C²Metadata: Automating the Capture of Data Transformations from Statistical Scripts in Data Documentation

COMPUTER SCIENCE & ENGINEERING
UNIVERSITY OF MICHIGAN

Jie Song (jiesongk@umich.edu)[1] · George Alter[2] · H. V. Jagadish[1]

[1]Computer Science Engineering, University of Michigan
[2]Inter-university Consortium for Political and Social Research (ICPSR), University of Michigan

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.

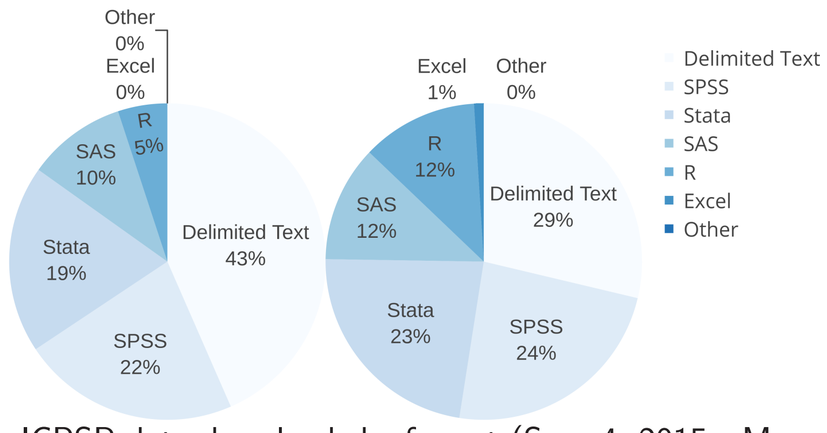


Figure 1. ICPSR data downloads by format (Sep. 4, 2015 - Mar. 4, 2016)

A Motivating Example

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

Original Data

CaseID	Age	V520041	V520042	V520043
~~1~~	~~66~~	~~1~~	~~0~~	~~0~~
2	70	1	1	1
3	71	1	0	1
...	...	...	...	...

Original Data Summary

Name	Type	Label	...
CaseID	Num	INTERVIEW NUMBER	...
Age	Num	AGE	...
V520041	Num	CARE MUCH WHICH PTY WINS	...
V520042	Num	XCARE WHO WINS ST ELCTN	...
V520043	Num	XCARE WHO WINS LCL ELCTN	...

SPSS Script

```
compute Partycare1 = V520041+V520042+V520043.
variable labels Partycare1 'Care who wins elections - Index 1'.
select if not (Age<67).
sort cases by Age(D) CaseID(A).
```

or functionally equivalent Stata Script

```
gen Partycare1 = V520041+V520042+V520043
label variable Partycare1 'Care who wins elections - Index 1'
drop if Age<67
gsort -Age CaseID
```

Revised Data

CaseID	Age	V520041	V520042	V520043	Partycare1
3	71	1	0	1	2
2	70	1	1	1	3
9	70	1	0	0	1
...	...	...	...	...	...

Revised Data Summary

Name	Type	Label	...
CaseID	Num	INTERVIEW NUMBER	...
Age	Num	AGE	...
V520041	Num	CARE MUCH WHICH PTY WINS	...
V520042	Num	XCARE WHO WINS ST ELCTN	...
V520043	Num	XCARE WHO WINS LCL ELCTN	...
Partycare1	Num	Care who wins elections - Index 1	...

Figure 2. SPSS and Stata scripts for data transformation

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.

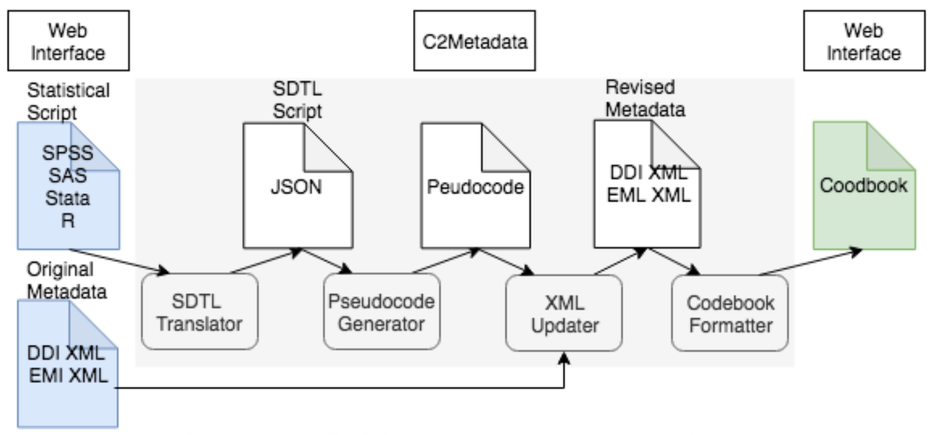


Figure 3. C2Metadata system workflow

- SDTL Translator
  - ☐ parses the input statistical script (one parser per statistical language) and translates commands into Abstract Syntax Trees (ASTs) using Antlr rules
  - ☐ maps ASTs into functions defined by SDTL, a standard data transformation representation (see Figure 4 and 5 for more details)
- Pseudocode Generator
  - ☐ converts SDTL functions into human readable text for a more user-friendly illustration of the transformations included
- XML Updater
  - ☐ updates the original metadata in DDI or EML standards (both XML-based) with both file level and data element level transformations including the original transformation script, the natural language description of transformations and the SDTL equivalent in XML format
- Codebook Formatter
  - ☐ generates an HTML codebook from the revised metadata describing the contents, structure and layout of the revised data

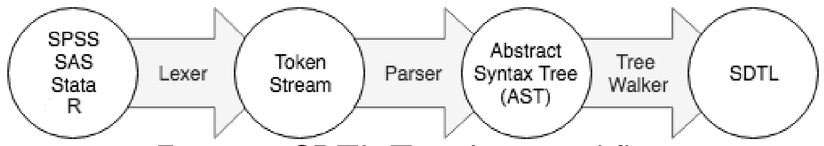


Figure 4. SDTL Translator workflow

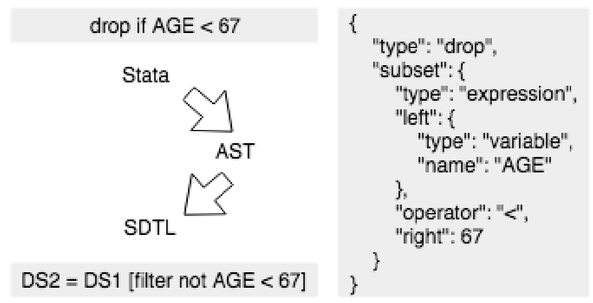


Figure 5. SDTL Translator translation of a Stata syntax

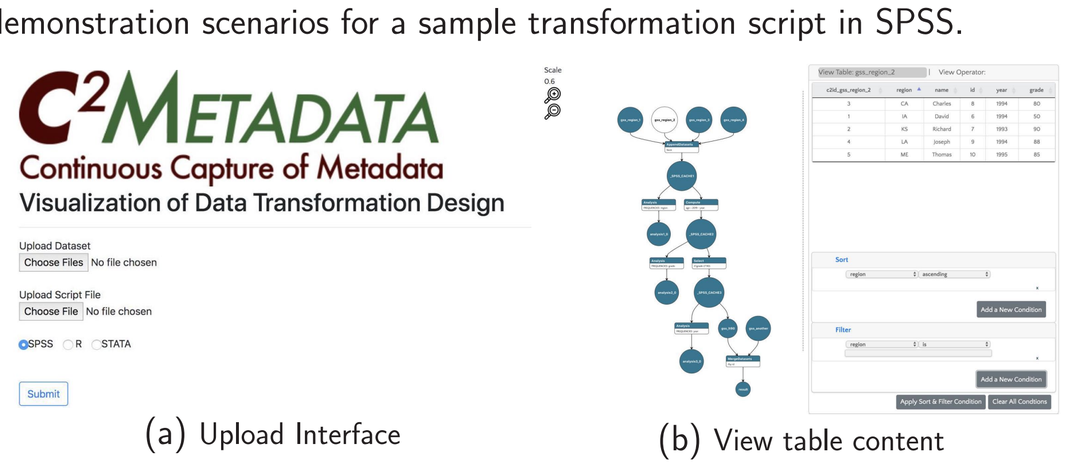
SDTA and SDTL as the Bridge

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