CSSI Element: Continuous Capture of Metadata for Statistical Data

PI: George Alter, Co-Pis: Jack Gager, Pascal Heus, Jeremy Iverson, Hosagrahar V Jagadish, Jared Lyle, Ornulf Risnes, Dan Smith
Institutions: University of Michigan, Algenta Technologies, Metadata Technologies North America, Norwegian Centre for Research Data

Award #: ACI-1640575

C²METADATA

Continuous Capture of Metadata

c2metadata.org

c2metadata@umich.edu

C²Metadata automates the documentation of data transformations performed by statistics software (SPSS, SAS, Stata, R, Python). Scientists can create variable-level provenance metadata from the command scripts used to modify the data.

1. Move data provenance to the variable level

The C²Metadata Project captures provenance metadata at the variable level, which is how transformations are specified. The lineage of each variable is described as a sequence of data transformation commands.

2. Structured Data Transformation Language (SDTL)

C²Metadata's Structured Data Transformation Language (SDTL) is an independent intermediate language for representing data transformation commands. Commands in four software packages (SPSS, Stata, SAS, R (tidyverse), and Python (Pandas) are translated into machine actionable JSON schemas.

3. Extract data transformations from command scripts

C²Metadata's Script Parsers create SDTL from scripts that are used to manage and modify data.

4. Update existing metadata

Updaters insert variable provenance into metadata files in standard formats: Data Documentation Initiative (DDI), Ecological Markup Language (EML), and JSON-LD

5. Create documentation: codebooks, data catalogs

Structured metadata files (DDI, EML) are used in data catalogs, codebooks, and other forms of documentation. Each variable is presented with its provenance. Transformation steps are translated into natural language text.

Recode in SPSS

```
RECODE V520131 (0=1) (1,2=2) (3 thru 6=3) (7,8=4) into EDUC2.
```

Recode in Stata

```
recode V520131 (0=1) (1 2=2) (3/6=3) (7 8=4), gener(EDUC2)
```

Recode in R

```
mutate(df0,  EDUC2 = cut(V520131, c( 0, 1, 3, 7, 8), include.lowest = TRUE , right=FALSE, labels=FALSE) )
```

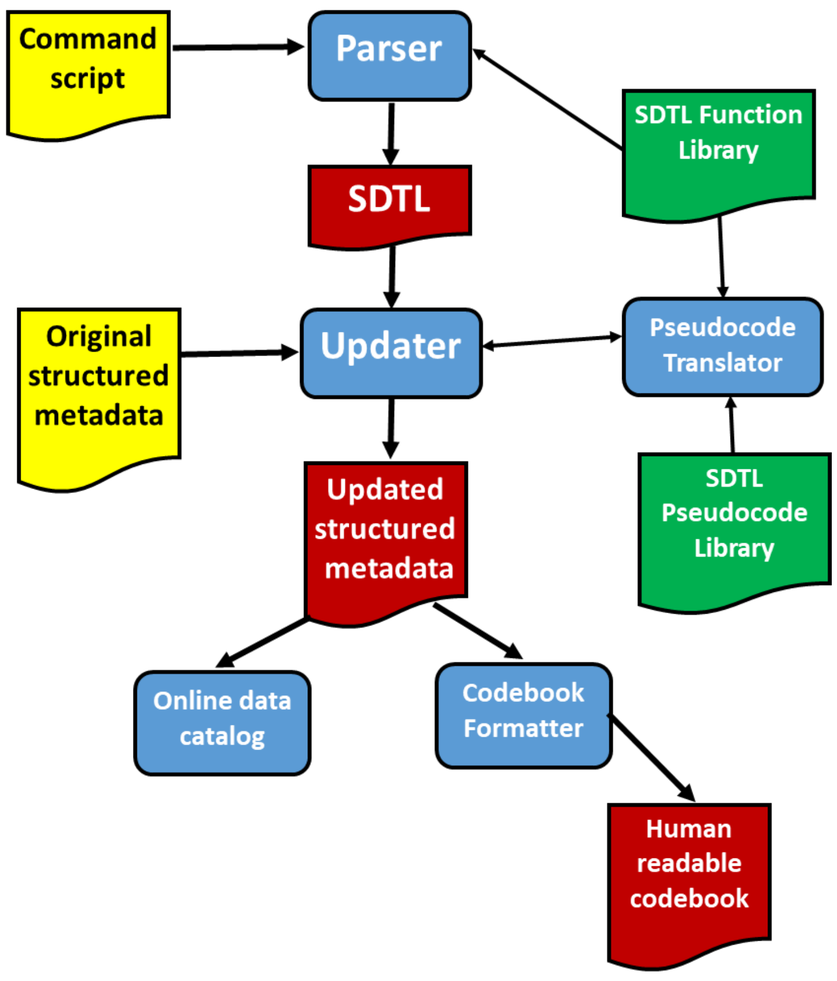
Recode in SDTL

```
"command" : "recode",
  "recodedVariables" : [ {
   "source" : "V520131",
   "target" : "EDUC2"
         } ],
  "rules" : [
   {"to" : 1, "fromValue" : [ 0 ]  },
   {"to" : 2, "fromValue" : [ 1, 2 ] },
   {"to" : 3, "fromValueRange" :
      [ {"first" : 3,  "last" : 6      } ]    },
   {"to" : 4, "fromValue" : [ 7, 8 ]}
    ],
```

Recode in natural language

Description: Recode variable V520131 into new variable EDUC2
so that 0 are coded as 1,
so that 1, 2 are coded as 2,
so that 3 through 6 are coded as 3,
so that 7, 8 are coded as 4.

C²Metadata Workflow



ICPSR

NSD NORWEGIAN CENTRE FOR RESEARCH DATA

mtna

colectica

ANES American National Election Studies

NORC 75 at the UNIVERSITY of CHICAGO