ISBN: 9789073592421

CPS Paper

Processing survey data with VTL

Author: Mr Thomas Dubois

Submission ID: 957

Reference Number: 957

Presentation File

abstracts/ottawa-2023_50a021fd301c3d24931e94dc758ace68.pdf

Brief Description

After renovating its data collection system for households surveys based on the concept of active metadata, INSEE has pursued technical investments for the post-collection processing.

The Validation and Transformation Language (VTL) proposed by the SDMX initiative is used for reconciling data collected from different modes and start the first processing.

VTL processing rules are used and interpreted thanks to Java and JavaScript implementations provided by the Trevas Open Source tool.

Abstract

INSEE has recently renovated its collection information system (with the so-called "Metallica" program) for households surveys based on the concept of active metadata: a single questionnaire specification expressed in DDI generates several collection instruments for the multimode data collection platforms. Several surveys by Internet and paper were thus operated in 2021, for example the "Daily life and health" survey, and by Internet and telephone in 2022, for example the "Housing" survey.
The survey data, once collected in different modes, needs to be processed to be integrated later in the statistical production process. The Metallica program has therefore pursued technical investments to reconcile the data from the different modes (GSBPM: "Process/Integrate data") and start the first processing (GSBPM: "Process/Classify and code"). In order to implement these features, Insee uses the Validation and Transformation Language (VTL) proposed by the SDMX initiative.
VTL is a standard language for defining validation and transformation rules for various kinds of statistical data.
It is intended to be used by statisticians and is at the "business" rather than technical level.
VTL processing rules are used and interpreted by the Metallica collection system thanks to Java and JavaScript implementations provided by the Trevas Open Source tool.
VTL is already used in the questionnaire design tool to specify logical expressions within the questionnaire (conditional expressions, checks and filters).
While the overhaul of the collection information system has led to a great deal of work on standardizing processing, there are still specificities to be taken into account for each survey. The use of the VTL processing language, dedicated to the designer and interoperable with the rest of the highly standardized system, has already made it possible to optimize the implementation and renovation of certain household surveys (all Insee household surveys will be migrated to the new system within the next 3-4 years) while preserving the specificities of each one. The VTL grammar makes it possible to cover the vast majority of needs in terms of post-collection processing specific to each survey, even in the case of complex protocols.
The next step will be to further develop the concept within complex panel and mutlimode processes, for example, to use VTL rules for the post-collection processing required for a new data collection cycle or change of mode, including through the use of paradata. It is also planned to develop a tool dedicated to the designer's work: the simplified specification of VTL rules for post-collection processing in a working environment, integrated with the one that already exists for the specification of questionnaires.