Continuous Capture of Metadata for Statistical Data
(NSF ACI-1640575)

PI George Alter, ICPSR, University of Michigan

Overview

Accurate and complete metadata is essential for data sharing and for interoperability across different data types. However, the process of describing and documenting scientific data has remained a tedious, manual process even when data collection is fully automated. Researchers in many fields use statistical software for data management as well as analysis, but the leading software packages do not document variable transformations. This project will greatly reduce the cost and increase the completeness of metadata by creating tools to capture and save data transformation metadata from general purpose statistical analysis packages.

Approach

Software developed by this project will extract variable transformation information from command scripts for statistical packages and update metadata files to reflect changes in the data. We have divided this process into two steps.

Step 1. Script Parser: Script Parsers will translate data transformation commands from four widely used statistical packages (SPSS®, SAS®, Stata®, R[1]) into the Validation and Transformation Language (VTL). Since every statistical package uses its own conventions, a Script Parser must be developed for each one. VTL, which is being developed under the auspices of seven international organizations, provides an independent, non-proprietary way to represent data transformations.

Step 2. Metadata Updater: The Metadata Updater will revise an existing metadata file by inserting information transferred to it in VTL. We will create Metadata Updaters for two internationally accepted metadata standards: the Data Documentation Initiative (DDI) standard for social science data and the Ecological Markup Language (EML) developed for the environmental and ecological sciences.

Since Script Parsers and Metadata Updaters communicate through VTL, the parser for R can be used to update both DDI and EML, and our approach can be easily extended to other statistical packages and metadata standards.

Project Partners

Inter-university Consortium for Political and Social Research (ICPSR)
Norwegian Centre for Research Data (NSD)
Metadata Technology North America (MTNA)
Colectica
American National Election Studies (ANES)
General Social Survey (GSS)
Data Documentation Initiative Alliance

---

[1] SPSS is a trademark of the IBM Corporation. SAS is a trademark of the SAS Institute Inc. Stata is a trademark of StataCorp LP. R is open source software R Core Team (2012).