The new Istat open source and standards based architecture for high quality web dissemination of official statistical data

Mr Carlo Boselli[1], Mr Alessio Cardacino[2]

[1]ISTAT - Italian National Institute of Statistics, Italy

[2]ISTAT - Italian National Institute of Statistics, Italy

Abstract

In the In the context of recent modernization processes, ISTAT has turned on the innovation of its web architectures for data dissemination, by designing and implementing its new corporate reference web architecture for the dissemination of official statistical data.

This architecture, based on international statistical standards for the availability of statistical data and metadata (in particular the SDMX standard - Statistical Data and Metadata Exchange) as well as highly interoperable and compliant to the open data paradigms at the highest levels, was implemented using the most recent and advanced web technologies: by this way it is also able to increase the quality statistical aggregated data dissemination and to foster the standardization and industrialization of statistical data dissemination processes.

Based on this new architecture, in the last two years ISTAT has published various corporate dissemination systems: the new Corporate DataWarehouse (IstatData), the data dissemination portal of the Permanent Population Census and the Public Statistics Hub system for the centralized dissemination of data coming from the institutions that are members of the National Statistical System, the new web system for the dissemination of Foreign Trade data, Have also been developed by ISTAT some systems, such as the new Territorial Statistical Atlas (AST), which have specialized some functionalities of the new dissemination architecture to highlight certain thematic aspects (in the case of the Atlas, the statistics about the territory.

This work aims to highlight the following specific features of the new platform, related to:

- Data modelling according to the SDMX information model;
- Data transformation and processing of no compliant SDMX data source;
- Industrialised solutions for the periodical data upgrade;
- Footnotes and flags (SDMX attribute) management;
- SDMX annotations as practical means to drive the user data visualisation experience;
- Dashboards for quick synthetic data analysis;
- Performance issues related to big size datasets and high volume of concurrent accesses;
- Functionalities facilitating organizational aspects related to data migration and corporate data warehouse management.

Keywords: Dissemination, Open-Source, Istat, SDMX, TOOLKIT

Although the paper is fruit of a common effort, paragraphs 1 and 2 should be attributed exclusively to Carlo Boselli, and paragraphs 3 and 4 to Alessio Cardacino.

1. Introduction

The choice of a new dissemination platform, for a National Statistical Institute (NSI), requires a multidisciplinary approach, considering different skills involved in the development of the system (experts in business organization, statisticians, data modellers, IT experts) and the high level of complexity due to pre-existing processes and new processes to activate in coherence with the data production pipeline. Moreover, the process involve different kinds of data like administrative data, survey data, micro and aggregate data.

All these aspects connect to internal and external user needs that must guide the definition of the requirements and the activity for the development of the platform.

In particular, is possible to underline that technology is not the leading factor in the definition of a new platform, but must follow the organizational aspects that define the ability of an organization to create and maintain a good dissemination system overtime. Furthermore, a new dissemination system must me adaptable with the processes and the production chain inside the organization.

Following Istat experience, considering technological aspects as innovative solutions, this study highlights first the main organizational issue and internal and external user needs for the dissemination of aggregated data, and then go to describe the solution proposed by Istat.

2. Main needs for the definition of a new dissemination platform requirements: The Istat case study

Inside National Statistical Institutes, considering the mission of the public body and relationships with key stakeholders, there are two main high-level needs for dissemination, often combined in the same organization: a need of stability and a need of flexibility.

For stability aspects, most of NSI are required to respond to other organization, like Eurostat, in a reporting format. From this point of view, a protocol is recommended that stabilize data provisioning and the entropy related with the possibility to develop different format or variation in the codification or data interpretation.

For flexibility aspects, every NSI has the mandate to collect data from institutional units and provide back to society aggregated data with a statistical value added, following a dissemination flexible approach. This process constantly redefine data structure, depending on new possibility to join data, new available technology, new needs from users.

The dissemination platform must be consistent with both these two needs. In particular, more for dissemination purposes than reporting, is important that changes in the data structure are possible without altering depending artefacts and interactions with end users. This is an

important constraint for the SDMX standard and for the technological platforms based on it, that are more consistent with the need of stability than flexibility. Therefore, technological platform based on the SDMX standard must allow tools useful to regain in terms of flexibility and guarantee the possibility, for example, to add new modalities in the codelists, update data structure definition (DSD), and in the future to modify the star schema and re-align all the depending artefacts reducing impact on external users.

Considering this, this study highlights the Istat context and specific needs, which however may present several points of similarity with that of other national statistical institutes.

As other European NSI, Istat disseminate aggregate data resulting from a large number of surveys under Eurostat Regulation. Internal processes reporting data for Eurostat are the same that produce aggregate data the national level dissemination, with only partial differences in the set of data disseminated. As internal need, new platform must be integrated with pre-existing processes and allow data migration form legacy systems without altering production flows or increase burden on the production sectors and reuse of existing procedures for populating and updating data and metadata.

In some cases, aggregate data derive from a combination of administrative and survey data that is possible to integrate for specific domains of analysis for a more granular data dissemination. Furthermore, another important internal need is a high data store capacity per single data cube. In particular, for the population and foreign trade aggregate data, there are datasets with more than 60 million of data points, and it is necessary to store, query and browse them on the web as well.

Considering external users, is crucial to have a homogeneous environment for browsing data on the web, to reduce access costs, with the possibility to create, eventually, different instances or nodes for different topics. The platform should have also solutions for synthetic data representation, as dashboards, to gain in terms of data comparison, linking data from different datasets or node of different organizations, without modifying data structure and underlying data production processes.

Another aspect regards performances for a very efficient data visualization that takes advantage of an adequate cache system and allows large tables to be displayed. In particular, it is necessary to browse tables with more than one million of data points in a single web visualization.

External users need an effective information search system combining a vertical thematic tree-based with a transversal browsing system, and a text search engine which in the future can take advantage from possibilities given by the use of artificial intelligence.

A final aspect related both for internal and external users is to consider an open source architecture that is based on a standard in order to take advantage of technological modularity (changes or improvement in each component of the platform) without changes in the data modelling and preserving data dissemination processes in the future. An architecture based on a standard allows machine-to-machine access using API, and avoid misalignments between what is stored in the database and what is published via web service. In order to maintain high performances in browsing data is important also to parallelize data acquisition on the web browser with respect to the machine-to-machine channel.

Last but not least, considering in particular other organization that need to install the platform, we need an open source software with an ease installation and configuration of the tools, based on plug-and-play and ready-to-use approaches, in which only few minutes are required for installation without particularly advanced technical skills.

3. SDMX Istat Toolkit : the technological solution developed by Istat for aggregate data dissemination

The SDMX Istat Toolkit (https://sdmxistattoolkit.github.io/) developed by ISTAT is an **open-source platform** based on international standards that facilitates the standardization and industrialization of business processes related to data dissemination and exchange. Initially designed to consolidate the experimental software created as part of the “Public Statistics Hub” project (also known as “Sistan Hub”), that involve several Italian entities, the toolkit integrated features and functionality useful for rationalization and evolution of the Institute's dissemination system. Furthermore, it represents Istat's offer within the National Statistical System, which entities can access to engineer or re-engineer its own dissemination systems.

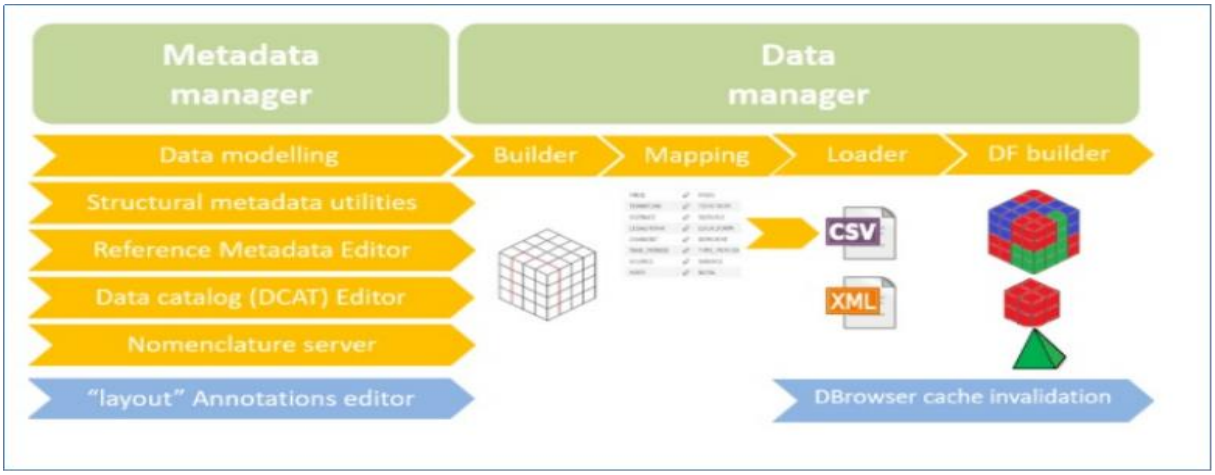
The Toolkit combines the two main modules "Meta and Data Manager" and "Data Browser" each of which is made up of a front-end, that allows interaction through a Web GUI, and a back-end that acts between the interactions of the user and the services that implement the various functionalities. The following paragraphs describe the functionalities of the two modules in relation to needs indicated in the first part of this study.

3.1.1 Meta and Data Manager

Meta and Data Manager is a web application that allows internal users to design and manage multidimensional datasets and the related structural and reference metadata

according to the SDMX Information Model. It allows publishing data and metadata in machine-to-machine mode via the SDMX web service (the NSI Web Service of the SDMX reference infrastructure of Eurostat) and the reference metadata API, through the following main functionalities (Figure 1).

Figure 1: Meta and Data Manager tools



The module guarantee consistency with the standard and the following functionalities are useful to obtain a high level of stability both for dissemination purpose and for reporting:

- Modelling of statistical tables through the definition of the relevant metadata, which specifies the role that each variable plays in the structure of the tables, as well as their representation through appropriate lists of codes or classifications. Creation and publication of descriptive metadata (conceptual, methodological, quality);
- Possibility to download and upload each artefact (as Data Structure Definition DSD, codelists, concept scheme) in different format, SDMX 21., SDMX 2.0 XML;
- Creation and management of statistical databases organized as multidimensional cubes in which to insert and update data starting from CSV, Excel and XML files;
- Display of multidimensional statistical tables through a specific API according to the SDMX standard;
- Creation and publication of catalogs of the datasets that you intend to disseminate. These catalogs are published using "DCAT vocabulary" according to the RDF standard for Linked Open Data. In particular, as default, everything necessary to implement the DCAT-AP_IT application profile created by AgID (Italian agency for digitalization) is provided;
- Creation, management and publication of thematic glossaries.

The architecture is completely oriented towards services activated and based on the actions performed by users by a client-side front-end that interacts with the back-end service (Node

Api). Furthermore, some services included in the architecture were developed by Eurostat as part of the SDMX Reference Infrastructure, and in particular: the NSI Web Service for managing and querying SDMX structural metadata and for data querying, and the Mapping Assistant Web Service for the implementation of data publication functions via Eurostat's SDMX web service (the previously indicated NSI Web Service).

The DM_API_WS service allows data management (hypercube creation, import, modification, deletion) for the purposes of their publication, and the METADATA_API is a front-end service that makes the reference metadata available in machine-to-machine mode in SDMX format (metadataset) and also the DCAT-AP catalogs of the datasets present in the system through appropriate public functions (methods) compliant with this standard: this service can be managed from an interactive client-side interface which is contained within the module front-end of the Meta&Data Manager, but which can also be used independently and separately from that application.

At the same time, the module guarantee several functionalities that allows to easily modify data structure, code list and other artefacts to obtain a good level of flexibility in particular for dissemination purpose:

- Ability to manage data structure definition (DSD), codelists and concepts through an intuitive interface that allows the internal user to easily edit, modify, upload or clone for easy reuse of the artefacts that allows you to create new versions without repeating all the steps;
- Ability to add modalities in codelists in a specific DSD and datacube, without modifying the versioning, guaranteeing external users a reuse of the pre-existing dataflows;
- Ability to add new attributes, with new footnotes or multiple encodings associated for a specific dimension of analysis, directly updating a pre-existing DSD, without destroying data cubes, dataflows and other depending artefacts;
- Using Meta-Data Handler tools, ability to make available data, already published according to a specific data structure, on the base of a different DSD defined by another international organization for reporting purposes;
- Ability to upload pre-existing csv files, with different formats, using file-mapping tools. This functionality allows internal user to upload in the Meta and Data Manager, without changes in the format, files already defined from production units to populate legacy dissemination systems, reducing the burden during data migration.

The application is also equipped with a very flexible and granular mechanism for managing user access profiles, which allows to define these profiles starting from the individual functions and selecting the domains or data sets accessible by users with an operational profile.

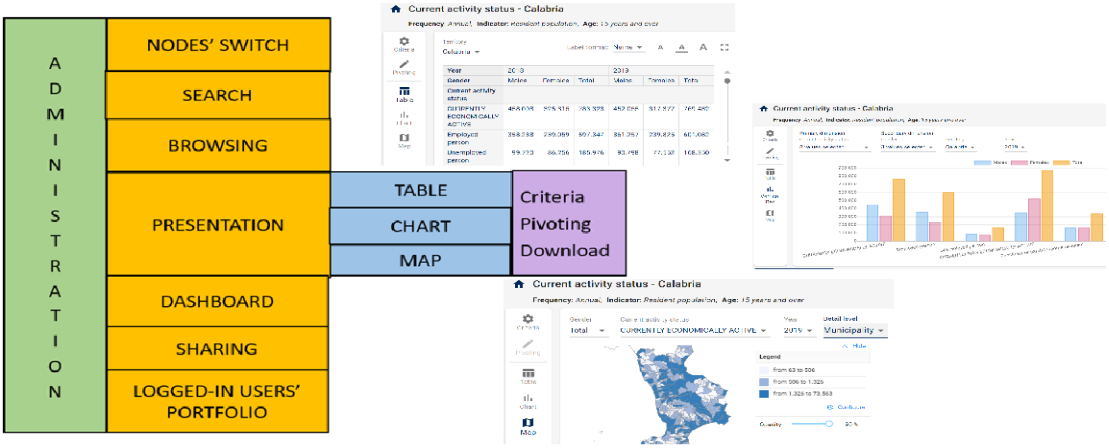
Through a single instance is possible to access more than one work environment for managing a data and metadata source (node), after having selected which one to operate on from the home page. Furthermore, it is possible to browse SDMX registries of structural metadata made publicly available by other organizations, via the machine-to-machine connection to their end-points.

The system has evolved with the possibility to insert in plug-in mode some procedures for transforming data structures and formats, which can be used both to load data and to be used as stand-alone functions inside the system.

3.1.2 Data Browser

Data Browser is a web application that interacts with SDMX API allowing users to browse, search, query, view statistical tables (datasets) and download them in various formats (xml, csv, json). The representation of the datasets can take place in the form of multidimensional tables, graphs, thematic maps (Figure 2).

Figure 2: Data Browser tools



This application can be used within a single organization in order to disseminate datasets stored in one or more databases (e.g. https://esploradati.istat.it), or within a distributed architecture on web servers reachable via the http protocol on the Internet (e.g. https://idp.sister.it). This aspect is important to compare data from different organizations, giving the possibility in a single instance of the application to browse, view and download data coming from multiple delocalized sources (nodes). This feature is possible both in "dispatcher" mode (i.e. browsing one node at a time) and in hub mode (i.e. accessing a node that allows to

query datasets coming from different delocalized data sources in real time as if these data were stored in the same database).

Another aspect related with data comparison and synthetic representation is the possibility to create dashboards for displaying tables, graphs, maps and text within the same web page whose data can come, for each individual dashboard, from different datasets and data sources.

It is possible to browse datasets in the format of multidimensional tables, associated with various graph modes, thematic maps, and customize the layout of all forms of visualization. It is also possible to download data in various formats, including open data such as XML (SDMX-ML), SDMX-CSV, SDMX-JSON, CSV (custom), Excel, image and pdf (for graphs and maps).

For the aspects related to performances, the module allows browsing a dataset with more than 300 million of data points, visualizing on the web a table with more than one million of data points per time and providing search criteria to change the selection to browse all stored data.

The system is also equipped with a sophisticated cache, in order to make queries and data visualizations highly efficient even in the presence of particularly large datasets. This solution allows storing and reusing, for subsequent queries, the results of all the queries carried out by users. This system is consistent with updating criteria of the datasets themselves, which invalidate the caches previously created.

Other features are the possibility to browse and search datasets within each single data source by typing keywords or directly selecting the dataset of interest with the help of specific hierarchical "theme trees", and the possibility for registered users to save their data views within a portfolio that can be consulted once logged into the system.

The system is also equipped with an administration module that allows to configure both the application and the connection to the various delocalized sources (nodes) via web interface.

The architecture of this application is also completely service-oriented, with a client-side front-end that queries a back-end service, which manages all aspects of data browsing and representation and interacts with the delocalized data sources by connecting to the respective SDMX web services.

4. Results and Conclusions

The new Istat architecture for dissemination has been implemented following the deep analysis of internal and external user needs. Both the possibility to reduce the impact on the production processes during the data migration from legacy systems and the ability to integrate into the data production and dissemination chain have been taken into consideration.

Particular attention has been paid to develop an open source architecture based on an international standard to guarantee stability in the connection with other organizations for reporting, machine-to-machine data and metadata transfer (based on recognized APIs) and possibility of reuse.

The new architecture presents also a high level of flexibility in terms of management of data and metadata, and the possibility for external users to browse information coming from different organizations and compare them within the same environment through synthetic and interactive representations such as dynamic dashboards.

Particular importance has been given to the performances, obtained with an innovative cache system implemented into the data browser module.

The new architecture is one of the best performing at an international level in terms of data representation, with the possibility of archiving large datasets, up to 300 million records, and browsing them via web tables with one million cells displayed in the same visualization.

References

Cardacino, A., Boselli, C., Coccia, S. (2023). La nuova architettura Istat per la diffusione web dei dati statistici. AISRE - XLIV Conferenza Scientifica Annuale 2023: https://www.aisre.it/wp-content/uploads/aisre/64b2bc0e324823.88632485/Cardacino.pdf

Cammarrota, M. (2023). Sistema integrato di diffusione delle statistiche dell'Istat: percorso e stato attuale. *Letture statistiche – Metodi*, ISBN: 978-88-458-2123-3 *https://www.istat.it/it/archivio/293006*

Boselli C. (2017). The dissemination process of the Frame-SBS: legislative and methodological aspects linked to increase information detail. R*ivista di statistica ufficiale n. 1-2-3/2017*, https://www.istat.it/it/files/2021/02/RSU_1-2-3_2017_Article_2.pdf

European Statistical System − ESS, and Eurostat. 2017. European Statistics Code of Practice. For the National Statistical Authorities and Eurostat (EU statistical authority). Luxembourg:Publication Office of the European Union. https://ec.europa.eu/eurostat/web/quality/europeanquality-standards/european-statistics-code-of-practice.

Nadezhda VLAHOVA (2017). Integration of INSPIRE & SDMX data infrastructures for the 2021 population and housing census. https://unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.58/2017/mtg3/2017-UNECE-Standards-Workshop-INSPIRE-SDMX-paper-V2__1_.pdf

Spence, R. 2007. Information visualization. Design for interaction (2nd Edition). London, UK: Pearson.