INTERSTAT – a project on Open Statistical Data

Interoperability

**Keywords:** Open Data, Linked Open Data, Statistical Open Data, framework

# Introduction

Statistical information is a key data source to support public policy definitions but also to enable utility services across different domains for a large range of stakeholders. At national and European level there is still lack of interoperability between several open data portals of the public sector and statistical portals, with different ways to access and reuse open datasets.

Funded in the framework of European CEF (Connecting Europe Facility - <https://ec.europa.eu/inea/en/connecting-europe-facility>), the INTERSTAT project will provide a set of solutions to enable interoperability among different national statistical portals and the European Data Portal. The main objective of the project is to achieve data harmonisation in the field of (linked) open data statistics and to provide standards, methodologies and a technical solution, deploying cross-border end-users services in statistical domains, reusing harmonised statistical data in combination with open dataset from European Data Portal and other national open data portals.

The INTERSTAT project is coordinated by Engineering – Ingegneria Informatica spa, an Italian IT company, and participated by the French National Statistical Institute INSEE, the Italian Statistical Institute ISTAT and FIWARE, a German foundation developing open platforms.

# Methods

INTERSTAT will achieve the above-mentioned objectives performing a set of scientific, technical but also legal-related activities, with the final aim of provisioning a coherent and holistic framework to enable and foster the deployment of new replicable cross-border services reusing European statistical and valuable open datasets.

Main INTERSTAT tasks and activities will be:

1. Requirements and guidelines for open statistical data reuse and exploitation. The analysis will consider the current European legal and technical framework, identifying the impact and the adoption of open data in the different European countries in relation to the statistical domain.
2. Technical framework for open statistical data interoperability. The overall objective of this activity is to design and implement the technical framework to achieve Open Statistical Data interoperability, enabling pilot service execution. The technical solution will integrate and make interoperable different Statistical Data systems with mature open source software, focusing on data sharing and interaction with external systems, in particular the European Data Portal.
3. Pilot services execution and assessment. The overall objective of this activity is to deploy, run, and assess the pilot services of INTERSTAT designed and implemented in the previous task. The activity includes the detailed description of the use cases, the elicitation of service requirements, the design and technical deployment of the pilot services and the impact evaluation.
4. Project dissemination and sustainability. The consortium will produce and execute a comprehensive communication and dissemination plan, in order to reach as large an audience as possible. The strategy will be to create public awareness, but also to contribute to the development of an open source community and culture.

The main activity is on the technical framework designed to achieve Open Statistical Data interoperability, enabling pilot service execution. The technical solution to be defined in this task will integrate and make interoperable different Statistical Data systems with mature open source software, focusing on the provisioning of API for data sharing and the interaction with external systems.



The INTERSTAT framework architecture, as depicted in the picture, is composed of different layers and components. The “Statistical data sources” layer includes the data coming from different statistical providers: in the INTERSTAT project pilot scenarios, statistical data will come from ISTAT and INSEE systems (Italian and French national statistical institutions). Statistical Data are originally based on heterogeneous standards and formats for metadata and data. The data collected from these systems, through specific data connectors, are then harmonised in terms of content performing specific data transformation and adopting a common set of shared ontologies for statistical data, defined during the project.

The Open Data Interoperability Layer is the core of INTERSTAT’s technical framework because it includes all the components necessary for the data harmonisation and the data provisioning via standard APIs. In particular the main components are:

* **Idra** [1] is an open source platform aiming to federate and harmonise open data coming from heterogeneous sources (i.e. open data portals). It will be a central component, able to communicate with the other ones, and it will have the main role of adapting and harmonising open statistical data to the main standards for their provisioning through open data portals, in particular the European Open Data portal. Idra will expose statistical data using DCAT-AP [2] metadata and de-facto standard APIs such as CKAN [3] ones. Idra will be also able to connect to the European Data Portal to get open datasets making them available through the INTERSTAT framework in order to be used in combination with the Open Statistical Data;
* **CEF Context Broker Building Block** [3] which will have the role of interoperability component for third party systems exposing statistical (linked) data and metadata via the NGSI-LD API [4]. External applications will consume this information based on a publish-subscribe approach allowing to be notified about changes on data or metadata. The NGSI-LD API will be used also to access to historical data;
* A module for translating **SPARQL** [6] **queries to SDMX** [5] **queries** and submitting them to the **Eurostat’s SDMX web service**: if a mapping among OWL Statistical Ontologies and SDMX metadata is already defined, the module will interpret the mapping in order to translate the SPARQL to SDMX queries, otherwise it will directly translate the SPARQL to SDMX queries. The SDMX web service's open data outputs can be also federated with external open data portals by the availability of CKAN API and DCAT-AP standards;
* **RSS Feed service**: a service that can be activated by users to be notified of new dataset publications through web feeds.

In addition to these components, the INTERSTAT framework will provide some end-users data visualisation tools, in particular:

* **Guided semantic search**: which allows users to search data using keywords and to formulate SPARQL queries through guided interfaces accessing data in open format, based on semantic web technologies and standards, without the necessity of having the technical knowledge;
* **Ontology visualization**: which allows users to visualize and navigate ontologies and metadata underlying the published data, and in general open statistical data, in a more effective way.

The upper part of the architectural picture shows the Open API layer, that includes a set of APIs that will be used to access the harmonised Open Statistical Data, in combination with additional open data (e.g. from European Data Portal), using different standards and technologies. In particular, the innovative pilot services, developed and piloted during the project, will access to the data, produced and provided by INTERSTAT, through the Open API layer.

# Results

Main outcomes of the projects will be the following:

* Legal and technical reports on the adoption of (Linked) Open Statistical Data in Europe
* Standards, methodologies and a technical framework to achieve data harmonisation in the field of (Linked) Open Statistical Data among different national statistical institutions
* Uniform technical interfaces for a standard and simple re-use of statistical information through the adoption of CEF Context Broker Building Block and the implementation of ETSI NGSI-LD API specification
* Tools to simplify statistical data visualisation and analysis for non-technical end-users
* Deployment and piloting of cross-border end-users services in the domain of Population and Households Census, reusing harmonised statistical data in combination with open dataset (i.e. city-related data) from European Data Portal and other national open data portals.

# Conclusions

INTERSTAT project started in September 2020: for the NTTS conference, some of the first results of the project will already be available. Particularly, the assessment of the legal and technical situation on statistical Open Data in Europe, and the use cases will be clearly defined and therefore they will be presented and discussed with the statistical community.

# References

|  |  |
| --- | --- |
| [1]  | “Idra - Open Data Federation Platform,” [Online]. Available: https://github.com/OPSILab/Idra. |
| [2]  | “DCAT Application Profile for data portals in Europe Version 2.0.1,” 2020. [Online]. Available: https://joinup.ec.europa.eu/sites/default/files/distribution/access\_url/2020-06/e4823478-4458-4546-9a85-3609867ad089/DCAT\_AP\_2.0.1.pdf. |
| [3]  | “CKAN Association,” [Online]. Available: https://ckan.org/. |
| [4]  | “ETSI GS CIM 009 V1.3.1 (2020–08) Context Information Management (CIM); NGSI-LD API,” 2020. [Online]. Available: https://www.etsi.org/deliver/etsi\_gs/CIM/001\_099/009/01.03.01\_60/gs\_CIM009v010301p.pdf. |
| [5]  | “CEF Context Broker Building Block,” [Online]. Available: https://ec.europa.eu/cefdigital/wiki/display/CEFDIGITAL/Context+Broker. |
| [6]  | “SDMX technical specifications,” [Online]. Available: https://sdmx.org/?page\_id=5008. |