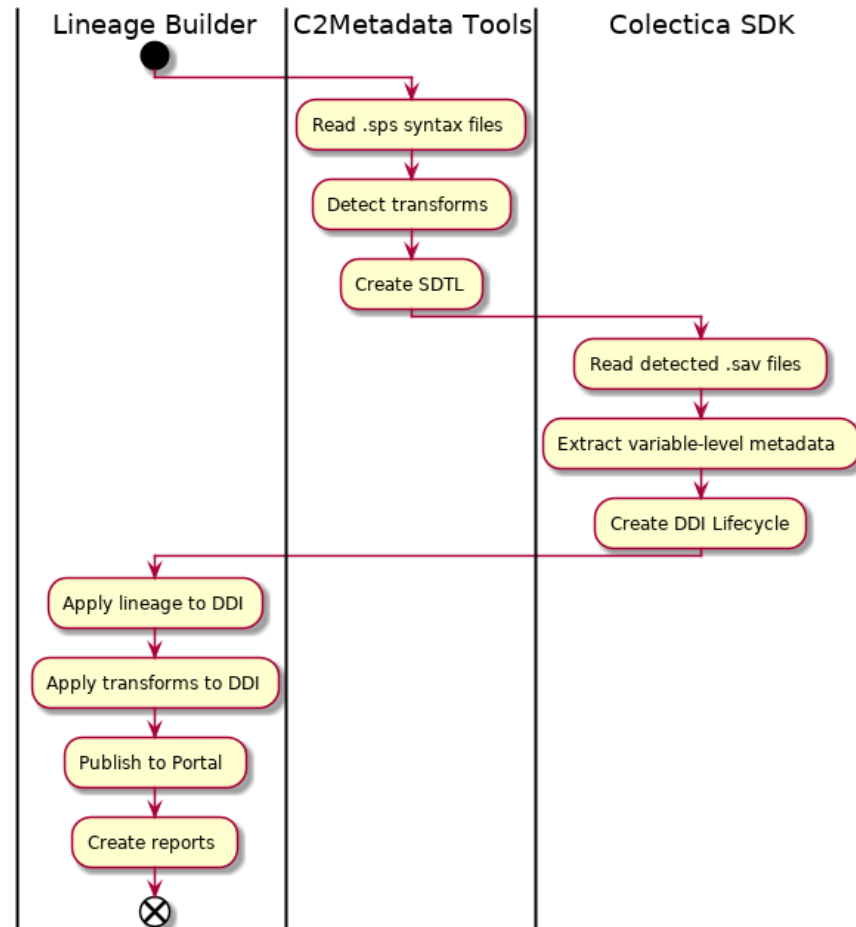
# MIDUS Overview

- ▣ Midlife in the United States
  - Longitudinal study, since 1995
  - MIDUS 3 Project 1
    - 2,339 questions
    - 2,575 public use variables
    - 21 transform files

# MIDUS 3 Transform Pipeline



# Applying C2Metadata to MIDUS



# Applying C2Metadata to MIDUS

1. Read .sps syntax files
2. Detect transforms, creating SDTL
3. Read .sav files, extracting DDI variable metadata
4. Apply lineage to DDI (set source variables)
5. Apply transforms to DDI
6. Publish to Portal
7. Create reports

# Lineage Builder Tool

- ▣ Dependencies

- C2Metadata.Common .NET library
- Colectica SDK

- ▣ Inputs

- 21 SPSS syntax files (*.sps*)
- 20 Data files (*.sav*)

# Detected Transforms

Transform	Count
analysis	353
comment	261
select	240
setValueLabels	151
recode	73

# Detected Transforms

Transform	Count
setDatasetProperty	42
compute	38
setMissingValues	28
load	23
save	19
setVariableLabel	17
rename	13
setDisplayFormat	11
keepVariables	9
delete	5
mergeDatasets	2

# Lineage Builder Outputs



- DDI for 11,032 variables in 19 intermediate datasets
- 2,575 public use variables with documented lineage



# Lineage in Portal (before)

## Lineage

---



**M3P1 - C1PB19**

Marital status currently



**C1PB19**

Are you married, separated, divorced, widowed, or never married?

# Lineage in Portal (after)





# Summary

# Summary



- ▣ Variable-level lineage helps understand data
- ▣ Statistical packages do not provide much help here
- ▣ DDI Lifecycle allows rich descriptions of lineage
- ▣ Tools allow manual and automated recording of lineage in great detail