Alessandro Capezzuoli (Italian National Institute of Statistics)
Emanuela Recchini (Italian National Institute of Statistics)

DataSTAT Hub: a tool for the automatic collection of administrative data to produce official statistics

Topic 1 – Bringing in information from where we can get it

Keywords: automatic data collection, administrative sources, REST data integration, official statistics

Introduction

The need for relevant, reliable and even more timely statistical data to support decision making process and scientific research has contributed to a growing demand for new statistical information to best analyze and rule, at various levels, the deep social, economic and environmental changes occurred at regional and global scale. Official statistics, characterized by the highest quality possible inasmuch as they are produced in compliance with the United Nations Fundamental Principles of Official Statistics and the European Statistics Code of Practice, are best suited to meet this need.

It is widely recognized the increasing role that administrative data are playing in the production of more timely, more disaggregated statistics at higher frequencies than traditional survey data. They offer further information on a wide range of issues, including some which cannot be answered cost-effectively from survey data. The efficient use of all available information to produce timely, accurate and high quality statistics is a challenge for National Statistical Offices (NSOs), which are even more committed to developing methods and suitable tools for the production, collection, standardization and integration of different types of statistical data.

Bringing together information from different sources makes it possible to fill information gaps or provide insights which cannot be gleaned from the unlinked data and to improve the knowledge and the understanding about specific phenomena

Methods / Problem statement

The production of statistics based on administrative data from different sources is closely related to the methods and techniques of collection and integration of archives. Problems related to automatic data collection are numerous as they involve the production, the harmonization and the standardization of output and information flows, in order to make them usable by web applications or be stored in one database to be connected (record linkage), processed by statistical software and/or visualized within ad hoc created web platforms.

Results / Proposed solution

DataSTAT Hub is a tool that takes advantage of the potential offered by HTTP 2.0 through which it is possible to create REST microservices and exploit the methods offered by the CRUD architecture (Create, Read, Update, Delete). DataSTAT HUB can be used through two different architectures: star or centralized. The former implies that each microservice (hub node) is automatically populated by data providing subject through a set of querystrings (Create, Update, Delete) and can be accessed in reading (GET) by the central institution that performs data collection.

The latter architecture implies the automatic population of the central hub that interfaces with the various institutions through the just mentioned querystrings (Create, Update, Delete) that allow users to store data and metadata in a NoSQL database (Cassandra) using the key-value data model for their representation. DataSTAT Hub allows users to standardize the outputs in various formats (XML, JSON, CSV) and models (JSON-STAT, SDMX, DDI).

Conclusions

DataSTAT Hub is a suitable and easy tool for administrative data collection, standardization and integration: it does not require knowledge of the internal data base since the update is performed through the HTTP querystrings and can be used with any programming language. By allowing us to overcome some critical issues related to the use of administrative data, including those connected with privacy and security, a tool such as DataSTAT Hub is time saving and cost-effective, while providing high quality information.

It is a user-friendly tool developed by making use of open source technologies (PHP, MySQL, Cassandra) and can be conveniently shared among NSOs, while it is extensible to any institution interested in the automatic collection and integration of administrative data.