Introducing Colectica Datasets

Dan Smith*[1]

[1]Colectica – United States

Abstract

Statistical data tools like SAS, SPSS, and Stata and programming languages such as R and Python give data users powerful capabilities for data analysis. However, these tools have limited metadata capabilities. Colectica Datasets is a new tool that takes a metadata first approach to dataset creation, embedding DDI metadata into the data files themselves.
Colectica Datasets can import SAS, SPSS, Stata, CSV, and Parquet files. Additional metadata about the dataset and its variables can be created by the user. Variable labels and value labels can be specified in multiple languages. The tool can then export the enhanced dataset to SPSS, Stata, CSV, Parquet, or Excel, and create summary statistics, weighted statistics, and codebook documentation. Users can run quality checks on their dataset which show alerts where additional metadata would be useful.

In addition to embedding enriched metadata, Colectica Datasets can be used as a transfer tool between various statistical file formats. It automatically handles different types of missing values, format mapping, and data type detection.
Colectica Datasets is free for personal use and available on Windows and macOS.

*Speaker