International Journal of Population Data Science

Journal Website: www.ijpds.org

Data Quality Automation: a Generic Approach for Large Linked Research Datasets

Elmessary, MA¹, Thayer, D¹, Rees, S¹, Kemp, L¹, and Rees, A¹

¹Swansea University

Introduction

When datasets are collected mainly for administrative rather than research purposes, data quality checks are necessary to ensure robust findings and to avoid biased results due to incomplete or inaccurate data.

When done manually, data quality checks are time-consuming. We introduced automation to speed up the process and save effort.

Objectives and Approach

We have devised a set of automated generic quality checks and reporting, which can be run on any dataset in a relational database without any dataset-specific knowledge or configuration.

The code is written in Python. Checks include: linkage quality, agreement with a population data source, comparison with previous data version, duplication checks, null count, value distribution and range, etc.

Where dataset metadata is available, checks for validity against lookup tables are included, and the output report includes documentation on data contents. An HTML report with dynamic datatables and interactive graphs, allowing easy exploration of the results, is produced using RMarkdown.

Results

The automation of the generic data quality check provides an easy and quick tool to report on data issues with minimal effort. It allows comparison with reference tables, lookups and previous versions of the same table to highlight differences. Moreover, this tool can be provided for researchers as a means to get more detailed understanding about their data.

While other research data quality tools exist, this tool is distinguished by its features specific to linked data research, as well as implementation in a relational database environment. It has been successfully tested on datasets of over two billion rows.

The tool was designed for use within the SAIL Databank, but could easily be adapted and used in other settings.

Conclusion/Implications

The effort spent on automating generic testing and reporting on data quality of research datasets is more than compensated by its outputs. Benefits include quick detection and scrutiny of many sources of invalid and incomplete data. This process can easily be expanded to accommodate more standard tests.