C²METADATA

Continuous Capture of Metadata

NSF ACI-1640575

George Alter, PI

University of Michigan

altergc@umich.edu

George Alter (ICPSR)
Darrell Donakowski (ANES)
Jack Gager (MTNA)
Pascal Heus (MTNA)
Carson Hunter (MTNA)
Sanda Ionescu (ICPSR)
Jeremy Iverson (Colectica)
H.V. Jagadish (University of Michigan)
Carl Lagoze (University of Michigan)
Jared Lyle (ICPSR)
Thomas Murphy (ICPSR)
Sigbjørn Revheim (NSD)
Matthew Richardson (ICPSR)
Ornulf Risnes (NSD)
Karunakara Seelam (ICPSR)
Dan Smith (Colectica)
Tom Smith (NORC)
Jie Song (University of Michigan)
Yashas Vaidya (University of Michigan)
Ole Voldsæter (NSD)

c2metadata.org **c2metadata@umich.edu**

The Problem: Many scientists use statistical analysis packages (SPSS, Stata, SAS, R) for data management and data transformations. Since these packages have limited metadata, changes to the data are not recorded. The C²Metadata Project automates the documentation of data transformations performed by statistics software. Scientists can create variable-level provenance metadata from the command scripts used to modify the data.

Solutions:

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 in an updated metadata file.

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, and R) are translated into JSON schemas, which are machine actionable.

3. Extract data transformations from command scripts

Most data transformation tasks are performed by command scripts. C²Metadata's Script Parsers create SDTL from these scripts.

4. Update existing metadata

Updaters insert variable provenance into metadata files in two widely used formats: Data Documentation Initiative (DDI) and Ecological Markup Language (EML)

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.

Example:

Recode in SPSS

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

Recode in Stata

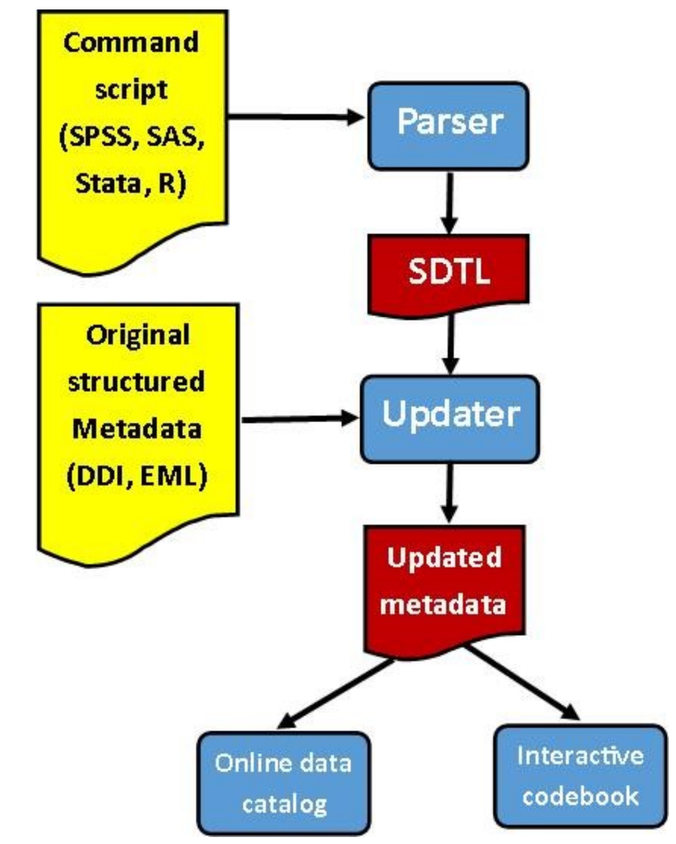
```
recode V520131 (0=0) (1 2=1) (3/6=2) (7
8=3), gener(EDUC2)
```

Recode in natural language

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

Recode in SDTL

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



ICPSR
NSD NORWEGIAN CENTRE FOR RESEARCH DATA
mtna
colectica
ANES American National Election Studies
NORC at the UNIVERSITY of CHICAGO