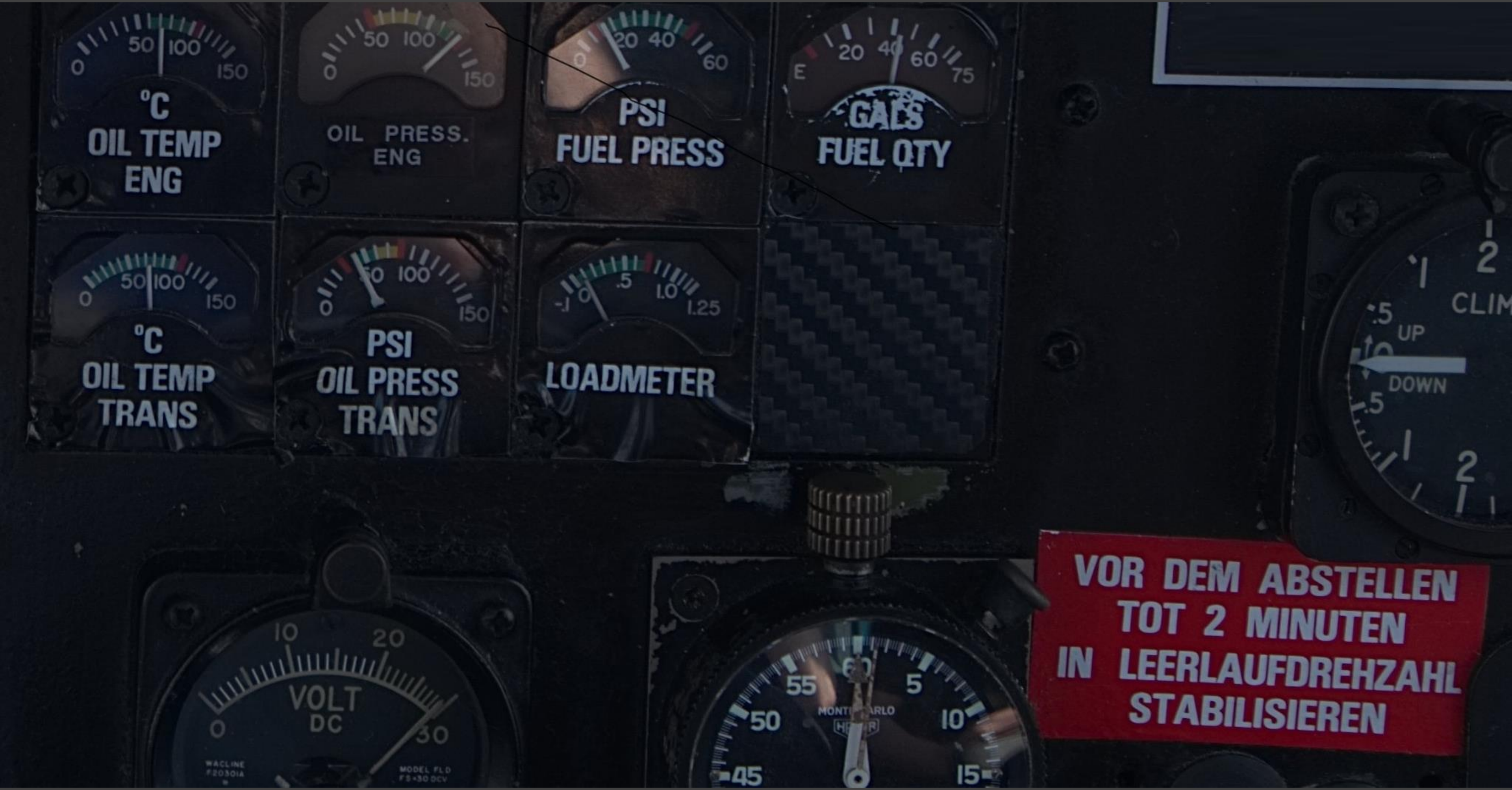
colectica

Preparing Data Files for Preservation with Colectica Datasets

IASSIST 2017 – University of Kansas

20
40
60

E
20
40
60
75
GAS
FUEL QTY

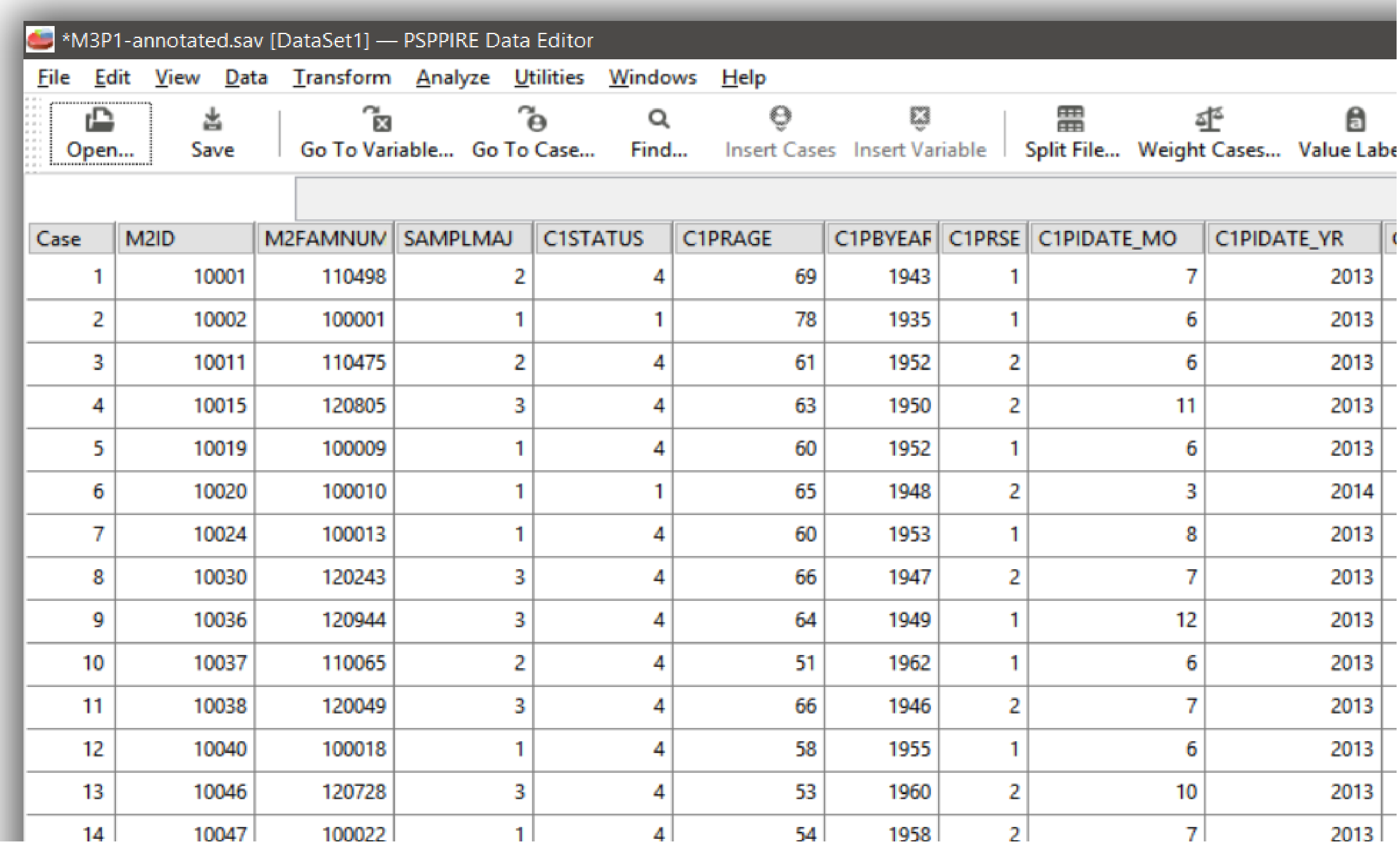

°C
OIL TEMP
ENG
OIL PRESS.
ENG
PSI
FUEL PRESS
GALS
FUEL QTY
°C
OIL TEMP
TRANS
PSI
OIL PRESS
TRANS
LOADMETER
CLIM
UP
DOWN
VOLT
DC
MONTE CARLO
HEUER
VOR DEM ABSTELLEN
TOT 2 MINUTEN
IN LEERLAUFDREHZAHL
STABILISIEREN

A/F FUEL
TRANS CHIP
T/R CHIP
FILTER
ENG CHIP
DET
DET
DET
Gen
SPARE
GEN FAIL
FUEL
FUEL
TRANS
TRANS
ENG
PRESS
OUT
°C OIL TEMP ENG
OIL PRESS. ENG
PSI FUEL PRESS
GALS FUEL QTY
°C OIL TEMP TRANS
PSI OIL PRESS TRANS
LOADMETER
VOLT DC
UP
DOWN
VOR DEM ABSTELLEN TOT 2 MINUTEN IN LEERLAUFDREHZAHL STABILISIEREN

Context Matters

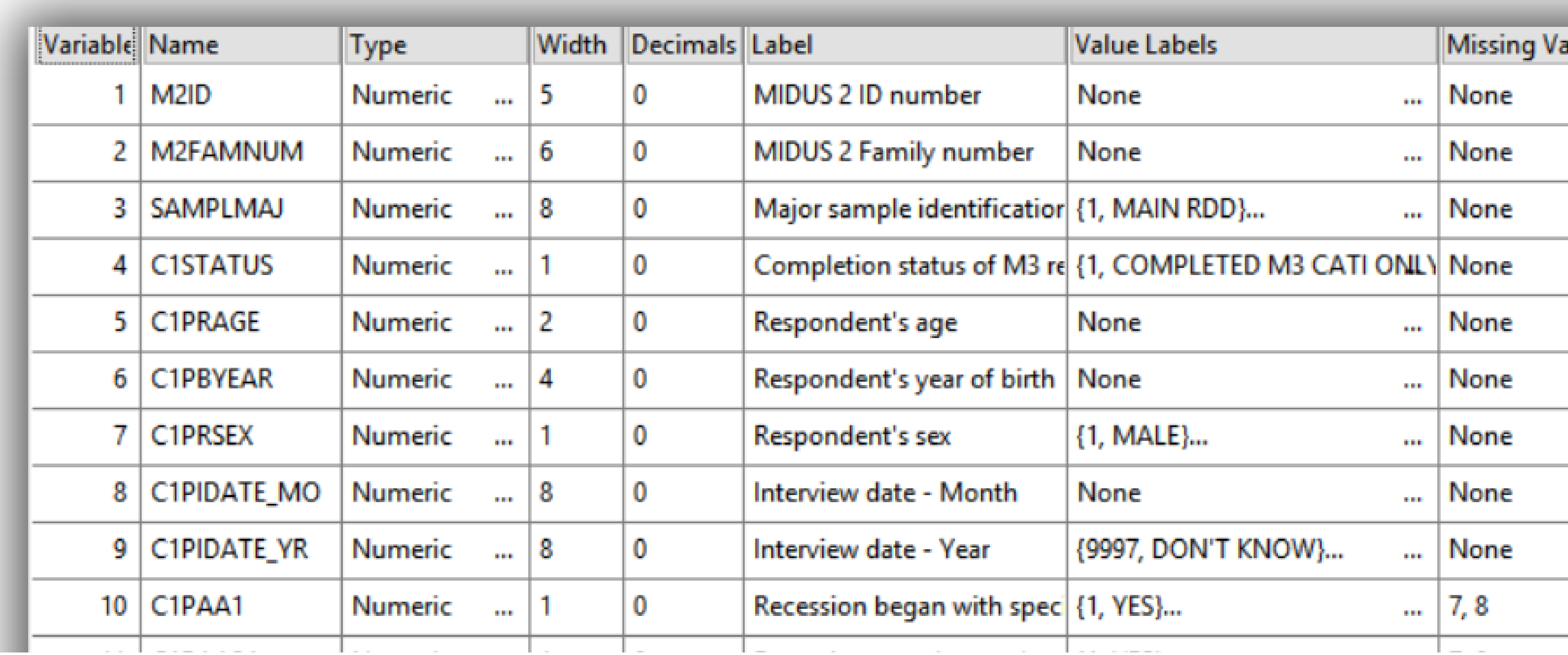
Metadata provides context for Data

Data



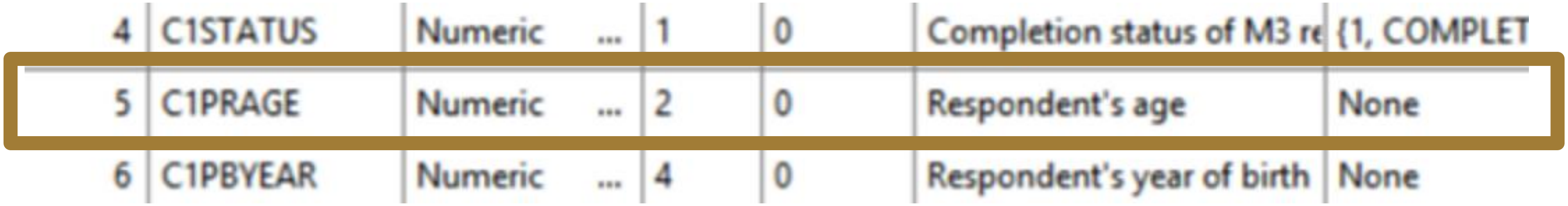
Case	M2ID	M2FAMNUM	SAMPLMAJ	C1STATUS	C1PRAGE	C1PBYEAR	C1PRSE	C1PIDATE_MO	C1PIDATE_YR
1	10001	110498	2	4	69	1943	1	7	2013
2	10002	100001	1	1	78	1935	1	6	2013
3	10011	110475	2	4	61	1952	2	6	2013
4	10015	120805	3	4	63	1950	2	11	2013
5	10019	100009	1	4	60	1952	1	6	2013
6	10020	100010	1	1	65	1948	2	3	2014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

Metadata



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

Metadata

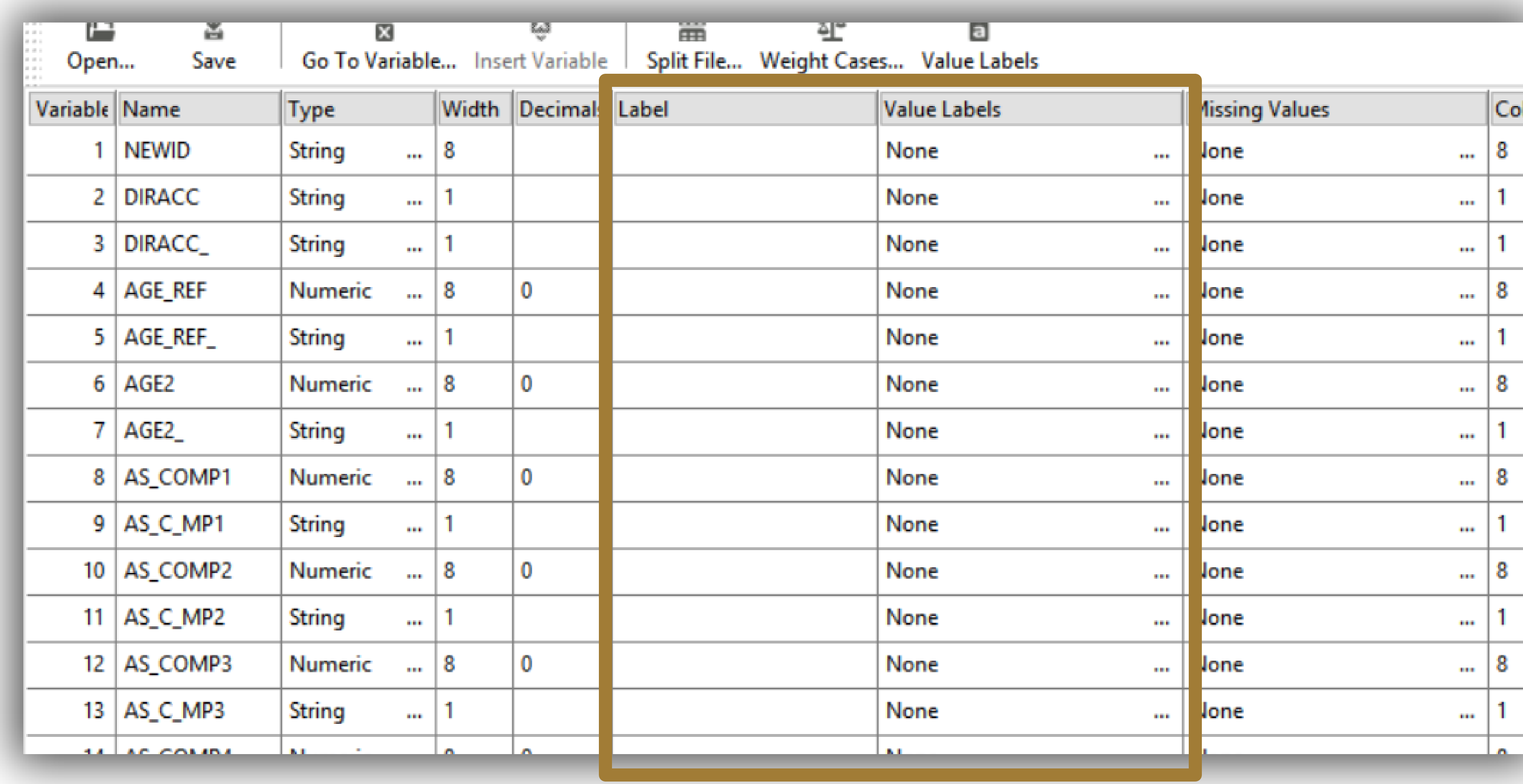


4	C1STATUS	Numeric	...	1	0	Completion status of M3 re	{1, COMPLET
5	C1PRAGE	Numeric	...	2	0	Respondent's age	None
6	C1PBYEAR	Numeric	...	4	0	Respondent's year of birth	None

Statistical Tools have Limited Metadata

- Data Types
- Variable Labels
- Value Labels

No Metadata ☹



Variable	Name	Type	Width	Decimals	Label	Value Labels	Missing Values	Co
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

The Metadata Problem

- Metadata is needed to understand data
- Statistical tools have limited metadata capabilities

The Goal

- Make a simpler way to add rich metadata to statistical files

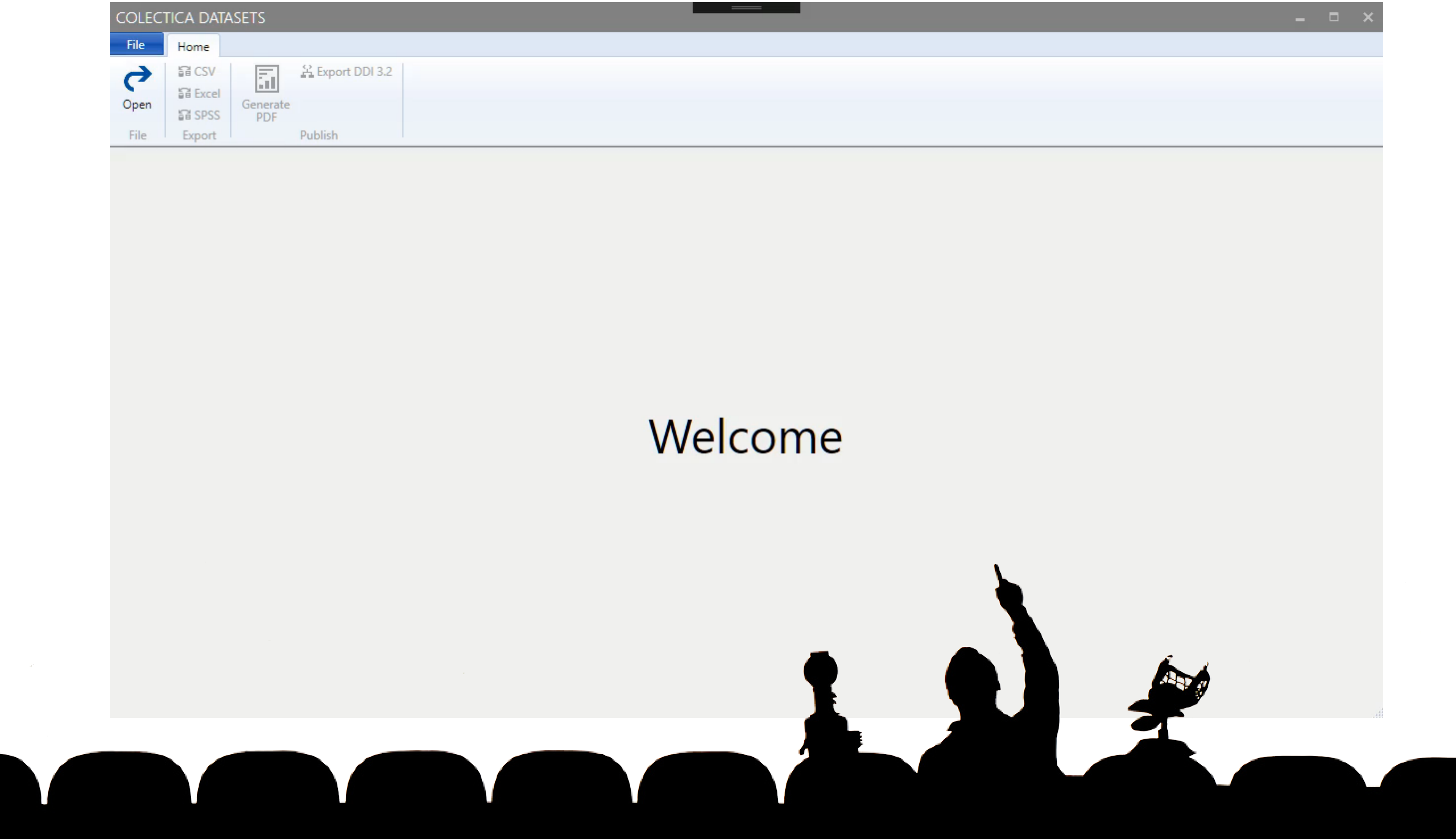
Existing Metadata Tools

- Not quite early enough and easy enough

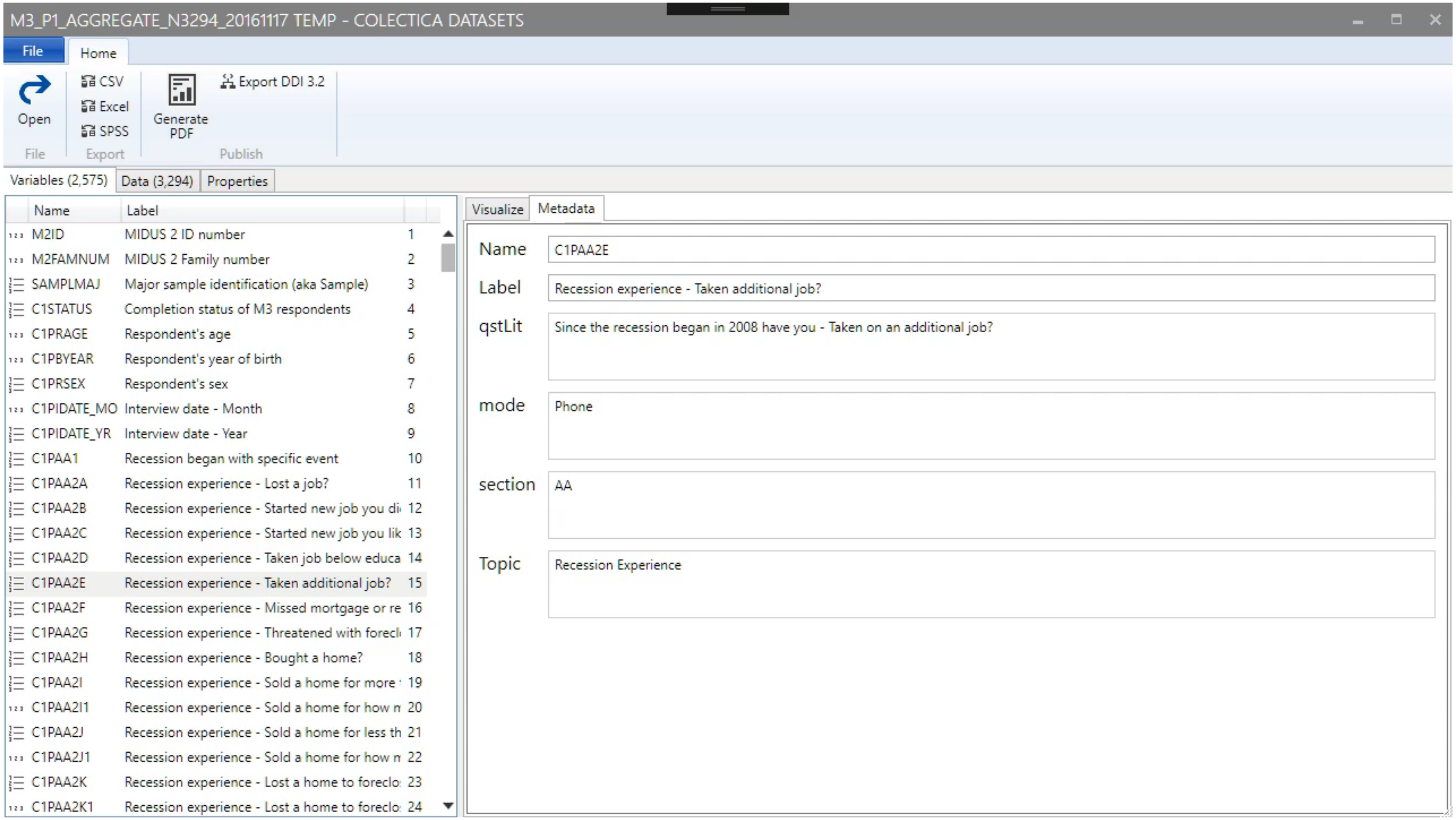
Colectica Datasets

- View
- Improve
- Publish

Explore


COLECTICA DATASETS
File
Home
Open
CSV
Excel
SPSS
Generate PDF
Export DDI 3.2
File
Export
Publish
Welcome

View Data



Name	Label	
M2ID	MIDUS 2 ID number	1
M2FAMNUM	MIDUS 2 Family number	2
SAMPLMAJ	Major sample identification (aka Sample)	3
C1STATUS	Completion status of M3 respondents	4
C1PRAGE	Respondent's age	5
C1PBYEAR	Respondent's year of birth	6
C1PRSEX	Respondent's sex	7
C1PIDATE_MO	Interview date - Month	8
C1PIDATE_YR	Interview date - Year	9
C1PAA1	Recession began with specific event	10
C1PAA2A	Recession experience - Lost a job?	11
C1PAA2B	Recession experience - Started new job you di	12
C1PAA2C	Recession experience - Started new job you lik	13
C1PAA2D	Recession experience - Taken job below educa	14
C1PAA2E	Recession experience - Taken additional job?	15
C1PAA2F	Recession experience - Missed mortgage or re	16
C1PAA2G	Recession experience - Threatened with forecl	17
C1PAA2H	Recession experience - Bought a home?	18
C1PAA2I	Recession experience - Sold a home for more	19
C1PAA2I1	Recession experience - Sold a home for how n	20
C1PAA2J	Recession experience - Sold a home for less th	21
C1PAA2J1	Recession experience - Sold a home for how n	22
C1PAA2K	Recession experience - Lost a home to foreclo	23
C1PAA2K1	Recession experience - Lost a home to foreclo	24

Name: C1PAA2E

Label: Recession experience - Taken additional job?

qstLit: Since the recession began in 2008 have you - Taken on an additional job?

mode: Phone

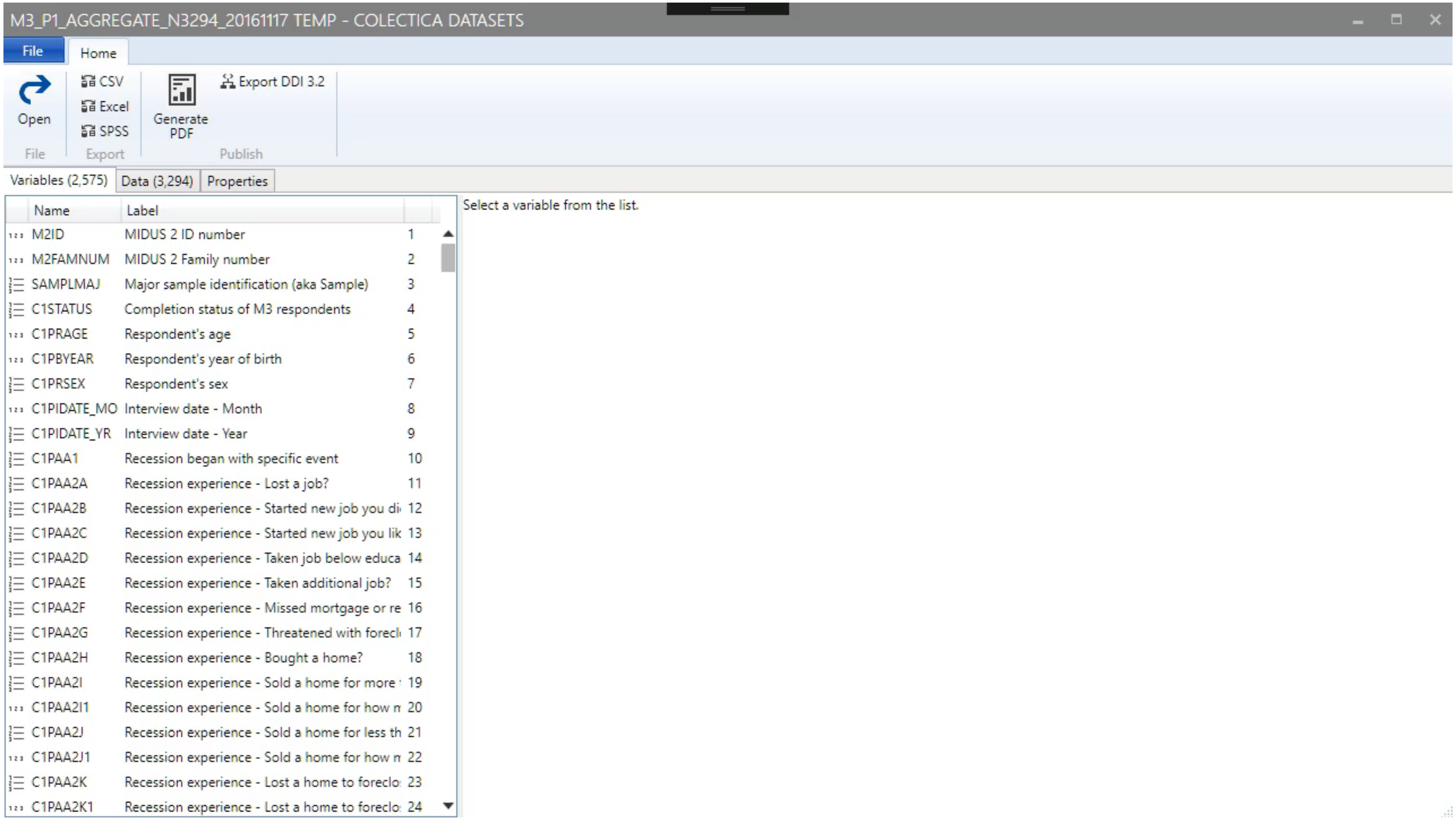
section: AA

Topic: Recession Experience

Soon: Metadata Quality Checks

- Data quality report card
  - Missing metadata
  - Spell checking
  - Extensible validations
  - Easily fix issues directly in the app

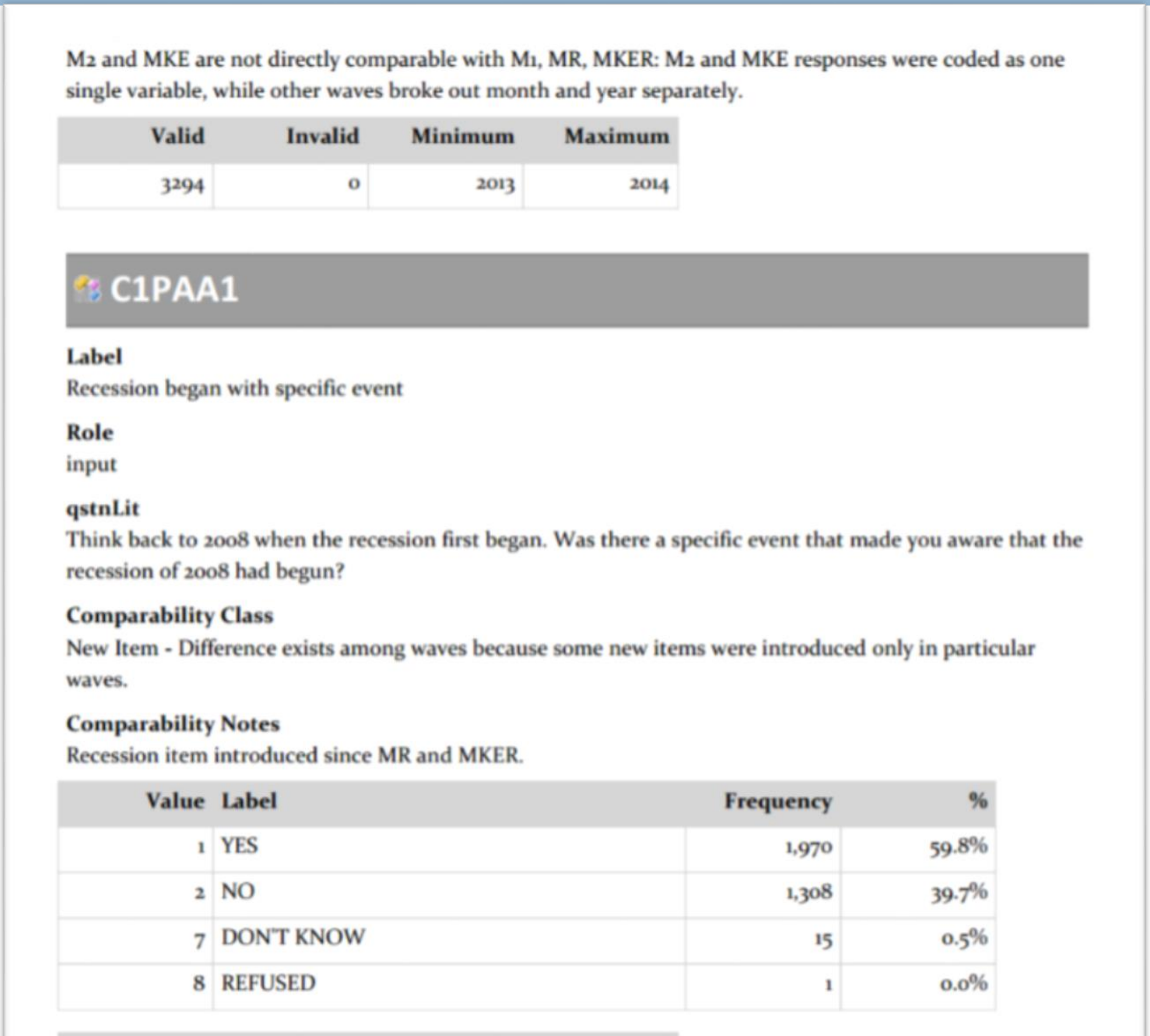
Publish



Name	Label	
M2ID	MIDUS 2 ID number	1
M2FAMNUM	MIDUS 2 Family number	2
SAMPLMAJ	Major sample identification (aka Sample)	3
C1STATUS	Completion status of M3 respondents	4
C1PRAGE	Respondent's age	5
C1PBYEAR	Respondent's year of birth	6
C1PRSEX	Respondent's sex	7
C1PIDATE_MO	Interview date - Month	8
C1PIDATE_YR	Interview date - Year	9
C1PAA1	Recession began with specific event	10
C1PAA2A	Recession experience - Lost a job?	11
C1PAA2B	Recession experience - Started new job you di	12
C1PAA2C	Recession experience - Started new job you lik	13
C1PAA2D	Recession experience - Taken job below educa	14
C1PAA2E	Recession experience - Taken additional job?	15
C1PAA2F	Recession experience - Missed mortgage or re	16
C1PAA2G	Recession experience - Threatened with forecl	17
C1PAA2H	Recession experience - Bought a home?	18
C1PAA2I	Recession experience - Sold a home for more	19
C1PAA2I1	Recession experience - Sold a home for how n	20
C1PAA2J	Recession experience - Sold a home for less th	21
C1PAA2J1	Recession experience - Sold a home for how n	22
C1PAA2K	Recession experience - Lost a home to foreclo	23
C1PAA2K1	Recession experience - Lost a home to foreclo	24

Select a variable from the list.

Data Dictionary



M2 and MKE are not directly comparable with M1, MR, MKER: M2 and MKE responses were coded as one single variable, while other waves broke out month and year separately.

Valid	Invalid	Minimum	Maximum
3294	0	2013	2014

C1PAA1

Label

Recession began with specific event

Role

input

qstnLit

Think back to 2008 when the recession first began. Was there a specific event that made you aware that the recession of 2008 had begun?

Comparability Class

New Item - Difference exists among waves because some new items were introduced only in particular waves.

Comparability Notes

Recession item introduced since MR and MKER.

Value	Label	Frequency	%
1	YES	1,970	59.8%
2	NO	1,308	39.7%
7	DON'T KNOW	15	0.5%
8	REFUSED	1	0.0%

Publish

- Publish to more locations
  - BagIt
  - Figshare
  - Zenodo

Availability

- 2017
- For Windows and macOS

Technical Previews

- Get in touch if you’d like to help test and give feedback colectica

Thank you

Web	colectica.com
Twitter	@Colectica
YouTube	youtube.com/colectica