C²METADATA
Continuous Capture of Metadata

George Alter[1], Darrell Donakowski[2], Jack Gager[3], Pascal Heus[3], Sanda Ionescu[1]
Jeremy Iverson[4], H.V. Jagadish[5], Carl Lagoze[5], Jared Lyle[1], Thomas Murphy[1]
Ornulf Risnes[6], Dan Smith[4], Tom Smith[7], Jie Song[5]

Supported by the Data Infrastructure Building Blocks (DIBBs) program of the National Science Foundation through grant NSF ACI-1640575

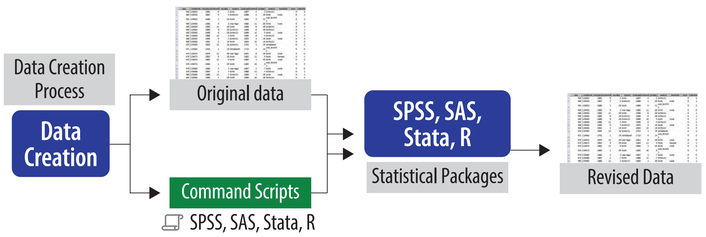
[1] ICPSR [2] ANES [3] MTNA [4] Colectica [5] University of Michigan [6] NSD [7] NORC

The Problem: Many scientists in a wide variety of disciplines use statistical analysis packages (SPSS, Stata, SAS, R) for data management and data transformations. These packages include limited metadata, and they do not record provenance information about data transformations. Consequently, metadata are created by hand after data transformations have been completed. This is a slow, inefficient process and much valuable information is lost. The C²Metadata Project will create a continuous workflow for metadata creation by automating the capture of data transformations performed by statistical analysis software.

c2metadata.org

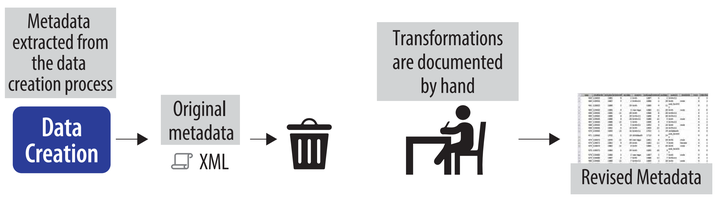
The Data Creation Pipeline

Most data are born digital and pass directly into data transformation and analysis. The leading statistical analysis packages (SPSS, Stata, SAS, R) are often used for data management and data transformation.



The Metadata Data Creation Pipeline before C²Metadata

Metadata can often be extracted directly from the data creation process. However, changes made by statistical analysis packages (SPSS, Stata, SAS, R) make the original metadata obsolete. Data transformations are documented manually and incompletely.



The C²Metadata Solution

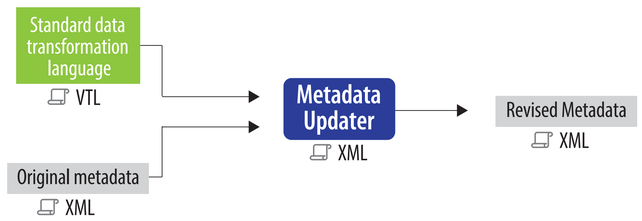
Researchers use command scripts to operate statistical analysis packages. We can extract data transformation information from command scripts and use that information to update metadata files.

Step 1. Parsing the command script and creating VTL

The Script Parser reads the command script, outputting a description of the data transformations in a standardized form – Validation & Transformation Language (VTL).

Step 2. VTL is used to update the metadata file

The Metadata Updater uses the VTL to add data transformations into the original metadata file.



The C²Metadata Pipeline: Data and metadata are changed in parallel.