



BEST PRACTICES FOR DOCUMENTING REPEATED STUDIES IN DDI

EDDI 2015 – Copenhagen

Goal



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

Overview



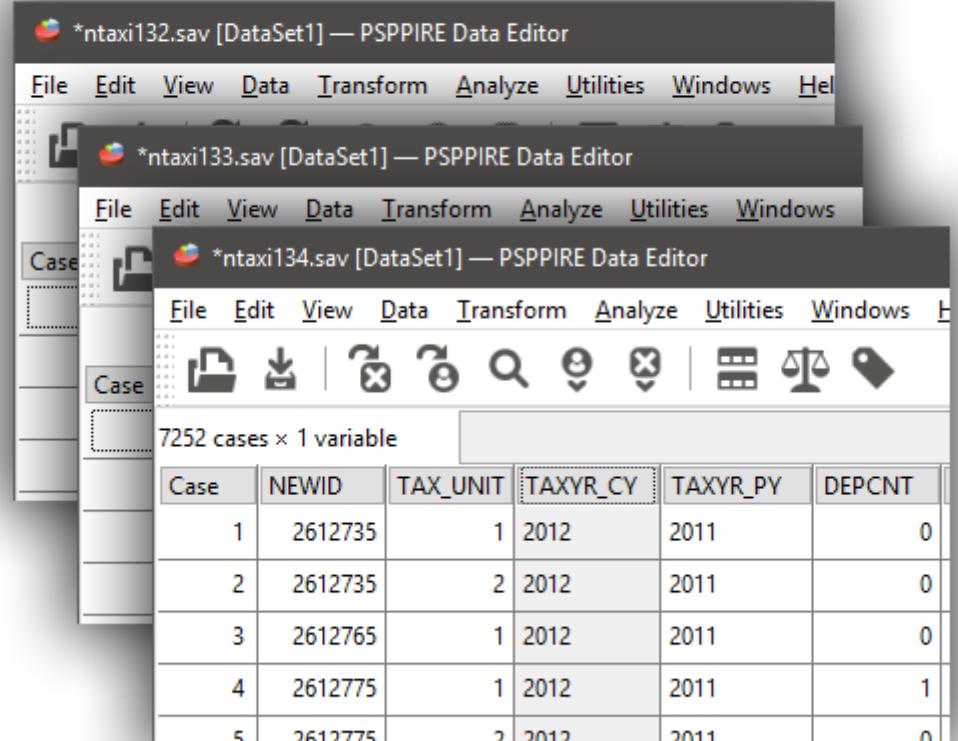
1. Background
2. Metadata standards for documenting repeated data
3. From zero documentation to actionable metadata in seven steps
4. From actionable metadata to information for researchers

Background

Official Statistics

Annual,
Quarterly,
Monthly

Same data
structures
released
regularly



The image shows three overlapping windows of the SPSS Data Editor. Each window has a title bar reading '*ntaxi132.sav [DataSet1] — PSPPIRE Data Editor'. The windows are arranged vertically, with the top one slightly offset to the right. Each window displays a data table with the following structure:

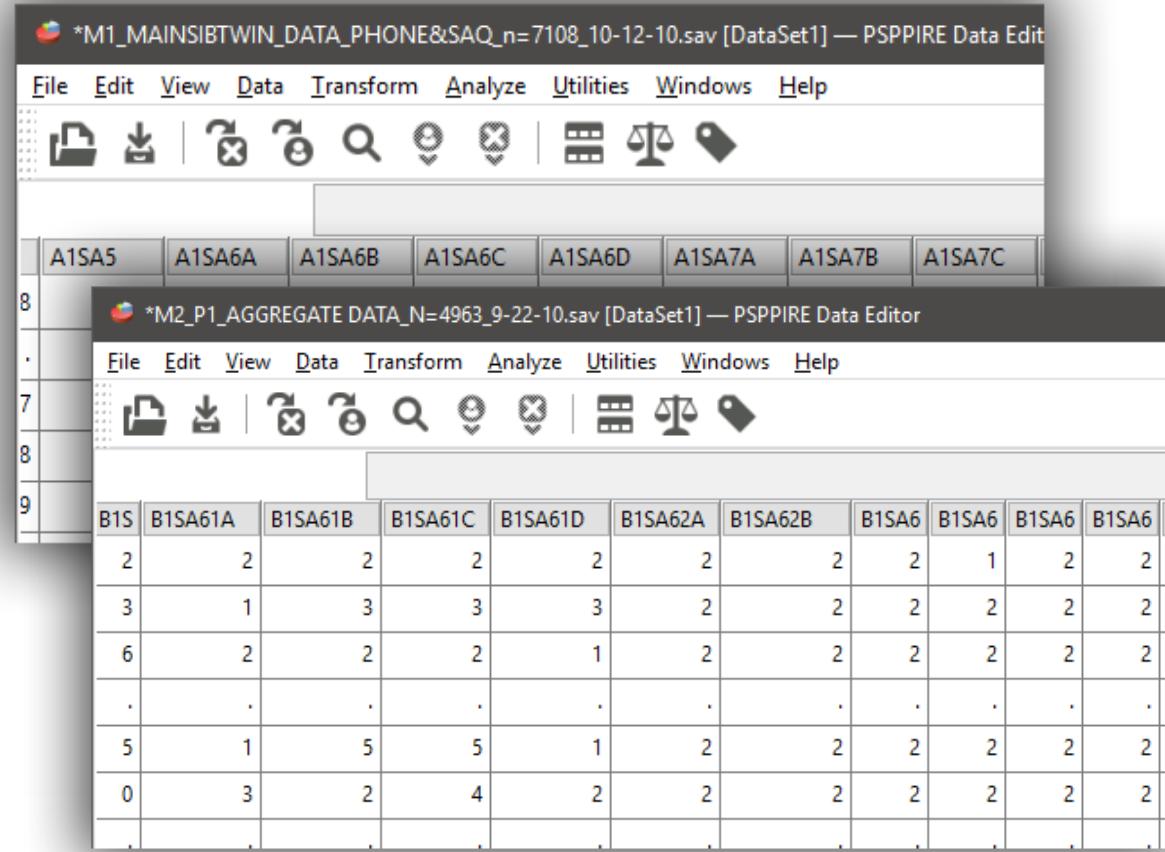
Case	NEWID	TAX_UNIT	TAXYR_CY	TAXYR_PY	DEPCNT
1	2612735	1	2012	2011	0
2	2612735	2	2012	2011	0
3	2612765	1	2012	2011	0
4	2612775	1	2012	2011	1
5	2612775	2	2012	2011	0

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

*ntaxi134.sav [DataSet1] — PSPPIRE Data Editor

File Edit View Data Transform Analyze Utilities Windows Help

7252 cases × 1 variable

Case	NEWID	TAX_UNIT	TAXYR_CY	TAXYR_PY	DEPCNT
1	2612735	1	2012	2011	0
2	2612735	2	2012	2011	0
3	2612765	1	2012	2011	0
4	2612775	1	2012	2011	1
5	2612775	2	2012	2011	0

*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	B1SA62A	B1SA62B	B1SA6	B1SA6	B1SA6	B1SA6
2	2	2	2	2	2	2	2	1	2	2
3	1	3	3	3	2	2	2	2	2	2
6	2	2	2	2	1	2	2	2	2	2
.
5	1	5	5	1	2	2	2	2	2	2
0	3	2	4	2	2	2	2	2	2	2

Comparing Data

Sweep 1 data

*M2_P1_AGGREGATE DATA_N=4963_9-22-10.sav [DataSet1] — PSPPIRE Data Editor

B1S	B1SA61A	B1SA61B	B1SA61	B1SA61D	B1SA62A	B1SA62B	B1SA6	B1SA6	B1SA6	B1SA6
2	2	2		2	2	2	2	2	1	2
3	1	3		3	2	2	2	2	2	2
6	2	2		2	1	2	2	2	2	2
.
5	1	5		5	1	2	2	2	2	2
0	3	2		4	2	2	2	2	2	2

Sweep 2 data

*M2_P1_AGGREGATE DATA_N=4963_9-22-10.sav [DataSet1] — PSPPIRE Data Editor

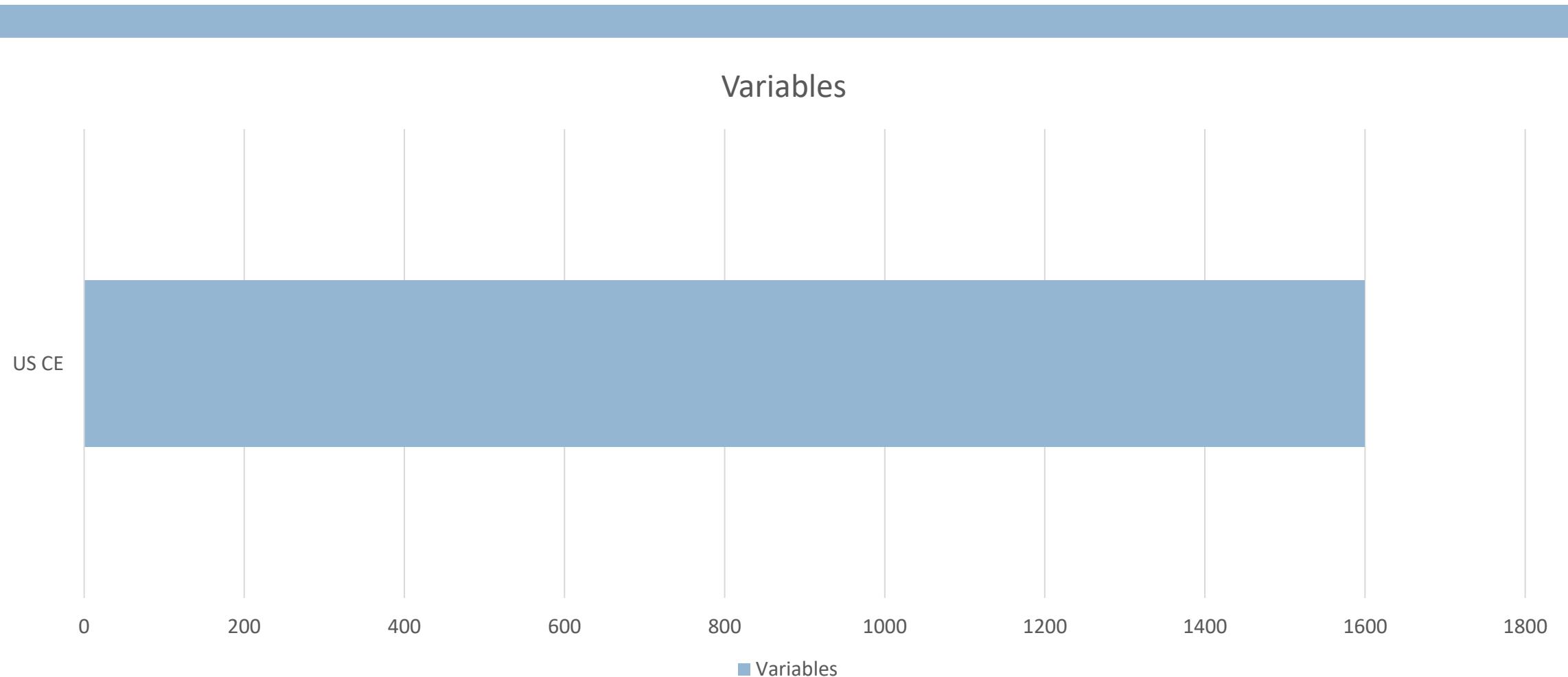
B1S	B1SA61A	B1SA61B	B1SA61C	B1SA61D	B1SA62A	B1SA62B	B1SA6	B1SA6	B1SA6	B1SA6
2	2	2	2	2	2	2	2	2	1	2
3	1	3	3	3	2	2	2	2	2	2
6	2	2	2	2	1	2	2	2	2	2
.
5	1	5	5	1	2	2	2	2	2	2
0	3	2	2	4	2	2	2	2	2	2

Scope

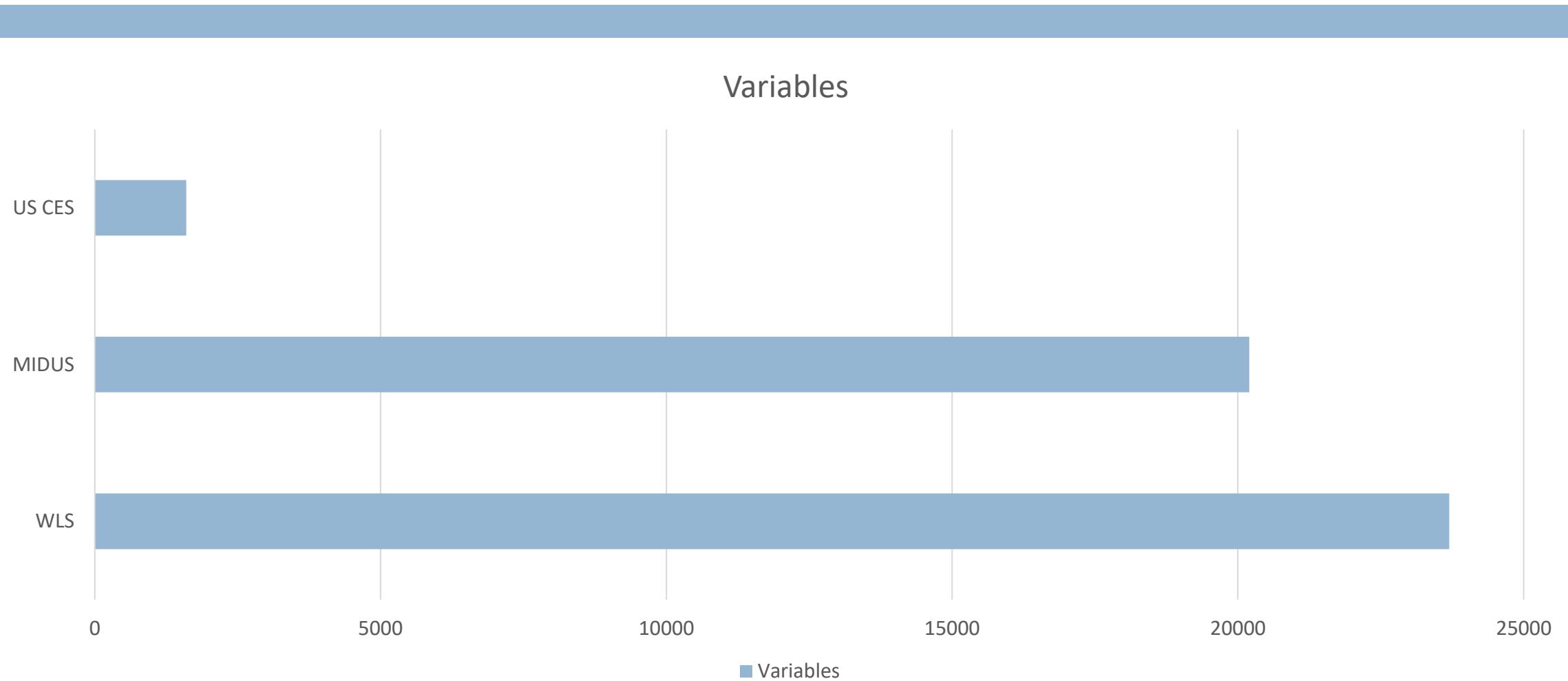


- Variable counts are huge, so rich documentation is critical

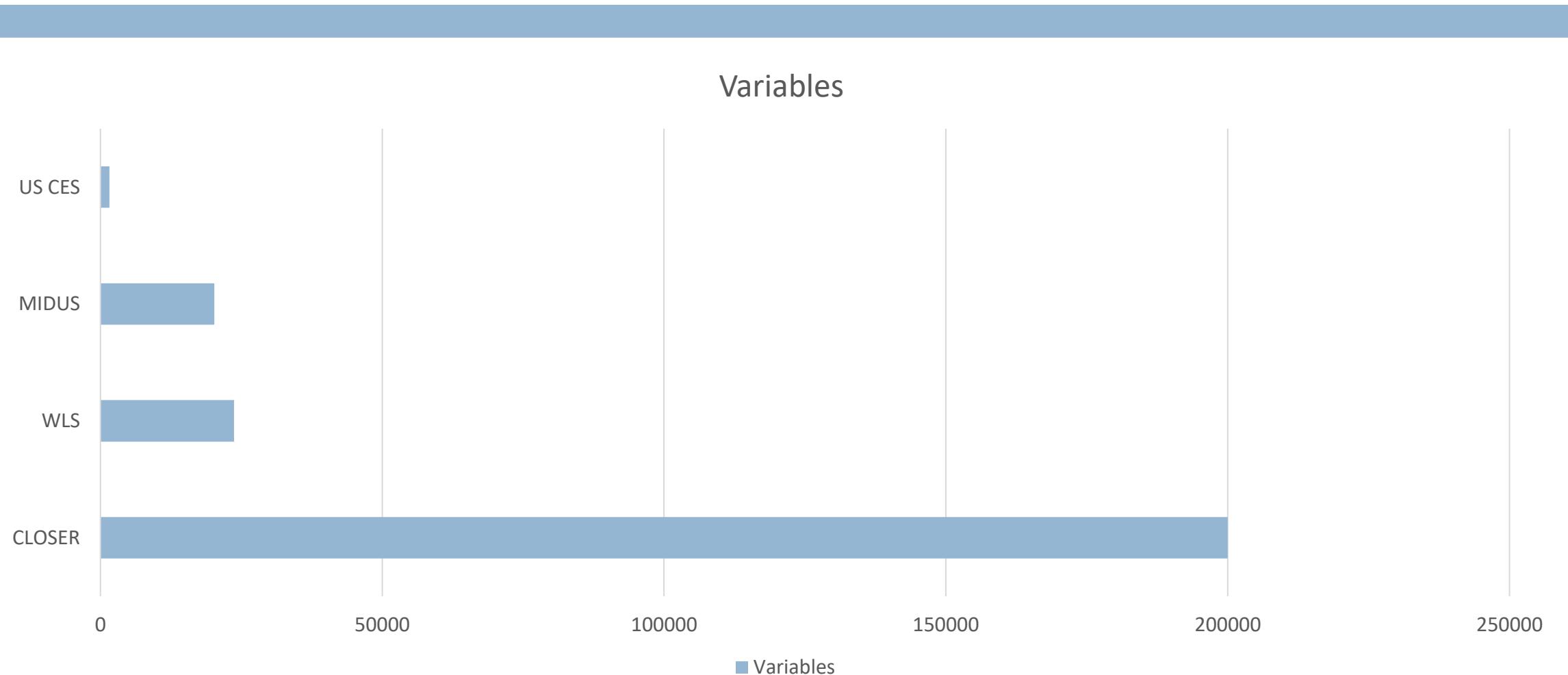
Scope: US Consumer Expenditure Survey



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

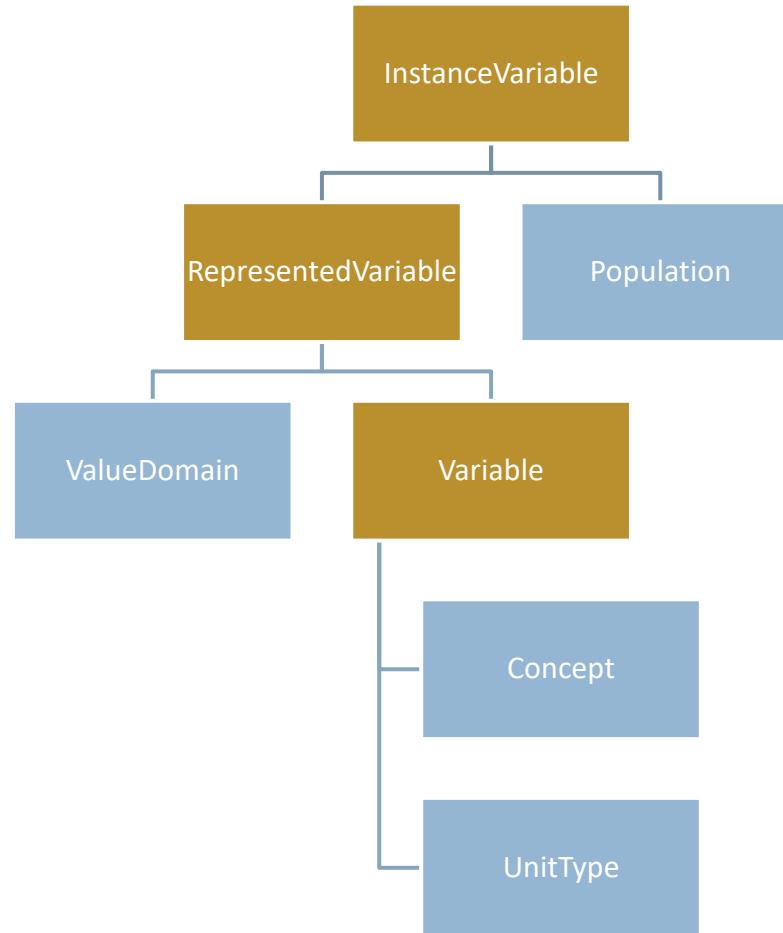
- 
- GSIM
 - DDI

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

The screenshot shows the PSPP Data Editor interface with a dataset titled "sample.sav [DataSet1]". The window title bar reads "*sample.sav [DataSet1] — PSPP Data Editor". The menu bar includes File, Edit, View, Data, Transform, Analyze, Utilities, Windows, and Help. The toolbar contains icons for Open..., Save, Go To Variable..., Go To Case..., Find..., Insert Cases, and Insert Variable. A dropdown menu labeled "Variable" is open over the toolbar. The main area displays a data table with columns: Case, RespondentID, Name, Age, Gender, and an empty column. The data rows are:

Case	RespondentID	Name	Age	Gender	
1	1.00	Darryl	19.00	1.00	
2	2.00	MacKensi	52.00	2.00	
3	3.00	Benjamin	67.00	1.00	
4	4.00	Prescott	44.00	1.00	
5	5.00	Jamal	6.00	1.00	
6	6.00	Bianca	39.00	2.00	
7					

At the bottom, there are tabs for Data View and Variable View, with Variable View selected. Other buttons include Filter off, Weights off, and No Split.

*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 B1SA6 B1SA6 B1SA6 B1SA6

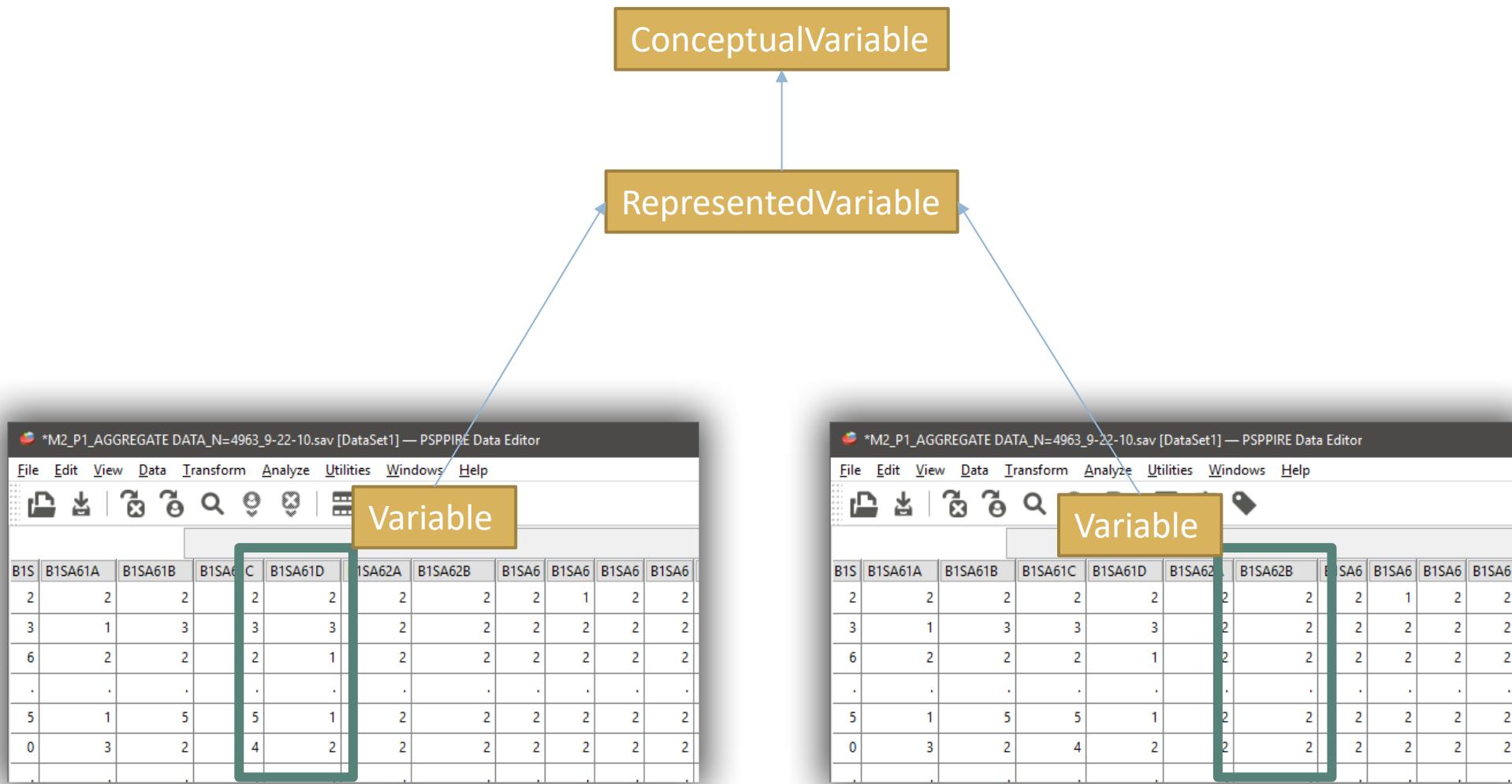
2	2	2		2	2	2	2	2
3	1	3		3	3	2	2	2
6	2	2		2	1	2	2	2
.	2	.
5	1	5		5	1	2	2	2
0	3	2		4	2	2	2	2
.
.
.
.

*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 B1SA62 B1SA62B B1SA6 B1SA6 B1SA6

2	2	2		2	2	2	2	2	2
3	1	3		3	3	2	2	2	2
6	2	2		2	2	1	2	2	2
.
5	1	5		5	1	2	2	2	2
0	3	2		4	2	2	2	2	2
.
.
.
.



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

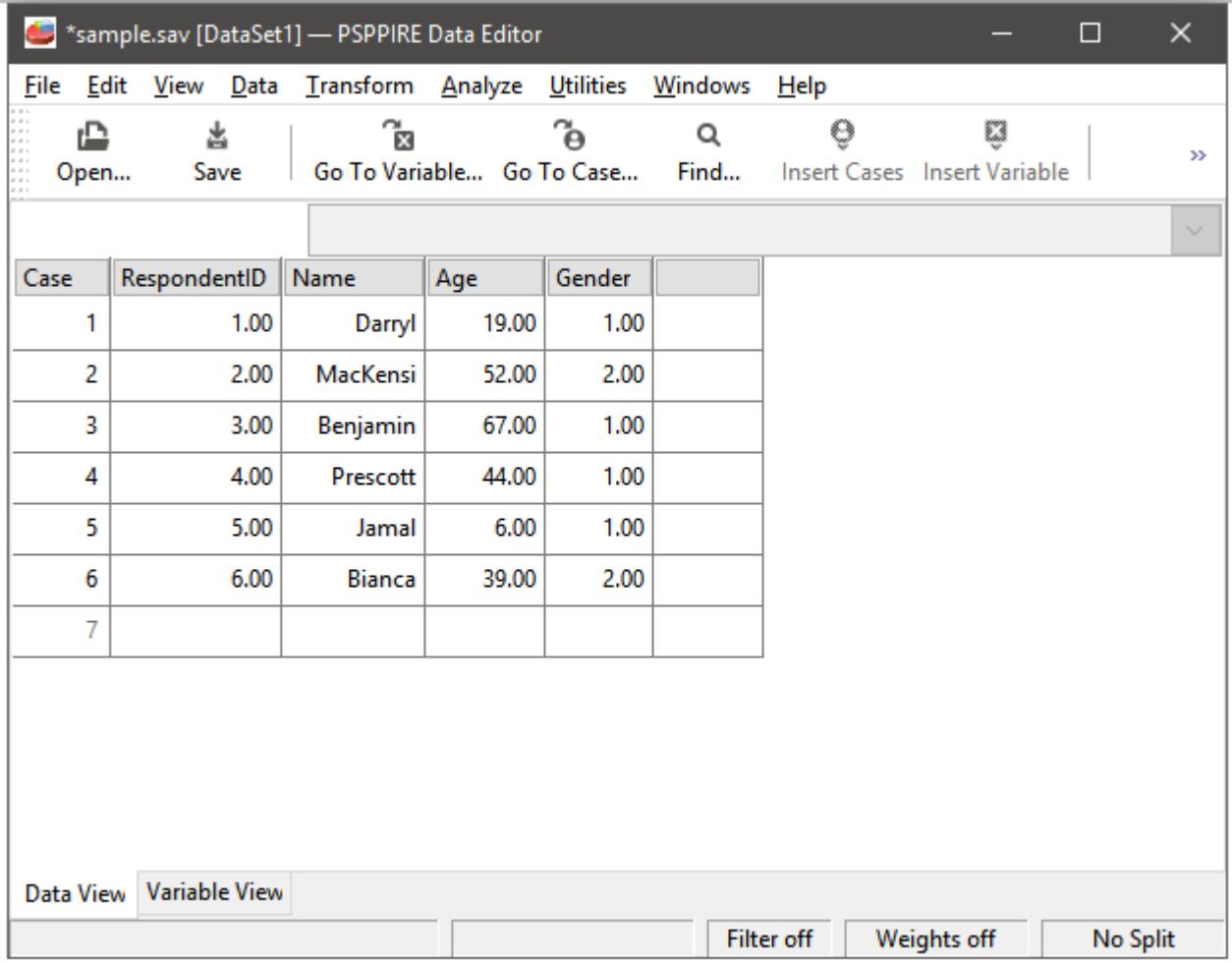
Seven Steps



1. Describe a single dataset

□ Elements

- PhysicalInstance
- DataRelationship
- Variable
- CodeList
- Category



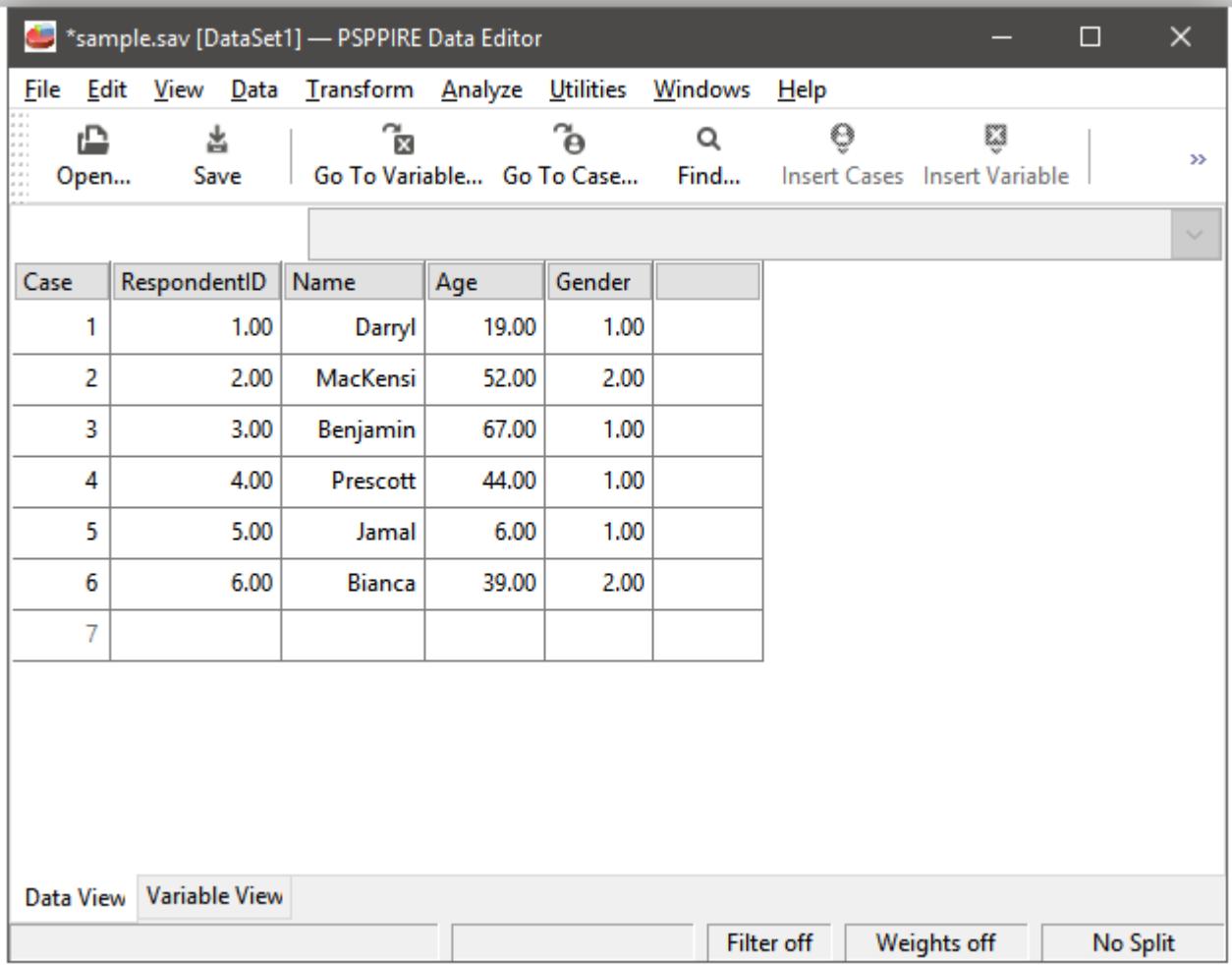
The screenshot shows the PSPPIRE Data Editor interface. The title bar reads "*sample.sav [DataSet1] — PSPPIRE Data Editor". The menu bar includes File, Edit, View, Data, Transform, Analyze, Utilities, Windows, and Help. Below the menu is a toolbar with icons for Open..., Save, Go To Variable..., Go To Case..., Find..., Insert Cases, and Insert Variable. The main area displays a data grid with 7 rows (Case numbers 1 to 7) and 5 columns (RespondentID, Name, Age, Gender, and an empty column). The data is as follows:

Case	RespondentID	Name	Age	Gender	
1	1.00	Darryl	19.00	1.00	
2	2.00	MacKensi	52.00	2.00	
3	3.00	Benjamin	67.00	1.00	
4	4.00	Prescott	44.00	1.00	
5	5.00	Jamal	6.00	1.00	
6	6.00	Bianca	39.00	2.00	
7					

At the bottom, there are tabs for Data View (selected) and Variable View, and buttons for Filter off, Weights off, and No Split.

2. Create a concept hierarchy

- Elements
 - Concept



The screenshot shows the PSPPIRE Data Editor interface. The title bar reads "*sample.sav [DataSet1] — PSPPIRE Data Editor". The menu bar includes File, Edit, View, Data, Transform, Analyze, Utilities, Windows, and Help. Below the menu is a toolbar with icons for Open..., Save, Go To Variable..., Go To Case..., Find..., Insert Cases, and Insert Variable. The main area displays a data grid with the following columns: Case, RespondentID, Name, Age, Gender, and an empty column. The data rows are:

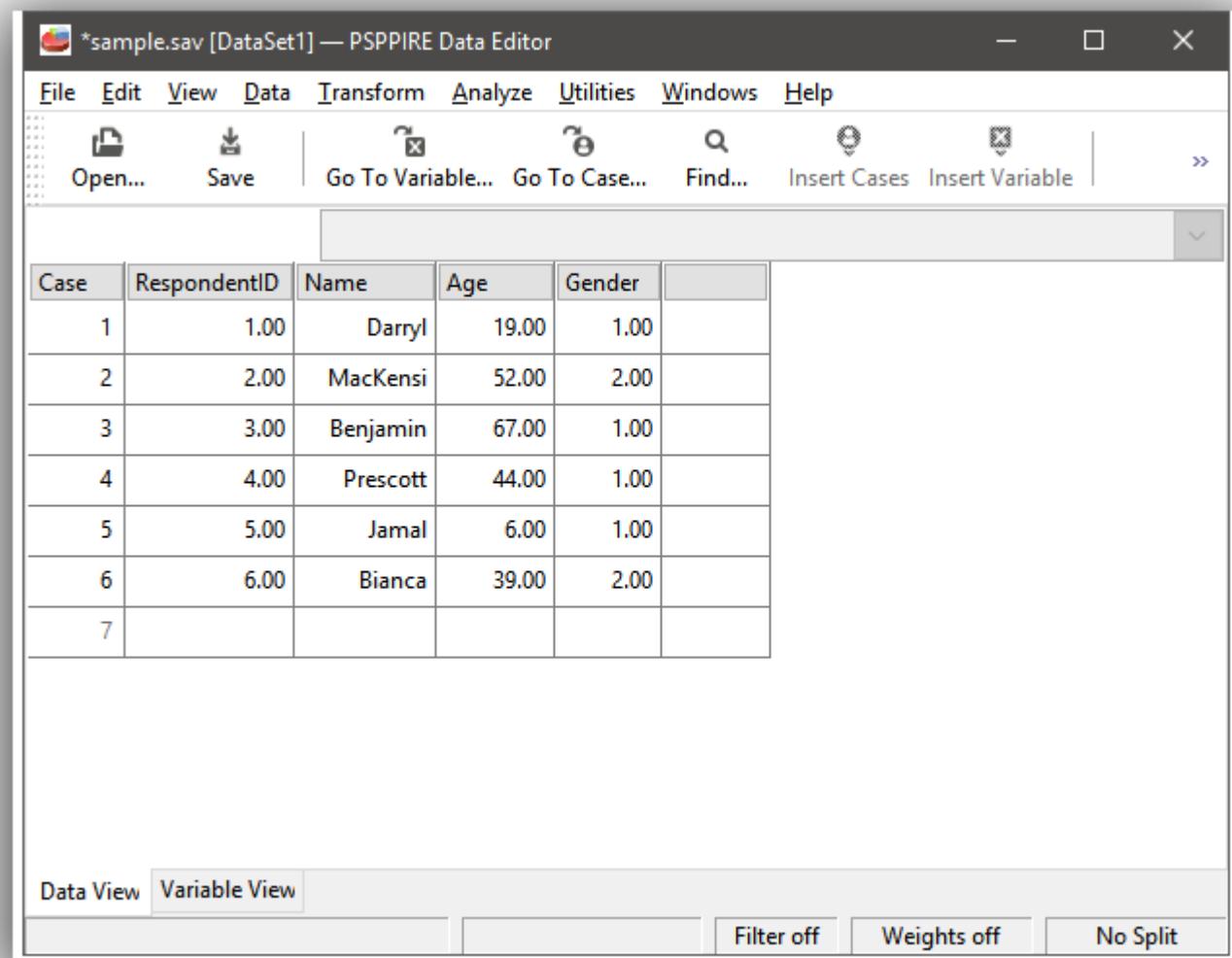
Case	RespondentID	Name	Age	Gender	
1	1.00	Darryl	19.00	1.00	
2	2.00	MacKensi	52.00	2.00	
3	3.00	Benjamin	67.00	1.00	
4	4.00	Prescott	44.00	1.00	
5	5.00	Jamal	6.00	1.00	
6	6.00	Bianca	39.00	2.00	
7					

At the bottom, there are tabs for Data View (selected) and Variable View, and buttons for Filter off, Weights off, and No Split.

3. Describe a Second dataset

□ Elements

- PhysicalInstance
- DataRelationship
- Variable
- CodeList
- Category



The screenshot shows the PSPPIRE Data Editor interface. The title bar reads "*sample.sav [DataSet1] — PSPPIRE Data Editor". The menu bar includes File, Edit, View, Data, Transform, Analyze, Utilities, Windows, and Help. The toolbar contains icons for Open..., Save, Go To Variable..., Go To Case..., Find..., Insert Cases, and Insert Variable. The main area displays a data table with 7 rows (Case numbers 1 to 7) and 5 columns (RespondentID, Name, Age, Gender, and an empty column). The data is as follows:

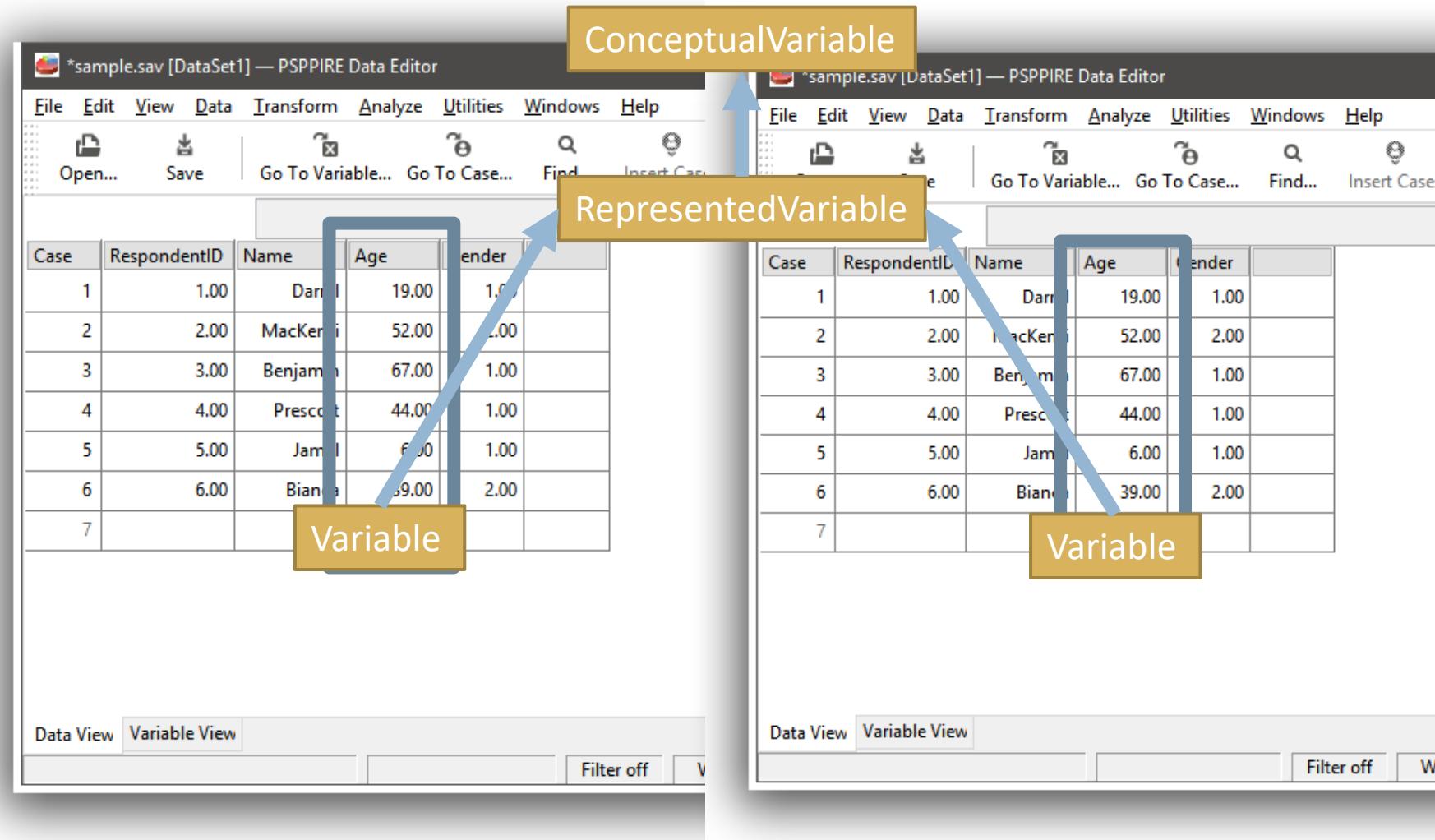
Case	RespondentID	Name	Age	Gender	
1	1.00	Darryl	19.00	1.00	
2	2.00	MacKensi	52.00	2.00	
3	3.00	Benjamin	67.00	1.00	
4	4.00	Prescott	44.00	1.00	
5	5.00	Jamal	6.00	1.00	
6	6.00	Bianca	39.00	2.00	
7					

At the bottom, there are tabs for Data View (selected) and Variable View, and buttons for Filter off, Weights off, and No Split.

4. Create common variable definitions

□ Elements

- RepresentedVariable
- ConceptualVariable



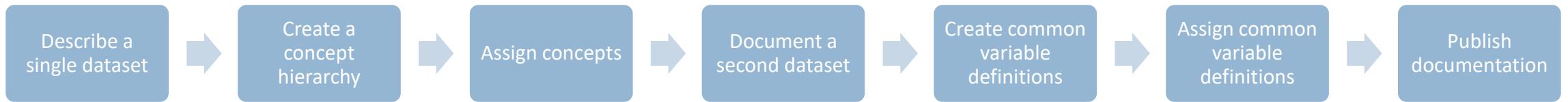
5. Assign common variable definitions

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

6. Assign concepts

- Relationships
 - *Variable -> Concept

Repeat as Necessary



7. Publish documentation



From actionable metadata to information for
researchers

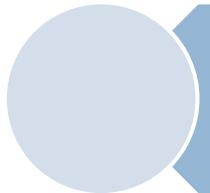
What can we do with this actionable
metadata?

People don't want this

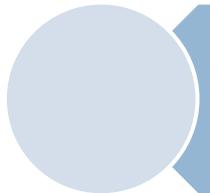
But machines
do

```
<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>
```

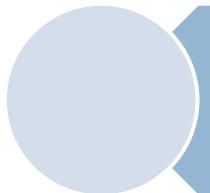
Documentation Ideas



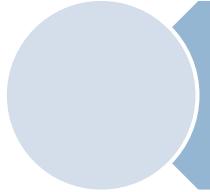
PDF data dictionaries



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

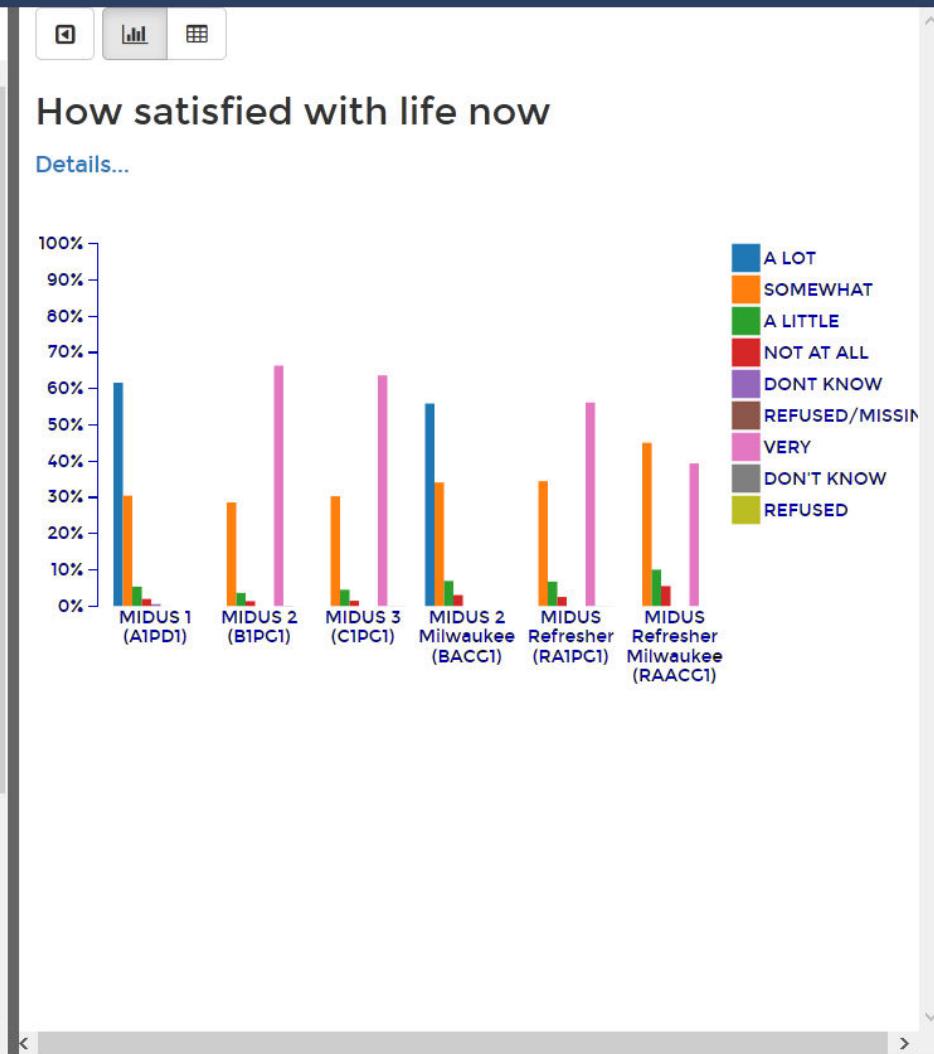


Help



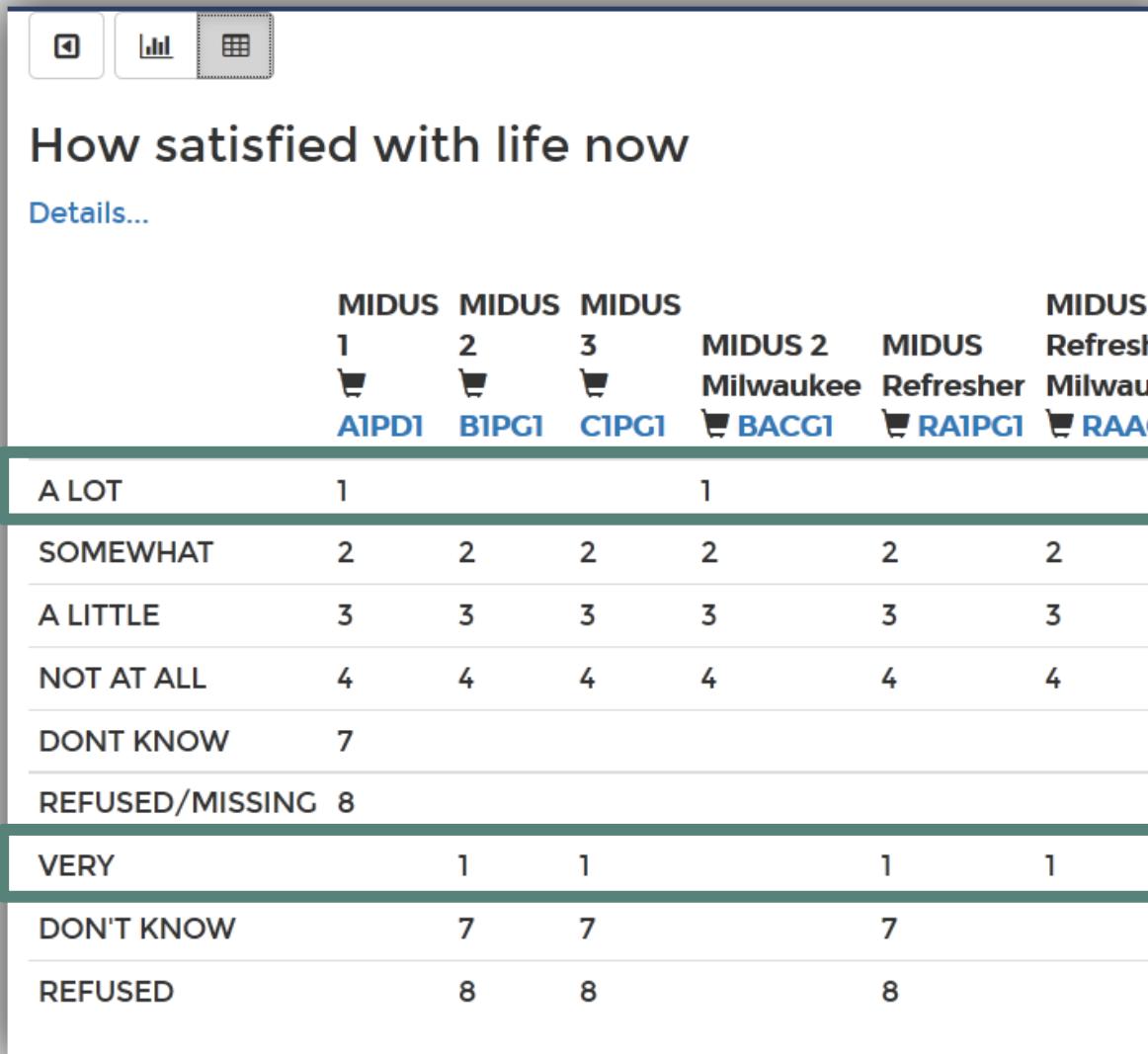
jeremy@colectica.com

	MIDUS 1	MIDUS 2	MIDUS 3	MIDUS 2 Milwaukee	MIDUS Refresher	MID Refresher Milwaukee
How satisfied with life now	A1PD1	B1PG1	C1PG1	BACG1	RA1PG1	RAA1PG1
Satisfied with life before recession			C1PC1A		RA1PG1A	RAA1PG1A
Control of life in general	A1PD2	B1PG2	C1PG2	BACG2	RA1PG2	RAA1PG2
Satisfied with self	A1PD3	B1PG3	C1PG3	BACG3	RA1PG3	RAA1PG3
Rate contribution to others	A1PD8	B1PG4	C1PG4	BACG4	RA1PG4	RAA1PG4
Disappointed with achievements	A1PD9	B1PG5	C1PG5	BACG5	RA1PG5	RAA1PG5
Level of agreement to D9	A1PD9A	B1PG5A	C1PG5A	BACG5A	RA1PG5A	RAA1PG5A



- + Admin
- + Recess
- + Health
- + Education
- + House
- + Caregiv
- + Living A
- + Race A
- + Life Sat
- + Your He
- + Health
- + Parent
- + Person
- + Work
- + Finance
- + Commu
- + Your Ne
- + Social F
- + Childre
- + Marria
- + Sexuali
- + Religio
- + Discrin
- + Life Ov
- + Childh
- + Images

Compare Data Types



The screenshot shows a data visualization interface with a blue header bar. In the top left corner of the main area, there are three icons: a square with a dot, a bar chart, and a grid. Below these icons, the title "How satisfied with life now" is displayed in bold black font. Underneath the title is a link labeled "Details...". The main content is a grid table with seven columns and eight rows of data. The columns are labeled at the top: MIDUS 1, MIDUS 2, MIDUS 3, MIDUS Milwaukee, MIDUS Refresher, MIDUS Refresh, and MIDUS Milwaukee. The rows represent different response categories: A LOT, SOMEWHAT, A LITTLE, NOT AT ALL, DONT KNOW, REFUSED/MISSING, VERY, and REFUSED. The data values are represented by small blue icons with labels like "A1PD1", "B1PG1", "C1PC1", "BACG1", "RA1PG1", and "RAAC". The first and last rows ("A LOT" and "REFUSED") are highlighted with a green border.

	MIDUS 1 A1PD1	MIDUS 2 B1PG1	MIDUS 3 C1PC1	MIDUS Milwaukee BACG1	MIDUS Refresher RA1PG1	MIDUS Refresh 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](#)

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

8 Studies

33 Sweeps

27,302 Variables

77 Instruments

12,047 Questions

Copyright © 2015 CLOSER. [View licence agreement](#).

MRC | Medical Research Council

E·S·R·C
ECONOMIC & SOCIAL RESEARCH COUNCIL

Statistics Denmark

The screenshot shows the Statistics Denmark website with a blue header bar. In the top right corner, there are links for CONTACT, PRESS, INFORMATION SERVICES, and DANSK. Below the header, there is a navigation menu with four tabs: STATISTICS, UNIT TYPES, CLASSIFICATIONS, and REGISTERS AND VARIABLES. The STATISTICS tab is currently selected. The main content area has a title "Population". Under this title, there are three sections: "Abstract", "Purpose", and "Concordance". The "Abstract" section contains the text "Here is the abstract.". The "Purpose" section contains the text "Here is the purpose.". The "Concordance" section contains a table comparing variables across four years: 2012, 2013, 2014, and 2015. The table has five columns: Population 2012, Population 2013, Population 2014, and Population 2015. The rows represent different conceptual variables: "Person - age at onset of disability conceptual variabnle", "Address", and "Person gender conceptual variable". The table shows that the "Age" variable was introduced in 2013, and the "Person gender instance variable" was introduced in 2014. The bottom left corner of the page displays the address and contact information for Statistics Denmark.

	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