Extending data validation with standardised metadata from SDMX registries

Olav ten Bosch (presenter), Mark van der Loo
Statistics Netherlands, the Netherlands

Standardisation of metadata is crucial in official statistics. A harmonized statistical picture of society and economy can only be produced if metadata such as classifications, code lists and variables are agreed among organisations. Internationally these metadata elements are stored in Statistical Data and Metadata Exchange (SDMX) registries. Examples are the ***Global Registry*** and the ***Eurostat Registry.*** They provide the concepts, variable definitions, data flows, structures and code lists underlying international official statistics. Moreover these registries provide a standardized application programming interface (API) to automatically access specific versions of the metadata from any programming environment. This makes them a key element in international validation processes and hence important for data validation practices in general.

The ***R-package validate*** is a popular tool in official statistics to validate data. Based on a set of predefined validation rules, varying from variable checks to multivariate or statistical checks, the software can execute analyses on possibly large datasets, providing the user with extensive feedback on the health of their data. The software supports the majority of checks that are actually needed in today's practical International data validation exercises. Validation results are presented graphically and in a machine readable standardized validation report, which can be used as input for consecutive processes. Common rules can be re-used among multiple validation processes. For more details we refer to the online validation ***cookbook*** offering recipes for the most common validation scenario's.

In the ***ValidatFOSS2***[1] project we extended the R validate package with data validation based on SDMX metadata. The aim was to make data validation as easy as possible by re-using metadata already provided in SDMX registries. The SDMX metadata such as the dimensions, attributes, code lists and data representations are automatically retrieved from the respective registry based on the identifiers specified and cached for consecutive validation runs. The architecture presented is generic and can be used on any SDMX 2.1 compliant SDMX registry from any international organisation. Alternatively the approach can be used on local DSD files or an organisation-internal registry.

In this presentation we present the results of this endeavor. We address the main approach, our experiences working with the international SDMX registries, some examples of the use in practice and we show the setup of a new chapter of the validation cookbook on connecting to SDMX.

Links

Global registry: https://registry.sdmx.org
Eurostat registry: https://webgate.ec.europa.eu/sdmxregistry
R-package Validate: https://cran.r-project.org/package=validate
Validation cookbook: https://cran.r-project.org/web/packages/validate/vignettes/cookbook.html
Awesome official statistics software: http://awesomeofficialstatistics.org

[1] VadlidatFOSS2: Validation with Free and Open Source Software, 2020.
Supported by Eurostat under GrantAgreement (GA) No: 882817