BEST PRACTICES FOR DOCUMENTING REPEATED STUDIES IN DDI

EDDI 2015 – Copenhagen

Goal

- Provide a recipe for documenting repeated data
  - Metadata elements
  - Steps

Overview

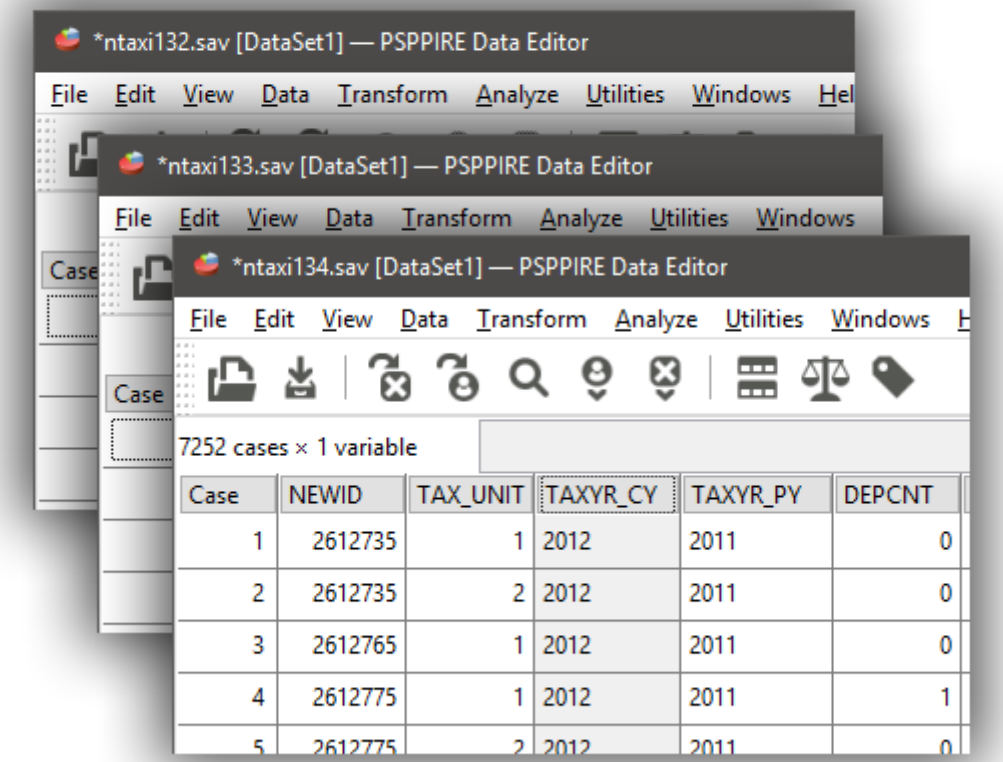
1. Background
2. Metadata standards for documenting repeated data
3. From zero documentation to actionable metadata in seven steps
4. From actionable metadata to information for researchers

Background

Official Statistics

Annual, Quarterly, Monthly

Same data structures released regularly

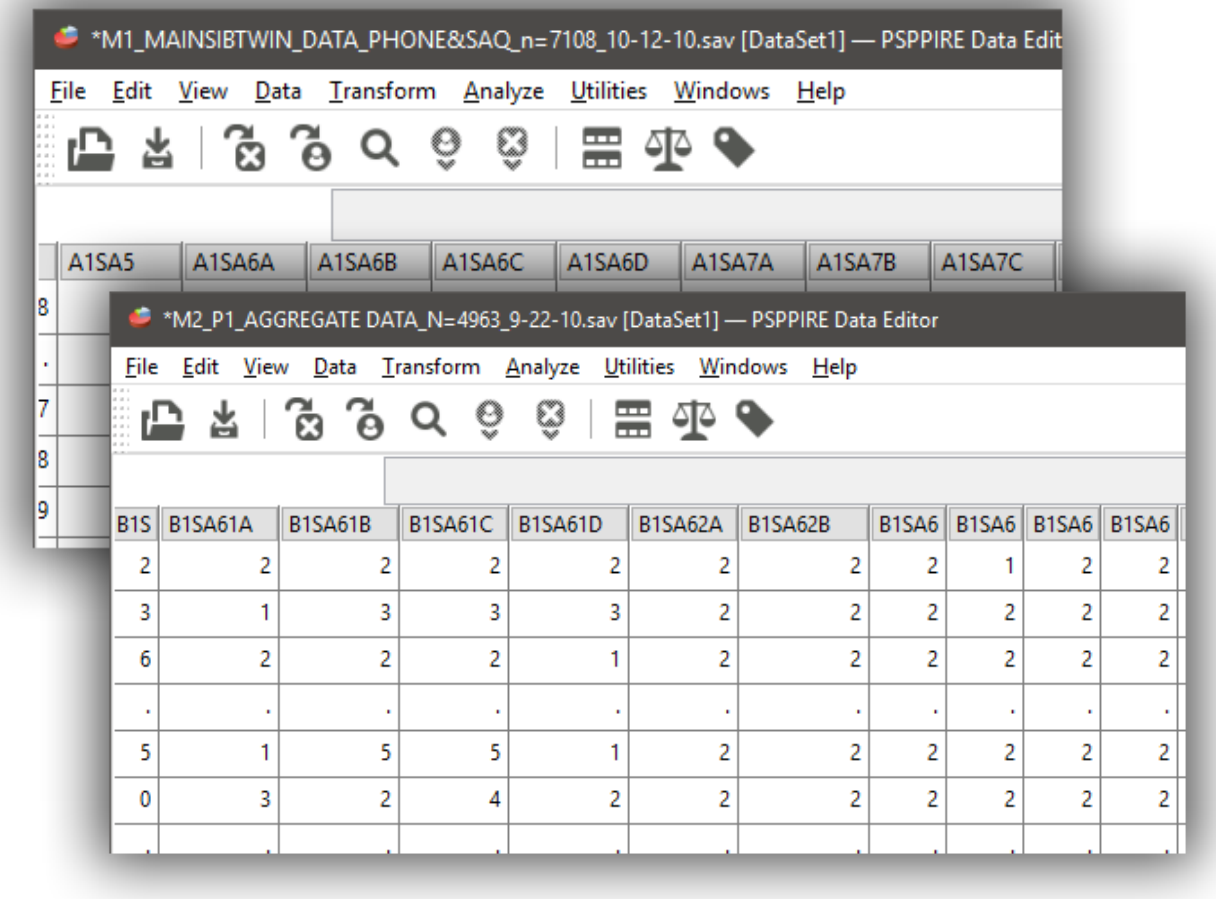


Longitudinal Studies

Collection is often irregular

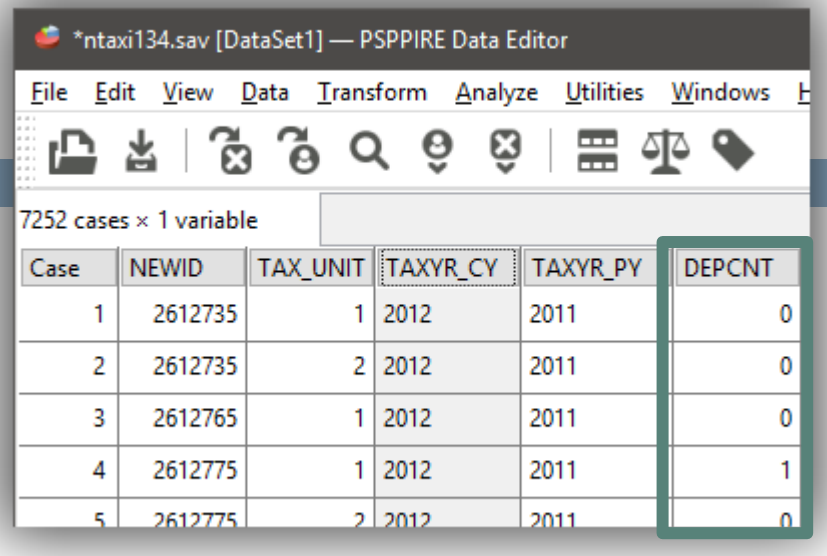
Wide datasets

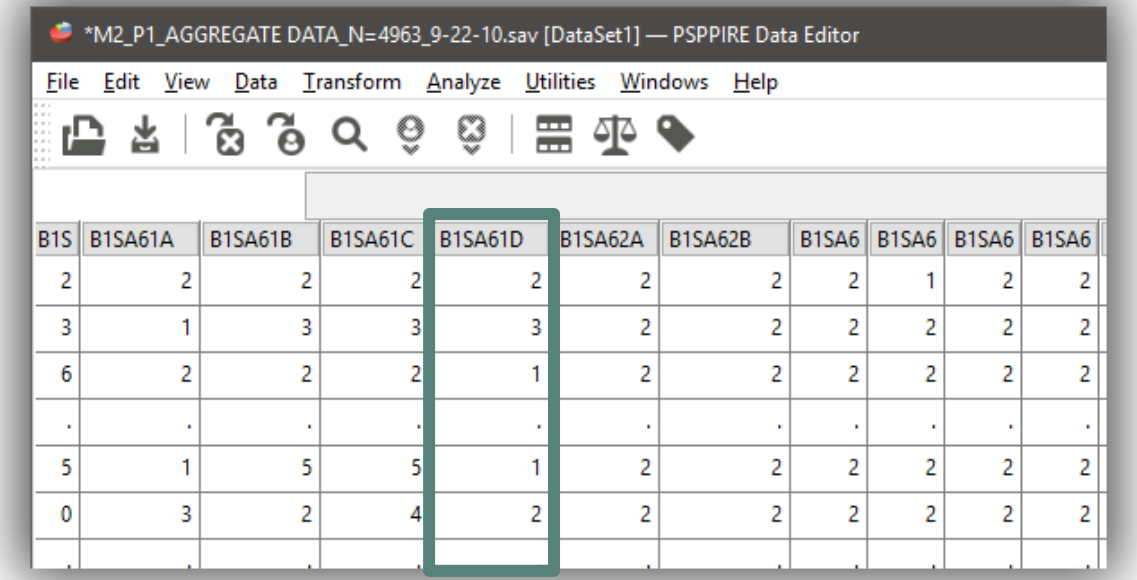
Different variable names represent the same information at different times



Understanding Repeated Data

- Researchers care about variables

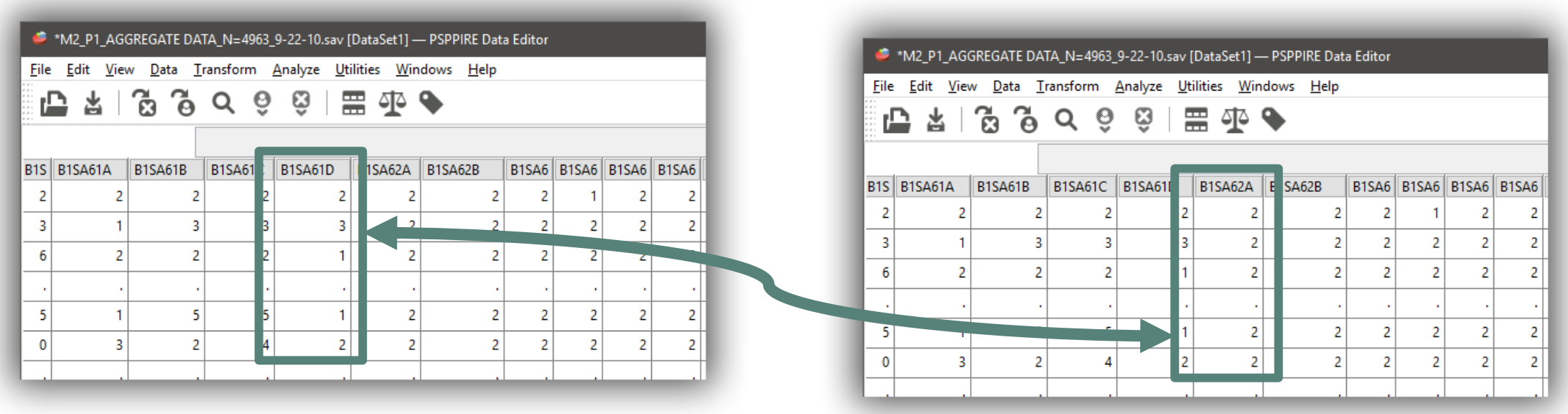




Comparing Data

Sweep 1 data

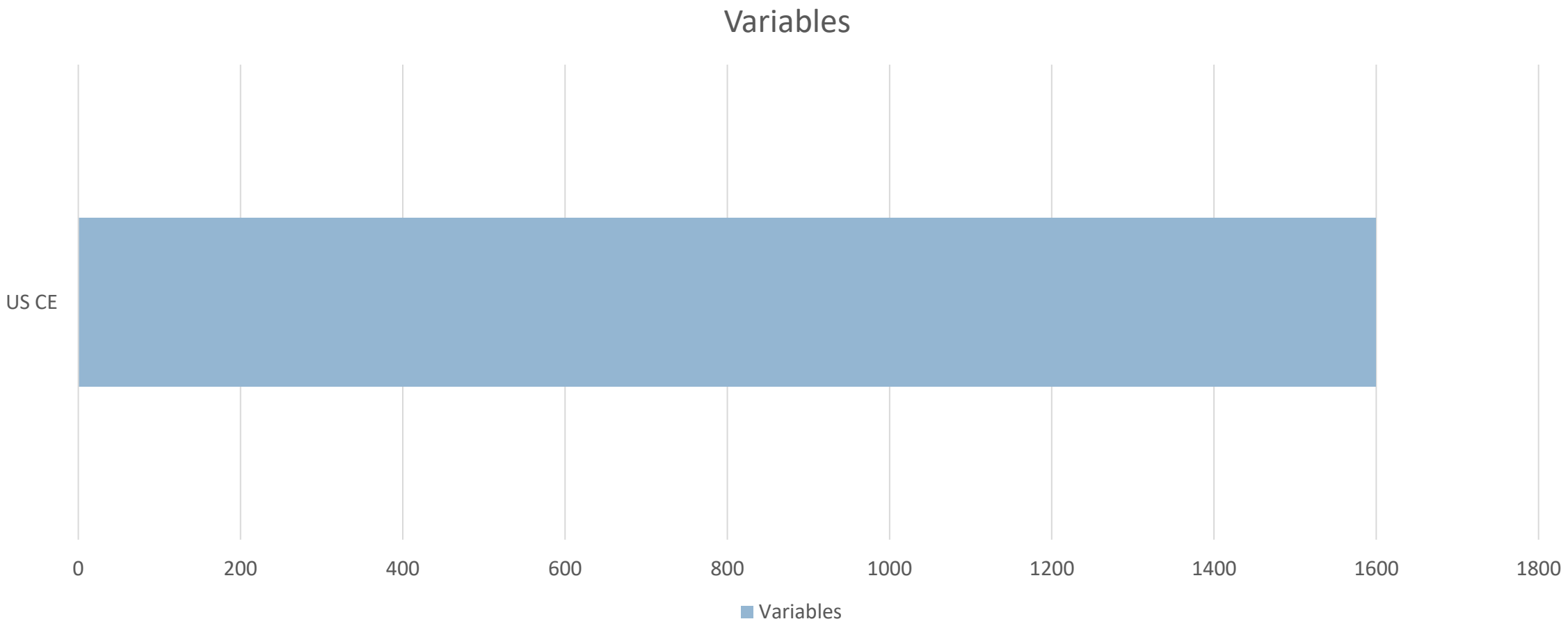
Sweep 2 data



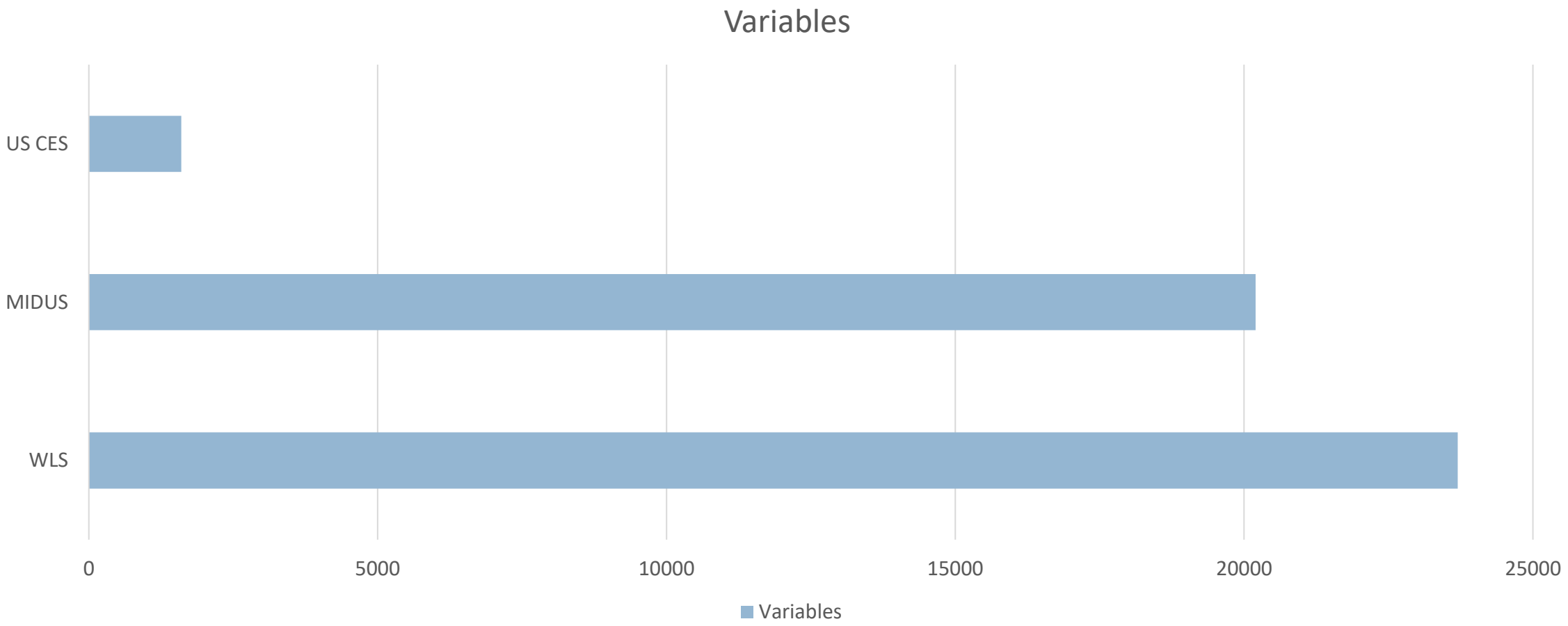
Scope

- Variable counts are huge, so rich documentation is critical

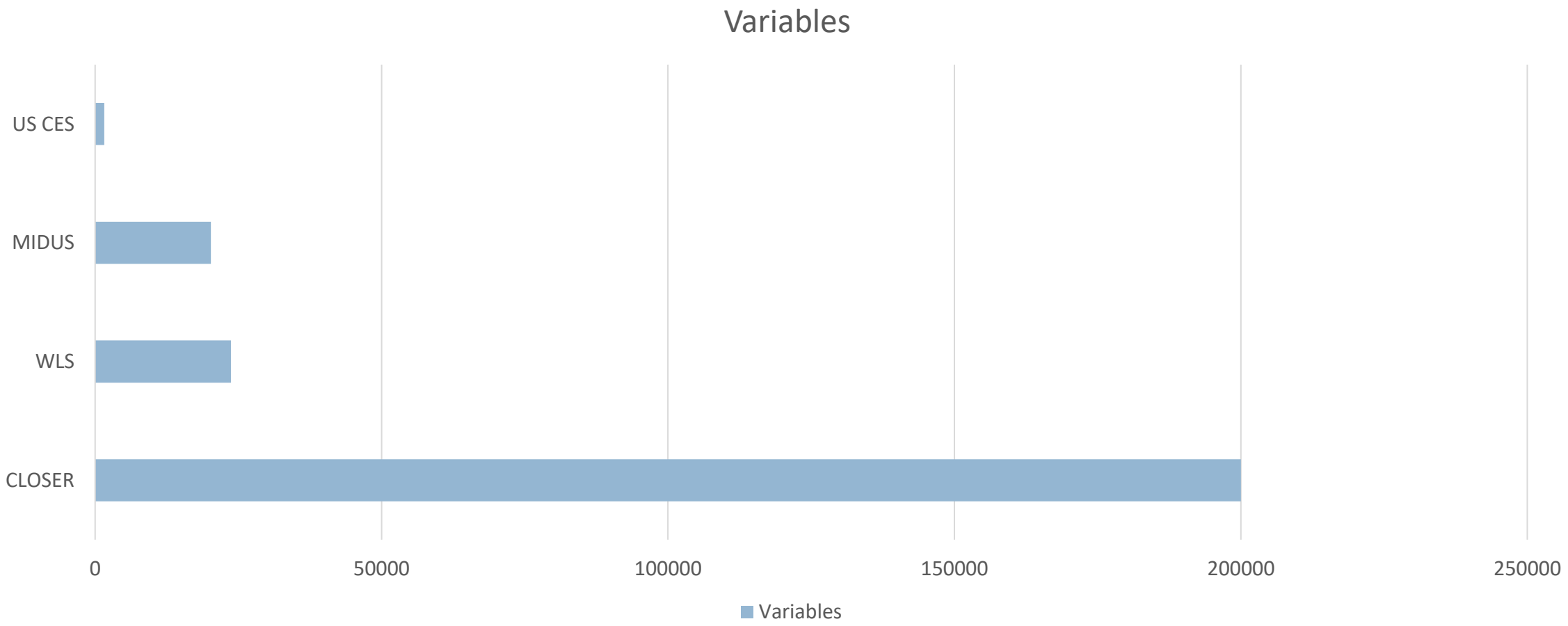
Scope: US Consumer Expenditure Survey



Scope: WLS and MIDUS



Scope: CLOSER



Why Metadata?

Researchers Need Documentation

1. Find information about variables
2. Find comparable data

Why Metadata Standards?

- Share tools
- Share funding
- Benefit from other organizations' investments

Metadata Standards for Repeated Data

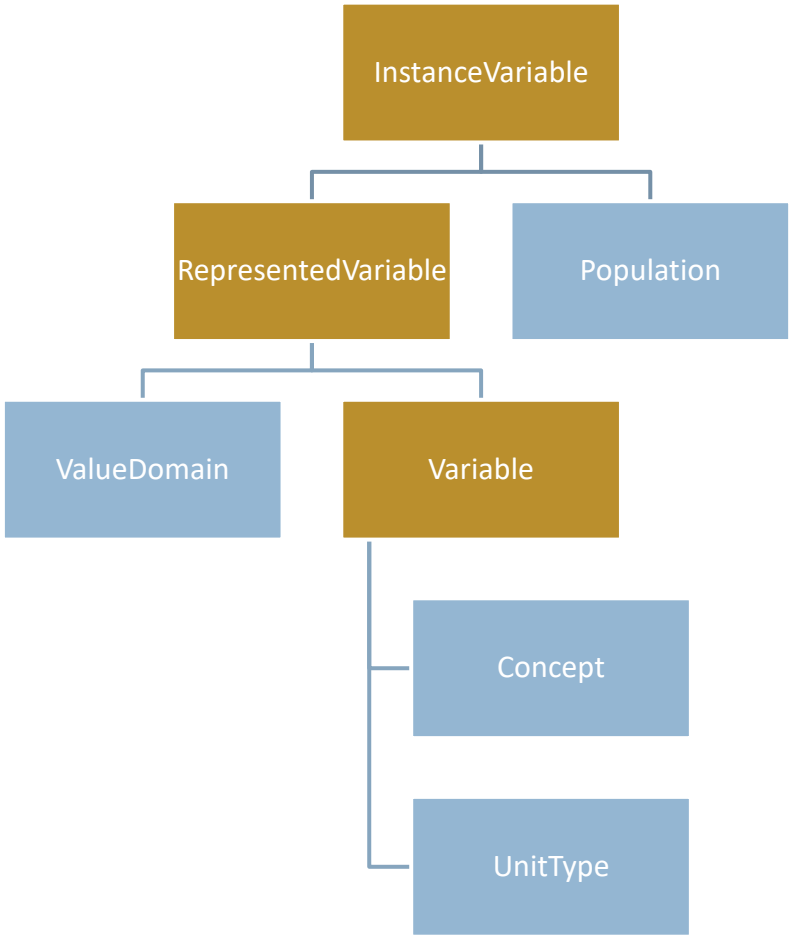
Metadata Standards for Repeated Data

- GSIM
- DDI

Generic Statistical Information Model

- Conceptual model for statistical information
- Internationally agreed definitions, attributes, and relationships
- Industrialization of official statistics

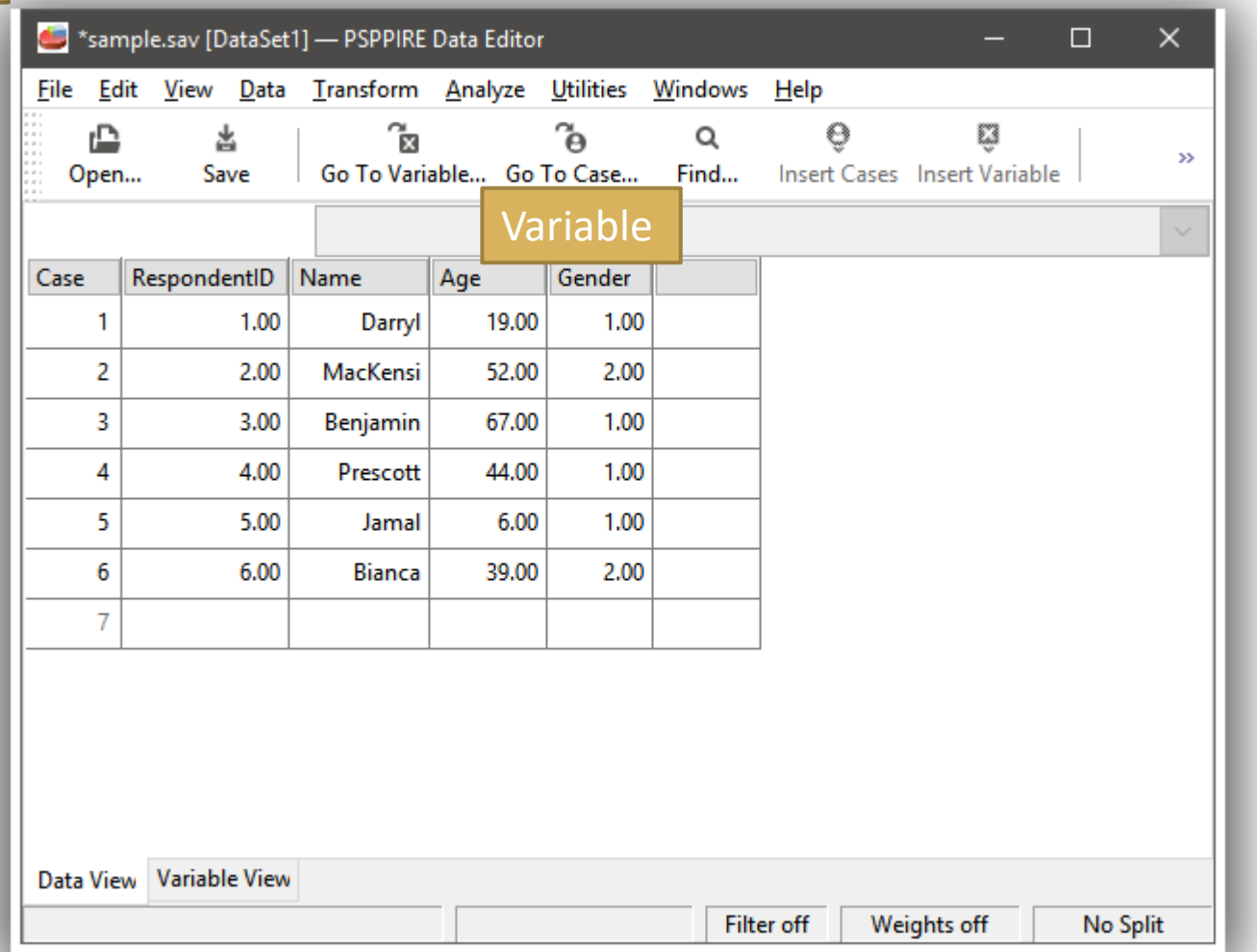
GSIM Model for Repeated Data

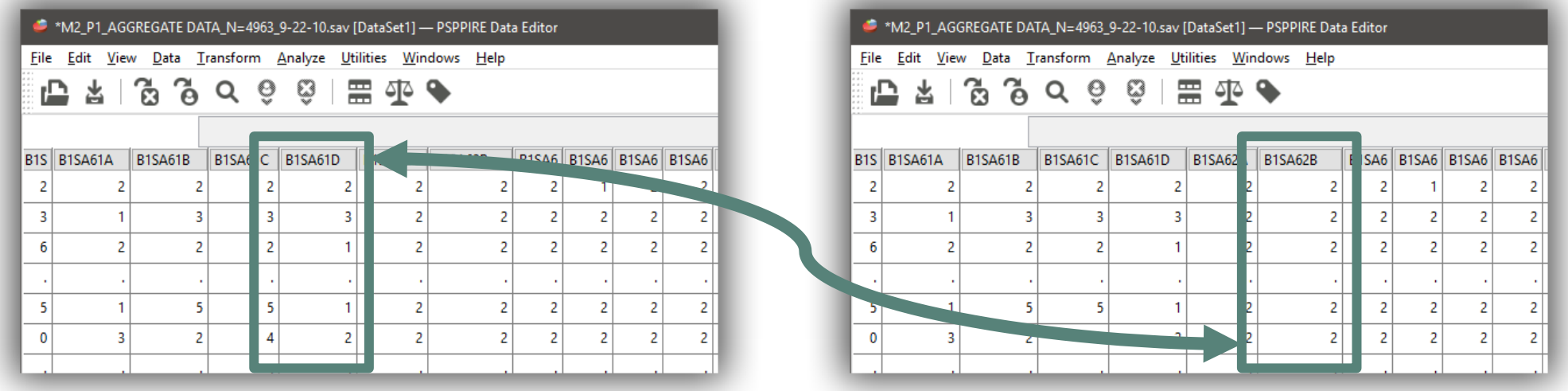


DDI Lifecycle Model for Repeated Data

- It’s pretty much the same model

PhysicalInstance

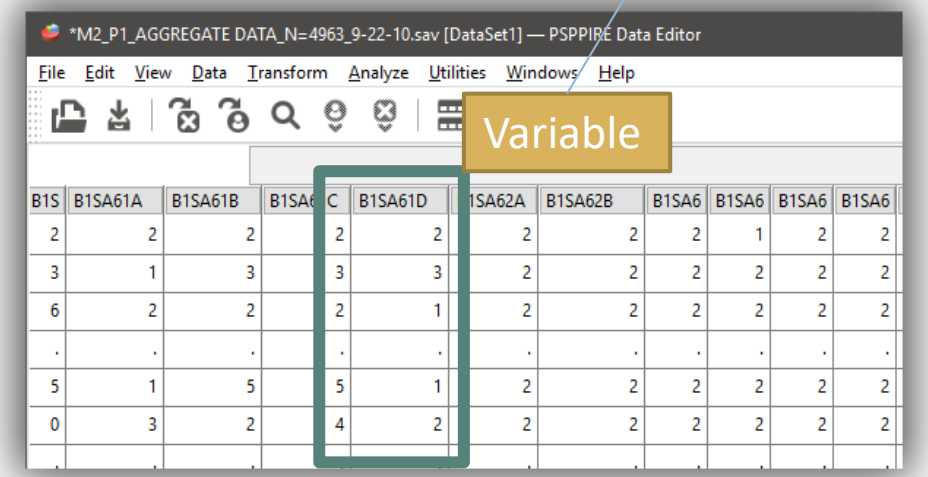



*M2_P1_AGGREGATE DATA_N=4963_9-22-10.sav [DataSet1] — PSPPIRE Data Editor
File Edit View Data Transform Analyze Utilities Windows Help
B1S B1SA61A B1SA61B B1SA61C B1SA61D
2 2 2 2 2
3 1 3 3 3
6 2 2 2 1
5 1 5 5 1
0 3 2 4 2
*M2_P1_AGGREGATE DATA_N=4963_9-22-10.sav [DataSet1] — PSPPIRE Data Editor
File Edit View Data Transform Analyze Utilities Windows Help
B1S B1SA61A B1SA61B B1SA61C B1SA61D B1SA62B
2 2 2 2 2 2
3 1 3 3 3 2
6 2 2 2 1 2
5 1 5 5 1 2
0 3 2 2

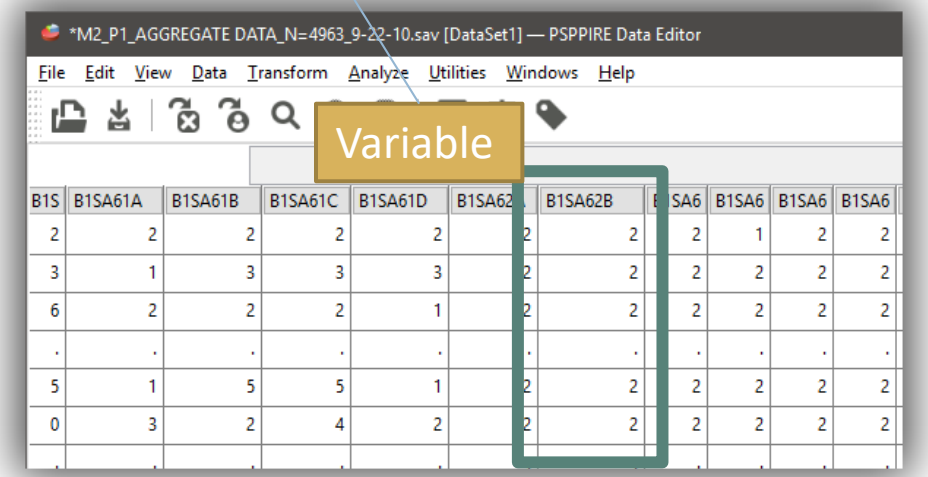
ConceptualVariable

RepresentedVariable

Variable



Variable



Three Core Elements

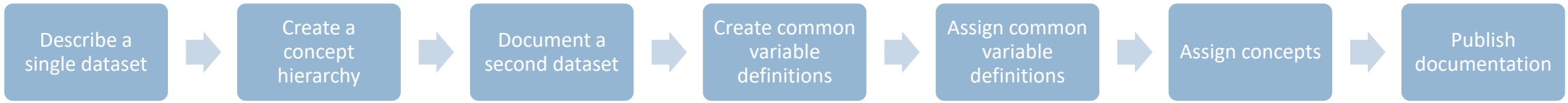
- Variable
- RepresentedVariable
- ConceptualVariable

Nine Elements

- 1. PhysicalInstance
- 2. DataRelationship
- 3. VariableStatistics
- 4. Variable
- 5. RepresentedVariable
- 6. ConceptualVariable
- 7. Concept
- 8. CodeList
- 9. Category

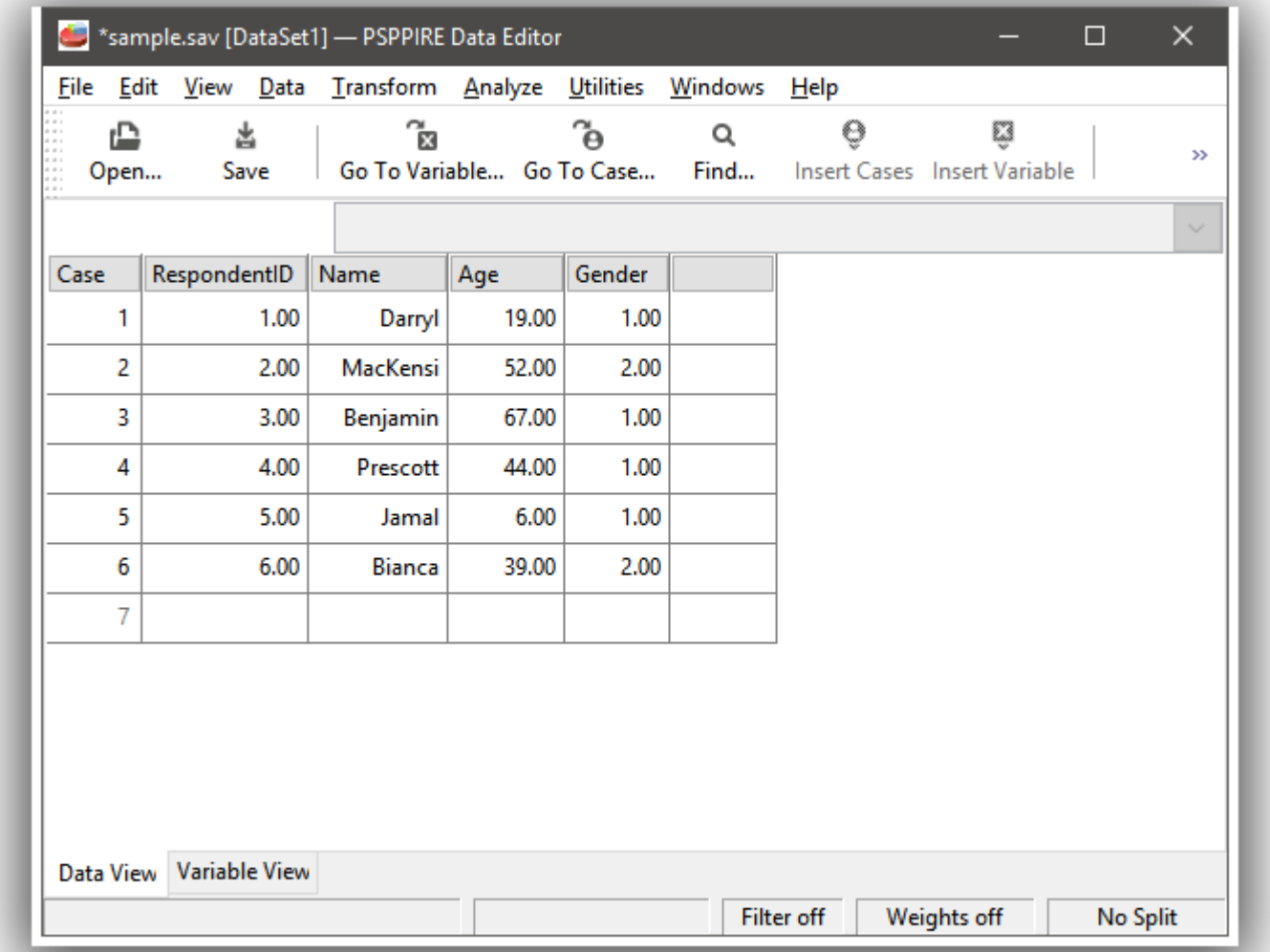
From zero documentation to actionable metadata in seven steps

Seven Steps


Describe a single dataset
Create a concept hierarchy
Document a second dataset
Create common variable definitions
Assign common variable definitions
Assign concepts
Publish documentation

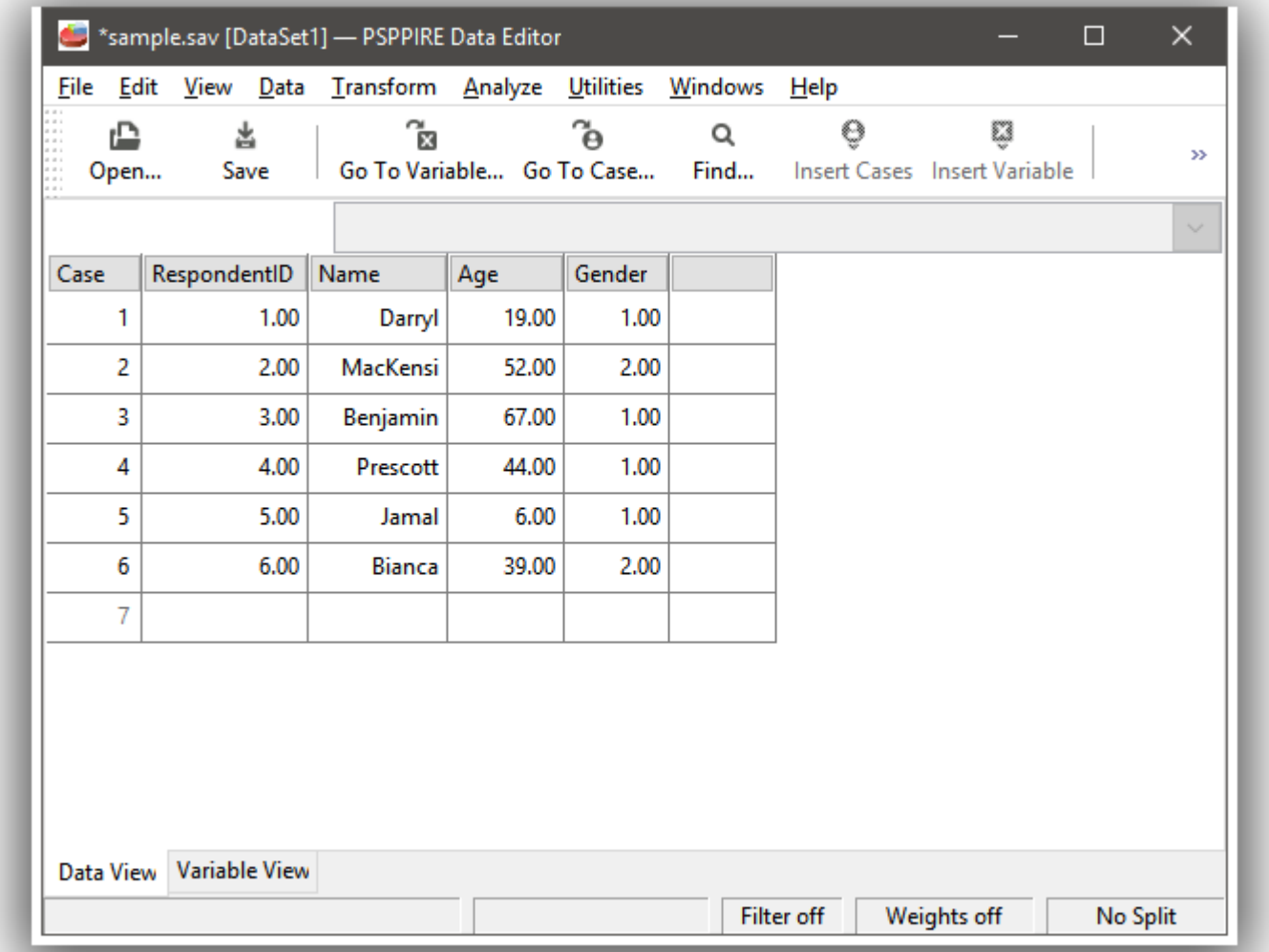
1. Describe a single dataset

- Elements
  - PhysicalInstance
  - DataRelationship
  - Variable
  - CodeList
  - Category



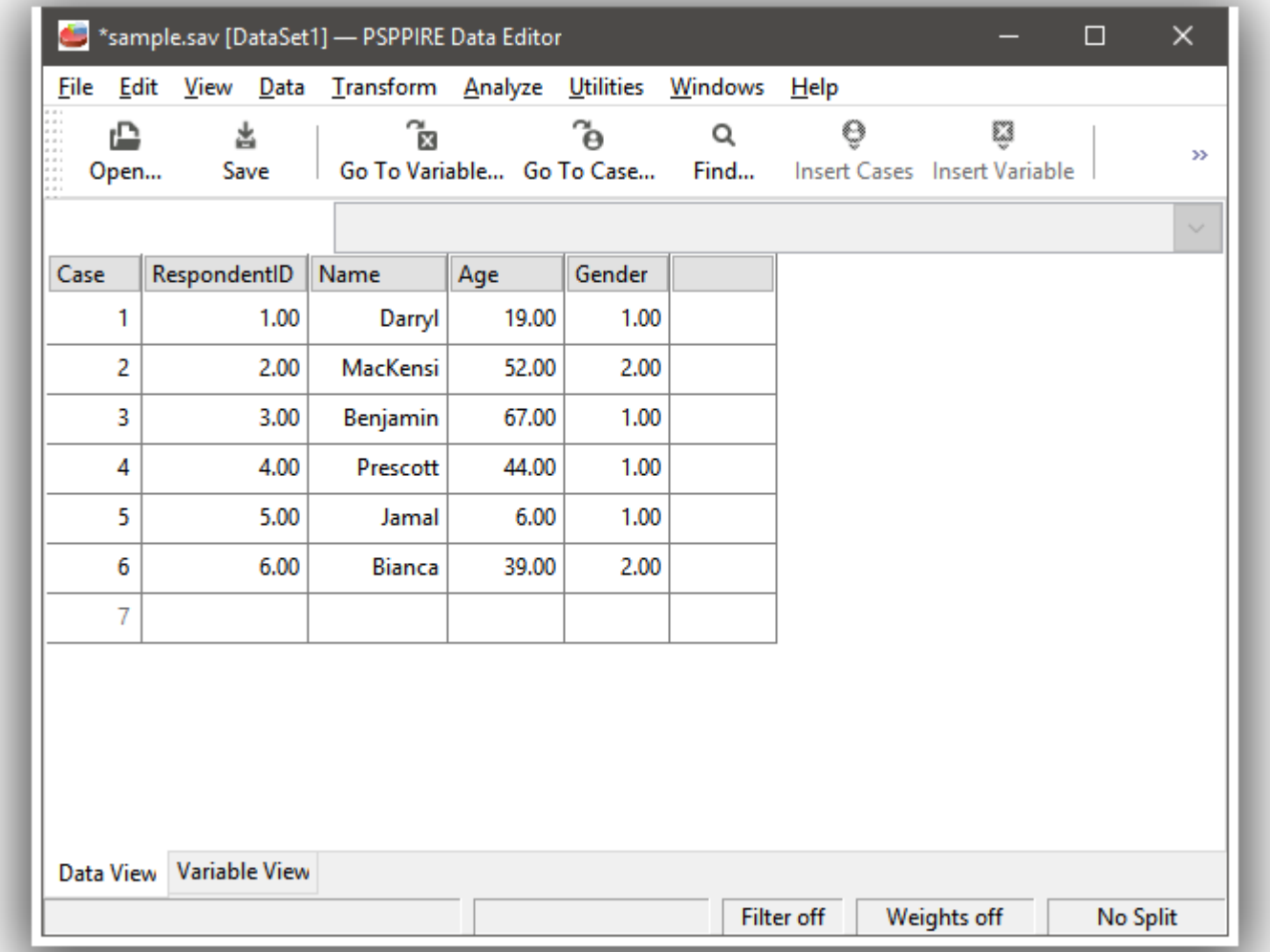
2. Create a concept hierarchy

- Elements
  - Concept



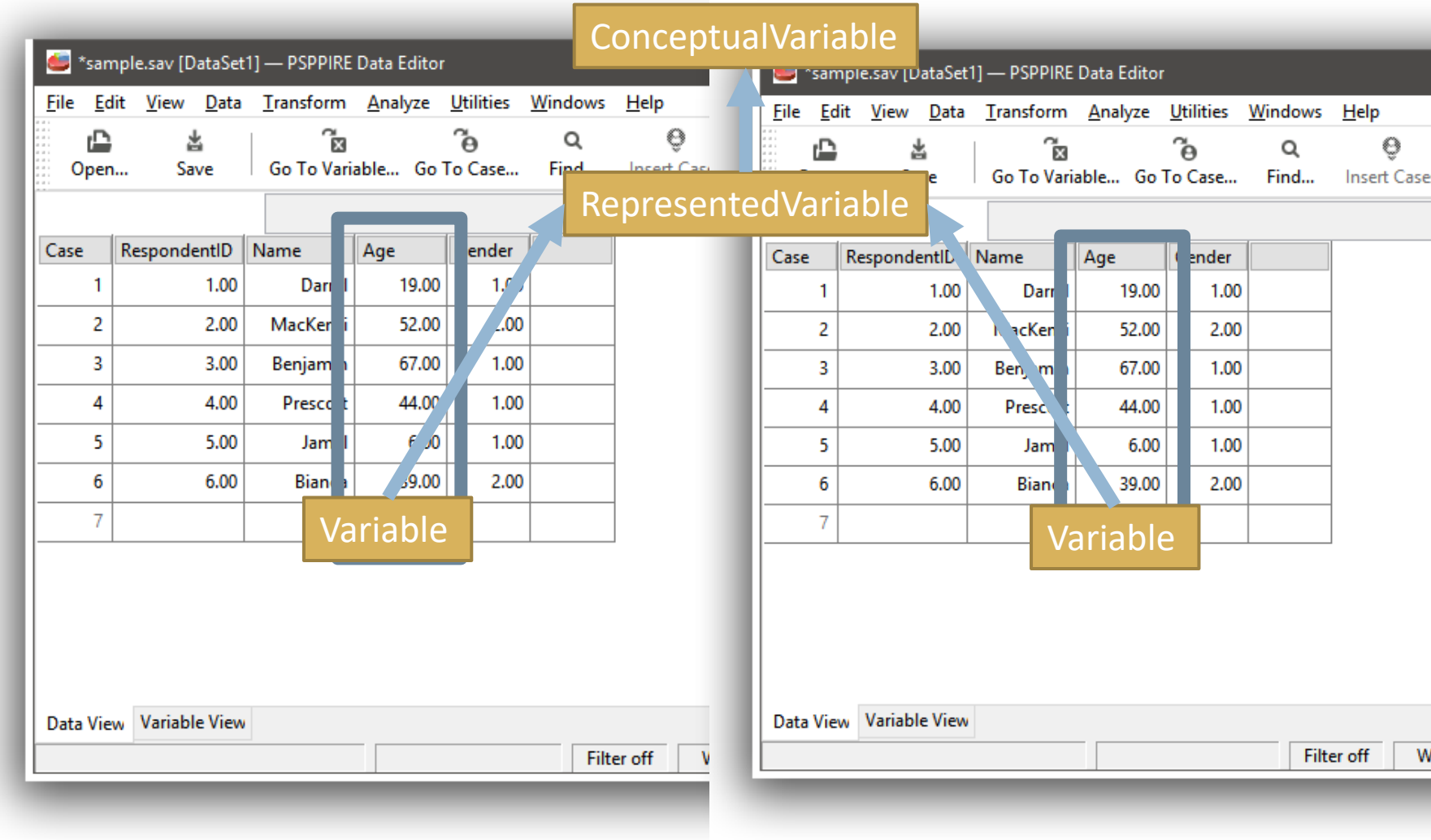
3. Describe a Second dataset

- Elements
  - PhysicalInstance
  - DataRelationship
  - Variable
  - CodeList
  - Category



4. Create common variable definitions

- Elements
  - RepresentedVariable
  - ConceptualVariable



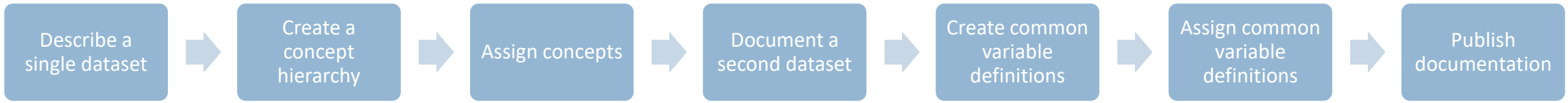
5. Assign common variable definitions

- Relationships
  - Variable -> RepresentedVariable
  - RepresentedVariable -> ConceptualVariable

6. Assign concepts

- Relationships
  - *Variable -> Concept

Repeat as Necessary



7. Publish documentation

From actionable metadata to information for researchers

What can we do with this actionable metadata?

People don’t want this

But machines do

```
<VariableName>
  <String xml:lang="en" xmlns="ddi:reusable:3_2">conferenceCount</String>
</VariableName>
<Label xmlns="ddi:reusable:3_2">
  <Content xml:lang="en">How many times, excluding this year, have you attended EI
</Label>
<Description xmlns="ddi:reusable:3_2">
  <Content xml:lang="en">This is the longer description.</Content>
</Description>
<RepresentedVariableReference>
  <Agency xmlns="ddi:reusable:3_2">int.example</Agency>
  <ID xmlns="ddi:reusable:3_2">82d7e42f-8bc0-450c-b49a-b0dec45070a0</ID>
  <Version xmlns="ddi:reusable:3_2">1</Version>
  <TypeOfObject xmlns="ddi:reusable:3_2">RepresentedVariable</TypeOfObject>
</RepresentedVariableReference>
```

Documentation Ideas

- PDF data dictionaries
- Static web sites
- Dynamic, searchable web applications
- Annotated data files

Information for Researchers

- Variables are at the center
- How can I find the variables I want?
- How do things relate to variables?
  - Other variables
  - Questions
  - Additional information

MIDUS

- MIDUS Portal

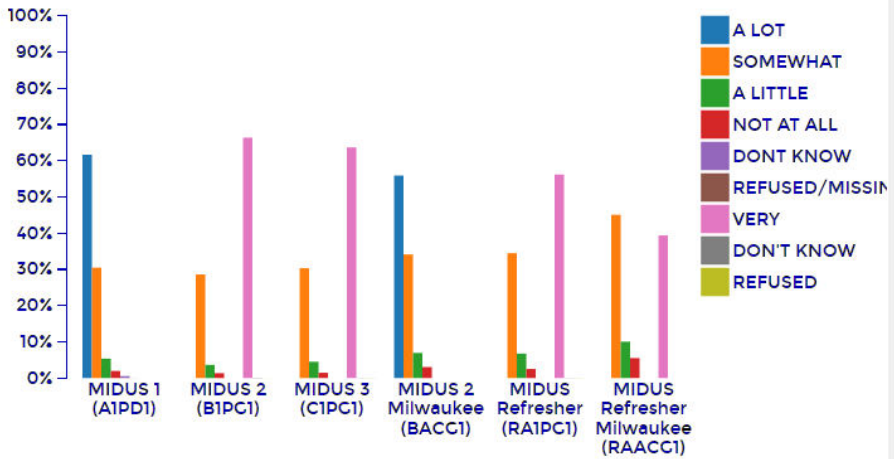
Home | Search | Explore | Basket 0 | Admin

Feedback | Help | jeremy@colectica.com

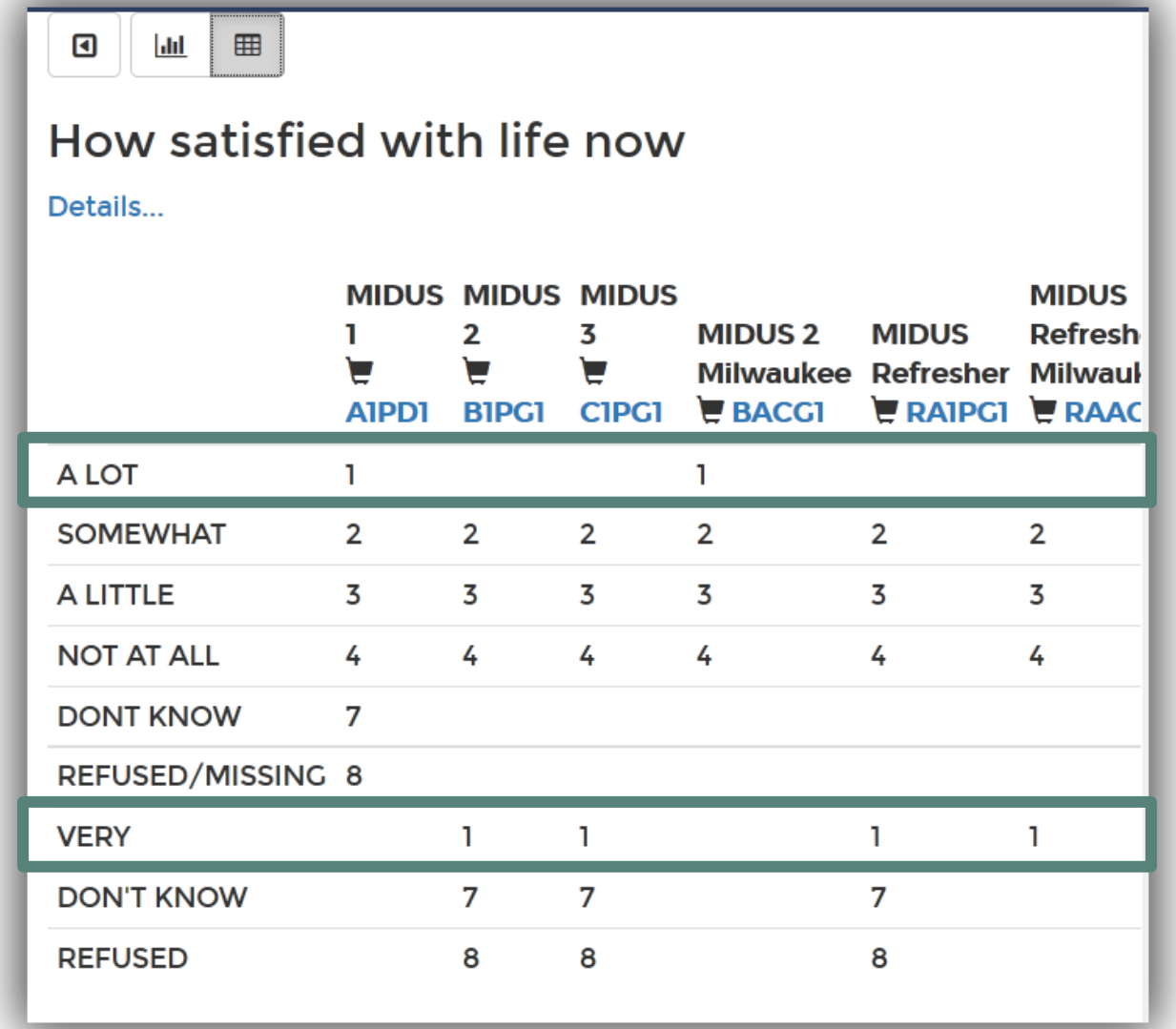
	MIDUS 1	MIDUS 2	MIDUS 3	MIDUS 2 Milwaukee	MIDUS Refresher	MID Refr Milw
How satisfied with life now	A1PD1	B1PG1	C1PG1	BACG1	RA1PG1	RAA
Satisfied with life before recession			C1PG1A		RA1PG1A	RAA
Control of life in general	A1PD2	B1PG2	C1PG2	BACG2	RA1PG2	RAA
Satisfied with self	A1PD3	B1PG3	C1PG3	BACG3	RA1PG3	RAA
Rate contribution to others	A1PD8	B1PG4	C1PG4	BACG4	RA1PG4	RAA
Disappointed with achievments	A1PD9	B1PG5	C1PG5	BACG5	RA1PG5	RAA
Level of agreement to D9	A1PD9A	B1PG5A	C1PG5A	BACG5A	RA1PG5A	RAA

How satisfied with life now

Details...



Compare Data Types



How satisfied with life now

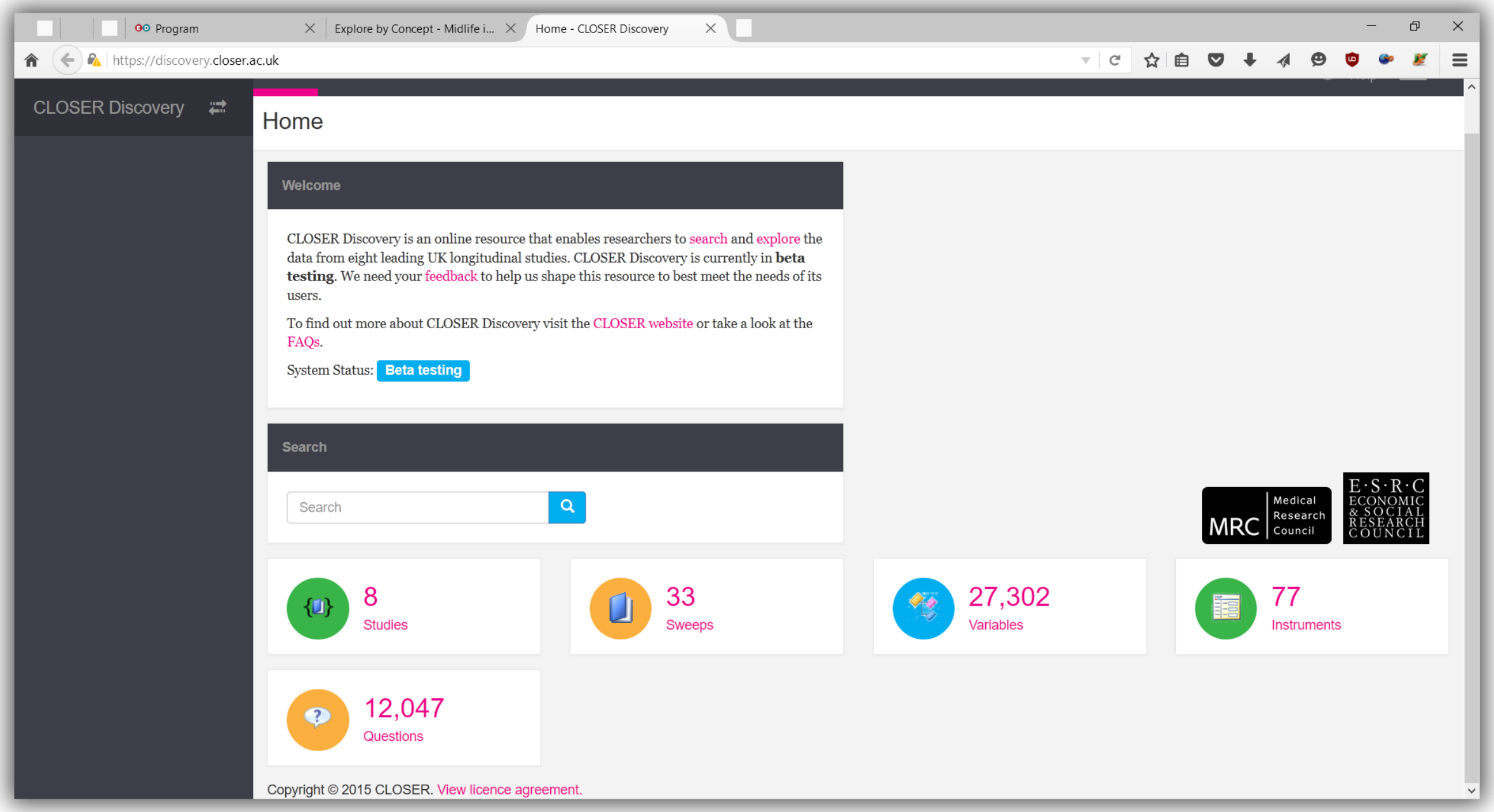
Details...

	MIDUS 1 A1PD1	MIDUS 2 B1PG1	MIDUS 3 C1PG1	MIDUS 2 Milwaukee BACG1	MIDUS Refresher RA1PG1	MIDUS Refresh Milwaul RAAC
A LOT	1			1		
SOMEWHAT	2	2	2	2	2	2
A LITTLE	3	3	3	3	3	3
NOT AT ALL	4	4	4	4	4	4
DONT KNOW	7					
REFUSED/MISSING	8					
VERY		1	1		1	1
DON'T KNOW		7	7		7	
REFUSED		8	8		8	

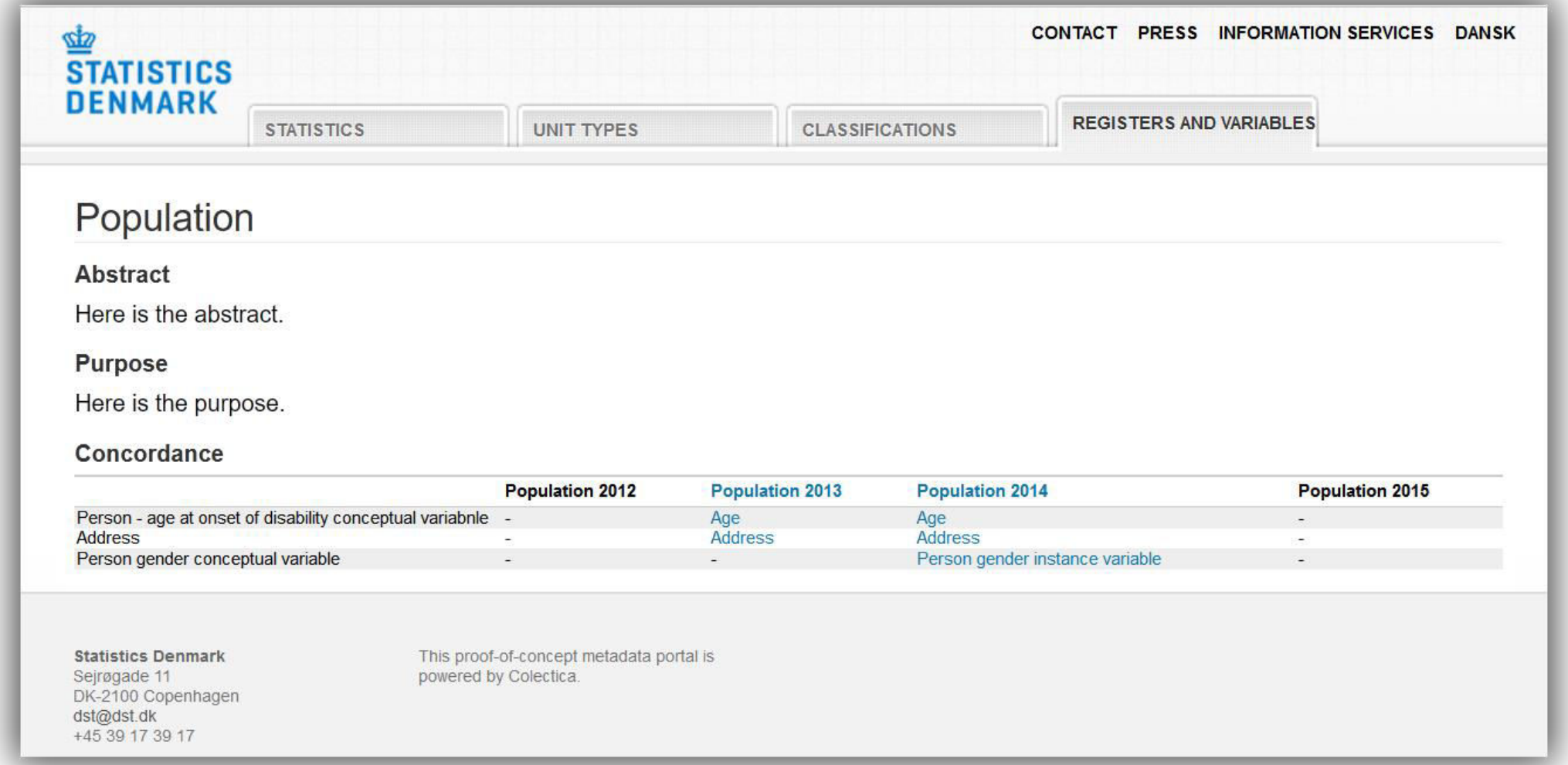
CLOSER

- 9 independent studies
- Data from 1930 – 2014
- 200,000+ variables
- Purpose: Maximize the use, value, and impact of the UK's longitudinal studies
- Document and harmonize all variables and questionnaires

- CLOSER Portal


Program
Explore by Concept - Midlife i...
Home - CLOSER Discovery
https://discovery.closer.ac.uk
CLOSER Discovery
Home
Welcome
CLOSER Discovery is an online resource that enables researchers to search and explore the data from eight leading UK longitudinal studies. CLOSER Discovery is currently in beta testing. We need your feedback to help us shape this resource to best meet the needs of its users.
To find out more about CLOSER Discovery visit the CLOSER website or take a look at the FAQs.
System Status: Beta testing
Search
Search
MRC Medical Research Council
E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL
8
Studies
33
Sweeps
27,302
Variables
77
Instruments
12,047
Questions
Copyright © 2015 CLOSER. View licence agreement.

Statistics Denmark



STATISTICS DENMARK

CONTACT PRESS INFORMATION SERVICES DANSK

STATISTICS | UNIT TYPES | CLASSIFICATIONS | REGISTERS AND VARIABLES

Population

Abstract

Here is the abstract.

Purpose

Here is the purpose.

Concordance

	Population 2012	Population 2013	Population 2014	Population 2015
Person - age at onset of disability conceptual variabnle	-	Age	Age	-
Address	-	Address	Address	-
Person gender conceptual variable	-	-	Person gender instance variable	-

Statistics Denmark
Sejrøgade 11
DK-2100 Copenhagen
dst@dst.dk
+45 39 17 39 17

This proof-of-concept metadata portal is powered by Colectica.

Multiple Data Providers, One Metadata Standard

Variable

RepresentedVariable

ConceptualVariable

THANK YOU

jeremy@colectica.com