INTERSTAT

Open Statistical Data Interoperability Framework

www.cef-interstat.eu

D3.2 - Report on pilots' development and deployment

Co-financed by the Connecting Europe Facility of the European Union

The contents of this publication are the sole responsibility of INTERSTAT consortium and do not necessarily reflect the opinion of the European Union

Project full title

INTERSTAT - Open Statistical Data Interoperability Framework

Grant Agreement No.

INEA/CEF/ICT/A2019/2063524

Project Document Number

Deliverable 3.2 (Activity 3)

Project Document Delivery Date

08.06.2022 (v1.1)

Deliverable Type and Security

Report – Public

Authors

Franck Cotton (Insee), Romain Tailhurat (Insee), Dubois Thomas (Insee), Adele Maria Bianco (ISTAT), Paolo Francescangeli (ISTAT), Roberta Radini (ISTAT), Michele Karlovic Riccio (ISTAT), Giuseppina Ruocco (ISTAT), Francesca D'Agresti (ENG), Fernando Lopez (FIWARE).

Contributors

Cristiano Maione (ISTAT)

Reviewers

Martino Maggio (ENG)

Co-financed by the Connecting Europe Facility of the European Union

Table of Contents

Co-financed by the Connecting Europe Facility of the European Union

List of figures

List of tables

Co-financed by the Connecting Europe Facility of the European Union

The contents of this publication are the sole responsibility of INTERSTAT consortium and do not necessarily reflect the opinion of the European Union

Executive Summary

The main goal of this report is to document the implementation and deployment of the services related to the use cases described in deliverable "D3.1 - Report of the use cases to demonstrate the cross-border benefits of the proposed solution" [1] and further detailed in "D2.1 - Ontologies and tools to enable cross-border semantic interoperability" [2], conceived to test cross-border and cross-domain interoperability. More in detail, the first chapter provides a short summary of the use cases, and a quick overview of the functional and technical aspects of the production environment.

Chapter 2 briefly describes the client applications developed for data dissemination, and focuses on the workflow implementing the data pipelines, based on the following approaches:

- The ETL approach for the "Geolocalize Facilities" pilot
- The Domain knowledge approach, applied in the other two use cases: "The School for You", and "Support for Environment Policies".

Chapter 3 details the main steps of the pilot development, documenting the input data sources, the target data and metadata models, as well as the different services implemented for the dissemination of linked data.

Chapter 4 covers a specific transversal activity aimed to achieve interoperability between statistical data (SDMX) and Context Information Management (ETSI NGSI-LD specification).

In the end, the last chapter summarizes the main lessons learned during the pilot implementation, and the future steps to improve and continue the use cases development until the end of the project.

Co-financed by the Connecting Europe Facility of the European Union

1 Introduction

1.1 Scope and objectives of the document

This document describes the INTERSTAT pilot services, how they were developed and how they can be used. It is part of the Milestone 6 (M6) of the project, which consists in having the "pilot services deployed and working in real environments", and in particular in the availability of the pilot services applications. M6 pertains to Activity 3 of the project ("Pilot services execution and assessment"), which extends until the end of the project with additional milestones dedicated to the monitoring and assessment of the pilots. Therefore, the services and applications will continue to evolve by continuous improvement in the future. It was thus decided to keep this accompanying document relatively short and high-level and to reference where appropriate to the online resources, much more detailed and which will be kept up to date as the project continues to evolve.

As already mentioned, the document starts with a short recap about the different pilots and the technical production environment set up for INTERSTAT. General considerations on the development of pilots are then exposed, with a particular focus on the approaches followed for creating the data pipelines and on the technical stack used for the client applications. More detail is then provided for each of the three pilots, including a reminder of the business case, the description of the relevant models and of the data, metadata, pipeline process and client application. A summary of the lessons learned, the remaining problems and the next steps is given in conclusion.

1.1.1 One-liners on the pilots

- *Support for Environmental Policies (SEP):* One of the main goals of this use case is to enrich air quality data with demographic analysis, to support local public authorities responsible for environmental policies.
- *Geolocalized facilities (GF):* dissemination of information about facilities and integration with related sources
- *School for You (S4Y):* This pilot allows users, such as citizens and political decision-makers, to discover aggregated data resulting from the integration of several sources about school attendance and the distribution of students in Italy and France.

Co-financed by the Connecting Europe Facility of the European Union

1.2 Services’ environment

The client applications developed are one for each pilot service; they are web applications that consume and display data using SPARQL language. This section briefly describes the information relating to the production environment, the programming language used for the implementation and the type of repository used to collect the data.

1.2.1 Programming language adopted for development: ReactJS

React [3] is a JavaScript library for building user interfaces. The key features are mainly two: the first one, is the concept of UI components. The React code is composed of entities called components that are reusable; this means that it is necessary to simply define a component only once, and it can be used multiple times. This makes it much quicker, convenient, and easier to handle the production of a large-scale website. This allows the ability to breakdown complex UI and allows to work on individual components.

The other key feature is that React finds out what changes have been made, and modifies only what needs to be changed. It creates an in-memory data-structure cache, computes the resulting differences, and then updates the browser's displayed DOM (Document Object Model) efficiently. This selective rendering provides a major performance boost. It saves the effort of recalculating the CSS style, the layout for the page and the rendering for the entire page.

1.2.2 Data repository: GraphDB

GraphDB [4] is a highly efficient and robust graph database with RDF and SPARQL support. In particular, it is an enterprise ready Semantic Graph Database, compliant with W3C Standards. Semantic graph databases (also called RDF triplestores) provide the core infrastructure for solutions where modelling agility, data integration, relationship exploration and cross-enterprise data publishing and consumption are important.

This scalable RDF database streamlines the load and use of linked data cloud datasets, as well as the own resources. For easy use and compatibility with the industry standards, GraphDB implements the RDF4J framework interfaces, the W3C SPARQL Protocol specification, and supports all RDF serialization formats. It is one of the few triplestores that can perform semantic inferencing at scale, allowing users to derive new semantic facts from existing facts. It handles massive loads, queries, and inferencing in real time.

In GraphDB, data is organized in repositories; once the one of interest has been selected, the data is extracted through the SPARQL query language. GraphDB also supports SPARQL Federated Query which is an extension of the basic Query Language. Using it, it is possible to combine a query on a repository in the current instance with a remote call to a different SPARQL endpoint.

The strength of GraphDB is that it is also an HTTP service endpoint that can receive requests and can process SPARQL queries; this service is queried directly by client applications and the requests results are displayed in them.

1.2.3 Sharing and exchange of project files: SFTP server

As part of the INTERSTAT project, within the different project activities, an SFTP [5] (SSH File Transfer Protocol) server was used to exchange data, transfer, share and manage files produced by the intermediate stages of the development pipeline. Transferring files through an SFTP server is a secure way to transmit data, as the data to establish the connection and finalize the authentication, as well as the data transmitted are encrypted. With an SSH FTP server, it is possible to transfer files securely over an encrypted SSH connection. The SFTP server of the project contains both the files used as input to some phases of the pipeline, and the output files resulting from the activities also.

2 Pilots architecture

In this section, we describe how the pilot services were created, with a specific focus on the development of the data pipelines and on the client applications.

2.1 Data pipelines

Two different approaches were used for the implementation of the data pipelines: a classical ETL (Extract, transform, load) pattern for the GF pilot, and an approach based on domain knowledge for SEP and S4Y.

2.1.1 ETL approach

The GF ETL process was designed with the following principles in mind:

- Openness
- Maximal automation
- Reproducibility
- Efficiency.

Openness leads to developing all code on the specific repository in the INTERSTAT GitHub [6] starting from the first line, and to using only open source tools.
Maximal automation avoids manual treatments, which saves time and improves traceability. It is often a difficult principle to follow, in particular with messy data, since it requires rigor and a bigger development effort, but it largely pays off in the end, especially if source data changes frequently.
Reproducibility results from automation and from detailed documentation inside and outside code.
Efficiency is ensured by the selection of a technical framework that provides for execution of the pipeline in a distributed environment.
Regarding the tooling, the following choices were made:

- Use of Python3 [7] as a programming language
- Use of Prefect [8] as a build, run, and monitor framework.

Prefect allows for good modularity and readability of the code, and provides process visualization tools for the conception and execution stages (see the example of the "Geolocalized Facilities" pilot

Co-financed by the Connecting Europe Facility of the European Union

[9]). Prefect pipelines can be executed locally, for example for test purposes, or on a cloud platform [10] (which can be installed on premises).

More details on the technical environment for the ETL Python implementation is available [11].

2.1.2 Domain knowledge approach

The unique pipeline, designed for both pilots and based on Ontology, is described in the following picture. This methodology is generalized and applicable to other use cases, as well as the mixed approach based on both ETLs and data integration through domain ontology (Ontology Based Data Management – OBDM approach).

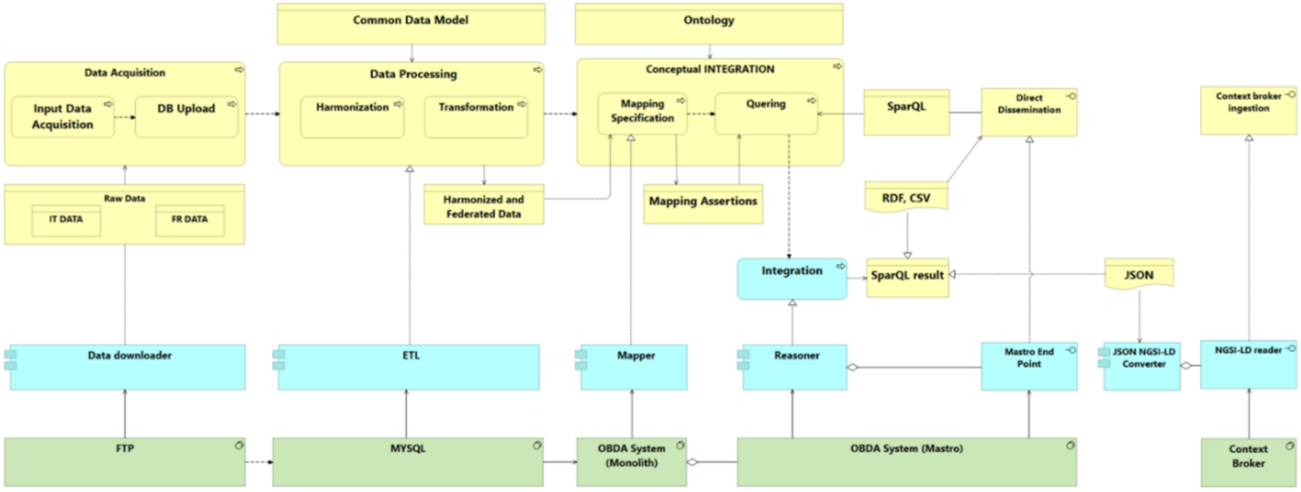


Figure 1 - Layered data flow outline in the domain knowledge approach

Main steps of the pipeline to be implemented for SEP and S4Y pilot:

1. **Data Acquisition** - Input Data are downloaded from the source websites into a staging area. 1.1. *Input Data Acquisition* - Data are downloaded from the web sources either by procedure or manually 1.2. *DB Upload* – Data are uploaded into a relational database for data processing.

2. **Data Processing** - Data are harmonized to a common data model through ETL, or transformed to create new variables required for data matching. Several transformations are performed on uploaded data at this stage. 2.1. *Harmonization* - Data are harmonized to a common data model through ETL. 2.2. *Transformation* - Data aggregation and standardization, unit identification, common variables creation according to target data

Co-financed by the Connecting Europe Facility of the European Union

models, creation of new variables. Queries and views implement **ETL Logical Level Processes** in the underlying host **MYSQL database**. Harmonization and Transformation processes can be run in parallel, and the result is stored into **harmonized datasets**.

3. **Conceptual Integration** – Once data are harmonized, they can be integrated. Data are integrated on a conceptual level through the domain ontology. Data are not linked physically but through a SPARQL query, thus the integration is virtual. 3.1. *Mapping* – Virtualization process associating physical data to ontology concepts. 3.2. *Querying* - Virtual integration by SPARQL queries. Results can be exported into the desired format. Data are **federated**, they can be viewed as a single coherent set, even when actual data sources vary in format and storage technology. The components implementing the **mapper** and the **reasoner** are submodules of an Ontology Based Management System (**OBDA System**).

4. **Direct dissemination** – End point can be used to query and disseminate data in table format or send data to specific applications. They represent the communication interface with the external world. End users can query linked data through predefined queries, or by writing new queries, or by selecting the concepts of interest modelled in the ontology, realizing conceptual integration by design (through the Sparqling tool part of the INTERSTAT framework). 4.1. *SPARQL* – Queries are provided as input to the system through the endpoint interface 4.2. *SPARQL result* – Results are provided as output through the endpoint interface.

5. **Context Broker Ingestion** – Data exported in JSON can be sent to context broker via a converter module into NGSI-LD format. 5.1. Queries are provided at design time by the designer when using this endpoint. 5.2. Result set must be converted into a specified format JSON NGSI-LD through a dedicated converter module.

2.1.3 Comparison of approaches

The main differences between the two approaches (ETL and Domain knowledge) concern the different tools used for data harmonization and the methodology applied to obtain and convert integrated data in RDF format. More in detail, in the first approach ETL procedures have been developed in Python, while in the second approach in SQL. In relation to data integration through ontologies, the first approach generates RDF triples from a CSV dataset through Python procedures, while the second approach uses relational DB and Monolith [12]. In addition, while the first approach uses meta-ontologies (e. g.: SKOS [13], Data Cube vocabulary), the second approach is based on domain ontologies and then links the concepts of the domain ontology to meta-ontologies concepts. The two approaches have converged in the dissemination step. In both cases, RDF triples, generated through the different pipelines have been uploaded on GraphDB repository, as well as the metadata already modelled according to SKOS meta-ontology (air pollutants code list, geographic code lists). RDF triples related to different domains, (e.g.: census and air pollution observations) have linked through SPARQL queries.

Computing Management and Optimization Tasks can be managed and executed either physically or virtually, to balance the resources for data processing. While physical elaboration is quicker but static, the virtualization is more dynamic but more complex. One example is the conversion of the Geo coordinates in Administrative Units (LAU). If Virtualization is chosen, GeoSPARQL [14] queries can be integrated into the Ontological framework, and the corresponding LAU can be derived virtually, but this is rather heavy on the reasoner because it must be calculated on the fly at each query. So, one can just statically convert the coordinates through a dedicated service and create a materialized new variable to store LAU and then virtualize it without the need to reference GeoSPARQL.

2.2 Client applications

At the end of the development of the data pipelines, the data produced is suitably consumed by *client applications*, which are web applications exposing services in which the information obtainable from the underlying data is organized and displayed. Three applications have been developed, one for each pilot, and the related services are described in detail in the next chapter. The description of the exposed services will be inserted into tables in which, in particular, they are classified into:

- *Cross-border service*: it compares a specific indicator or a specific variable in selected Italian and French areas.
- *Cross-domain service*: it allows to link different domains (such as Census and Air Quality) in order to produce new useful information from their combination.

A service can also be both cross-border and cross-domain.

Co-financed by the Connecting Europe Facility of the European Union

3 Use cases development

3.1 Support for Environment Policies (SEP)

3.1.1 Business case

One of the main goals of this use case is to enrich air quality information, produced to support local public authorities responsible for environmental policies. More in detail, several decision makers could get insights from the combination of:

- Sensor data, measuring the concentration of air pollutants
- Statistical data, describing the structure and the main characteristics of the resident population.

Linking air quality indicators and demographic data could allow decision makers to prioritize target areas of intervention. As an example of data integration benefits, a set of focused actions could be planned according to:

- The resident population living in areas where air pollutants exceed air quality thresholds
- The assessment of the effects of air pollution on vulnerable population groups.

3.1.2 Models

The target data model for census and air quality data, exported in Web Ontology Language (OWL [15] format), combines several existing vocabularies, such as SOSA [16] for sensor description and AQD [17] model for Air pollution. An overview of the ontology structure of census and air quality data is depicted in the following figure:

Co-financed by the Connecting Europe Facility of the European Union

The contents of this publication are the sole responsibility of INTERSTAT consortium and do not necessarily reflect the opinion of the European Union

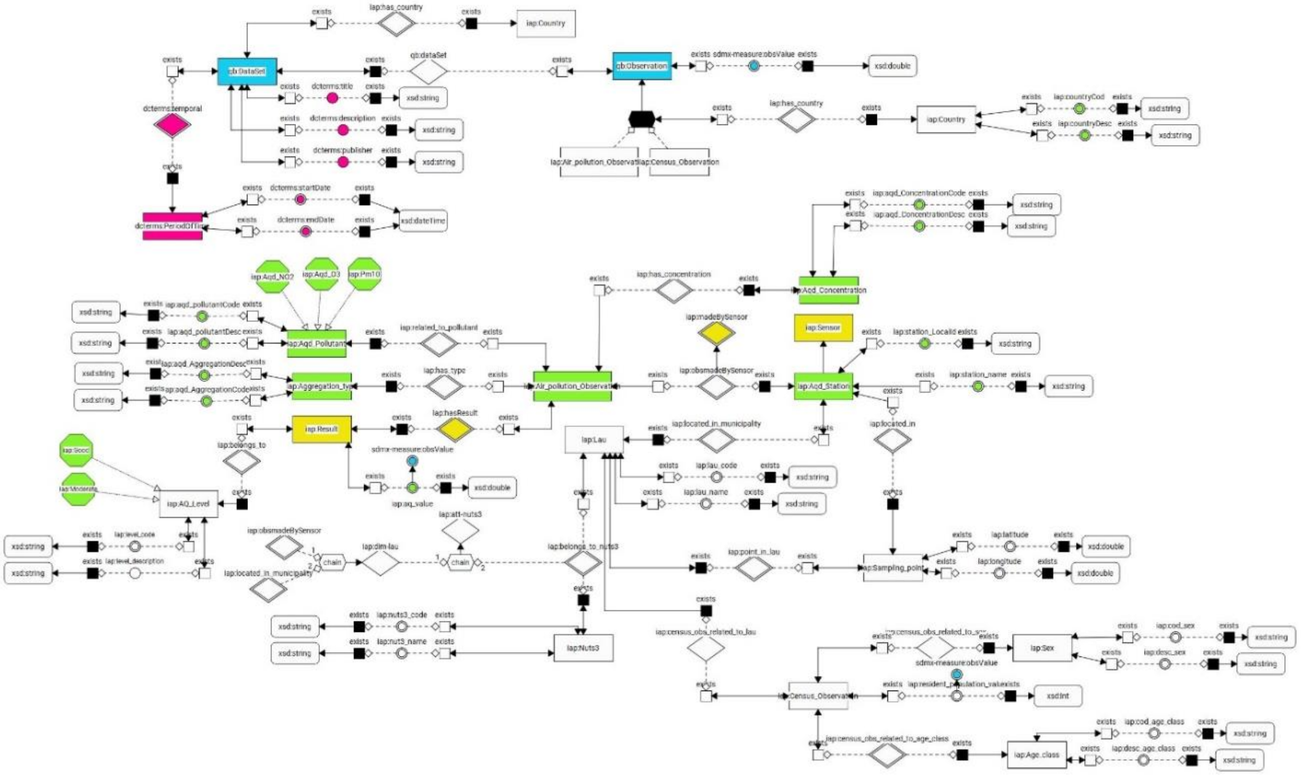


Figure 2 - SEP Ontology

The following figure shows the subset of SKOS concepts integrated into the SEP domain ontology.

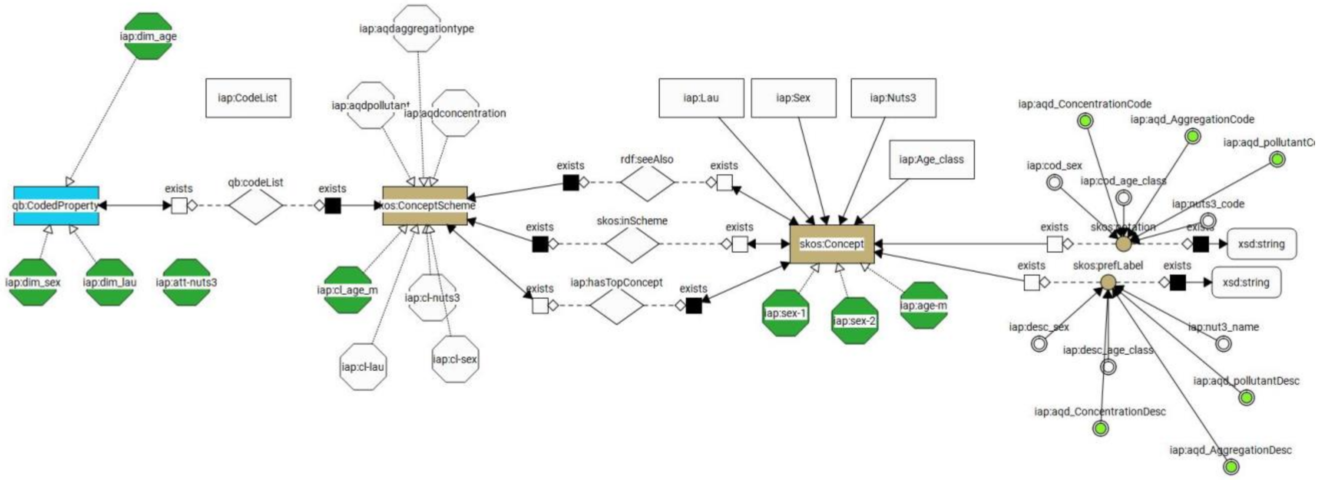


Figure 3 Subset of SKOS concepts integrated into the SEP ontology

In the modelled ontology, the information objects are colour-coded as follows:

Co-financed by the Connecting Europe Facility of the European Union

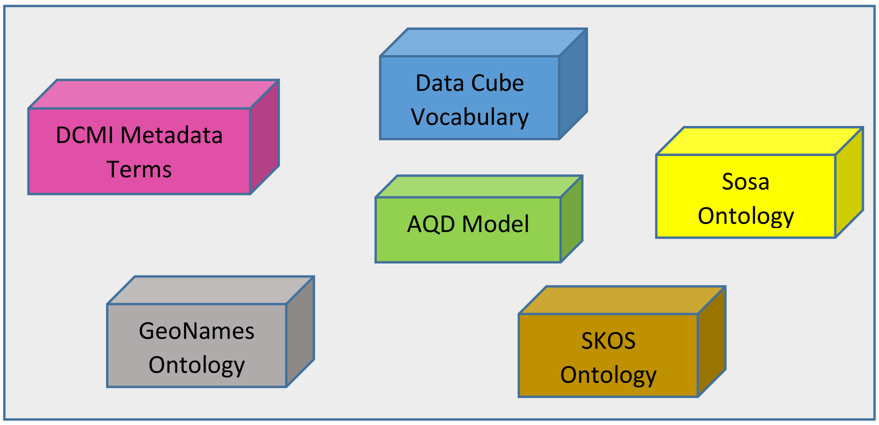


Figure 4 - Ontologies involved in the Air Quality ontology

The updated version of the ontology contains some metamodels, such as Data Cube Vocabulary and SKOS.

In the near future, the ontology will be further completed through the addition of the concepts related to Dimensions and Data Structure Definition and the definition of their instances.

Concerning the metadata, the air quality level will be modelled according to the SKOS ontology, while the country concept showed in Figure 4, will be replaced by the level 0 of the NUTS classification.

3.1.3 Data

This paragraph provides an overview of structural metadata describing the different data sources to be linked.

Census data

The SEP pilot plans to combine air quality data with demographic data from the French and Italian censuses. Census data whose metadata are defined by European legislation have been selected in order to minimize interoperability issues and ensure reproducibility at the European level. The explanatory notes for the 2021 census round give details on this subject. In particular, they present a new feature of the 2021 round: the dissemination of population data at the 1 km² grid level, for which Eurostat will provide Inspire metadata and which will be particularly interesting to combine with air quality data.

For testing purposes, it is easier to start with simple data, for example the breakdown of population by age range, sex and geographic local administrative unit. In DDI-CDI terms, census data corresponds to a "Dimensional" (actually "Cube") data structure. The definition of this data

structure according to the SDMX model is described in the specific section of the INTERSTAT GitHub repository [18]. Italian census data have been extracted from the section of ISTAT website [19] disseminating the main results of permanent censuses, based on the integration of administrative sources and data collected on a representative sample of municipalities and households.
The following table reports the list and the description of the fields extracted and transformed in the data processing step.

Field name	Description	Data type
ITTER107	LAU codes	String
Territory	LAU names	String
SEXISTAT1	Gender code	Code list
ETA1	Age class code	Code list
Age class	Age class description	Code list
TIME	Reference year	Year (always '2019')
Value	Value of the resident population in the reference period	Float

Table 1 - Fields extracted and transformed in the SEP Census data processing

For France, the data is collected from the Insee website [20]. The reference year is 2018 for the population counts and 2020 for the reference geography.

Italian Air pollution data

The data related to air pollutants have been extracted from the annual reports published by ISPRA, the reference authority for monitoring and assessing air quality in Italy. More in detail, It is available the endpoint to extract data concerning PM10 pollutant [21].

The subset of fields extracted from the original data source are described below.

Field name	Description	Data type
Regione	Region Name	Text
Provincia	Department name	Text
Comune	Municipality name	Text
Nome della stazione	Name of the Sensor station	Text
Valore medio annuo[3] [µg/m^3]	Average annual value	Float

Table 2: Fields extracted and transformed in the SEP Italian air pollution data processing

French Air pollution data

Co-financed by the Connecting Europe Facility of the European Union

French data concerning PM10 pollutant, collected in 2019 as reference year, are available [22]. Air quality data is available from the European Environment Agency (EEA) at the Air Quality e-Reporting web page. More precisely, the "AIDE F" data flow seems in first approach to be the most relevant for the SEP pilot. In particular, the description of variables for AIDE F is reproduced below. The table below lists the fields extracted from the original data source to be transformed in the following steps.

Field name	Description	Data type
CountryOrTerritory	Country or territory name	String
ReportingYear	Year for which primary data have been reported	Numeric
StationLocalId	Inspire identifier (LocalId) of air quality measurement station, given by data provider	String
SamplingPoint_Latitude	Latitude of sampling point (decimal degrees)	Numeric
SamplingPoint_Longitude	Longitude of sampling point (decimal degrees)	Numeric
Pollutant	Air polluting substance, level of which is measured and reported to the EEA (see notation in Data Dictionary)	String
AggregationType	Information about process of data aggregation into annual values (see in Data Dictionary)	String
Unit	Unit of concentration or level of air polluting substance (see in Data Dictionary)	String
AQValue	Concentration or level of air polluting substance, here given as an aggregation of air pollutant concentration values from primary observation time series	Numeric

Table 3 - Fields extracted and transformed in the SEP French air pollution data processing

3.1.4 Metadata

Concerning geographic location codes, the classification of Local Administrative Units (LAUs) and European regions according to NUTS (Nomenclature of territorial units for statistics) system is published in Eurostat website [23]. The description of the code lists of categorical variables extracted from the original data sources are available in the specific section of the INTERSTAT GitHub repository [24]. Most of these code lists are based or compliant with the official statistical classifications available in RAMON, the Eurostat's metadata server. The code lists used for Air quality data are documented in the Eionet Data Dictionary. They are available in SKOS form, with additional information. For example the AQD - Air Quality Pollutants scheme contains also data on recommended unit or measurement equipment for the pollutant.

3.1.5 Process

Step 1: Data acquisition

Italian and French Air quality data have been extracted from the websites mentioned above. The extracted datasets were uploaded to the SFTP area of the project.

Step2: Data processing

Census data

Italian census data have been transformed according to the requested Data Structure through a script in R language [25]. More in detail, data have been filtered and NUTS3 variable (Third level of NUTS classification) has been added using a dataset with LAUs codes published on Eurostat website. Concerning French census data, an R script allows to obtain the CSV file directly from the data published on Insee's web site. The script uses auxiliary CSV files containing reference data about age groups [26] and French LAU/NUTS which are also described in the CSV metadata [27].

Air quality data

The Data transformation phase was applied only to the dataset related to the PM10 pollutant. The French dataset about the PM10 taken from the European Environmental Agency and uploaded to the SFTP server, in its initial version contains the geographic coordinates; it has been enriched with the Municipality codes through a script in java using the specific service/API. NUTS3 codes have been added to classify both Italian and French territories, while metadata regarding pollutant type, data reference time and aggregation type have been added in the Italian dataset. Data transformation has been implemented in R language. Once input files have been uploaded, data harmonization has been done via SQL union queries wrapped in a single view.

Step3: Conceptual integration

Data integration is realized on a conceptual level through an Ontology Base Data Access (OBDA) architecture. MySQL is used as data repository for Monolith, the tool implementing this approach and used for data mappings. Specifically, Monolith associates mappings with SQL queries on MySQL database, so that SPARQL queries can be rewritten automatically into SQL queries.

Step4: Direct dissemination

The tool Monolith can export the queries based on the ontology concepts in XML format. The SPARQL result set can be formatted in CSV, JSON and RDF and sent to the subsequent stages of the pipeline. RDF triples can be uploaded to INTERSTAT GraphDB for data querying through the client application.

3.1.6 SEP Client application

One of the main goals of this pilot application [28] is to enrich air quality information, produced to support local public authorities responsible for environmental policies. More in detail, several decision makers could get insights from the combination of:

- *Sensor data* measuring the concentration of air pollutants.
- *Statistical data* describing the structure and the main characteristics of the resident population.

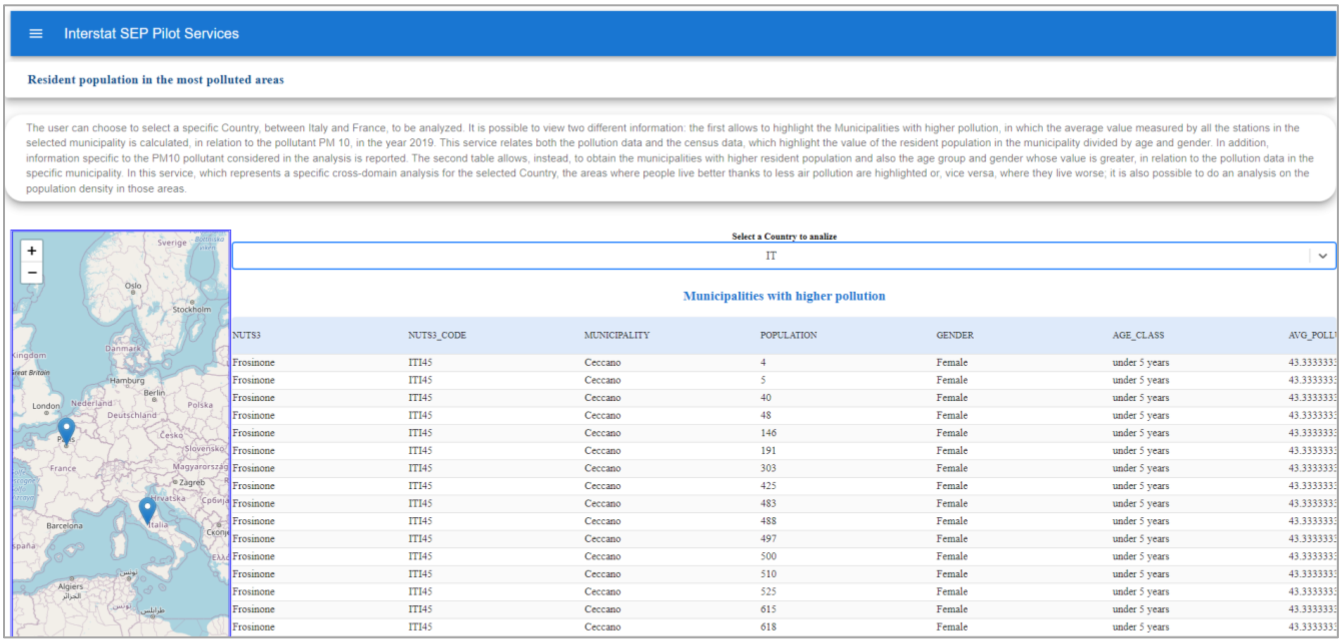
Linking air quality indicators and demographic data could allow decision makers to prioritize target areas of intervention.

Service Name	Description	Type of data visualization	Cross-border service	Cross-domain service
Resident population in the most polluted areas	Through this service the user can obtain the resident population value living in areas where air pollutants exceed the air quality thresholds and also the most populated municipalities, in relation with the air quality data.	The user can select a specific Country, between Italy and France, to be analysed. It is possible to view two different types of information: the first one allows to highlight the Municipalities with higher pollution level, based on the average value measured by all the stations in the selected municipality. This service linking pollution and census data highlights the value of the resident population in the Municipality grouped by age and gender. In addition, specific information related to the PM10 pollutant considered in the analysis, is reported. The second table allows, instead, to obtain the Municipalities with higher resident population and also the age group and gender whose value is greater, in relation to the pollution data in the selected Municipality. In this service, which represents a specific cross-domain analysis for the selected Country, the areas with the highest or the lowest levels of air pollution, are highlighted. It is also possible to do an	X	✓

		analysis on the resident population in those areas.		
Evaluation of the pollution effects (considering PM10) on specific population groups in Italy and France	Through this service it will be possible to evaluate the effects of pollution on specific population groups (for example on more vulnerable population groups) by comparing specific areas of Italy and France (for example Rome and Paris) and visualize specific details about the pollutant values (detection station, unit of measurement of the pollutant, aggregation type and information about the source dataset).	The user can select a specific NUTS region and a Municipality related to it (both Italian and French) and a specific age group (for example from 80 to 85 years). For each of the selected Municipalities, the different pollutant registration stations are highlighted with the detailed information and the value registered by the station. In the tabular view, it is possible to view and compare French and Italian data air pollution data (considering PM10) relating to the resident population value of the selected age group. With this service it is possible to observe, for example, in which municipalities the most vulnerable population groups are most at risk from pollution.	✓	✓
Evaluation of the pollution effects (considering PM2.5 and NO2) on specific population groups in Italy and France	Through this service it is possible to obtain the values of the PM2.5 and NO2 pollutants and specific details (detection station, unit of measurement of the pollutant, aggregation type and information about the source dataset).	The user can select a specific NUTS region and a related Municipality, both Italian and French. Is than possible to highlight the different pollutant registration stations considering PM 2.5 (Particulate matter <2.5 µm) and NO2 (Nitrogen dioxide). In the tabular view, it is possible to obtain and compare French and Italian air pollution data and information about the stations in the selected Municipalities, in relation with the two considered pollutants.	✓	✓

Table 4 - SEP client application: services description

Co-financed by the Connecting Europe Facility of the European Union



Resident population in the most polluted areas

The user can choose to select a specific Country, between Italy and France, to be analyzed. It is possible to view two different information: the first allows to highlight the Municipalities with higher pollution, in which the average value measured by all the stations in the selected municipality is calculated, in relation to the pollutant PM 10, in the year 2019. This service relates both the pollution data and the census data, which highlight the value of the resident population in the municipality divided by age and gender. In addition, information specific to the PM10 pollutant considered in the analysis is reported. The second table allows, instead, to obtain the municipalities with higher resident population and also the age group and gender whose value is greater, in relation to the pollution data in the specific municipality. In this service, which represents a specific cross-domain analysis for the selected Country, the areas where people live better thanks to less air pollution are highlighted or, vice versa, where they live worse; it is also possible to do an analysis on the population density in those areas.

Select a Country to analize

IT

Municipalities with higher pollution

NUTS3	NUTS3_CODE	MUNICIPALITY	POPULATION	GENDER	AGE_CLASS	AVG_POLL
Frosinone	ITI45	Ceccano	4	Female	under 5 years	43.333333
Frosinone	ITI45	Ceccano	5	Female	under 5 years	43.333333
Frosinone	ITI45	Ceccano	40	Female	under 5 years	43.333333
Frosinone	ITI45	Ceccano	48	Female	under 5 years	43.333333
Frosinone	ITI45	Ceccano	146	Female	under 5 years	43.333333
Frosinone	ITI45	Ceccano	191	Female	under 5 years	43.333333
Frosinone	ITI45	Ceccano	303	Female	under 5 years	43.333333
Frosinone	ITI45	Ceccano	425	Female	under 5 years	43.333333
Frosinone	ITI45	Ceccano	483	Female	under 5 years	43.333333
Frosinone	ITI45	Ceccano	488	Female	under 5 years	43.333333
Frosinone	ITI45	Ceccano	497	Female	under 5 years	43.333333
Frosinone	ITI45	Ceccano	500	Female	under 5 years	43.333333
Frosinone	ITI45	Ceccano	510	Female	under 5 years	43.333333
Frosinone	ITI45	Ceccano	525	Female	under 5 years	43.333333
Frosinone	ITI45	Ceccano	615	Female	under 5 years	43.333333
Frosinone	ITI45	Ceccano	618	Female	under 5 years	43.333333

Figure 5 SEP client application: Service 1

3.2 Geolocalized facilities (GF)

3.2.1 Business case

The main objective of this pilot is to disseminate information about facilities or equipment so that it can be contextualized in space and integrated with other sources of data.

Two specific user stories are defined for the GF pilot:

- In the “visitor” case, we consider a user visiting a place she does not know and wondering where the nearest facilities of different types are located. He also would like to know what events are programmed in the nearby cultural facilities. From the description of locations or events, it should be simple to navigate on the web for further detail (e.g., history of places, links to the locations’ web sites, etc.).

- The “local decider” story is about a person in charge of an investment decision at a local level. It can be the manager of a bus company wondering if he should replace an old vehicle, an employee of an educational public service assessing the creation of a new class in a community school, or a young couple thinking of moving to a rural place, etc. He needs information about the level and capacity of the equipment in the neighborhood, linked with data on the demographic evolution at a fine level. He will probably need to combine that information with other sources more specifically relevant to his specific problem.

Co-financed by the Connecting Europe Facility of the European Union

3.2.2 Models

The target model for the data on facilities is expressed in OWL (see also WebVOWL [29] visualization). The overall structure of the ontology is represented in the following figure:

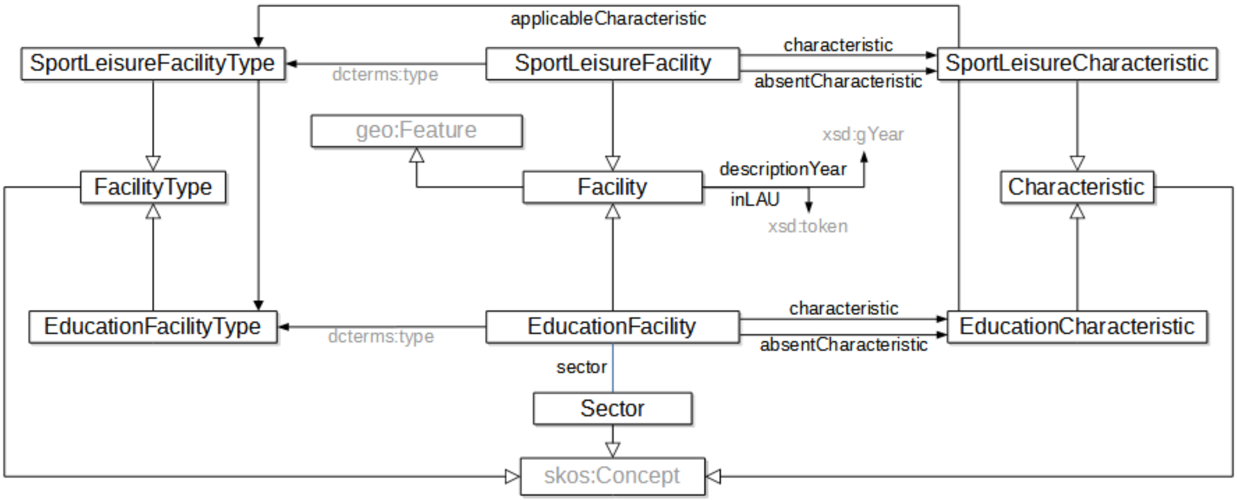


Figure 6 Target data model of GF pilot

The facility coordinates are represented using the GeoSPARQL ontology [14]. In the BPE, the quality of the geocoding is documented according to a 3-star-like system. This is rendered in RDF using quality annotations defined in the DQV vocabulary [30]. The articulation of these different elements is shown in the following figure.

Co-financed by the Connecting Europe Facility of the European Union

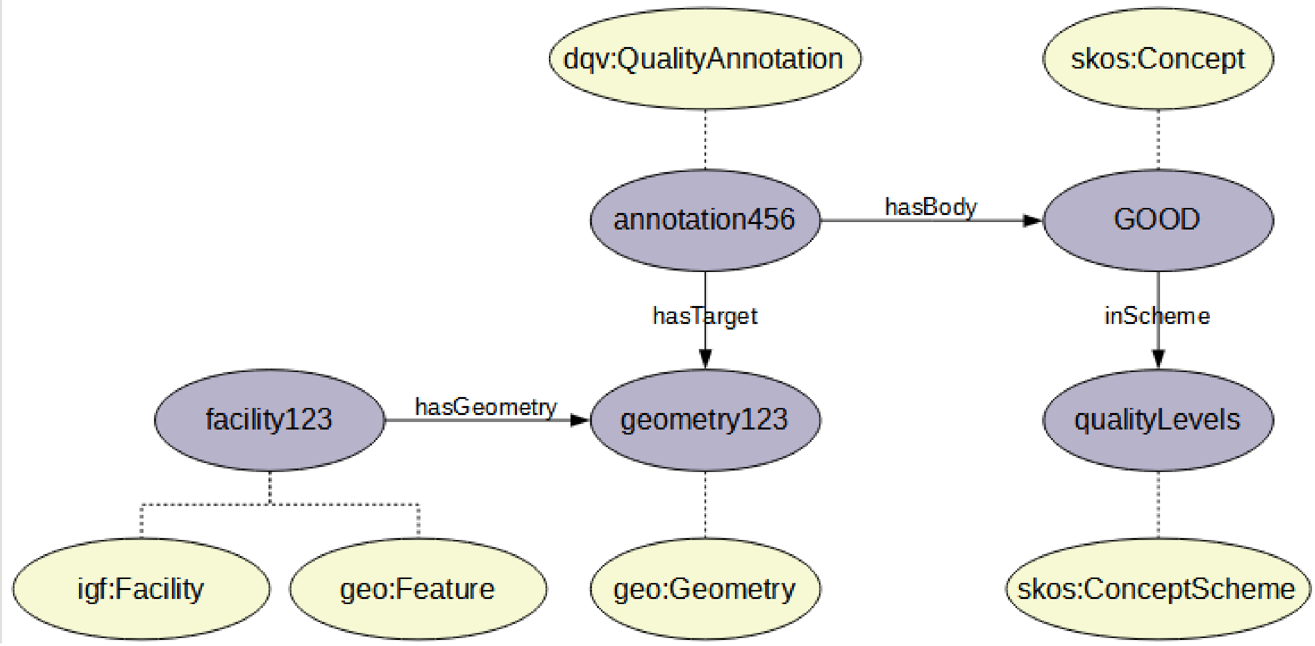


Figure 7: Representation of BPE facility coordinates

3.2.3 Data

French data

The central source of French data for this pilot is the Permanent database of facilities [31] (BPE in French) published by Insee. For a working example, we extract a list of columns from the CSV file containing the data from the 2020 edition of the database.

BPE data and metadata are available in CSV formats from the BPE [32]. More specifically, the example uses an extract of the following geocoded facilities:

- Dataset 1: Exposition venues and heritage, the file is available [33].
- Dataset 2: Education, file available [34].

The extraction is performed directly from the online CSV files by a Python script [35].

The list of columns extracted is given in Table 5.

Co-financed by the Connecting Europe Facility of the European Union

Field name	Description	Data type	Data availability
Facility_ID	Facility identifier	String	Datasets 1 & 2
Year	Reference year	Year (always '2020')	Datasets 1 & 2
LAU	Municipality	Code list	Datasets 1 & 2
Coord_X	Latitude	Float	Datasets 1 & 2
Coord_Y	Longitude	Float	Datasets 1 & 2
Quality_XY	Quality of geocoding	Code list	Datasets 1 & 2
Facility_Type	Type of facility	Code list	Datasets 1 & 2
CL_PELEM	Presence or absence of a pre-elementary class in primary schools	Code list	Dataset 2
CL_PGE	Presence or absence of a preparatory class for the high schools in upper secondary	Code list	Dataset 2
EP	Membership or not in a priority education scheme	Code list	Dataset 2
Sector	Membership of the public or private education sector	Code list	Dataset 2

Table 5: Fields extracted from BPE

French geographic coordinates are expressed using the Lambert 93 coordinate system.

Italian data

Museums

The data on Italian museums is extracted from the MiBACT web site [36] published by the Ministero della Cultura, and more precisely from the SPARQL endpoint [37]. The columns extracted are:

Field name	Description	Data type	Property path
subject	Museum	URI	(a cis:CulturalInstituteOrSite)
Nome_Istituzionale	Institutional name	String	cis:institutionalCISName
Descrizione	Description	String	lo:description
Latitudine	Latitude	String	geo:lat
Longitudine	Longitude	String	geo:long
Disciplina	Discipline	String	cis:hasDiscipline/l0:name
Indirizzo	Address	String	cis:hasSite/cis:siteAddress/clvapit:fullAddress
Codice_postale	Postal code	String	cis:hasSite/cis:siteAddress/clvapit:postCode
Comune	Municipality name	String	cis:hasSite/cis:siteAddress/clvapit:hasCity/rdfs:label
Provincia	Province name	String	cis:hasSite/cis:siteAddress/clvapit:hasProvince/rdfs:label
WebSite	Web site	String	smapit:hasOnlineContactPoint/smapit:hasWebSite/smapit:URL

Table 6: Fields extracted from the Italian Ministry of Culture - Museums

The "Property path" column refers to the RDF property or path of properties giving the field value in reference to the Cultural-ON ontology and the vocabularies it relies on. There are some adjustments made, for example institutionalCISName instead of institutionalName. An example of RDF data provides additional detail, along with the namespaces associated to the prefixes used in the table.

Events

The data on Italian cultural events is also extracted by a query on the MiBACT SPARQL endpoint [38]. The columns extracted are:

Field name	Description	Data type
EVENTO	Event identifier	URI
NOME	Name of the event	String
DATA_INIZIO_EVENTO	Starting date of the event	Datetime
DATA_FINE_EVENTO	Ending date of the event	Datetime
CATEGORIA	Type of event	Code list
SITO_WEB	Web site	HTTP URI or domain name
EMAIL	Email address	Mailto URI
VIA	Street name	String
NUMERO_CIVICO	Number in the street	String
CAP	Postal code	String (5 digits)
COMUNE	Municipality name	String
PROVINCIA	Province name	String
REGIONE	Region name	String

Table 7: Fields extracted from the Italian Ministry of Culture - Events

There again, the source RDF data is conformant to the Cultura-Ontology ontology [39], and thus is actually more structured and expressive than the flat transformed format described in the table above. In consequence, it could be interesting for the client application to test the option of distributed SPARQL queries on both the INTERSTAT and "Dati cultura" endpoints.

Schools

Data concerning Italian schools are described in the S4Y Pilot (3.3.3).

3.2.4 Metadata

Apart from the ontology describing the data, metadata about the data is available in different forms:

Co-financed by the Connecting Europe Facility of the European Union

- The CSV data extracted from the BPE is described using the CSV on the web [40] (CSVW) vocabulary. CSV is a notoriously sloppy standard, and CSVW is a powerful way to describe tabular data available online so that they can be understood easily by humans and machines, thus dramatically improving its usability. A CSV on the web description of the CSV distribution of GF data is produced semi-automatically by the ETL pipeline [41].

- The Cross-Domain Integration model is a development of the DDI Alliance aiming at improving coherence and interoperability of metadata [42]. In particular, DDI-CDI allows the description of a wide range of data structures. In DDI-CDI terms, the BPE data corresponds to a "wide" data structure. A tentative DDI-CDI description [43] of the BPE file is provided with the GF data.

- Finally, descriptive metadata using the DCAT standard are provided for the source data [44]

3.2.5 Process

The ETL process of the Geolocalized Facilities pilot is described in detail and is available [45]. The process is organized according to the usual steps:

- Extraction is performed on data which are all available online, in various formats: CSV for French facilities and Italian schools, and RDF for Italian museums (MIBACT data available via SPARQL). Note that the latter also contains information about cultural events for Italy: those are not extracted, but they might be queried directly from the client application.

- The main transformation steps are made on French metadata in order to transform them into CSV on the Web. Regarding data, the main steps are conversion of the coordinates from Lambert 93 to WGS 84 for the French facilities. For Italian schools, addresses are geocoded using the Nominatim API provided by OpenStreetMap (with application of the usage policy). Both sources are then converted to RDF and merged.

- CSV files are finally uploaded to the INTERSTAT SFTP server and the RDF/Turtle files to the GraphDB triple store. Note that uploading the French facilities to SFTP is not useful for the pipeline itself, but it gives the possibility to describe two different distributions in the DCAT metadata.

It should be noted that the data part of the process is fully automated and reproducible at will. Regarding the metadata, the part concerning structural metadata (specifically code lists) is also automated, but some aspects of the production of descriptive metadata still require manual intervention.

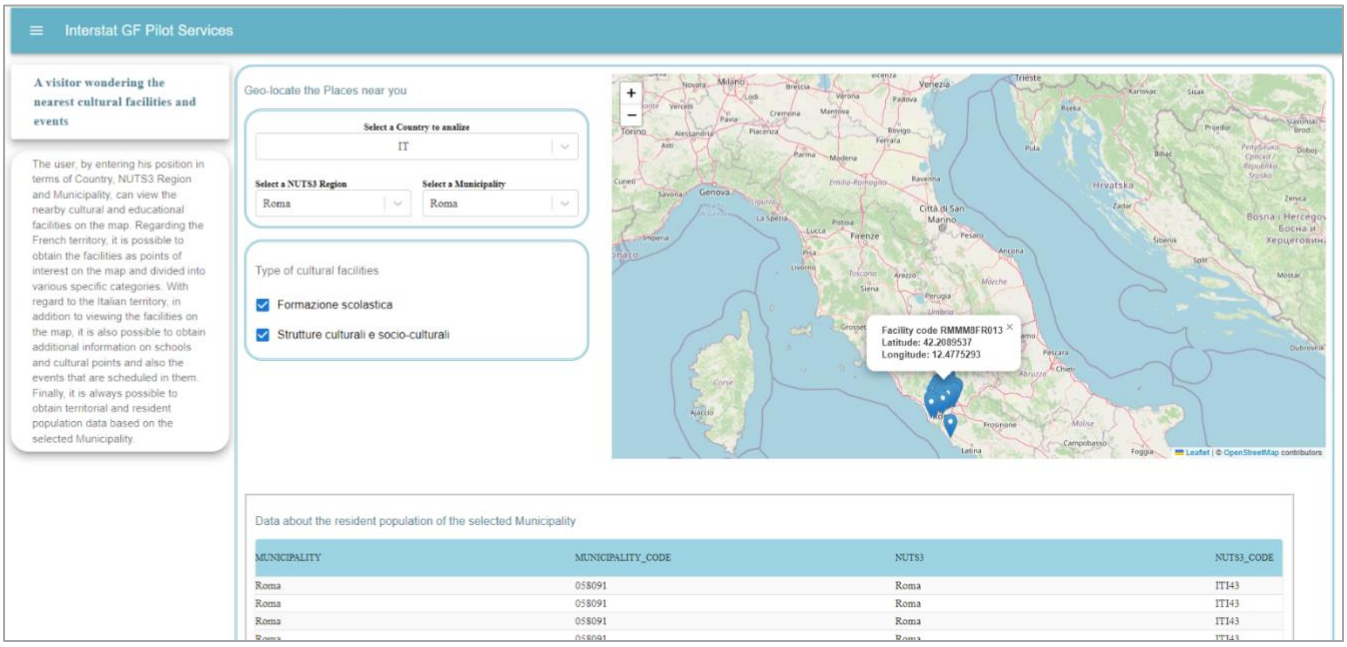
3.2.6 GF client application

The main objective of this pilot application [46] is to set up a mechanism for the dissemination and use of information about facilities or equipment, so that the information is contextualized in space and can be integrated with other sources of data. The *facilities* are understood as points of services which are accessible to the public and operate in domains like education, health, social services, transport, sports, leisure, culture, or tourism. The following table contains the service developed for this client application. This service satisfies the user stories defined in the paragraph 3.2.1 of the document.

Service Name	Description	Type of data visualization	Cross-border service	Cross-domain service
A citizen wondering the nearest cultural facilities and events	The user is visiting an unknown place, Italian or French, and needs to know the nearest facilities what events are programmed within them.	The user, by entering his position in terms of Country, NUTS3 Region and Municipality, can view the nearby cultural and educational facilities on the map. Regarding the French territory, it is possible to obtain the facilities as points of interest on the map and divided into various specific categories. With regard to the Italian territory, in addition to viewing the facilities on the map, it is also possible to obtain additional information on schools and cultural points and also the events that are scheduled in them. Finally, it is always possible to obtain territorial and resident population data based on the selected Municipality.	✓	✓

Table 8 - GF client application: services description

Co-financed by the Connecting Europe Facility of the European Union



A visitor wondering the nearest cultural facilities and events

The user, by entering his position in terms of Country, NUTS3 Region and Municipality, can view the nearby cultural and educational facilities on the map. Regarding the French territory, it is possible to obtain the facilities as points of interest on the map and divided into various specific categories. With regard to the Italian territory, in addition to viewing the facilities on the map, it is also possible to obtain additional information on schools and cultural points and also the events that are scheduled in them. Finally, it is always possible to obtain territorial and resident population data based on the selected Municipality.

Geo-locate the Places near you

Select a Country to analize: IT

Select a NUTS3 Region: Roma

Select a Municipality: Roma

Type of cultural facilities

☑ Formazione scolastica

☑ Strutture culturali e socio-culturali

Facility code RMMM8FR013
Latitude: 42.2089537
Longitude: 12.4775293

Data about the resident population of the selected Municipality

MUNICIPALITY	MUNICIPALITY_CODE	NUTS3	NUTS3_CODE
Roma	058091	Roma	ITI43
Roma	058091	Roma	ITI43
Roma	058091	Roma	ITI43
Roma	058091	Roma	ITI43

Figure 8 GF client application: Service 1

3.3 The School for You (S4Y)

3.3.1 Business case

One of the main objectives of this use case is to respond to the need of citizens and political decision-makers to know the distribution of students or potential students in the territory and the services addressed to them, especially the distribution of schools by educational level and for structural characteristics. Parents who must choose the school for their children need to know not only the location, but also the educational services that a specific area offers. Finally, the comparison with the main aspects of the resident population could provide useful information to ensure adequate investments in educational services.

3.3.2 Models

The domain concepts about school attendance and school units are reported in the following ontology, exported as owl and as image.

Co-financed by the Connecting Europe Facility of the European Union

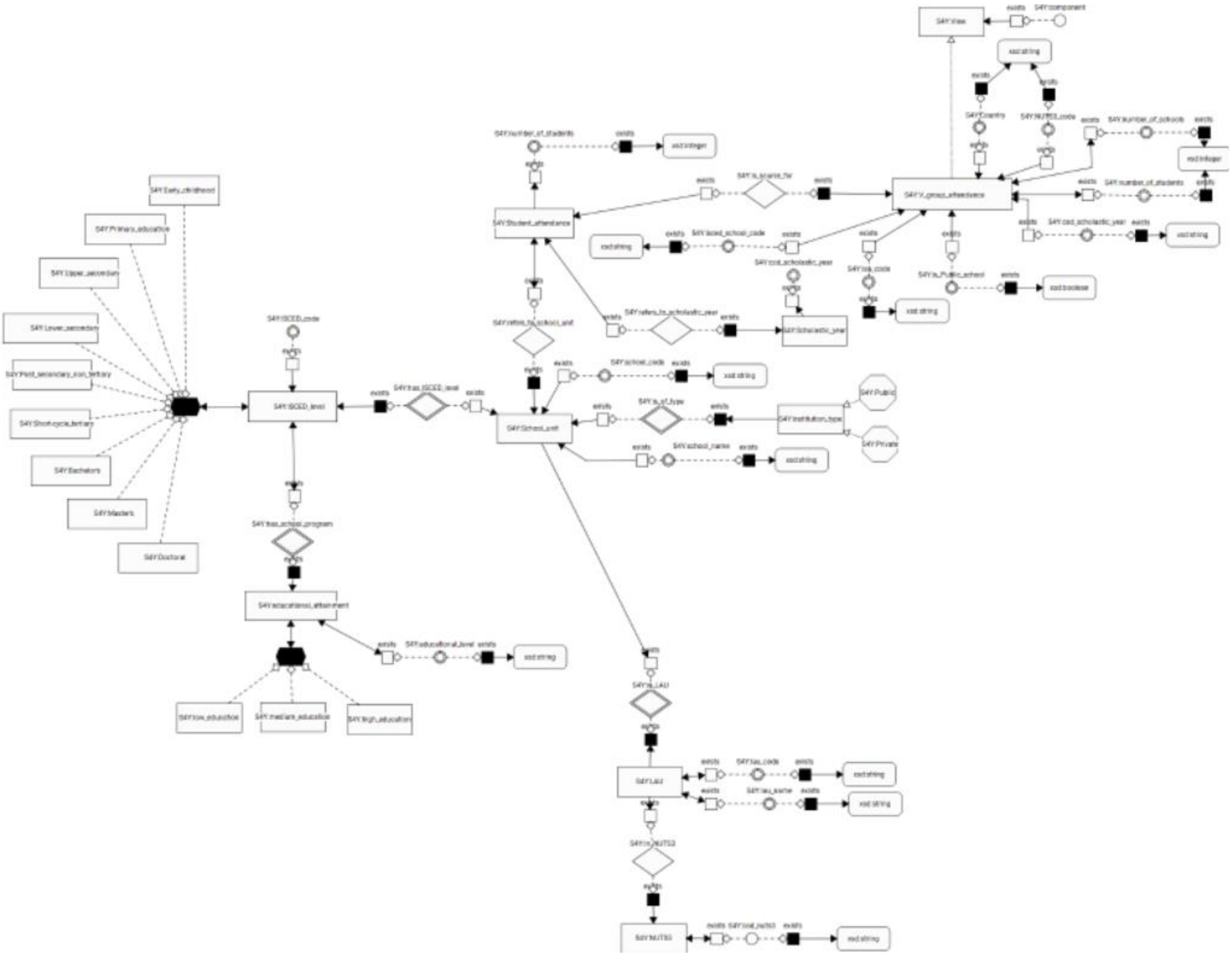


Figure 9 - S4Y domain ontology

This ontology contains the class *V_group_attendance* as istance of metaclass View. An ontological view on the domain allows to aggregate data about the domain for the following measures: number of students, number of schools and by the following dimensions: ISCED school level, Scholastic year, Public school, Municipality (LAU). In the first version of the pilot, only a subset of the concepts have been mapped over the data.

3.3.3 Data

The following paragraph provides also an overview of Italian and French data sources linked to test cross-border and cross-domain interoperability. Data extracted for the pilot focus mainly on students' enrollment and school location.

Italian School data

Co-financed by the Connecting Europe Facility of the European Union

The Italian data used for the pilot belong to the MIUR (The Ministry for Education) catalogue [47]. The datasets, available also in RDF format and concerning both public and private schools, provide the following information:

- Registry information on schools [48] (Informazioni anagrafiche scuole statali, Informazioni anagrafiche scuole paritarie);
- Students by course year and age group [49] (Studenti per anno di corso e fascia di eta'. Scuola statale, Studenti per anno di corso e fascia di età. Scuola paritaria); The reference dataset for the list and location of schools of all levels is available [50] and is also considered for the GF pilot. The coordinates of the schools have been added to this dataset and the output file is present in the specific SFTP area of the project [51].

The following table reports the subset of fields extracted from the Registry information on schools.

Field name	Description	Data type
ANNOSCOLASTICO	Scholastic year	Numeric
CODICESCUOLA	School identifier	Text
DENOMINAZIONESCUOLA	School name	Text
INDIRIZZOSCUOLA	School address	Text
CODICECOMUNESCUOLA	School cadastral code	Text
DESCRIZIONETIPOLOGIAGRADOISTRUZIONESCUOLA	Level of educational attainment	Text

Table 9 - Description of the fields extracted from the Registry information on schools

For the available scholastic years, the number of students has been extracted from the following dataset:

- Students by course year and age group.

The variables extracted for data integration are reported in the table below.

Field name	Description	Data type
ANN_SCO_RIF	Scholastic year	Numeric
COD_PLE_UTE	School identifier	Text
QTY_NUM_ALU	Number of students	Text

Table 10 - Description of the variables extracted for data integration

French School data

The following fields have been extracted from the school registry dataset [52] (called "Address and geolocalization of primary and secondary educational institutions"). Data extraction is performed using the API to retrieve a CSV file.

The number of students for the three last scholastic year (2019, 2020 and 2021) has been extracted from four datasets:

- Student enrollment by grade and number of classes per school [53] (only primary and nursery school).
- Number of students by grade, gender, most frequent modern languages 1 and 2, by lower secondary school [54].
- Number of students by grade, gender, most frequent modern languages 1 and 2, by upper secondary school [55] (school of general and technological education).
- Number of students by grade, gender, most frequent modern languages 1 and 2, by upper secondary school [56] (vocational school).

Data extraction is performed based on CSV files directly available online. Variables finally extracted are:

Field name	Description	Data type	Comment
school_id	Unique school identifier	String	variable "numero_uai" in the dataset n°1
name	Official name of the school	String	variable "appellation_officielle in the dataset n°1
latitude	Latitude of the school (WGS84)	Float	variable "latitude" in the dataset n°1
longitude	Longitude of the school (WGS84)	Float	variable "longitude" in the dataset n°1
lau	Lau code	Code list	variable "code_commune" in the dataset n°1
institution_type	Type of institution	Code list (Coded as "PR" for "Private" and "PU" for "Public")	variable "secteur_public_prive_libe" in the dataset n°1
code_nature	Nature code	Code list	variable "nature_uai" in the dataset n°1
ISCED_level	Level of ISCED classification	Code list	coded based on code_nature (see mapping file here)
scholastic_year	Scholastic year	Year	Datasets n°2 to n°5

Co-financed by the Connecting Europe Facility of the European Union

students_number	Number of students (measure) per scholastic year	Integer	Datasets n°2 to n°5

Table 11 - Description of the variables extracted for French school data

3.3.4 Metadata

ISCED (International Standard Classification of Education) refers to the national (and sub-national) education programme and the related recognised educational qualification.

The ISCED mappings for different countries [57] and the ISCED mappings for French are available [58].

Educational attainment level: low education, medium education, high education **ISCED attainment level** *Low education*

- *ISCED 0*: Early childhood education ('less than primary' for educational attainment)
- *ISCED 1*: Primary education
- *ISCED 2*: Lower secondary education *Medium education*
- *ISCED 3*: Upper secondary education
- *ISCED 4*: Post-secondary non-tertiary educ *High education*
- *ISCED 5*: Short-cycle tertiary education
- *ISCED 6*: Bachelor's or equivalent level
- *ISCED 7*: Master's or equivalent level
- *ISCED 8*: Doctoral or equivalent level

Institution type: { public = 1, private = 0 }

Metadata and harmonization outline are available in the specific Excel file [59].

3.3.5 Process

The main steps of the data pipeline are:

Step 1: Data acquisition

Italian and French data have been extracted from the websites mentioned above. The extracted datasets were uploaded to the SFTP area of the project. At the moment the datasets referring to the year 2019 have been loaded, but all the school years could be considered and added in order to allow the user to make an analysis of the data over time. The ETL process producing French data file is organized according to the usual steps: Extraction is performed on data which are all available online: CSV retrieved using API for school registry and CSV directly downloaded for four other datasets.

The main transformation steps are:

- recoding institution type and coding ISCED level based on a mapping file
- merging school registry dataset and students number datasets
- producing CSV file CSV data file is finally uploaded to the INTERSTAT SFTP server [60].

Step2: Data processing

Some variables have been created or extracted by ad-hoc services/API. More in detail:

- Italian coordinates have been processed through a geo converter service from Address and LAU or ZIP-CODE
- Italian School Year has been converted to a standard format
- Italian information about the school being public or private has been added
- French School Year has been extracted from other sources.

Step3: Conceptual and logical integration

On a conceptual level, Italian and French data have been integrated through ontology concepts. Starting from the conceptual integration achieved through ontology, the main goal of the common logical model is to harmonize cross domain and cross border data sources. The common logical model has the following structure:

Table	Field	Is part of key	Description
V_school_unit	school_id	True	Code to identify one school unit. The code is unique on different countries
	school_name		
	In_lau		Municipality (LAU code) of school The code is unique on different countries
	Institution_type		For public schools is 1, otherwise 0
	ISCED_school_code		ISCED code of prevalent Education

Co-financed by the Connecting Europe Facility of the European Union

			level provided by the school Prevalent because a school could provide many education levels
V_students_attendance	school_id	True	Identifier of School unit. Attendances of students for one School unit and for one scholastic year
	scholastic_year	True	Identifier of scholastic year (first of year couple)
	number_of_students		Number of students attending

Table 12 - Description of the structure of the common logical model

Data from the original data sources have been transformed and harmonized according to the common logical model. The diagram below describes how data have been transformed and loaded in the following objects: **V_school_unit**, **V_students_attendance** and **V_group_attendance**, which is the information object resulting from the conceptual integration.

Co-financed by the Connecting Europe Facility of the European Union

The contents of this publication are the sole responsibility of INTERSTAT consortium and do not necessarily reflect the opinion of the European Union

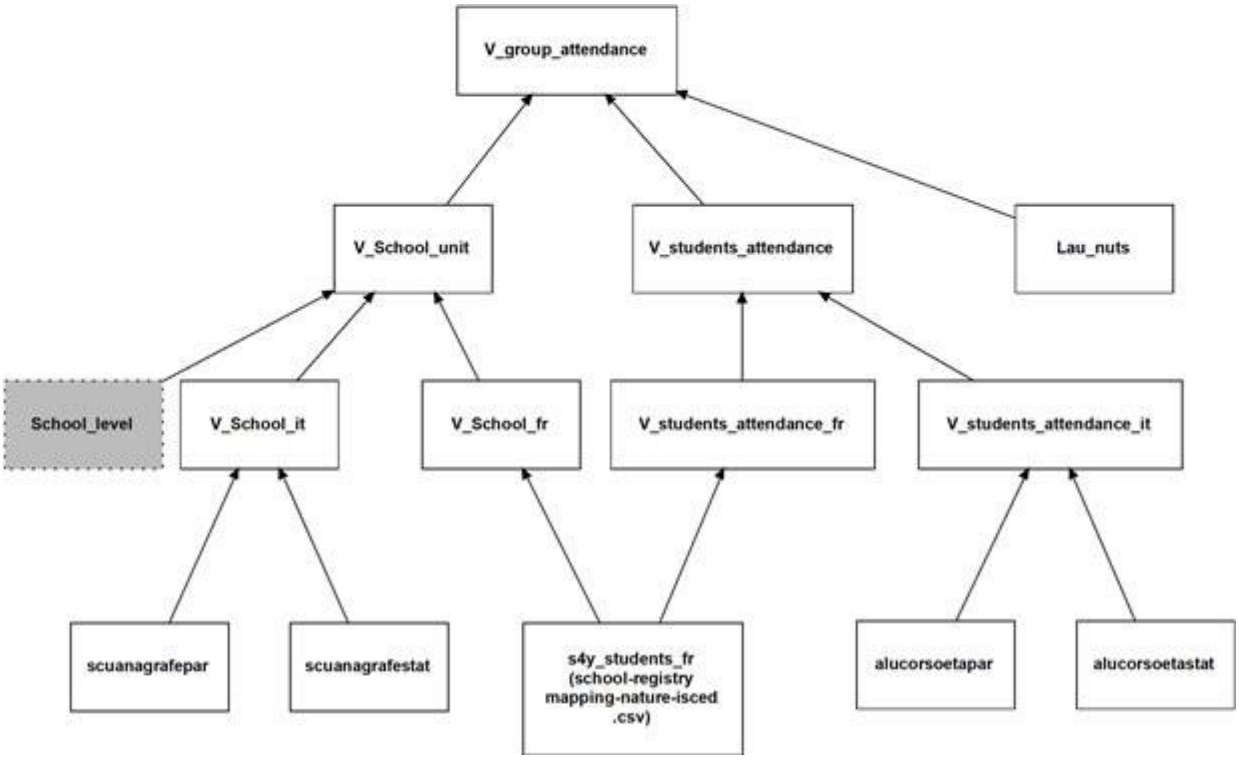


Figure 10 - Hierarchical relationship between tables and data views

Step4: Direct dissemination

See the description of SEP data pipeline.

3.3.6 S4Y client application

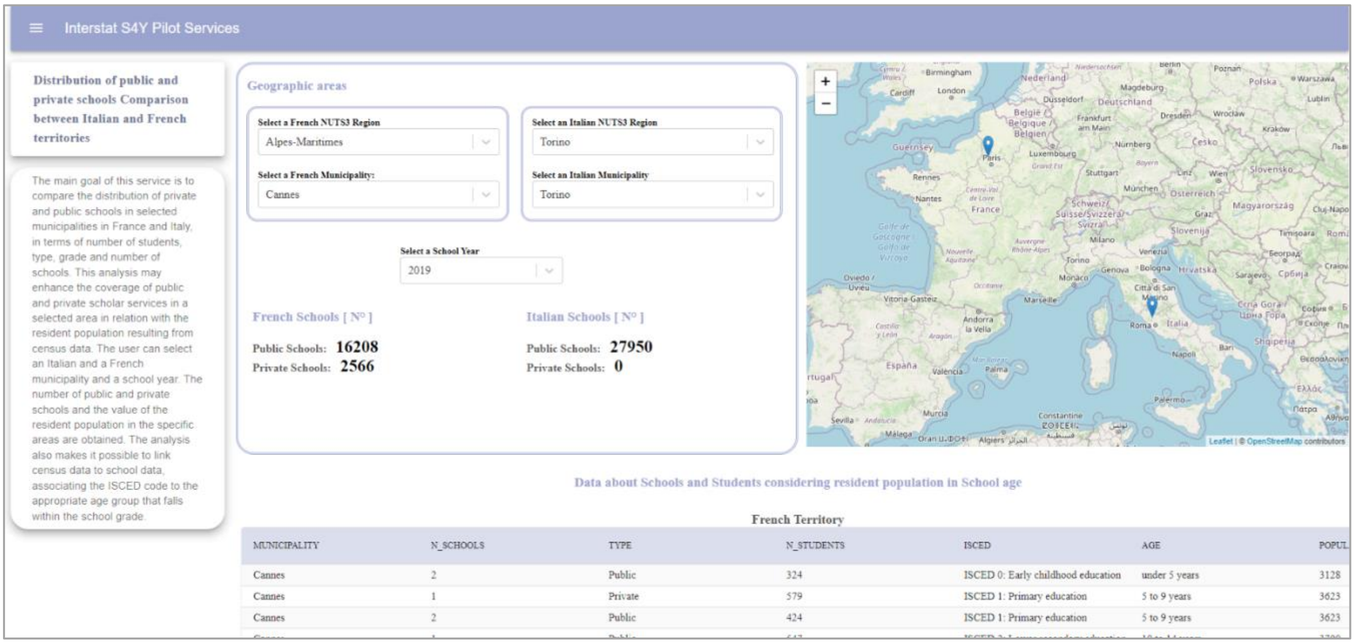
This pilot [61] allows users such as citizens and political decision-makers to discover aggregated data resulting from the integration of several sources about school attendance and the distribution of students in Italy and France.

Service Name	Description	Type of data visualization	Cross-border service	Cross-domain service
Analysis of resident population and schools in a specific municipality	This service allows to explore the number of schools and students in a selected area, according to the main level of educational attainment. This analysis may enhance the coverage of scholar	This service allows to explore the number of schools and students in a selected area, according to the main level of educational attainment. This analysis may enhance the coverage of scholar services in a selected municipality and the relationship between school attendance and the	X	✓

	services in a selected municipality and the relationship between school attendance and the population structure resulting from census data. This Analysis may support policy makers to improve access to education trainings and facilities in specific areas.	population structure resulting from census data. The result of data integration may support policy makers to improve access to education trainings and facilities in specific areas. The user can select a municipality, an Educational Level (ISCED) and a School Year. The number of schools, the number of students and the resident population (divided by age group) are obtained, in tabular form. Furthermore, the total number of schools present in the selected Municipality is displayed, considering schools of all types and grades.		
Distribution of public and private schools: comparison between Italian and French territories	This analysis may enhance the coverage of public and private scholar services in a selected area in relation also with the population resident resulting from census data. The aim is also to provide the comparison between public and private schools in those specific areas in terms of number of schools.	The main goal of this service is to compare the distribution of private and public schools in selected municipalities in France and Italy, in terms of number of students, type, grade and number of schools. This analysis may enhance the coverage of public and private scholar services in a selected area in relation with the resident population resulting from census data. The user can select an Italian and a French municipality and a school year. The number of public and private schools and the value of the resident population in the specific areas are obtained. The analysis also makes it possible to link census data to school data, associating the ISCED code to the appropriate age group that falls within the school grade.	✓	✓

Table 13 - S4Y client application: services description

Co-financed by the Connecting Europe Facility of the European Union



Interstat S4Y Pilot Services
+
−
Leaflet | © OpenStreetMap contributors

Distribution of public and private schools Comparison between Italian and French territories

The main goal of this service is to compare the distribution of private and public schools in selected municipalities in France and Italy, in terms of number of students, type, grade and number of schools. This analysis may enhance the coverage of public and private scholar services in a selected area in relation with the resident population resulting from census data. The user can select an Italian and a French municipality and a school year. The number of public and private schools and the value of the resident population in the specific areas are obtained. The analysis also makes it possible to link census data to school data, associating the ISCED code to the appropriate age group that falls within the school grade.

Geographic areas

Select a French NUTS3 Region: Alpes-Maritimes

Select a French Municipality: Cannes

Select an Italian NUTS3 Region: Torino

Select an Italian Municipality: Torino

Select a School Year: 2019

French Schools [N°]

Public Schools: 16208
Private Schools: 2566

Italian Schools [N°]

Public Schools: 27950
Private Schools: 0

Data about Schools and Students considering resident population in School age

French Territory

MUNICIPALITY	N_SCHOOLS	TYPE	N_STUDENTS	ISCED	AGE	POPUL
Cannes	2	Public	324	ISCED 0: Early childhood education	under 5 years	3128
Cannes	1	Private	579	ISCED 1: Primary education	5 to 9 years	3623
Cannes	2	Public	424	ISCED 1: Primary education	5 to 9 years	3623

Figure 11 S4Y client application: Service 2

4 SDMX to ETSI NGSI-LD

FIWARE Context Broker is designed to provide Context Information Management based on different timestamp. It means that different measurements of an entity (e.g., a sensor) are differentiated by the different timestamp of those measurements. In case of statistical information, we have a complete dataset, in which we can have or not a timestamp associated to the acquisition of the data.

It creates a challenge in the design of a solution to cover this statistical information, with only three possible implementations and the consequences of architectural design of the solutions:

1. Put all the data into a simple entity (e.g., Dataset). It resolves the problem of the timestamp of the data but on the other size it is only limited to small-medium size datasets due to the limitation of 2MB data into an Entity. Besides, the ETSI NGSI-LD queries over the data are not possible.

2. We deal each of the row of a data set in separate entities all of them with a pseudo manage timestamp to guarantee that all of them are created at the same time. It facilitates the management of the data due to it is possible that the Entities' size is less than 2MB. On the other side, the searching process around the data is very complex or almost impossible with a medium or large sized datasets to realize using the ETSI NGSI-LD API.

3. We manage the metadata information of the dataset into a specific data model and we make reference to an external source in which we can find the complete information. This approach resolves the issue of the Entity Size as well as delegate the management of the data to the corresponding tools.

The third option is the selected one to be implemented due to facilitate the metadata sharing and therefore, allowing finding similar datasets and eventually connect them. Additionally, the publication of the metadata involves the data value creation around the statistical data in which future services can be developed using the information about how to access the data.

Once that we selected the best approach for our solution, we need to develop the corresponding data models to represent this metadata information in JSON-LD format, in order to be used by the FIWARE Context Broker. In this point, we will adopt the DCAT-AP specification and the specific

Co-financed by the Connecting Europe Facility of the European Union

application profile developed based on it, StatDCAT-AP, to define the metadata information of any statistical dataset.

Next sections describe the data models created within the Smart Data Models initiative [62] both in DCAT-AP and StatDCAT-AP as well as the grammar that we have developed in order to parse any RDF Turtle file to translate into JSON-LD compatible with ETSI NGSI-LD API.

4.1 DCAT-AP data models

This is the basic for the corresponding translation of the metadata information into the JSON-LD that will be controlled by the FIWARE Context Broker (in this case Orion-LD). Metadata are descriptors of the data sources (or services), including for example Last Update, Creator, Creation date, or Size of the dataset. Besides, sharing metadata allows finding similar datasets and, eventually, connect them.

We adopt Data Catalog Vocabulaty (DCAT) Application Profile (AP) as the source of our data model definition because it is used at the EU open data portal with around 1.2M datasets. DCAT-AP was created at Joinup Initiative at EU commission as a result of an agreement on a common format for data exchange to support the sharing of public sector information with the intention to discover them and re-use of this data for developing services. The DCAT Application Profile data models, defined in this document, are mapped from version 2.0.1 of the DCAT-AP standard [63]. It is based on the specification of the Data Catalog Vocabulary (DCAT) of 16 January 2014 [64], and the Data Catalog Vocabulary (DCAT) - Version 2, W3C Recommendation 04 February 2020 [65].

The list of data models created can be found in dataModel.DCAT-AP repository [66] inside the Smart data Models organization in GitHub and involves the definition of the corresponding Entity Types. Entity Types (Classes) are divided into mandatory, recommended, and optional. Not all the DCAT-AP classes have been mapped, only those that are relevant for the ongoing requirements:

- **Dataset** [67]. A conceptual entity that represents the information published. This is a mandatory class. Dataset Schema meeting DCAT-AP 2.0 specification.
- **CatalogueDCAT-AP** [68]. A catalogue or repository that hosts the Datasets being described. This is a mandatory class. Catalogue of datasets compliant with DCAT-AP specification.
- **AgentDCAT-AP** [69]. An entity that is associated with Catalogues and/or Datasets. This is a mandatory class. Agent Schema meeting DCAT-AP 2.0 specification.

Co-financed by the Connecting Europe Facility of the European Union

- **DistributionDCAT-AP** [70]. A physical embodiment of the Dataset in a particular format. This is a recommended class. This is a distribution belonging to a dataset according to the DCAT-AP standard 2.0.1.
- **CatalogueRecordDCAT-AP** [71]. description of a Dataset's entry in the Catalogue. This is an optional class. This is a Catalogue Record belonging to a dataset according to the DCAT-AP standard 2.0.1.
- **DataServiceDCAT-AP** [72]. A collection of operations that provides access to one or more datasets or data processing functions. This is an optional class. Data Service is adapted from DCAT-AP 2.0 specification, but extended with additional properties and compatible with ETSI NGSI-LD standard.

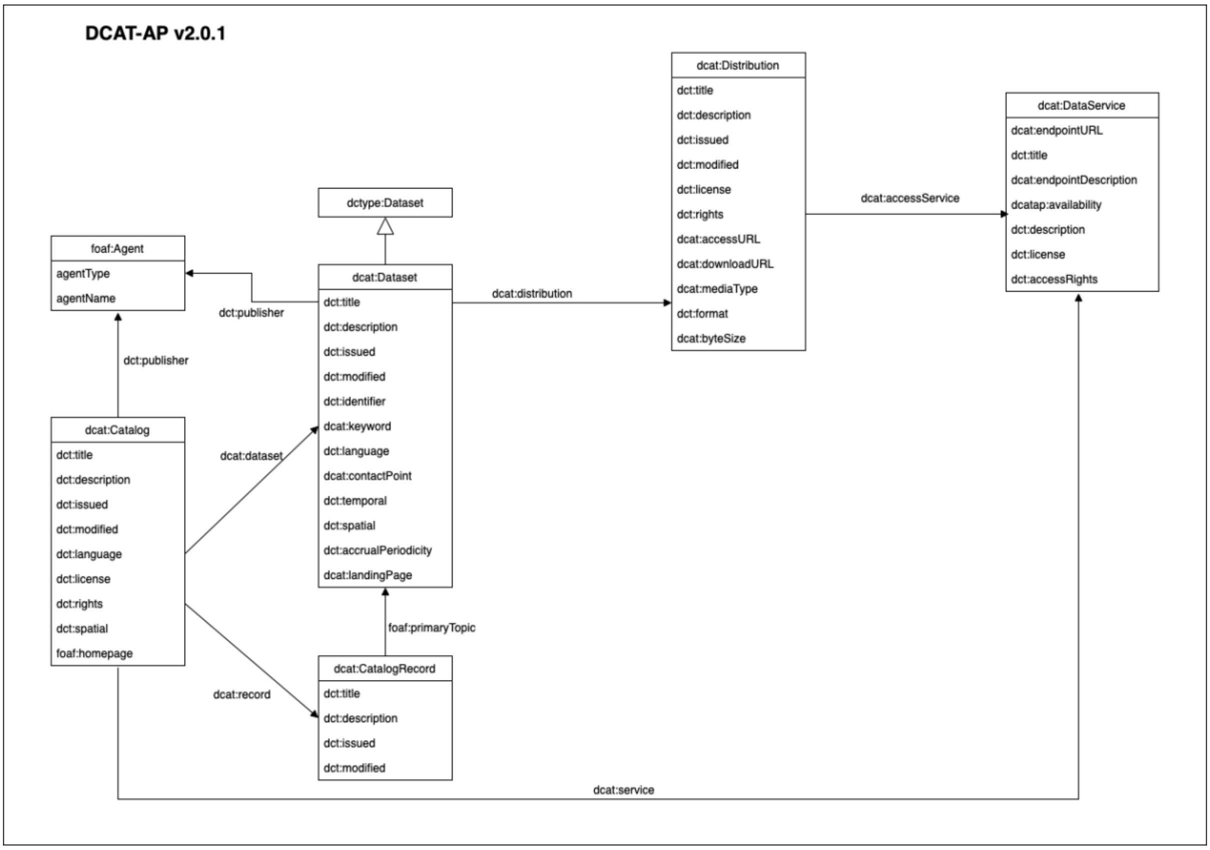


Figure 12 DCAT-AP schematic data model

The reason of using "*DCAT-AP*" in the name of some Classes or Entity Types is because we need to resolve the ambiguity with other Entities defined in the Smart Data Models initiative (e.g., DatasetDCAT-AP and DatasetSTAT-DCAT-AP).

Nevertheless, there are a set of limitations in this specification that we need to address like the impossibility to describe the structure of a dataset as well as the normalized license used on them. For this purpose, we decide to move forward and adopt the StatDCAT Application Profile to provide this information in our statistical data models.

4.2 StatDCAT-AP data models

DCAT Application Profile for description of statistical datasets (StatDCAT-AP) is the extension of the DCAT Application Profile, version 1.1, for data portals in Europe for describing statistical datasets, dataset series and services, with a common agreed vocabulary for statistical open data. Basically, StatDCAT-AP model defines a set of additions to the DCAT-AP model to describe datasets in any format, for example in statistical data and metadata Exchange (SDMX). The idea behind was to know which elements in statistical data standards can be manage by DCAT-AP and extends it in order to help in the discovery and use of statistical datasets.

The StatDCAT-AP data model includes four classes that are already available in DCAT-AP (Catalogue, Catalogue Record, Dataset, and Distribution). Basically, the Smart Data Models initiative has worked in the extension of the corresponding **Dataset** data model [73] in order to include the following properties:

- **stat:attribute**: This property links to a component used to qualify and interpret observed values, e.g., units of measure, any scaling factors, and metadata such as the status of the observation (e.g., estimated, provisional). Attribute is a conceptual entity that applies to all distribution formats, e.g., in case a dataset is provided both in SDMX and in Data Cube.
- **stat:dimension**: This property links to a component that identifies observations, e.g., the time to which the observation applies, or a geographic region which the observation covers. Dimension is a conceptual entity that applies to all distribution formats, e.g., in case a dataset is provided both in SDMX and in Data Cube.
- **stat:statUnitMeasure**: This property links to a unit of measurement of the observations, for example Euro, square kilometre, purchasing power standard (PPS), full- time equivalent, percentage. Unit of measurement is a conceptual entity that applies to all distribution formats, e.g., in the case when a dataset is provided both in SDMX and in Data Cube.

Co-financed by the Connecting Europe Facility of the European Union

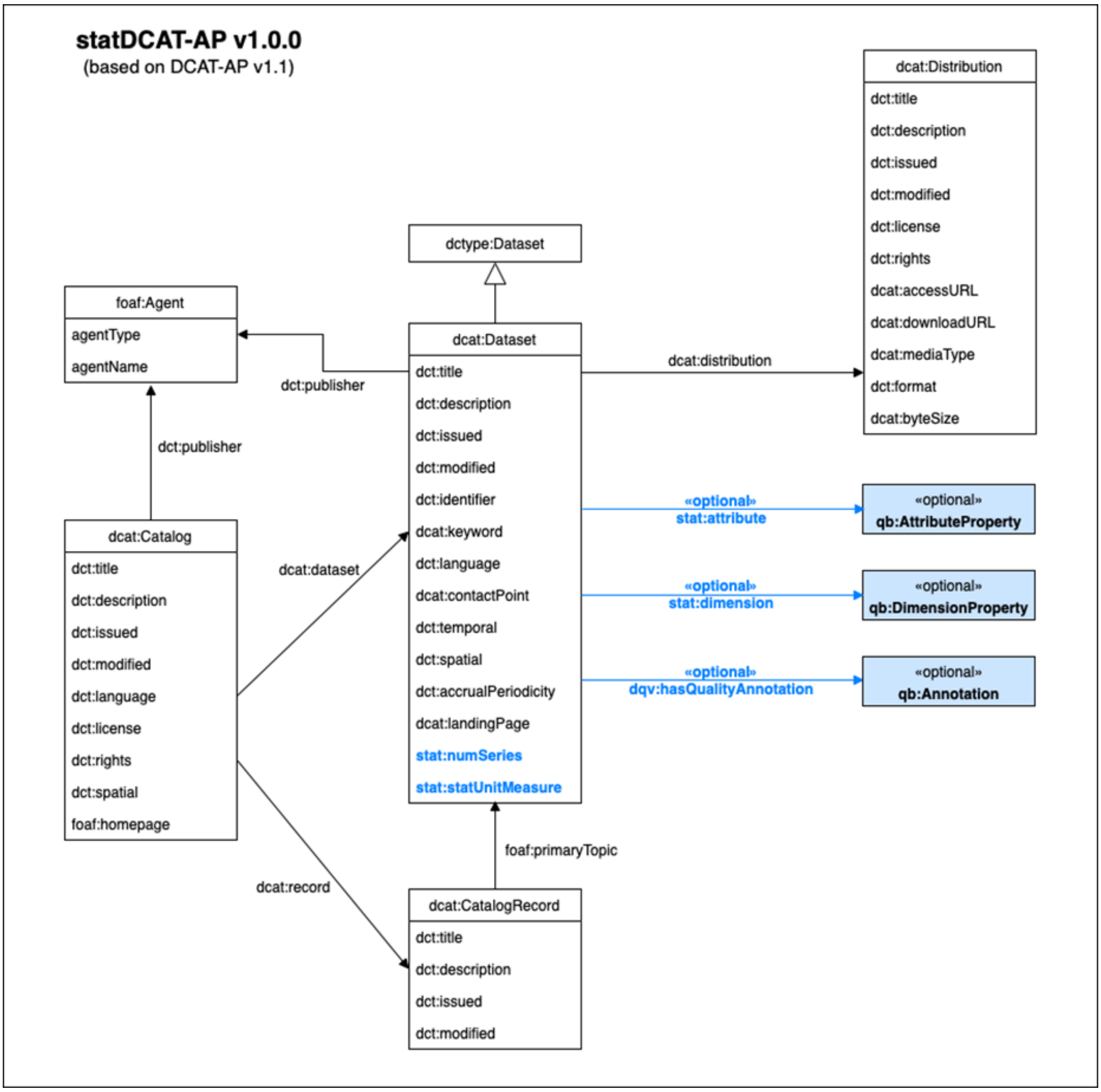


Figure 13 - StatDCAT-AP schematic data model

The rest of properties (**stat.numSeries**) and relations (**dqv:hasQualityAnnotation** and **dct:type** from Distribution class) are not created.

4.3 Mapping SDMX information model to StatDCAT-AP

The scope of this subsection is to explain how we can map from SDMX content to StatDCAT-AP. This allows us to know which metadata information have to be translated into the corresponding DCAT-AP classes and/or properties. For this purpose, the StatDCAT-AP defines a schematic operation of the high level classes in the SDMX Information Model and how they are translated into the StatDCAT-AP.

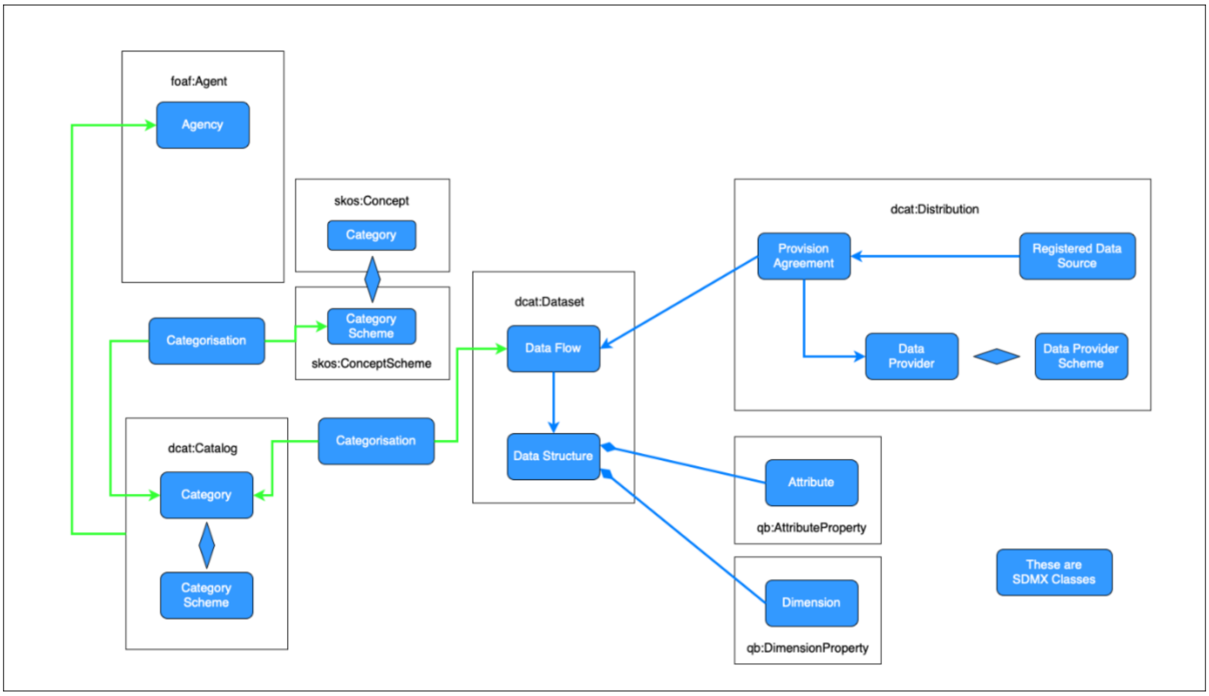


Figure 14 – Schematic map of StatDCAT-AP classes to SDMX

A narrative explanation of the previous schematic figure is the following:

1. A SDMX Category Schema is mapped to the DCAT Catalogue. Categorisation allows us to link both structural metadata object in SDMX with Dataflows in StatDCAT Dataset as well link to the topic themes used in the SDMX Dataflow.

2. The SDMX Dataflow and the corresponding SDMX Data Structure (DataStructureDefinition) are mapped into StatDCAT Dataset.

3. Dimension and Attribute in the SDMX Data Structure are mapped into the DimensionProperty and AttributeProperty classes in StatDCAT-AP.
4. The SDMX Category Scheme is mapped to the DCAT Category Scheme. The Categories in this Category Scheme are the topics that categorise the data type. Therefore, Categories are linked to Dataflows which are relevant for the topic of a Categorisation.
5. The SDMX Provision Agreement is mapped to the DCAT Distribution.
6. The SDMX Agency, the Maintenance Agency for the metadata, is mapped to the DCAT Agent and is maintained in a different scheme from the Data Provider. We need to keep in mind that the SDMX Agency is not the same entity as SDMX the Data Provider, even although they can get the same id.

Finally, we have created a EBNF LALR(1) grammar to facilitate the transformation mechanism from a RDF Turtle representation of the SDMX into our JSON-LD models based on StatDCAT-AP and publish it to the FIWARE Context Broker (Orion-LD). This grammar is based on the RDF 1.1 Turtle specification [74]. Some notes have to be considered in this specification:

1. When tokenizing the input and choosing grammar rules, the longest match is chosen.
2. The Turtle grammar is LL(1) and LALR(1) when the rules with uppercased names are used as terminals.
3. The entry point into the grammar is start.
4. The WS token [75] allows any number of white spaces.
5. The COMMENT token allows disregard comments in text.
6. The token LCASE_LETTER allows us identifying any lowercase character. The token UCASE_LETTER allows us identifying any uppercase character. The token LETTER represents any character either in lowercase or in uppercase.
7. The DIGIT token allows to identify any number between 0 and 9.
8. The token HEXDIGIT allows to identify any hexadecimal number where characters can be uppercase or lowercase.
9. The token ESCAPED_STRING allows us identifying any strings between double quotes.

Co-financed by the Connecting Europe Facility of the European Union

10. The strings '@prefix' [76] match the pattern for LANGTAG [77], though "prefix" is registered language subtags [72]. This specification does not define whether a quoted literal followed by either of these tokens (e.g., "A"@prefix) is in the Turtle language.

```
// RULES

start                 ::= statement*
statement             ::= directive | triples "."
directive             ::= prefixid
prefixid              ::= "@prefix" prefixname_ns iri "."
prefixname_ns         ::= PREFIXNAME_PREFIX? ":"
triples               ::= subject predicateobjectlist
subject               ::= iri
iri                   ::= "<" uriref ">" | prefixedname
uriref                ::= URIREF
predicateobjectlist         ::= verb objectlist (";" (verb objectlist)?)*
objectlist            ::= object ("," object)*
object                ::= iri | literal | blanknodepropertylist
literal               ::= rdfliteral
rdfliteral            ::= string (langtag | "^^" iri)?
string                ::= ESCAPED_STRING
verb                  ::= predicate | VERB
predicate             ::= iri
prefixedname          ::= PNAME_LN | PNAME_NS
blanknodepropertylist       ::= "[" predicateobjectlist "]"
langtag               ::= LANGTAG

// TOKENS

PREFIXNAME_PREFIX           ::=  PREFIXNAME_CHARS_BASE  ((PREFIXNAME_CHARS  |
".")* PREFIXNAME_CHARS)?
PREFIXNAME_CHARS_BASE     ::= LETTER
PREFIXNAME_CHARS            ::= PREFIXNAME_CHARS_U | "-" | DIGIT
PREFIXNAME_CHARS_U          ::= PREFIXNAME_CHARS_BASE | "_"
PNAME_LN                    ::= PNAME_NS PN_LOCAL
PNAME_NS                    ::= PREFIXNAME_PREFIX? ":"
PN_LOCAL                    ::=  (PREFIXNAME_CHARS_U  |  ":"  |  DIGIT  |  PLX)
((PREFIXNAME_CHARS | "." | ":" | PLX)* (PREFIXNAME_CHARS | ":" | PLX))?
PLX: PERCENT
PERCENT                     ::= "%" HEXDIGIT HEXDIGIT
LANGTAG                     ::=  "@"  [LCASE_LETTER  |  UCASE_LETTER]+  ("-"
[LCASE_LETTER | UCASE_LETTER | DIGIT]+)*
URIREF                      ::= /[^<>{}]+/
COMMENT                     ::= /#[^\n]*/
VERB                        ::= "a"
```

Figure 15 EBNF LALR grammar

5 Conclusions

5.1 Lessons learned

It is common to say that data harvesting and cleaning is the biggest part of the job when building data pipelines. Estimates of the effort spent on these preliminary phases typically vary between 50% and 80%, or even more. The work carried out on the INTERSTAT pilots can only confirm this fact, and it appears that we are rather at the top of the range. Several comments can be made about this:

- Some important data are still not easily accessible. For example, a manual action is still required to obtain the results of the Italian census, and it is probably the case in other countries as well. This means of course that it is not possible to fully automate processes using these data: for each update, a human intervention is necessary. This also means that traceability and reproducibility of the processes are greatly impaired.

- When available online, data are published in a great variety of formats: different versions of Excel, different flavors of CSV, XML, etc. It is rare to find data at the four- or five-star level of the Berners-Lee scheme [78]. Even at the two-star level, data structures are ad hoc, poorly documented and, non-interoperable. A blatant example is provided by the air quality measures from ISPRA, which use a different Excel structure for each pollutant, whereas the European Environment Agency does a great job of harmonizing and lifting the data to RDF [79].

- Even at the very lowest level, technical standardization is not assured: it is really deplorable to see that some data publishers still use local character sets instead of UTF-8.

- It is still extremely rare to see usage licenses attached to data, not to mention standard or machine-understandable licenses.

- The cross-border dimension of the INTERSTAT pilots adds of course to the complexity. Whereas in some domains like the population census, interoperability exists up to the semantic or even legal level, it is not the case for others. For example, a lot of exchanges between ISTAT and Insee were needed in order to come up with a common content for the data on schools. In view of this observation, we can only hope for a rapid deployment of the strategy proposed by the European Commission in terms of open data, and in particular with regard to high value datasets. It should be remembered, however, that even at the

Co-financed by the Connecting Europe Facility of the European Union

best level of automatable access, the data will remain unusable until interoperable, accessible and coherent metadata are linked to them.

In comparison to the creation of the data pipelines, the development of client applications greatly benefited from previous investments made by certain members of the consortium and went rather smoothly.

5.2 Next steps

The work will continue in Activity 3 until the end of the project, along several axes:

- Continuous improvement of the pilots: enhancement of the documentation, better automation, continued effort on metadata harmonization and standardization, etc. The integration of new data in some pilots will also be studied, for example income and poverty indicators for the SEP pilot [80]. Development of the client applications will also be pursued.
- The bulk of the work will now shift to the assessment framework, which constitutes the next milestone. This will give the opportunity to compare the two approaches used for the development of the pilots, which are based on different technical paradigms and also different guiding principles.
- Finally, more resources will be dedicated to a better integration of the Context Broker in the pilots, especially at the model level with more effort on achieving interoperability between the SDMX and NGSI-LD data and metadata models.
- Dissemination of the pilots' datasets via open data catalogues to be harvestable by the European Data Portal. This will be done by the usage of Idra platform and its interoperability with the Context Broker that will be able to provide the dataset metadata in DCAT-AP/ StatDCAT-AP models

References

[1] INTERSTAT, "D3.1 - Report of the use cases to demonstrate the cross-border benefits of the proposed solution," 2020. [Online]. Available: https://cef-interstat.eu/resources/.

[2] "D2.1 - Ontologies and tools to enable cross-border semantic interoperability," 2021. [Online]. Available: https://cef-interstat.eu/resources/.

[3] "ReactJS," 2022. [Online]. Available: https://reactjs.org/.

[4] Ontotext, "GraphDB," 2022. [Online]. Available: https://www.ontotext.com/products/graphdb/.

[5] "SFTP," 2022. [Online]. Available: https://www.sftp.net/servers.

[6] INTERSTAT, "GitHub GF code repository," 2022. [Online]. Available: https://github.com/INTERSTAT/Statistics-Contextualized/tree/main/code/Python/gf.

[7] "Python," 2022. [Online]. Available: https://www.python.org/.

[8] "Prefect," 2022. [Online]. Available: https://www.prefect.io/.

[9] "Geolocalized Facilities pilot example," 2022. [Online]. Available: https://raw.githubusercontent.com/INTERSTAT/Statistics-Contextualized/main/img/gf-flow-design.png.

[10] "Prefect cloud platform," 2022. [Online]. Available: https://www.prefect.io/cloud/.

[11] INTERSTAT, "Technical environment for the ETL Python implementation," 2022. [Online]. Available: https://interstat.github.io/Statistics-Contextualized/code/Python/.

[12] "Monolith tool," 2022. [Online]. Available: https://www.monolith.obdasystems.com/.

[13] "SKOS," 2022. [Online]. Available: https://www.w3.org/TR/2008/WD-skos-reference-20080829/skos.html.

[14] “GeoSPARQL,” 2022. [Online]. Available: https://www.ogc.org/standards/geosparql.

[15] "OWL," 2022. [Online]. Available: https://www.w3.org/OWL/.

[16] "SOSA," 2022. [Online]. Available: https://www.w3.org/TR/vocab-ssn/.

[17] "AQD model," 2022. [Online]. Available: https://dd.eionet.europa.eu/documentation.

[18] "SDMX model description," 2022. [Online]. Available: https://github.com/INTERSTAT/Statistics-Contextualized/blob/main/pilots/sep/sep-dsd.md.

[19] "ISTAT census data," 2022. [Online]. Available: http://dati-censimentipermanenti.istat.it/?lang=en&SubSessionId=e260034c-92f8-438f-b9f7-737286737689.

[20] "Insee website," 2022. [Online]. Available: https://www.insee.fr/fr/statistiques/5395878?sommaire=5395927.

[21] ISPRA, "Endpoint to extract PM10 pollutant data," 2022. [Online]. Available: https://annuario.isprambiente.it/sites/default/files/sys_ind_files/indicatori_ada/448/TABELLA%201_PM10_2019_rev.xlsx.

[22] "European Environmental Agency - PM10 data," 2022. [Online]. Available: http://aidef.apps.eea.europa.eu/?source=%7B%22query%22%3A%7B%22bool%22%3A%7B%22must%22%3A%5B%7B%22term%22%3A%7B%22CountryOrTerritory%22%3A%22France%22%7D%7D%2C%7B%22term%22%3A%7B%22ReportingYear%22%3A%222019%22%7D%7D%2C%7B%22term%22%3A%7B%22Pollutant%22%3.

[23] Eurostat, "Geographic location codes," 2022. [Online]. Available: https://ec.europa.eu/eurostat/web/nuts/local-administrative-units.

[24] INTERSTAT, "Description of the SEP code lists," 2022. [Online]. Available: https://github.com/INTERSTAT/Statistics-Contextualized/blob/main/pilots/sep/sep-dsd-1.ttl.

[25] INTERSTAT, "R script for the Italian census data," 2022. [Online]. Available: https://github.com/INTERSTAT/Statistics-Contextualized/files/7539489/Pilot.A.-.census.data.processing.txt.

[26] INTERSTAT, "Age groups definition," 2022. [Online]. Available: https://github.com/INTERSTAT/Statistics-Contextualized/blob/main/pilots/resources/age-groups.csv.

Co-financed by the Connecting Europe Facility of the European Union

The contents of this publication are the sole responsibility of INTERSTAT consortium and do not necessarily reflect the opinion of the European Union

[27] INTERSTAT, "SEP CSV metadata," 2022. [Online]. Available: https://github.com/INTERSTAT/Statistics-Contextualized/blob/main/pilots/sep/sep-census.csv-metadata.json.

[28] INTERSTAT, "SEP Client Application," 2022. [Online]. Available: https://interstat.eng.it/SEP-pilot-client/.

[29] "WebVOWL," 2022. [Online]. Available: http://vowl.visualdataweb.org/webvowl.html.

[30] "DQV vocabulary," 2022. [Online]. Available: https://www.w3.org/TR/vocab-dqv/.

[31] Insee, "BPE," 2022. [Online]. Available: https://www.insee.fr/en/metadonnees/source/serie/s1161.

[32] Insee, "BPE landing page," 2022. [Online]. Available: https://www.insee.fr/fr/statistiques/3568638?sommaire=3568656.

[33] Insee, "BPE Exposition venues and heritage," 2022. [Online]. Available: https://www.insee.fr/fr/statistiques/fichier/3568638/bpe20_sport_Loisir_xy_csv.zip.

[34] "BPE Education file," 2022. [Online]. Available: https://www.insee.fr/fr/statistiques/fichier/3568638/bpe20_enseignement_xy_csv.zip.

[35] INTERSTAT, "Python script for GF French data," 2022. [Online]. Available: https://github.com/INTERSTAT/Statistics-Contextualized/blob/main/code/Python/gf/bpe_extraction.py.

[36] MiBACT, "MiBACT web site," 2022. [Online]. Available: https://dati.cultura.gov.it/.

[37] MIUR, "MIUR SPARQL endpoint," 2022. [Online]. Available: https://dati.cultura.gov.it/sparql.

[38] MiBACT, “MiBACT SPARQL endpoint,” 2022. [Online]. Available: https://dati.cultura.gov.it/sparql.

[39] MiBACT, "Cultura-ONtology," 2022. [Online]. Available: https://dati.cultura.gov.it/cultural-ON/ENG.html.

[40] "CSVW," 2022. [Online]. Available: https://www.w3.org/TR/tabular-data-primer/.

[41] INTERSTAT, "CSVW description about GF data," 2022. [Online]. Available: https://interstat.eng.it/files/gf/output/gf_data_fr.csv-metadata.json.

[42] "DDI-CDI," 2022. [Online]. Available: https://ddialliance.org/Specification/ddi-cdi.

[43] "DDI-CDI description of the BPE file," 2022. [Online]. Available: https://github.com/INTERSTAT/Statistics-Contextualized/blob/main/pilots/gf/gf-cdi.ttl.

[44] INTERSTAT, "DCAT-AP metadata description for GF pilot," 2022. [Online]. Available: https://github.com/INTERSTAT/Statistics-Contextualized/blob/main/pilots/gf/gf-dcat.ttl.

[45] INTERSTAT, "The ETL process of the Geolocalized Facilities pilot," 2022. [Online]. Available: https://app.diagrams.net/#HINTERSTAT%2FStatistics-Contextualized%2Fmain%2Fimg%2Fgf-flow.drawio.

[46] INTERSTAT, "GF Client Application," 2022. [Online]. Available: https://interstat.eng.it/GF-pilot-client/.

[47] MIUR, "MIUR opendata catalogue," 2922. [Online]. Available: https://dati.istruzione.it/opendata/.

[48] MIUR, "Registry information on schools Dataset," 2022. [Online]. Available: https://dati.istruzione.it/opendata/ricerca/?searchinput=SCUANAGRAFE&lg=%24lang.

[49] MIUR, "Students by course, year and age group Dataset," 2022. [Online]. Available: https://dati.istruzione.it/opendata/ricerca/?searchinput=ALUCORSOETA&lg=%24lang.

[50] MIUR, "Location of schools Dataset," 2022. [Online]. Available: https://dati.istruzione.it/opendata/ricerca/?searchinput=EDIANAGRAFESTA&lg=.

[51] INTERSTAT, "GF files in the SFTP server," 2022. [Online]. Available: https://interstat.eng.it/files/gf/.

[52] Ministère de l'éducation, "Address and geolocalization of primary and secondary educational institutions - Dataset," 2022. [Online]. Available: https://data.education.gouv.fr/explore/dataset/fr-en-adresse-et-geolocalisation-etablissements-premier-et-second-degre.

[53] Ministère de l'éducation, "Student enrollment by grade and number of classes per school - Dataset," 2022. [Online]. Available: https://data.education.gouv.fr/explore/dataset/fr-en-ecoles-effectifs-nb_classes.

[54] Ministère de l'éducation, "Number of students by grade, gender, most frequent modern languages 1 and 2, by lower secondary school - Dataset," 2022. [Online]. Available: https://data.education.gouv.fr/explore/dataset/fr-en-college-effectifs-niveau-sexe-lv.

Co-financed by the Connecting Europe Facility of the European Union

[55] Ministère de l'éducation, "Number of students by grade, gender, most frequent modern languages 1 and 2, by upper secondary school - Dataset," 2022. [Online]. Available: https://data.education.gouv.fr/explore/dataset/fr-en-lycee_gt-effectifs-niveau-sexe-lv.

[56] Ministère de l'éducation, "Number of students by grade, gender, most frequent modern languages 1 and 2, by upper secondary school - Dataset," 2022. [Online]. Available: https://data.education.gouv.fr/explore/dataset/fr-en-lycee_pro-effectifs-niveau-sexe-lv.

[57] I. f. s. Unesco, "ISCED Mappings," 2022. [Online]. Available: http://uis.unesco.org/en/isced-mappings.

[58] I. d. S. d. l'Unesco, "ISCED mapping for French," 2022. [Online]. Available: http://uis.unesco.org/sites/default/files/documents/isced-2011-fr.pdf.

[59] INTERSTAT, "Schools metadata and harmonization outline," 2022. [Online]. Available: https://github.com/INTERSTAT/Statistics-Contextualized/files/8471721/S4Y-SchoolRegistry.meta.xlsx.

[60] INTERSTAT, "SFTP area for S4Y output files," 2022. [Online]. Available: https://interstat.eng.it/files/s4y/output/.

[61] INTERSTAT, "S4Y Client Application," 2022. [Online]. Available: https://interstat.eng.it/S4Y-pilot-client/.

[62] "Smart Data Models," 2022. [Online]. Available: https://smartdatamodels.org/.

[63] "DCAT-AP standard release 2.0.1," 2022. [Online]. Available: https://joinup.ec.europa.eu/collection/semantic-interoperability-community-semic/news/dcat-ap-release-201.

[64] W3C, "Data Catalog Vocabulary (DCAT)," [Online]. Available: https://www.w3.org/TR/2014/REC-vocab-dcat-20140116/.

[65] W3C, "Data Catalog Vocabulary (DCAT) - Version 2," 2022. [Online]. Available: https://www.w3.org/TR/vocab-dcat-2/.

[66] "dataModel.DCAT-AP repository," 2022. [Online]. Available: https://github.com/smart-data-models/dataModel.DCAT-AP.

[67] "Dataset DCATP-AP," 2022. [Online]. Available: https://swagger.lab.fiware.org/?url=https://smart-data-models.github.io/dataModel.DCAT-AP/Dataset/swagger.yaml.

[68] "Catalogue DCAT-AP," 2022. [Online]. Available: https://swagger.lab.fiware.org/?url=https://smart-data-models.github.io/dataModel.DCAT-AP/CatalogueDCAT-AP/swagger.yaml.

[69] "Agent DCAT-AP," 2022. [Online]. Available: https://swagger.lab.fiware.org/?url=https://smart-data-models.github.io/dataModel.DCAT-AP/AgentDCAT-AP/swagger.yaml.

[70] "Distribution DCAT-AP," 2022. [Online]. Available: https://swagger.lab.fiware.org/?url=https://smart-data-models.github.io/dataModel.DCAT-AP/DistributionDCAT-AP/swagger.yaml.

[71] "Catalogue Record DCAT-AP," 2022. [Online]. Available: https://smart-data-models.github.io/dataModel.DCAT-AP/CatalogueRecordDCAT-AP/swagger.yaml.

[72] "DataService DCAT-AP," 2022. [Online]. Available: https://swagger.lab.fiware.org/?url=https://smart-data-models.github.io/dataModel.DCAT-AP/DataServiceDCAT-AP/swagger.yaml.

[73] “Dataset STAT-DACT-AP,” 2022. [Online]. Available: https://swagger.lab.fiware.org/?url=https://smart-data-models.github.io/dataModel.STAT-DCAT-AP/DatasetSTAT-DCAT-AP/swagger.yaml.

[74] W3C, "RDF 1.1 Turtle specification," 2022. [Online]. Available: https://www.w3.org/TR/turtle/#sec-grammar-grammar.

[75] W3C, "WS token," 2022. [Online]. Available: https://www.w3.org/TR/turtle/#grammar-production-WS.

[76] W3C, "Prefix," 2022. [Online]. Available: https://www.w3.org/TR/turtle/#grammar-production-prefixID.

[77] W3C, "LANGTAG," [Online]. Available: https://www.w3.org/TR/turtle/#grammar-production-LANGTAG.

[78] "Berners-Lee scheme," 2022. [Online]. Available: https://5stardata.info/en/.

[79] INTERSTAT, "ISPRA data file harmonization," 2022. [Online]. Available: https://github.com/INTERSTAT/Statistics-Contextualized/issues/17#issuecomment-1029201619.

[80] Insee, "Income and poverty indicators," [Online]. Available: https://www.insee.fr/fr/statistiques/4507225.

[81] "5 Star OPEN DATA," 2022. [Online]. Available: https://5stardata.info/en/.

Co-financed by the Connecting Europe Facility of the European Union