Metadata-based synthetic data generation

Erik-Jan Van Kesteren*[1,2]

[1]Utrecht University – Netherlands
[2]ODISSEI (Open Data Infrastructure for Social Science and Economic Innovations) – Netherlands

Abstract

The Dutch Central Bureau of Statistics (CBS) possesses extremely valuable sensitive datasets with information on the entire population of the Netherlands. Researchers in the Netherlands may gain secure access to these datasets after submitting a detailed proposal with information on which datasets are required to answer their research question. However, it is difficult, time-consuming, and costly to manually comb through the metadata to understand whether and how the available data may answer their question of interest.
To help researchers in this initial exploration step, we are developing a software program that generates synthetic example data (public use files) based on variable-level information from public metadata – e.g., means, standard deviations, range, variable names and types, and dataset size. Using this synthetic data, researchers can define their requirements and even write analysis code before formally requesting access.
In this talk, I will show off a prototype app (https://github.com/sodascience/ddi-synth), discuss the implications, and talk about how this concept can grow in the future. We intend for this software to integrate with the Dataverse instance in development at ODISSEI, the the Dutch Social Science research infrastructure, and we are considering the DDI metadata format for this task.

*Speaker