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 C²-Metadata 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 C²-Metadata 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)
C²-Metadata’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.

Example:

Recode in SPSS
RECODE VS20131 (0=0) (1,2=1) (3 thru 6=2) (7,8=3) into EDUC2.

Recode in Stata
recode VS20131 (0) (1 2=1) (3/6=2) (7 8=3), gener(EDUC2)

Recode in natural language
Description: Recode variable VS20131 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" : "VS20131",
"target" : "EDUC2"
} ],
"rules" : [ {
"to" : 0, "label" : null, "fromValue" : [ 0 ] },
{"to" : 1, "label" : null, "fromValue" : [ 1, 2 ] },
{"to" : 2, "label" : null, "fromValueRange" : [ 3, 6 ] },
{"to" : 3, "label" : null, "fromValue" : [ 7, 8 ] }
],
3. Extract data transformations from command scripts
Most data transformation tasks are performed by command scripts. C²-Metadata’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.

c2metadata.org

 NSF ACI-1640575
 George Alter, PI
 University of Michigan
 altergc@umich.edu

 c2metadata@umich.edu

Recode in SPSS
RECODE VS20131 (0=0) (1,2=1) (3 thru 6=2) (7,8=3) into EDUC2.

Recode in Stata
recode VS20131 (0) (1 2=1) (3/6=2) (7 8=3), gener(EDUC2)

Recode in natural language
Description: Recode variable VS20131 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.