Introduction

Datasets are often derived by manipulating raw data with statistical software packages. These packages, however, lack tools for documenting variable transformations in the manner of a workflow system or even a database. At best, the operations performed by the statistical package are described in a script, which more often than not is not even available to future data users. Different statistics packages differ in data model, transformation representation and scope of transformations covered; thereby further complicating the understanding of the transformation process.

A Motivating Example

Here we show two functionally-equivalent data transformation scripts written in SPSS and Stata to further illustrate the disparities.

C2Metadata Overview

To reduce the cost and increase the completeness of metadata, we aim to work with common statistical packages to automate the capture of metadata at the granularity of individual data transformations in a simple yet expressive representation regardless of the original languages used. C2Metadata is such a system, the workflow of which is depicted in the figure below.

SDTA and SDTL as the Bridge

In addition to a generic data model for statistical data transformation, we develop a generic transformation model coupled with the data model. We define SDTA and SDTL as two realizations of the transformation model, both of which can be adopted as the bridge for communication between statistical languages and for standardization of statistical data transformation for documentation.

– SDTA: Structured Data Transformation Algebra
  □ inspired by Relational Algebra
  □ defines statistical data transformation using a small set of primitive operators
  □ simplifies and optimizes execution leveraging the benefit of algebraic expressions

– SDTL: Structured Data Transformation Language
  □ inspired by Query Language for relational databases
  □ defined by the Convention-based Ontology Generation System (COGS) information model providing multiple representations under one specification
  □ presents a declarative description of commonly used statistical data transformation operations

C2Metadata Functionalities

C2Metadata allows automatic documentation of transformation in metadata as well as visualization of the changes in data. We show in the figure below four demonstration scenarios for a sample transformation script in SPSS.

(a) Upload Interface
(b) View table content
(c) View operation
(d) Track tuples

Figure 6. Snapshots of C2Metadata functionalities