Curating Historical Statistics Data on Baltic Countries in 1897-1939: Providing Data with Rich Metadata

kaunas university of technology
1922

Problem

Historical statistics is, in general, rather scarce (data is available for a limited number of countries and for limited periods of time) and poorly documented (sources are only generally described). Historical researchers usually have to rely on the data collected in the "Western world" or analyze rather limited time series (ranging for only 30-50 years).

Historical statistics for Baltic countries is even more problematic: statistical data were collected and published during the interwar period, however, these data still are not properly digitized as to become readily available for research purposes (in machine-readable formats).

The need for proper data digitalization is not just a problem of accurate transfer of available historical data tables into machine-readable formats. Extensive and precise description of data (each cell, if necessary) is essential. Therefore, we aim to develop a solution for curating historical data sets.

Presently, most of tabular data are stored in CSV files with no metadata. When available, metadata is provided either separately (DDI, SDMX etc.) or integrated into the CSV file (CVSY, CDS/MRT, IPAC etc.). However, historical data are usually assembled from quite diverse available sources and need to be extensively documented, even on the individual cell level, which is not easily possible with currently available standards and solutions.

Solution: + Dataverse

https://dataverse.harvard.edu/dataverse/HistStatsBaltic

Solution: Extended tabular (CSV) data + Shiny

- Dataverse provides user-friendly solution for storing and publishing data with a possibility to remotely download data. Different types of data can be uploaded and stored on the Dataverse.
- Extended tabular format was developed (still needs to be developed further), which contains basic dataset-level metadata before the header (column names), and extended cell-level metadata (data source and notes) in the columns beside the data columns (see below). This format can be easily digested by R/Python.

- Shiny applications (for example) can be conveniently used for reporting, browsing and visualizing this type of data: vaidasmo.shinyapps.io/RDA15

Authors

Vaidas Morkevičius
Kaunas University of Technology, Lithuania

Supported by the funding under the project "RDA Europe Node - Lithuania"

Giedrius Žvaliauskas
Vilnius University, Lithuania

Supported by the funding from the European Social Fund (project number 09.3.3-LMT-K-712-01-0006) under grant agreement with the Research Council of Lithuania (LMT)

RDA RESEARCH DATA ALLIANCE EUROPE | LT NATIONAL NODE