

Automated Capture and Description of Data Transformations

Jeremy Iverson, Colectica

Dan Smith, Colectica

Pascal Heus, Metadata Technology North America

Ørnulf Risnes, Norwegian Centre for Research Data

Jared Lyle, ICPSR

George Alter, ICPSR (PI)

NSF Data Infrastructure Building Blocks (DIBBs) (ACI-1640575)

C²METADATA

Continuous Capture of Metadata



<http://c2metadata.org/>

Why Metadata?

- Data are useless without Metadata – “data about data”
- Metadata should:
 - Include all information about data creation
 - Describe transformations to variables
 - Be easy to create
- Our goal: Automated capture of metadata

Search Variables

Variables

- **c1_a4 c1_a4.** What is the percent chance that you will vote in the Congressional elections this November?
- **c1_state c1_state.** Respondent's state
- **c1_inserts_b1 c1_inserts_b1.** DATA ONLY: order of candidates in B1
- **c1_b1 c1_b1.** Who did you vote for in the election for the U.S. House of Representatives?
- **c1_inserts_b2 c1_inserts_b2.** DATA ONLY: order of candidates in B2
- **c1_b2 c1_b2.** Who did you vote for in the election for Governor of [STATE]?
- **c1_inserts_b3 c1_inserts_b3.** DATA ONLY: order of candidates in B3
- **c1_b3 c1_b3.** Who did you vote for in the election for the U.S. Senate?
- **c1_inserts_b4 c1_inserts_b4.** DATA ONLY: order of candidates in B4
- **c1_b4 c1_b4.** Who did you vote for in the election for the U.S. Supreme Court?

series: American National Election Study (ANES) Series > study: ANES 2010-2012 Evaluations of Government and Society Study > variable: c1_a4 >

c1_a4: c1_a4. What is the percent chance that you will vote in the Congressional elections this November?

What is the percent chance that you will vote in the Congressional elections this November?

Value	Label	Unweighted Frequency	%
Missing Data			
-7	No answer	-	-
-6	Not asked, unit non-response	-	-
-5	Not asked, terminated	-	-
-2	Missing, see documentation	-	-
Missing Data (System Missing)			
-1	Inapplicable	-	-
Total		1,241	100%

Based upon 1,160 valid cases out of 1,241 total cases.

Summary Statistics

- maximum: 100
- minimum: 0
- mean: 70.234
- mode: 100.0
- standard deviation: 39.793

Location: 322-324 (width: 3; decimal: 0)

Variable Type: numeric

Universe

The question in variable c1_a4 was asked if the question in variable c1_a1 was answered -6, -7, or 2.
The question in variable c1_a4 was asked if the question in variable c1_a2 was answered -1, or -6.

c1_a1: c1_a1. Have you already voted in the election being held [DAYS] days from now, or not?
-6: Not asked, unit non-response
1: Have already voted in that election



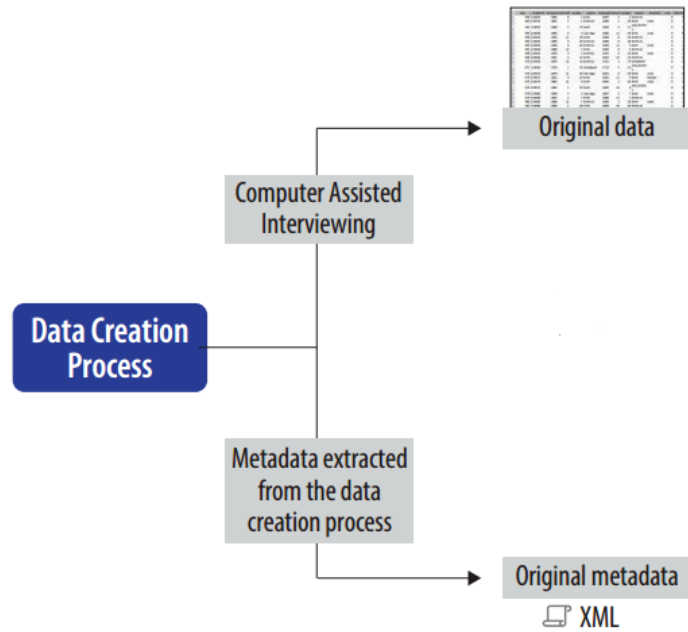
c1_a1: c1_a1. Have you already voted in the election being held [DAYS] days from now, or not? c1_a2: c1_a2. Which one of the following best describes how you voted?
-6: Not asked, unit non-response
-7: No answer
2: Have not voted in that election
-1: Inapplicable
-8: Not asked, unit non-response

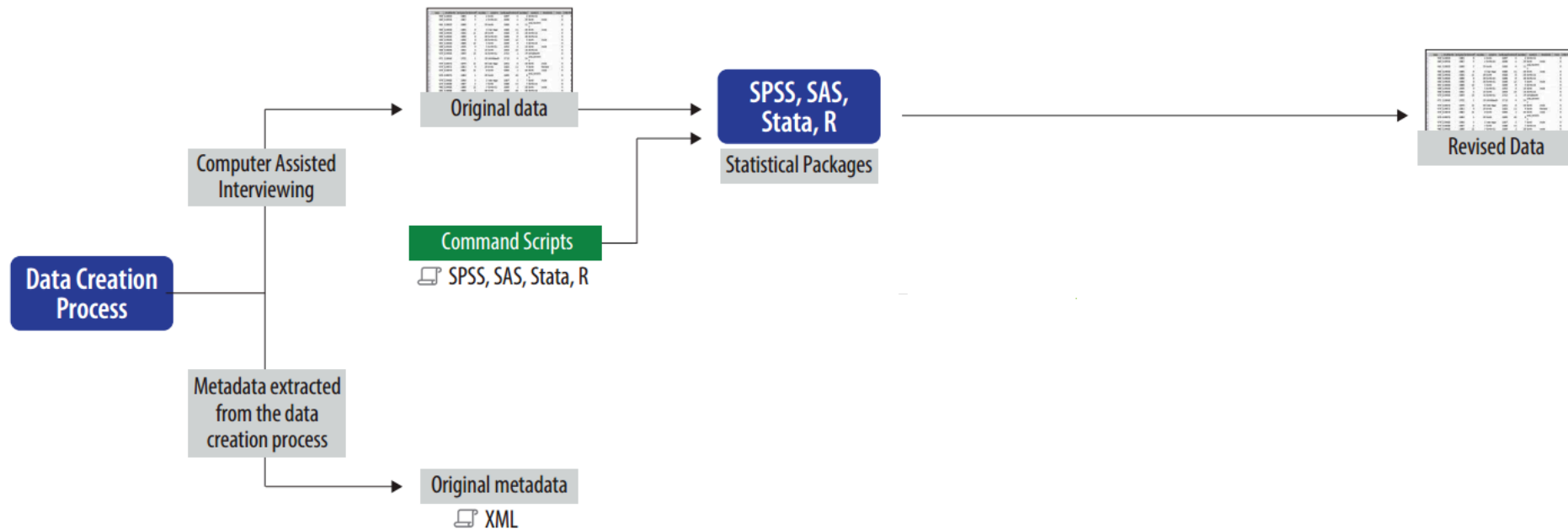


c1_a4
c1_a4. What is the percent chance that you will vote in the Congressional elections this November?

How are research data created?

- Most surveys are conducted with computer assisted interview software (CAI)
 - CATI – Computer-assisted Telephone Interview
 - CAPI – Computer-assisted Personal Interview
 - CAWI – Computer Aided Web Interview
- There is no paper questionnaire
- The CAI program is the questionnaire
 - **i.e. the program is the metadata**

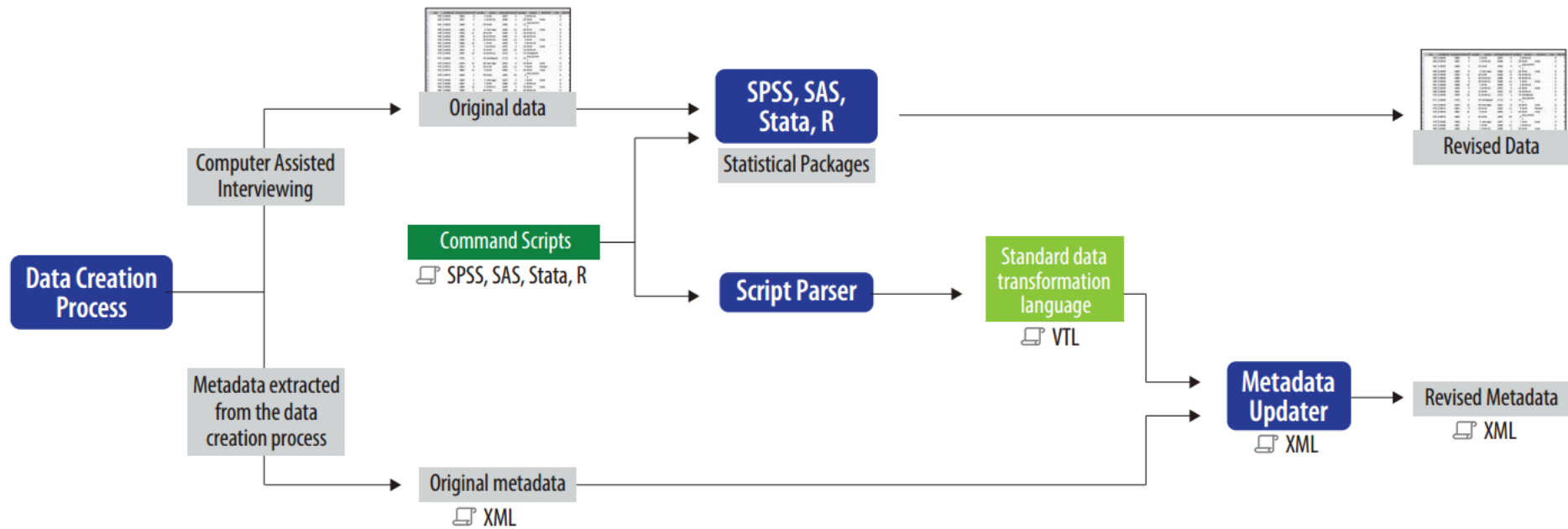




What's missing?

Statistics packages have limited metadata

- No question text
- No interview flow (question order, skip pattern)
- No variable provenance
- Data transformations are not documented.



Benefits of automated metadata capture

- Metadata will be better
 - All the information in the CAI can be included.
 - Variable transformations can be described
- Automation will lower costs
 - Metadata will not be discarded and re-created
- All metadata will be standardized and machine readable
 - Codebooks with rich information can be rendered at will
- If we make it easy and beneficial, researchers will use it.

Some Details

Project High Level Tasks

- **All**

- Collect scripts. Agree on VTL representations.
- Standards for incorporating VTL into DDI and EML

- **ICPSR**

- Test data and scripts
- Testing created DDI metadata
- Testing created EML metadata
- Active DDI codebook with VTL

- **NSD**

- Stata Script Parser
- SAS Script Parser

- **Colectica**

- SPSS Script Parser
- R data transformation package

- **MTNA**

- VTL to DDI Metadata Updater
- VTL to EML Metadata Updater
- DDI comparison and validation tool

Target Data Transformations for Year 1

Unconditional assignment (SPSS: COMPUTE; Stata: gener)

1.Arithmetic operations

2.List of mathematical functions (exp, ln, log, ...)

Conditional assignment (SPSS: IF; Stata: replace ...if)

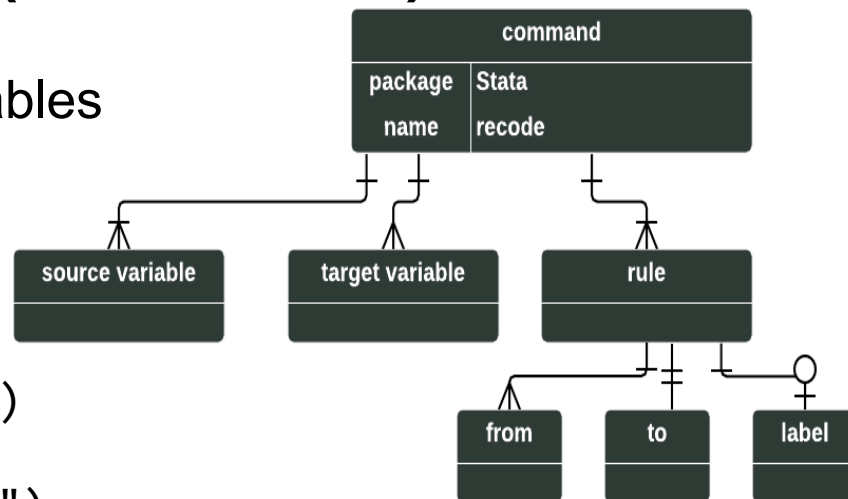
1.Logical expressions

Recode (SPSS: RECODE; Stata: recode)

SPSS	Stata	SAS	R	VTL
COMPUTE	generate replace	[assignment]		[assignment] :=
IF	replace ... if	IF...THEN/ ELSE		if ... then ... else
RECODE	recode	IF...THEN ... ELSE ... IF ... THEN		if ... then ... elseif ... then

What is **recode** (in Stata)?

recode -- Recode categorical variables



```
recode var1 var2 (1 2 = 2) (3/max = 5)
```

```
recode a b (1 = 2), prefix("modified_")
```

```
recode x y (1 = 2), generate(modified_x modified_y)
```

```
recode x y (1 = 2 "Labelfor2") (3 = 5 "Labelfor5")
```

```
recode total (0/140=0 F) (141/180=1 D) (181/210=2 C) (211/234=3 B)  
(235/300=4 A), gen(grade)
```

```
recode a (1 . .a 5/6 = 7) (nonmissing = 8) (missing = 9) (* = 2)
```

What is **recode** (in SPSS)?

RECODE **var1 var2 var3** (-1=7) (-2=8).

RECODE **AGE** (*MISSING=9*) (*18 THRU HI=1*) (*0 THRU 18=0*) INTO **VOTER**.

Apps to Detect Transformations

```
C:\svn\spss-sdtl-converter\src\C2Metadata.SpssToSdtl.Cli (master)  
A dotnet run -- c:\svn\spss-samples-public\basic\recode.sps -o d:\out\recode.json  
Recognized command count: 4
```


Recode in JSON

```
recode.json (C:\out) - GVIM2
1  [
2    {
3      "command": "recode",
4      "varlist": [
5        {
6          "source": "var1",
7          "target": "var1"
8        },
9        {
10         "source": "var2",
11         "target": "var2"
12       },
13       {
14         "source": "var3",
15         "target": "var3"
16       }
17     ],
18     "rules": [
19       {
20         "from": [
21           "-1"
22         ],
23         "to": "7",
24         "label": null
25       },
26       {
27         "from": [
28           "-2"
29         ],
30         "to": "8",
31         "label": null
32       }
33     ]
34   }
35 ]
```

JSON-schema for recode description

Published at

https://gitlab.nsd.uib.no/nsd-external/stata2vtl/blob/develop/resources/recode_schema.json

Functional recode parser available at

<http://ekstern.nsd.no/metacap/stata2vtl>

Continuing Work

- Describe more types of transformations
- Detect those transformations in SPSS, Stata, SAS
- R library for transforming data

Thanks!


<http://c2metadata.org/>

Following are slides not used in the main presentation...

Original
data

[illegible]

Computer Assisted Interviewing



We already have tools to convert CAI to machine-readable metadata.

<p>Convert to DDI: Collectica MQDS others</p>



CAI to DDI

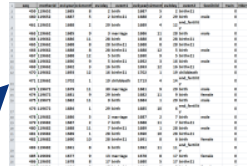
Original
metadata



DDI XML

What happens when a project modifies the data.

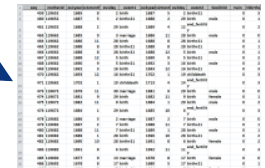
Original data



Statistical Packages

SPSS
SAS
Stata
R

Revised data



Computer Assisted Interviewing

CAI

Command scripts:

SPSS
SAS
Stata
R

CAI to DDI

Convert to DDI:
Collectica
MQDS
others

Original metadata

DDI XML

The modified data no longer match the metadata.



Metadata are re-created after the data are transformed.

