UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS
ModernStats World Workshop
21-22 October 2024, Geneva

Describing and Querying Data Transformation Scripts: SDTL and SDTH

Speaker: George Alter, University of Michigan

Author(s):

Abstract

One of the barriers to creating continuous, metadata-based data production workflows is the use of statistical software with minimal metadata capabilities. Even when data are "born-digital," information is usually lost, because complex XML files are too difficult to revise. Structured Data Transformation Language (SDTL) was created to provide common, machine actionable descriptions of data transformation scripts from multiple statistical analysis packages. We have shown that SDTL can be used to update existing metadata files after transformations have been performed. Structured Data Transformation History (SDTH) is an extension of the PROV model to include provenance relationships among data files, dataframes, and variables. Where SDTL provides all of the details required to reproduce data transformations, SDTH answers basic questions, such as "Which input variables affected the values of variable X?"