Open Statistical Data Interoperability Framework

www.cef-interstat.eu

D2.1 - Ontologies and tools to enable cross-border semantic interoperability

Co-financed by the Connecting Europe Facility 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 2.1 (Activity 2)

Project Document Delivery Date
30.06.2021 (v1.0)

Deliverable Type and Security
Report – Public

Authors
Francesca D'Agresti (ENG), Franck Cotton (Insee), Adele Maria Bianco (ISTAT), Paolo Francescangeli (ISTAT), Giuseppina Ruocco (ISTAT)

Contributors
Francesco Bosio (ISTAT), Cristiano Maione (ISTAT), Roberta Radini (ISTAT), Michele Karlovic Riccio (ISTAT)

Reviewers
Martino Maggio (ENG), Carlo Vaccari (ISTAT), Fernando Lopez (FF)

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

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

List of figures

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

List of tables

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

Executive Summary

This document reports the outcomes of the work performed, inside the activity 2 of the INTERSTAT project, in relation to ontologies and tools that allow the cross-border semantic interoperability between national statistical portals and that will enable the deployment of cross-border services.

Chapter 1 describes the technical components that have been proposed by different partners as possible reusable tools to enable semantic and technical interoperability. For each tool is presented a detailed description covering functional and technical aspects.

Chapter 2 is focused on the specific element of the Generic Statistical Business Process Model (GSBPM) whose main statistical processes and sub-processes were first described and subsequently analysed, identifying relation between them and the functionalities of the tools presented in chapter 1.

Chapter 3 describes the different ontologies and data models that can be reused and harmonise to enable a real cross-border and cross-domain interoperability in relation to the three project use cases: "The School For You", "Geolocalize Facilities" and "Support for Environment Policies".

The last chapter presents a generalized service pipeline, which can be adapted to several contexts, and which appears to be characterized by several steps that realize a modular chain of tools for publishing LOSD from heterogeneous sources and different data providers.

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

1 Interoperability Tools

This section will present different software tools related to semantic and technical interoperability that have been proposed and analysed by the INTERSTAT partners. The most of presented tools come from other research projects or have been developed, as open software, by statistical organisation to be used in their internal data process. These tools are the candidate ones to be integrated in the overall INTERSTAT framework, that will be developed in the following months, in order to be used in a coherent pipeline for linked data processing. For each tool are reported general information about the functionalities, the baseline technologies and references to licenses and source code.

1.1 Analysis and Visualization tools

1.1.1 Eddy

Name	**Eddy**
Description	Eddy [1] is aimed for the management and visualization of ontologies expressed in Graphol language; the application provides many features that support the user when editing an ontology. One aspect to consider is the interoperability: to allow interaction with other tools such as OWL 2 reasoners and editors like Protégé, Eddy can export the Graphol ontology in OWL2 and it also provides support for URIs.
Functionalities	Eddy is an open-source graphical editor for modelling and visualizing Graphol, a visual language for ontologies. Eddy offers ad-hoc functionalities to provide a user-friendly environment, based on a main drawing area and lateral widgets for editing, navigating, and inspecting diagrams. Eddy is compliant with W3C standard for ontology OWL 2 and allows to export the Graphol diagram into several formats, such as PDF, or Jpeg. The following figure shows Air quality ontology modelled in Eddy environment.
Owner/ Responsible	OBDA [2]
License	GNU GPL v3.0
Scope	Data management and visualisation
Technology	Desktop application developed in Python

Table 1 - Eddy Description

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

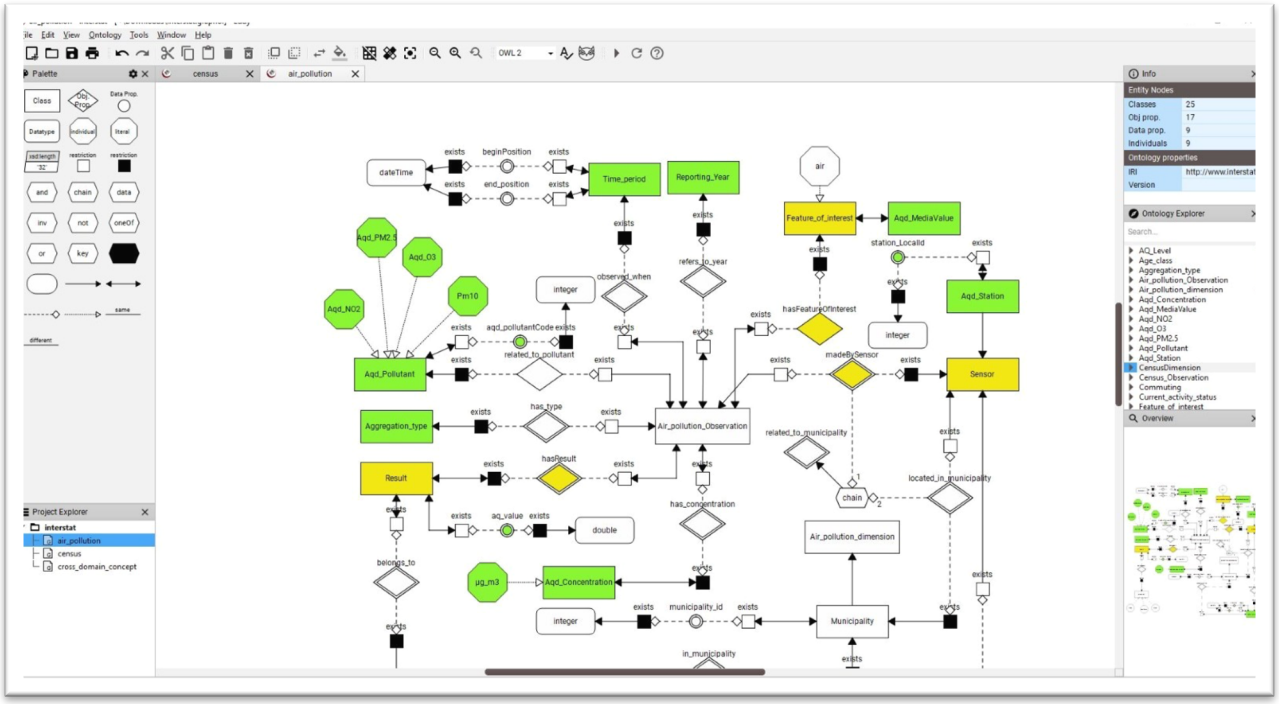


Figure 1 - Eddy

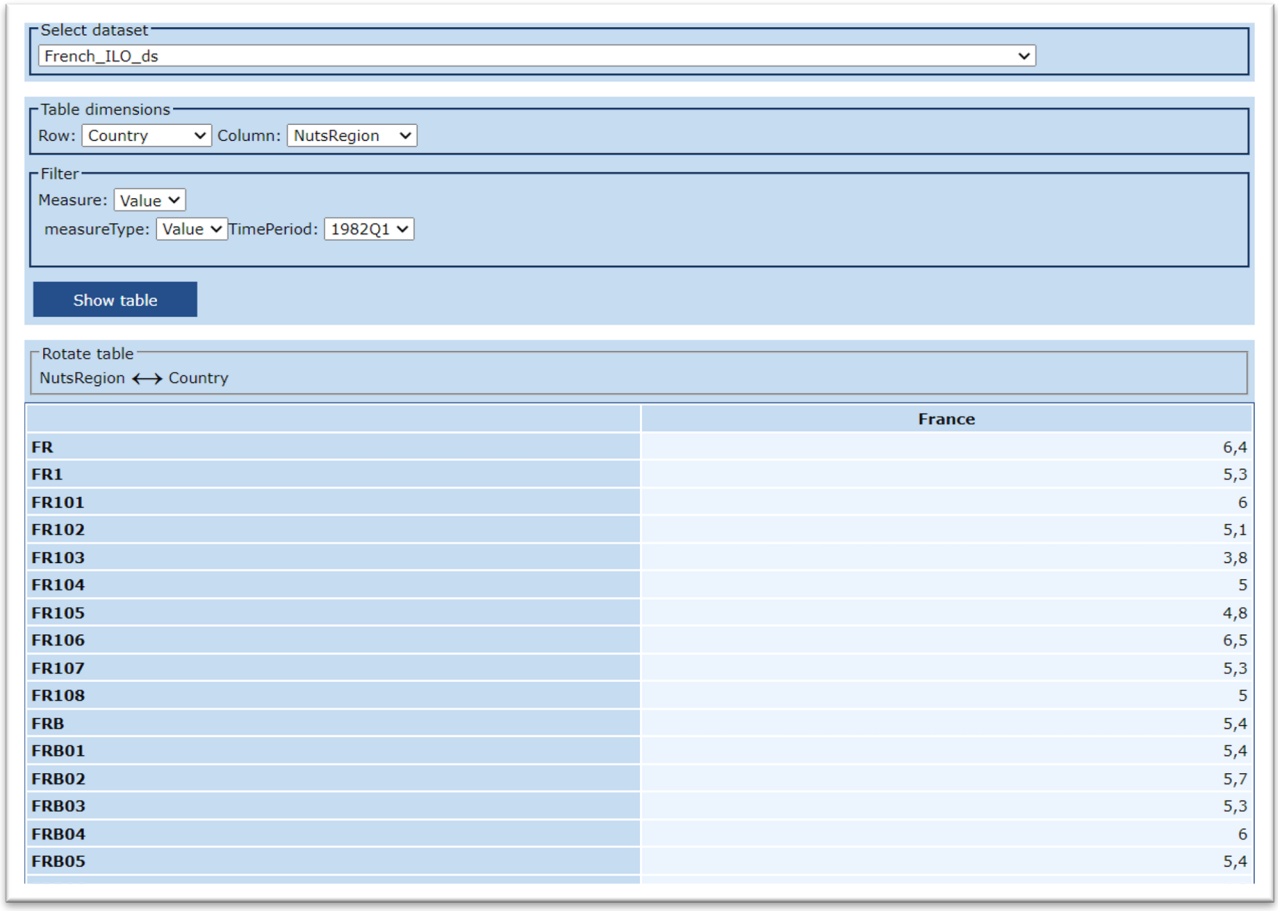
Co-financed by the Connecting Europe Facility of the European Union

1.1.2 Olap Browser

Name	Olap Browser
Description	Olap Browser [3] is a web application that allows to manage the datasets modelled and described as a Data Cube, giving the possibility to the user to explore the data. Specifically, it allows to view data in tabular form for browsing all aspects of a dataset. Through the using of OLAP Browser is possible to get views on specific data that a user wants to retrieve which would be more difficult for the user to obtain through a query. This tool is used as a data analysis and visualization tool. Olap Browser was developed for Linked Open Statistical Data (LOSD) project, delivered in the context of the European Statistical System's (ESS) DIGICOM project [4].
Functionalities	Users of the tool are data analysts who can view data on a two-dimensional slice of the cube as a table and enabling OLAP operations. For viewing the user can change the values of the fixed dimensions and thus select a different slice to be presented. The visualizations produced by OLAP browser are not sharable with other users. The tool can process Data Cubes by calling specific RESTful APIs.
Owner/ Responsible	Derilinx [5]
License	MIT
Scope	Data visualization
Technology	Web application developed in JavaScript

Table 2 - Olap Browser Description

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



	France
FR	6,4
FR1	5,3
FR101	6
FR102	5,1
FR103	3,8
FR104	5
FR105	4,8
FR106	6,5
FR107	5,3
FR108	5
FRB	5,4
FRB01	5,4
FRB02	5,7
FRB03	5,3
FRB04	6
FRB05	5,4

Figure 2 - Olap Browser

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

1.1.3 Cube Visualizer

Name	Cube Visualizer
Description	Cube Visualizer [6] is a web application that allows to manage the datasets made and described as a Data Cube, with the specific purpose of modelling datasets creating charts to display data. A user can thus get views on specific data that he wants to retrieve, which would be more difficult to obtain through building a query. Cube Visualizer provides functionality to explore the different aspects of a dataset to populate the charts which visualise the data. The URL produced through the application can be shared as a means of sharing specific visualizations of a dataset. The application was developed for Linked Open Statistical Data (LOSD) project, delivered in the context of the European Statistical System's (ESS) DIGICOM project [4].
Functionalities	The functionality of the web application is to create and present to the user graphical representations of an RDF data cube's one-dimensional slices. It is used as a data analysis and visualization tool. The user can choose several parameters which are then translated to appropriate API calls. It queries the RDF data store and the returned data are presented to the user in the form of a chart, the type of which can be also selected from bar chart, pie chart, sorted pie chart, area chart.
Owner/ Responsible	Derilinx [5]
License	MIT
Scope	Data visualization
Technology	Web application developed in JavaScript

Table 3 - Cube Visualizer Description

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

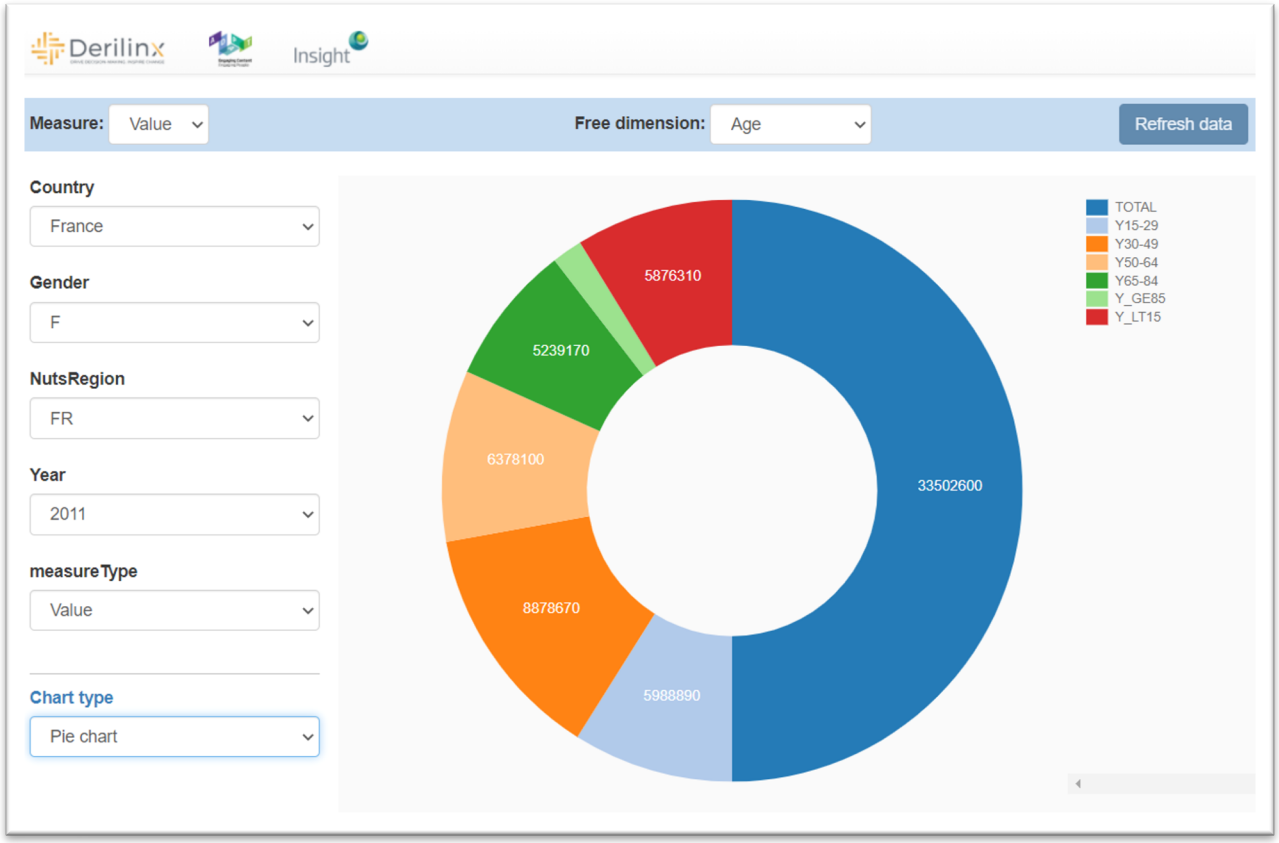


Figure 3 - Cube Visualizer

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

1.1.4 SparQLing

Name	SparQLing
Description	It is a tool [7] with the specific purpose to be a SPARQL graphical query editor and query builder. Its goal is to help users less familiar with the use of the SPARQL language, in creating the queries of interest in a visual and graphical way.
Functionalities	The tool needs as input an ontology expressed in Graphol, which is a visual language for ontologies that allows a completely visual representation of it, to help understanding for people not skilled in logic; in this way, the ontology is expressed as a graph with no formulas or textual syntax, only nodes and edges. The user has the possibility to select the properties and resources of interest directly from the graph and will automatically see the query in SPARQL language composed on the screen. This can then be saved in a file by the user.
Owner/ Responsible	OBDA [2]
License	MIT
Scope	Visual query builder
Technology	Web application developed in CoffeeScript

Table 4 - SparQLing Description

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

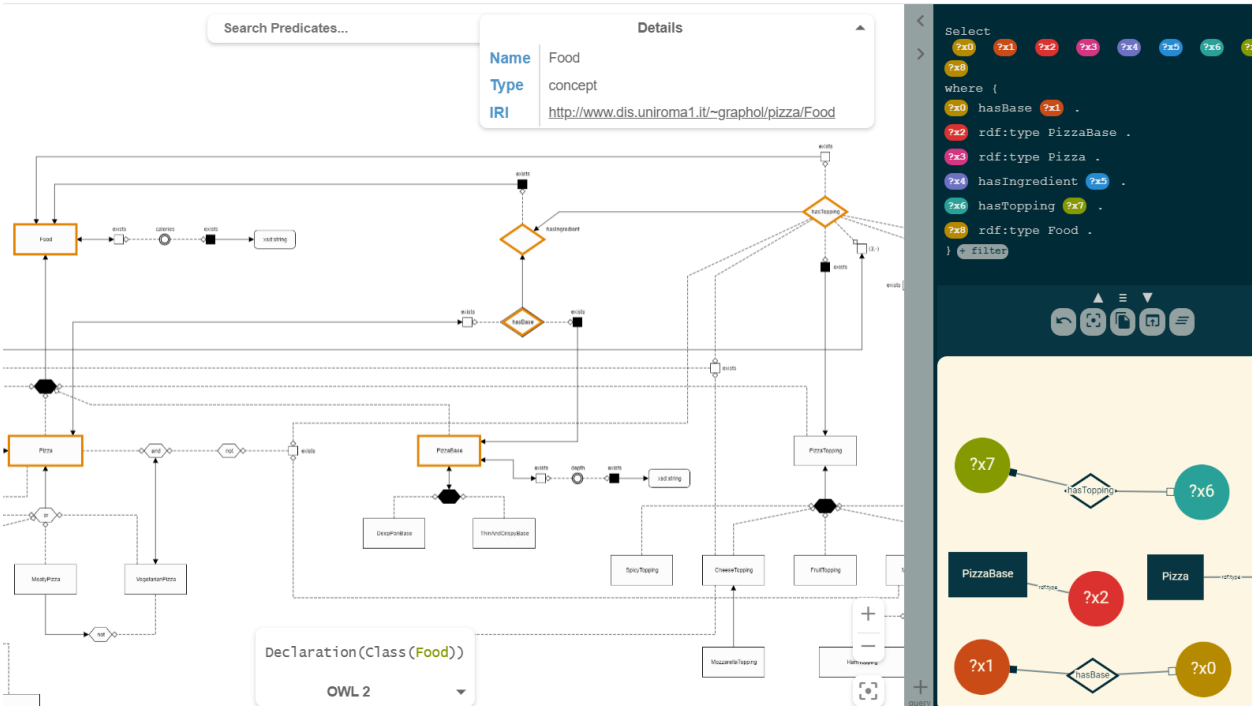


Figure 4 - SparQLing

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

1.1.5 Bauhaus

Name	Bauhaus
Description	Bauhaus [8] is a workbench for the management of linked metadata. It has a modular structure developed for the analysis of concept schemes, classifications, statistical objects, statistical operations and DCAT-AP catalogue entries.
Functionalities	Bauhaus is a modular application and the different modules allow to analyse the information depending on the type of metadata: there is a module for managing statistical operations and processes; a module for managing concept schemes, code lists, classifications and another module not yet fully completed about data structures definitions (DSD). Furthermore, the back-end consists of a set of APIs for managing statistical metadata.The front-end, instead, allows the user to select the module of interest on the main page and within it he can easily navigate between the various functions using a navbar. He also has the possibility to modify the data or export the analysis carried out.
Owner/ Responsible	INSEE
License	MIT
Scope	Linked metadata management
Technology	Web application developed in Java and JavaScript

Table 5 - Bauhaus Description

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

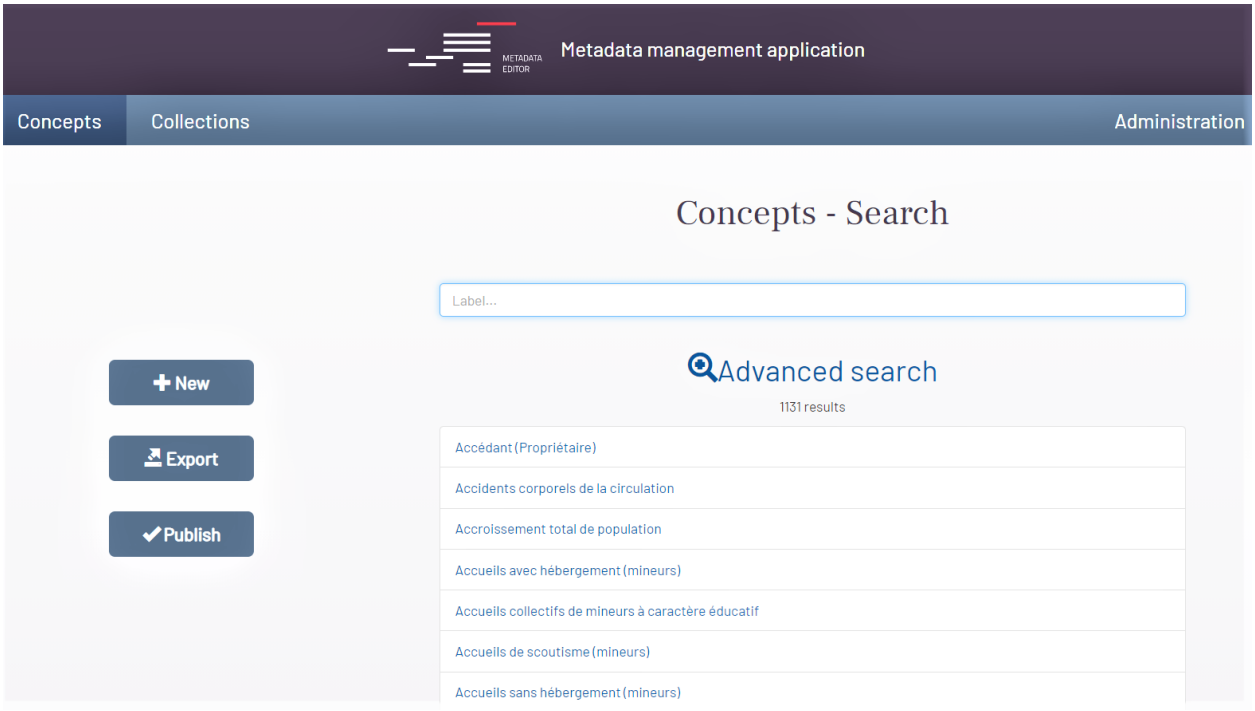


Figure 5 - Bauhaus

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

1.1.6 SPARQL React

Name	SPARQL React
Description	It is React Web application [9] to consume and display data; the application is a baseline for use cases design and development as it allows the data to be used for specific purposes. Specifically, it is a collection of demo applets showing JavaScript front-end on RDF data and for obtaining data of interest via SPARQL queries. It is possible to customize each section of the application based on use cases of interest.
Functionalities	The user, through a side menu, has the possibility to choose the specific section to access. Each section represents a set of activities that allow him to obtain answers to specific SPARQL queries on the underlying RDF data through a clear and easy to use user interface.
Owner/ Responsible	INSEE
License	MIT
Scope	Data analysis and visualization
Technology	Web application developed in JavaScript

Table 6 - SPARQL React Description

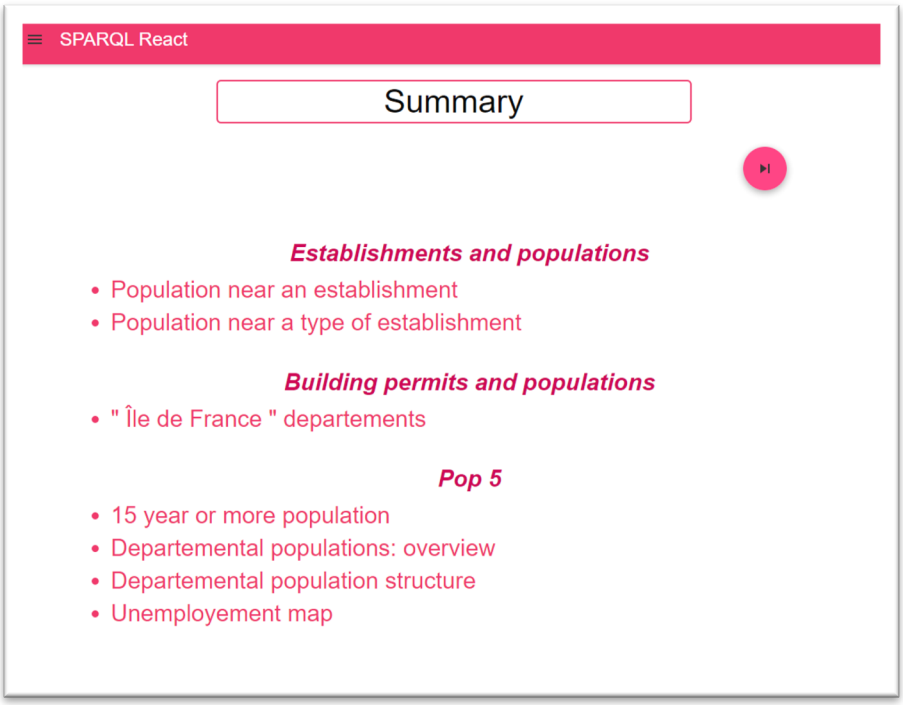


Figure 6 - SPARQL React

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

1.2 Mapping and data conversion tools

1.2.1 Juma

Name	Juma
Description	Juma [10] is an open-source editor for Graphical Mapping that allows to generate RDF triples. The standard adopted for the transformation is R2RML. It allows users with a poor knowledge of RDF to create their R2RML mappings using blocks through a visual interface for their management and creation. Juma was developed for Linked Open Statistical Data (LOSD) project, delivered in the context of the European Statistical System's (ESS) DIGICOM project [4].
Functionalities	Juma provides ad hoc solutions to facilitate the mapping stage. The graphical interface allows to create an R2RML mapping file, and then TTL files, starting from CSV data. The method used for the conversion is based on the block structure (or puzzle), a set of compatible blocks to be combined. Each block corresponds to a specific R2RML construct, thus facilitating the creation of consistent mappings, focusing on the conceptual layer, instead of the language syntax. In addition, Juma allows to import and use several vocabularies, such as FOAF, or SKOS.
Owner/ Responsible	Derilinx [5]
License	MIT
Scope	Visual ontology mapping
Technology	Web application developed in Java

Table 7 - Juma Description

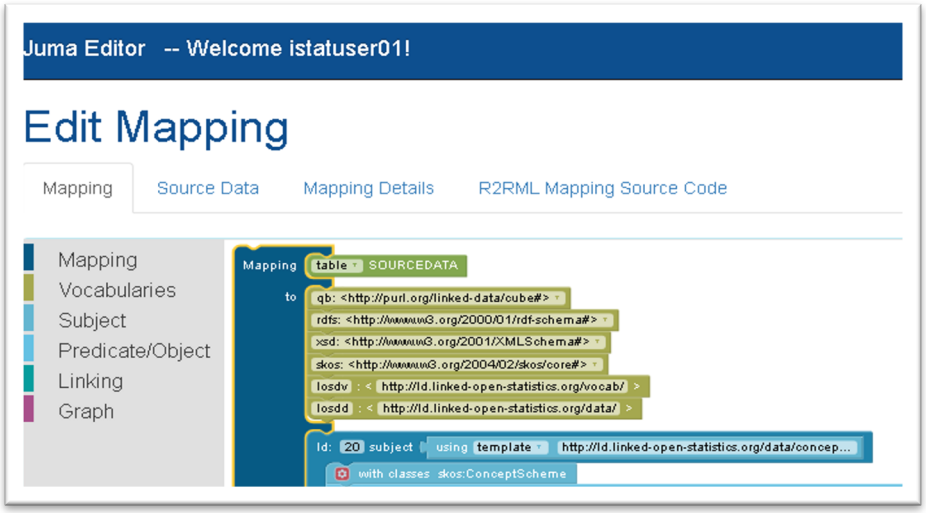


Figure 7 - Juma

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

1.2.2 Mapping Assistant

Name	Mapping Assistant
Description	The Mapping assistant application is part of the SDMX-Reference Infrastructure [11] and it allows to facilitate mapping between structural metadata and the data that reside in the dissemination database of a dissemination environment. The tool was developed as portable package or component which can be installed in other organisations.
Functionalities	Its main features concern the translation of the local database and nomenclatures to SDMX structure and SDMX codes and it facilitates the mapping between the structural metadata provided by an SDMX-ML Data Structure Definition (DSD) and those that reside in a database of a dissemination environment. Furthermore, the application maintains a Mapping Store for keeping the mappings between the SDMX and the local data storage scheme and in the SDMX Reference Infrastructure, provides mapping information to another component: the Data Retriever belonging to the Eurostat NSI WS. The Data Retriever module connects to the Mapping Store database and accesses the appropriate mappings to translate the SDMX-ML queries to SQL for the dissemination database.
Owner/ Responsible	Eurostat [12]
License	EUPL
Scope	Data mapping
Technology	Desktop application developed in C# for .NET Framework

Table 8 - Mapping Assistant Description

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

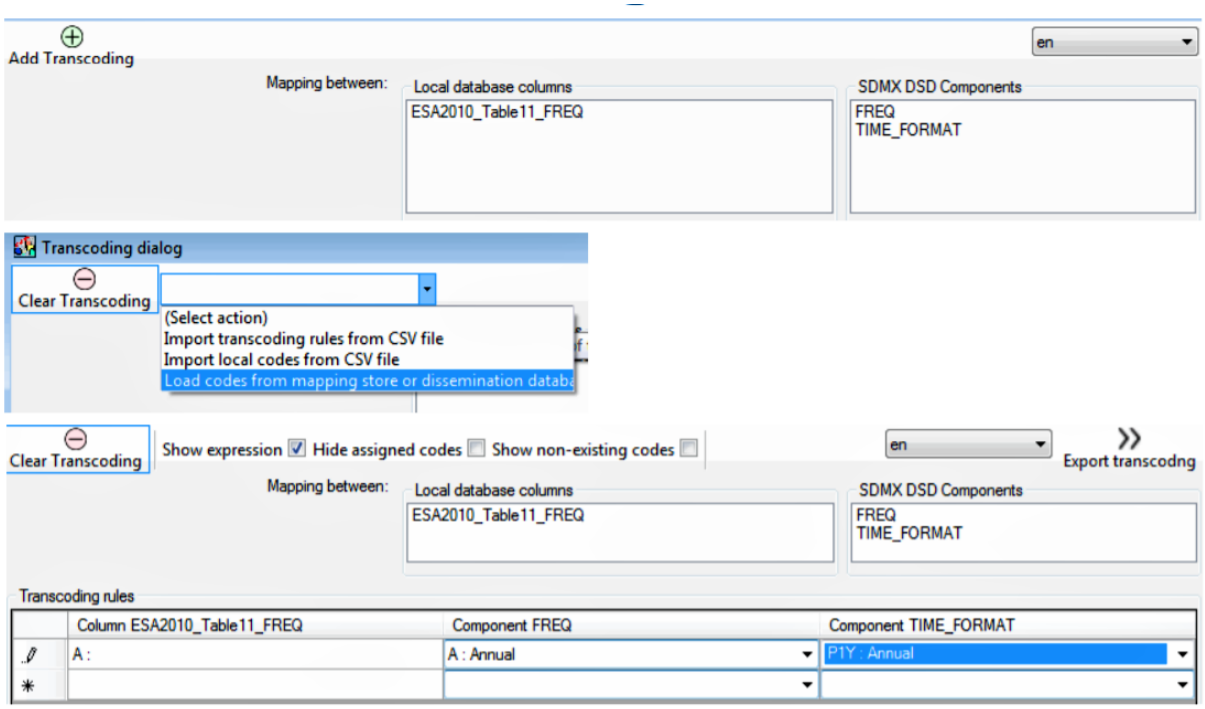


Figure 8 - Mapping Assistant

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

1.2.3 Excel/CSV to NGSI-LD

Name	Excel/CSV to NGSI-LD
Description	Web server [13] to automatically transform Excel or CSV files into ETSI NGSI-LD and upload directly into CEF Context Broker as Entities. This application allows the transformation of statistical data tables represented into MS Excel worksheets in CSV data files that can be imported into the database and then can be queried through the SDMX-RI's NSI web service. The data represented into the Excel worksheets can be already made available as open data and in in machine-to-machine mode through the web service. At present is foreseen the development of the porting into Meta and Data Manager as part of the Feature «Easy loading» designed in the AST (Statistical atlas of Territory) project [14].
Functionalities	The main features include the extraction of information from headers to create the attributes, the possibility to dismiss some columns from the Excel and/or CSV files and the possibility to define specific columns rename in configuration files. This is made possible by providing specific HTTP RESTful APIs to access linked data.
Owner/ Responsible	FIWARE
License	AGPLv3
Scope	Data conversion
Technology	Web server developed in Node.js

Table 9 - Excel/CSV to NGSI-LD Description

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

1.2.4 Excel2CSV

Name	Excel2CSV
Description	The Excel2Csv [15] application allows the transformation of the file format from Excel worksheet to CSV format, to allow the statistical tables to be correctly read and interpreted by data loading programs. To achieve this, the application must receive various information about the worksheet considered such as the precise location of the table inside the spreadsheet, the sheet number, the table boundaries, the boundaries of the axes and what cell the actual data start from. As future developments, is being planned to add a new feature that makes it possible to guess an SDMX Data Structure Definition from a given table and output it to a valid SDMX-ML file.
Functionalities	Among the features of the application, we find the possibility to detect empty rows and columns in a table and eventually read merged cells as well. Dates and times can be read directly from the Excel format or a string representing them. Should a time reference variable show non-standard values, its structure can be mapped or casted to a valid SDMX time format. Excel2Csv is also able to read tables representing n-dimensional data cubes or their subsets.
Owner/ Responsible	ISTAT
License	EUPL
Scope	Data conversion
Technology	Desktop application developed in C# for .NET Framework

Table 10 - Excel2CSV Description

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

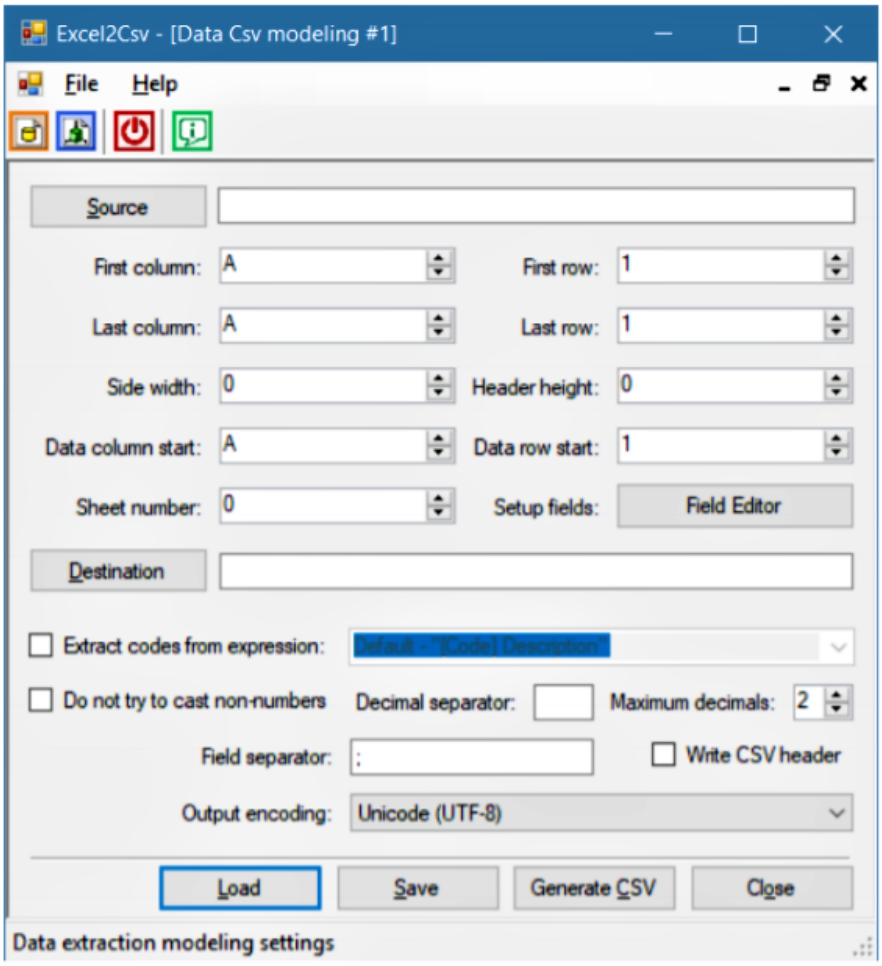


Figure 9 - Excel2CSV

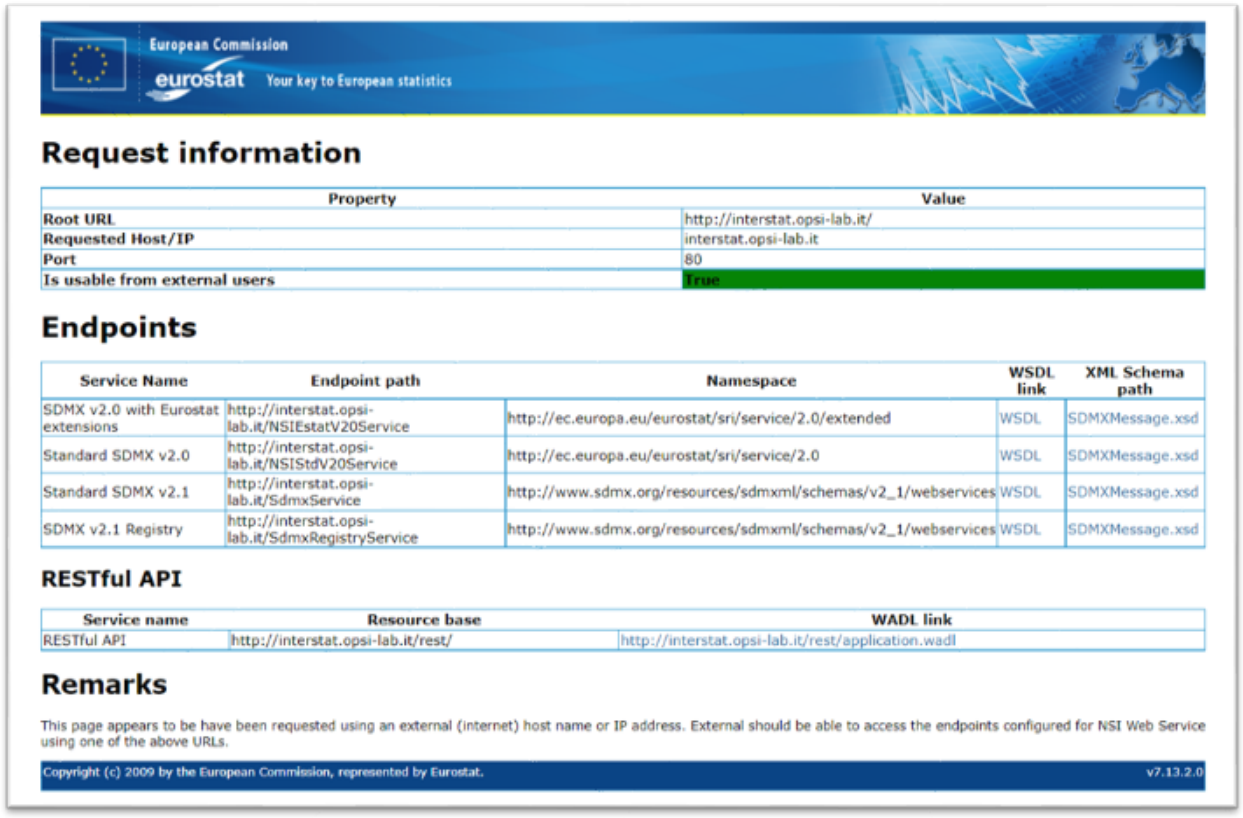
1.3 Dissemination tools

1.3.1 Eurostat NSI Web Service

Name	**Eurostat NSI WS**
Description	The Eurostat web service is part of the SDMX-Reference Infrastructure [11] and it allows National Statistical institutions to adopt the SDMX Reference Infrastructure, increasing the interoperability and visibility of their dissemination environment, simplifying the distribution of data to external systems and consumers. Furthermore, allows Member States to expose their dissemination database, using Web Service technologies, in SDMX-ML format. The web service consists of several modules: Web Service Wrapper: it implements the Web Service SOAP interface and is also responsible for the control of the order to which the other modules are invoked. Query Parser: it is responsible for the parsing of the SDMX Query into the Information Model. Data Retriever: it is responsible for the generation of an SQL query which is equivalent to the SDMX Query, it executes the query on the Dissemination Database and populates the Information Model with the returned dataset. Data Generator: it is responsible for the generation of SDMX-ML from the populated Information Model. Structure Retriever: it is responsible for retrieving structural metadata from the Mapping Store database. It can also return partial code lists based on the available data from the Dissemination Database.
Functionalities	The web service is the central element of the SDMX-Reference Infrastructure and its inputs come from two databases: the Dissemination database and the Mapping store database.Its main function is to receive a SOAP/REST request by a client, which can be a National statistical institution Client, gathers the data from the Dissemination database according to the structural metadata from the Mapping store database, creates the SDMX-ML data and sends it back to the client.
Owner/ Responsible	Eurostat [12]
License	EUPL
Scope	Data exposition and dissemination
Technology	Web application developed in C# for NET Core

Table 11 - Eurostat NSI WS Description

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



European Commission

eurostat Your key to European statistics

Request information

Property	Value
Root URL	http://interstat.opsi-lab.it/
Requested Host/IP	interstat.opsi-lab.it
Port	80
Is usable from external users	True

Endpoints

Service Name	Endpoint path	Namespace	WSDL link	XML Schema path
SDMX v2.0 with Eurostat extensions	http://interstat.opsi-lab.it/NSIEstatV20Service	http://ec.europa.eu/eurostat/sri/service/2.0/extended	WSDL	SDMXMessage.xsd
Standard SDMX v2.0	http://interstat.opsi-lab.it/NSIStdV20Service	http://ec.europa.eu/eurostat/sri/service/2.0	WSDL	SDMXMessage.xsd
Standard SDMX v2.1	http://interstat.opsi-lab.it/SdmxService	http://www.sdmx.org/resources/sdmxml/schemas/v2_1/webservices	WSDL	SDMXMessage.xsd
SDMX v2.1 Registry	http://interstat.opsi-lab.it/SdmxRegistryService	http://www.sdmx.org/resources/sdmxml/schemas/v2_1/webservices	WSDL	SDMXMessage.xsd

RESTful API

Service name	Resource base	WADL link
RESTful API	http://interstat.opsi-lab.it/rest/	http://interstat.opsi-lab.it/rest/application.wadl

Remarks

This page appears to be have been requested using an external (internet) host name or IP address. External should be able to access the endpoints configured for NSI Web Service using one of the above URLs.

Copyright (c) 2009 by the European Commission, represented by Eurostat. v7.13.2.0

Figure 10 - Eurostat NSI WS

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

1.3.2 Meta & Data Manager

Name	Meta & Data Manager
Description	SDMX based web application [16] for the management and publication of data and structural metadata by consuming Eurostat's SDMX-RI APIs, with a technologically advanced architecture composed by a completely service-based back end and a module for front-end. It is part of the SDMX Istat Toolkit [17] which can be used by a statistical organization for building stand-alone dissemination systems or a distributed data warehouse SDMX-based. This tool can model and disseminate statistical multidimensional tables; is possible to create and editing, through the Meta & Data Manager interfaces, specific SDMX annotations that can be used for defining, in the Data Browser tool, layout settings for browsing and visualizing datasets (SDMX dataflows).
Functionalities	The application interacts with a Metadata Repository via an SDMX Web service and provides a GUI for browsing, download, create and submit structural metadata (codelists, concept schemes, category schemes, dataflows, etc.). The application can interact with different SDMX Web Services, a user can browse metadata stored in different repositories; it allows to overcome some SDMX constraints and modify “finalized” SDMX item scheme artefacts. The application allows to manage and disseminate structural metadata, create dissemination and reporting databases, create, and disseminate reference metadata, Open Datasets Digital Catalogues (DCAT) and thematic glossaries.
Owner/ Responsible	ISTAT
License	EUPL
Scope	Data management and dissemination
Technology	Web application developed in C# for NET Core and REACT/REDUX

Table 12 - Meta & Data Manager Description

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

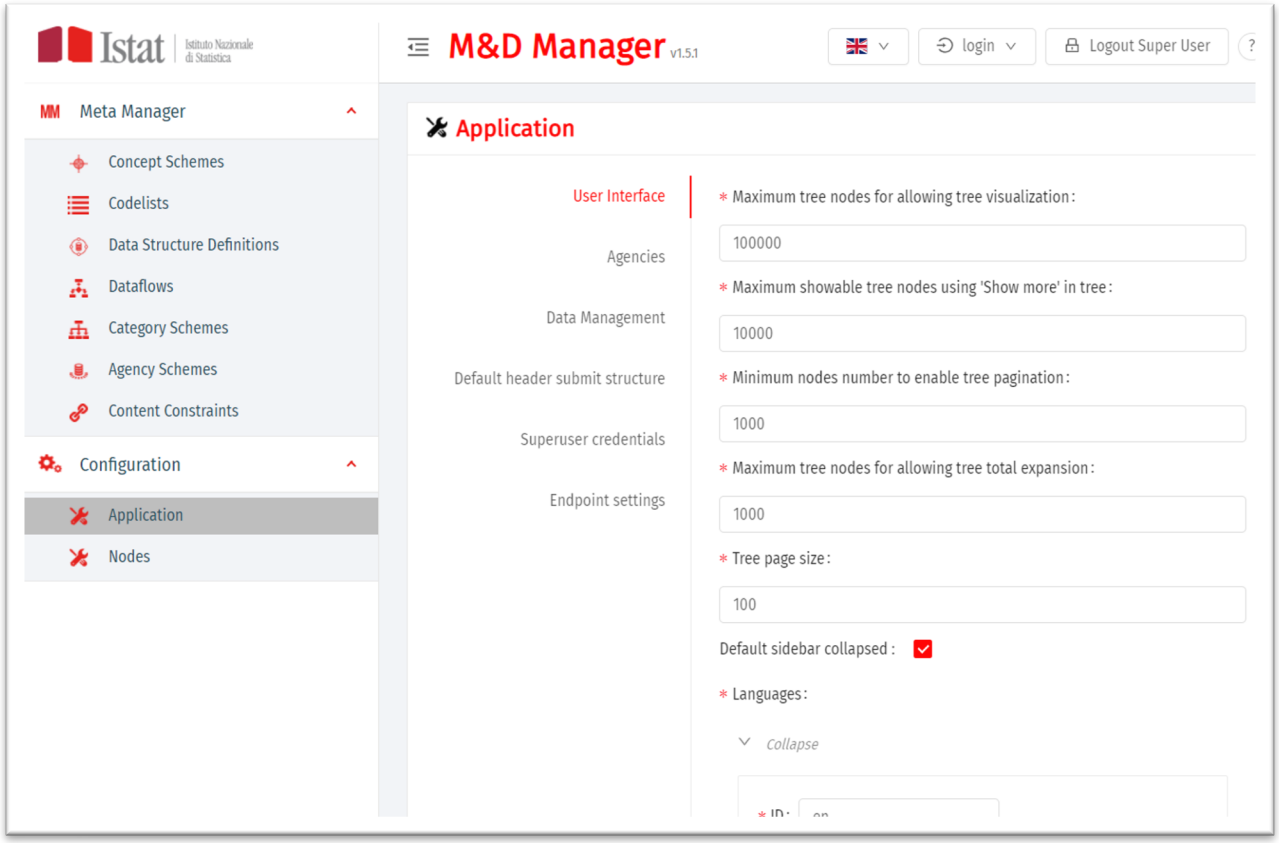


Figure 11 - Meta & Data Manager

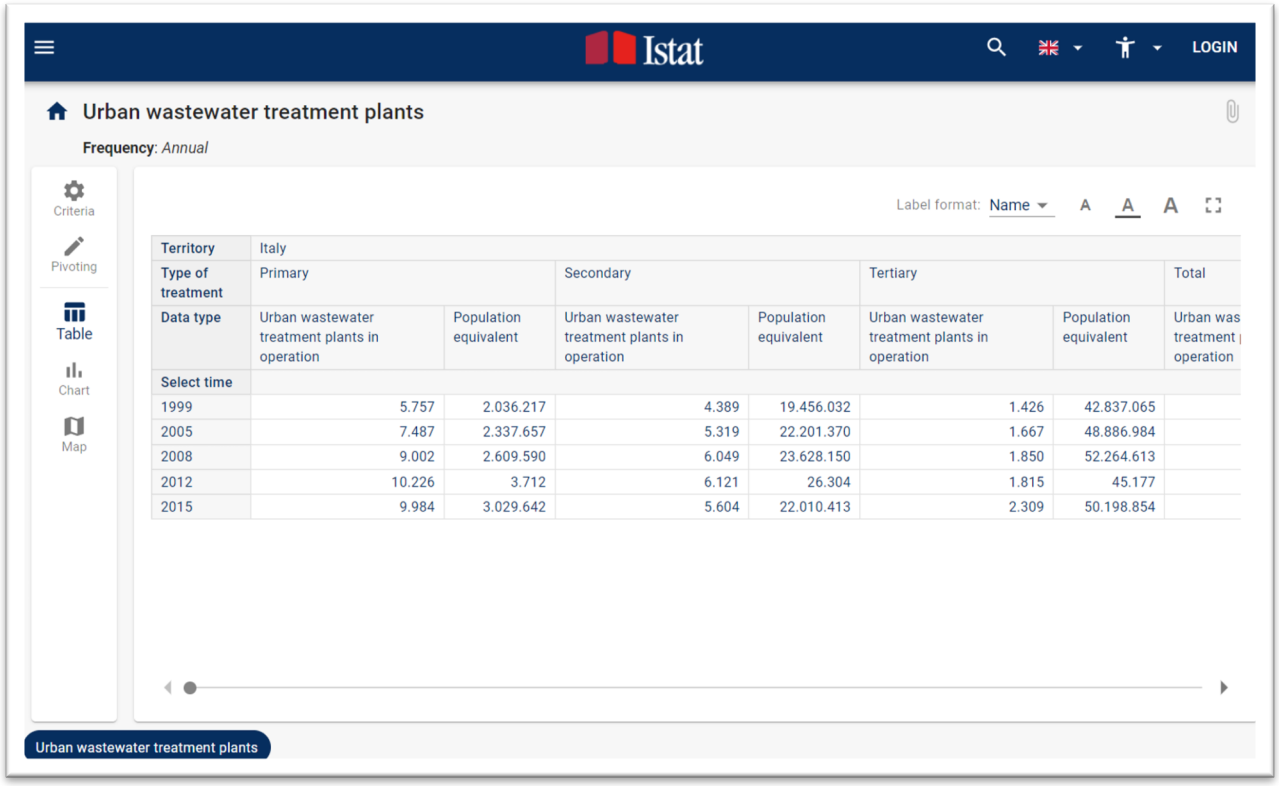
Co-financed by the Connecting Europe Facility of the European Union

1.3.3 Data Browser

Name	Data Browser
Description	Part of the SDMX Istat Toolkit [17], is a web application [18] that, connected with one or more SDMX web services, gives users the possibility to browse data and visualize datasets. It can be adopted within a single statistical organization or even in more than one (in a multi-source project), allowing them to expose their datasets through SDMX Web Services based on the SDMX-RI. Data exposure is possible thanks to its architecture organized in nodes: each node is managed by every individual entity participating in the project. The nodes thus create a network, allowing data to be published and available within each system node. The system implements a distributed data warehouse based on the SDMX standard which can be queried by external users via a web interface.
Functionalities	Data Browser is the ISTAT's reference tool for statistical data dissemination; it is the front end that must be used for browsing all the disseminated data (including open data and LODs). Specifically, this web portal aims to share, integrate and disseminate macro-data produced by statistical agencies. The strategic aim of this innovation is to create a "network" focusing on the web data distribution and make sure that the created network contains statistical data of good quality. It combines data and meta-data with a view to semantic interoperability and shares international best practices on statistical dissemination systems allowing users to switch between different distributed databases, select the one of interest, filter its data and store queries.
Owner/ Responsible	ISTAT
License	EUPL
Scope	Data browsing and dissemination
Technology	Web application developed in C# for NET Core and REACT/REDUX

Table 13 - Data Browser Description

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



Territory	Italy						
Type of treatment	Primary		Secondary		Tertiary		Total
Data type	Urban wastewater treatment plants in operation	Population equivalent	Urban wastewater treatment plants in operation	Population equivalent	Urban wastewater treatment plants in operation	Population equivalent	Urban was treatment operation
Select time							
1999	5.757	2.036.217	4.389	19.456.032	1.426	42.837.065	
2005	7.487	2.337.657	5.319	22.201.370	1.667	48.886.984	
2008	9.002	2.609.590	6.049	23.628.150	1.850	52.264.613	
2012	10.226	3.712	6.121	26.304	1.815	45.177	
2015	9.984	3.029.642	5.604	22.010.413	2.309	50.198.854	

Figure 12 - Data Browser

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

1.3.4 CEF Context Broker

Name	Context Broker
Description	The CEF Context broker [19] is a tool that allows to manage context information using a publish/subscribe approach. The information thus obtained is based on linked data standards following the ETSI NGSI-LD specification. It also provides the FIWARE NGSI-LD API or FIWARE NGSIv2 API, a simple and powerful Restful API enabling to perform updates, queries or subscribe to changes on context information.
Functionalities	An important feature of the tool is that it is small, fast and lightweight. The information is expressed in Entities that can be easily created, modified and deleted. Specifically, it is possible to create or update a set of Entities in a single request and query or retrieve Entities, with a rich set of filters and a powerful query language. Two interesting services are the Subscriptions and the Registrations: the former allow to get notified on changes in Entities, instead of continuously polling the broker; the latter allow to extend the broker with entities that live inside external context sources or brokers. This is made possible by providing specific HTTP RESTful APIs to access linked data.
Owner/ Responsible	FIWARE
License	AGPLv3
Scope	Management and dissemination of context information
Technology	Web application developed in C++

Table 14 - Context Broker Description

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

1.3.5 Idra

Name	Idra
Description	It is a web application [20] able to federate existing Open Data Management Systems (ODMS) based on different technologies providing a unique access point to search and discover open datasets coming from heterogeneous sources. Idra uniforms representation of collected open datasets, thanks to the adoption of international standards (DCAT-AP) and provides a set of RESTful APIs to be used by third party applications and to federate ODMS not natively supported. Idra supports natively ODMS based on CKAN, DKAN, Socrata, Orion Context Broker (NGSI v2) and many other technologies. The datasets remain in the original portals, the platform imports and manages only the metadata, updating it periodically to assure that the information provided is up to date. Idra is an open-source software developed inside the EU founded project FESTIVAL [21].
Functionalities	Web interface to search for Open Data in a federated and multi-language way providing the possibility for the user both to perform a simple search and to perform an advanced search on all the federated datasets. It provides a unique access point to open data regardless the heterogeneity of the underlying technologies. Public and open APIs to build smart and innovative applications are employed and it guarantees the compliance to DCAT-AP standard and European Data Portal specifications. Web scraping technology is used to federate open data portals that do not provide APIs and ease of performing SPARQL queries on 5 stars RDF linked open data collected from federated ODMS and allows to easily create charts based on federated open datasets, through DataIEt-Ecosystem Provider DEEP which was integrated with Idra to provide to users an open data visualization tool. It is a view WC, which is used to create rich, reusable visualization of data giving to the user the possibility to choose between different types of charts and to save these views in the environment.
Owner/ Responsible	ENG
License	AGPLv3
Scope	Data publishing
Technology	Web application developed in Java and Angular

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

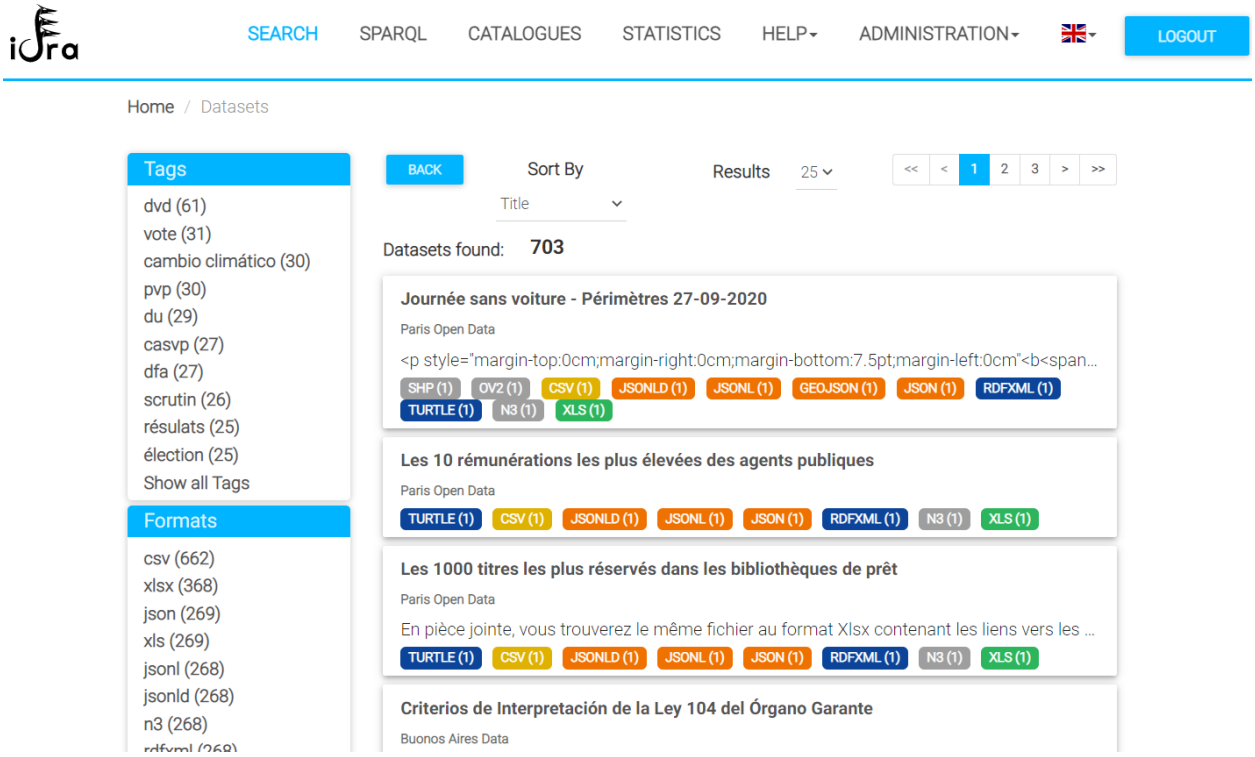


Figure 13 - Idra Tool

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

1.3.6 Datalift

Name	Datalift
Description	Datalift [22] is a web-based tool with the specific objective of helping statistical organizations to: • Publish data sets as RDF graphs. • Link these data sets together, by identifying equivalent resources in other data sources. • Describe these datasets through ontologies. Data publication is a two steps process: 1) The data must first be prepared by selecting the raw data to transform, the ontologies describing them and the possible external data they are linked to; 2) The data is then disseminated to allow access to users and external systems The Datalift application achieves this goal thanks to a modular architecture that allows the creation of various system configurations. The module regarding the data construction can also be excluded from the application configuration in order to prevent published data modifications.
Functionalities	The Datalift system is made of three types of components: 1) The core modules, that constitute a development framework and a runtime environment. They are characterized in general by the Datalift Framework and specifically by the Datalift Core which represents a concrete implementation of the framework. 2) The application modules. 3) A set of RDF stores (triple stores). The Datalift Framework builds onto existing and well-known APIs, to make the development of application modules as simple and fast as possible: the Java Enterprise Edition platform, the OpenRDF Sesame 2 API for RDF store access and data manipulation, the jQuery and jQuery UI JavaScript libraries for client user interfaces.
Owner/ Responsible	INSEE
License	MIT
Scope	Data conversion
Technology	Web application developed in Java

Table 15 - Datalift Description

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

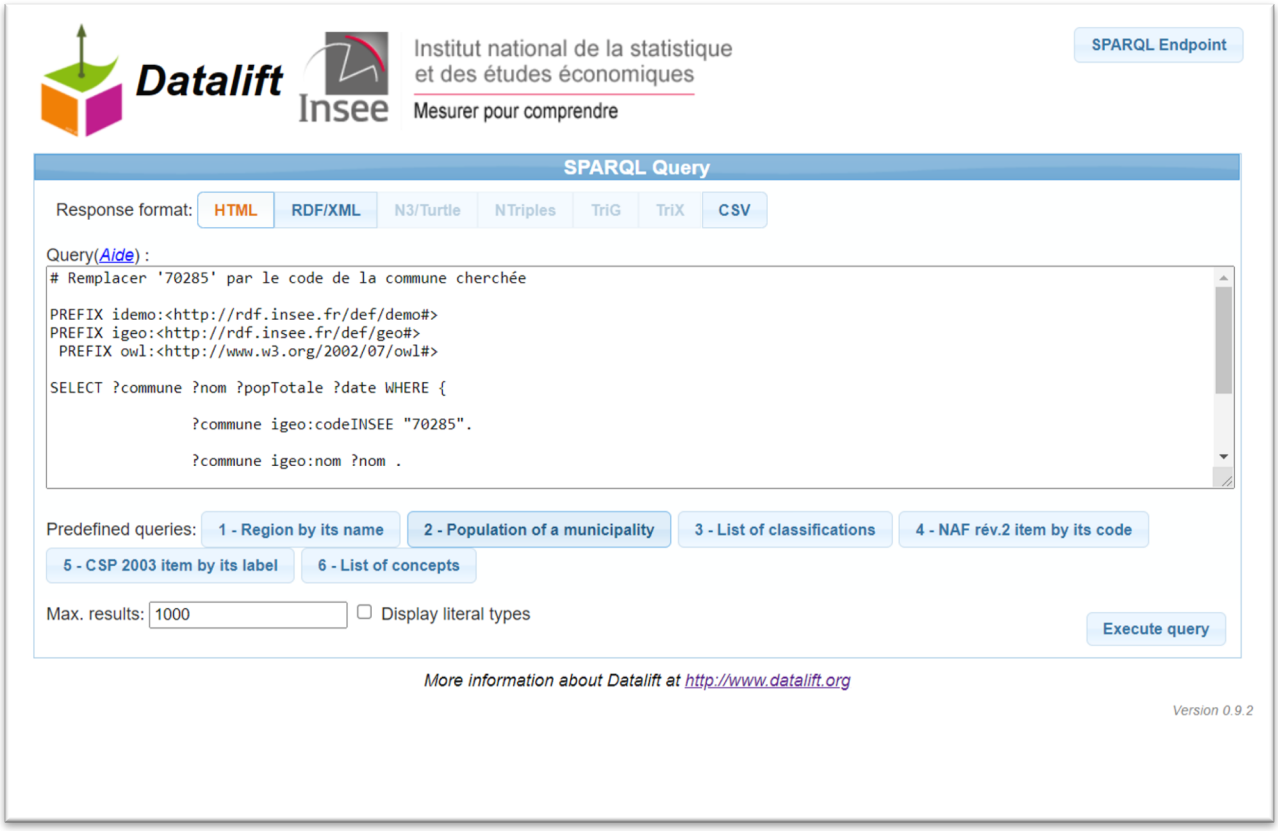


Figure 14 - Datalift

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

2 GSBPM Mapping

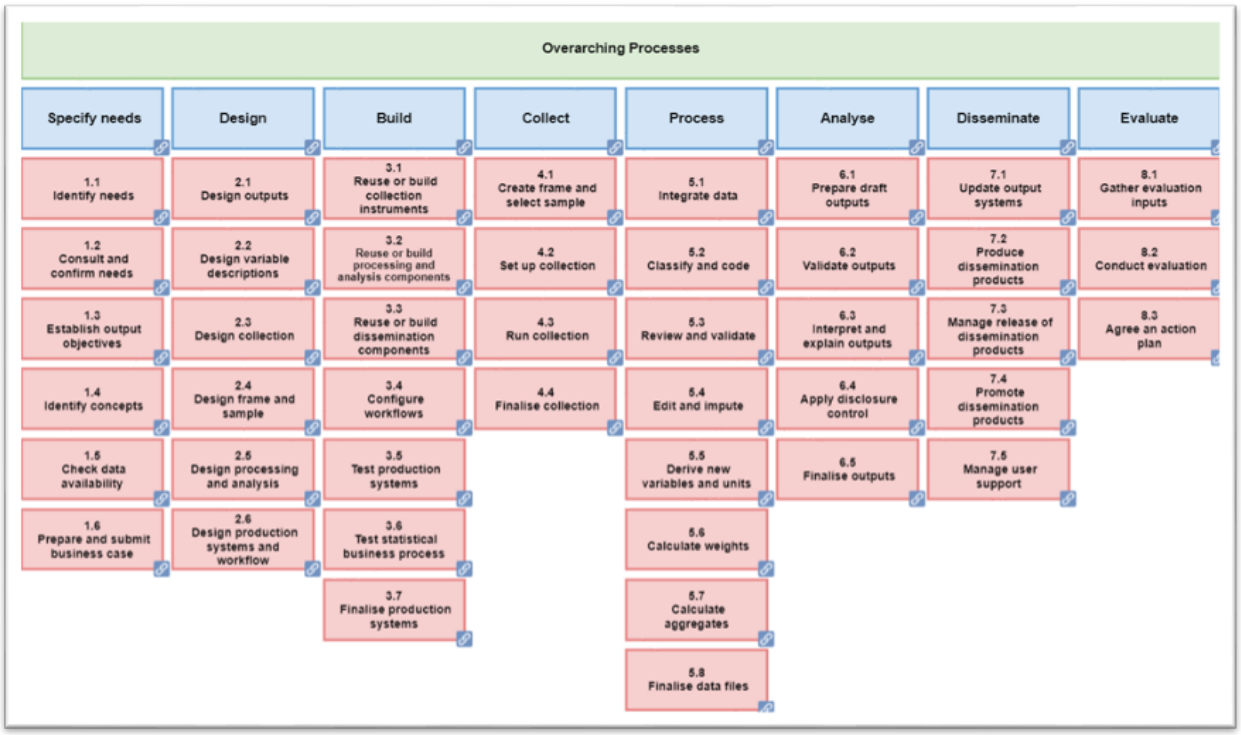


Figure 15 - GSBPM scheme with processes and sub-processes [23]

In the following chapter the Generic Statistical Business Process Model (GSBPM) [23] its purpose and its characteristics will be briefly illustrated; moreover, the different processes and sub-processes will be analysed and described. For each process are also identified possible tools, among the ones presented in the previous section, that can be directly related with it in terms of functionalities and scope.

It is important to remark that not all phases are covered by functionalities of specific tools, because many processes do not concern technical activities (e.g. requirement analysis or legal tasks).

In this section, the GSBPM grid should be understood in the light of the actual process covered by INTERSTAT, which is globally downstream compared to the main statistical process. For example, the "collection" phase for INTERSTAT corresponds to the gathering of open (aggregate) data published by statistical agencies, and this of course is not the same as the data collection conducted by these statistical agencies, which is generally done by surveys or administrative sources and concerns unit (individual) data.

The introduction of the GSBPM scheme responds to a need for classification and harmonization of the different phases of the processes implemented by the national statistical institutes; it also represents a model on which it is possible to base the assessment and improvement of the quality of these processes.

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

Specifically, it is a scheme that can be applied to any production process, from traditional surveys and the acquisition of administrative data to statistical processing, regardless of the thematic sector of reference, if there is an output in terms of statistical data and metadata. This universality is a direct consequence of the flexibility of the model, which is not constituted by a linear sequence of actions but by a matrix of phases and sub-processes, of different size and importance within the real processes. This allows to adapt the structure of the model to processes of different size and nature. Some steps could be applied to one process and not be applied to another; the sub-processes do not necessarily have to be followed according to a predetermined or hierarchical order, i.e., some can be skipped, others repeated several times, giving rise to iterative cycles.

In its most recent version, the GSBPM is made up of eight phases, each with a different number of threads within it (Figure 15). The phases cover the main steps in the development of a statistical process: from the identification of information needs to the dissemination and evaluation of results, through the design, collection, processing of data and various other intermediate steps.

2.1 Specify Needs

In this process the necessary subprocesses are created to identify the survey objectives and translate them into concepts which must be understandable and accessible to respondents and must be measurable and transformable into statistical variables, which will be designed in the next phase.

Below is a brief description of each of its subprocesses:

1.1) Identify needs: This sub-process includes the analysis and identification of the necessary statistics. It also includes consideration of the activities and actions undertaken by other statistical organizations producing similar data and the methods used by those organizations.
1.2) Consult and confirm needs: This sub-process focuses on consulting with all the stakeholders within the project, whose contribution is essential for achieving a specific goal of the organization, and the final definition of the needs for the statistics; a good understanding of user needs is required.
1.3) Establish output objectives: This sub-process identifies the statistical output objectives that are required to meet the user needs identified in the previous sub-process. It includes agreeing the suitability of the proposed outputs and their quality measures with users.
1.4) Identify concepts: This sub-process arises from the user's point of view with the aim of definitively specifying the required concepts to be measured.
1.5) Check data availability: This sub-process performs a check on the data sources: is verify if data meet user requirements and the conditions under which data would be available and takes in consideration any restrictions on their use.
1.6) Prepare and submit business case: This sub-process documents the findings of the previous sub-processes in the form of a business case; it must then be approved to implement the new or modified statistical business process.

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

2.2 Design Phase

This phase describes the research development and design activities necessary to produce statistical results, the concepts, methodologies, collection tools and operational processes. Organisations could reuse or adapt design elements from existing processes to enhance the usability and value of the statistical information.

Below is a brief description of each of its subprocesses:

2.1) Design outputs: This sub-process contains the design of the statistical outputs, products and services to be produced, including the related development work and preparation of the systems and tools used in the "Disseminate" phase. Wherever possible, outputs must follow existing standards.
2.2) Design variable descriptions: This sub-process defines the variables to be collected via the collection instrument and any statistical or geospatial classifications that will be used. Wherever possible, existing national and international standards will be followed.
2.3) Design collection: This sub-process concerns two different aspects: it aims to carry out the confirmation of any formal agreements and the identification of the most appropriate collection instruments and practices which may depend on the type of data collection and the available sources of data.
2.4) Design frame and sample: This sub-process only applies to processes which involve data collection based on sampling; It identifies and describes the population of interest, a sampling frame (and possibly the register from which it is derived) and determines the most appropriate sampling rule and methodology.
2.5) Design processing and analysis: This sub-process designs the statistical processing methodology to be applied during the "Process" and "Analyse" phases. This sub-process also includes design of specifications for data integration from multiple data sources, validation of data and estimation.
2.6) Design production systems and workflow: This sub-process takes care of checking all the required processes from data collection to dissemination within the whole production process, verifying that they cover all aspects with no gaps. Wherever possible, it is necessary to reuse existing production solutions and their processes and technologies.

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

2.3 Build Phase

In this process, the outputs of the previous phase are gathered and configured in order to create the complete environment to run the process. Below is a brief description of each of its subprocesses:

3.1) Reuse or build collection instruments: This sub-process describes the activities to build and reuse the collection instruments to be used during the "Collect" phase. A collection can receive data in various ways (e.g., SDMX web services) and collection instruments may also be data extraction routines used to gather data from existing statistical or administrative registers.
3.2) Reuse or build analysis components: This sub-process includes the practices that are needed for reuse existing components or build new ones that are necessary for the subsequent phases of the statistical process, in particular for the “Process” and “Analyse” phases.
3.3) Reuse or build dissemination components: This sub-process includes the practices that are needed for reuse existing components or build new ones that are necessary for the dissemination of statistical products such as those that provide web services, linked open data outputs, geospatial statistics, or maps.
3.4) Configure workflows: It configures the workflow, systems and transformations used within the business processes, from data collection through to dissemination.
3.5) Test production systems: This sub-process provides technical testing and checks of new programmes and routines developed and the confirmation that existing programmes from other statistical business processes are suitable for use and makes sure that the whole production solution works correctly.
3.6) Test statistical business process: This sub-process describes the activities to manage a field test or pilot of the statistical business process. It usually provides a study and various tests on collected data and on the collection instruments, to be sure that the statistical business process performs as expected.
3.7) Finalise production systems: This sub-process provides to insert correctly configured processes and services, new and modified ones, into a production environment, with the creation of the relative documentation.
The following table describes mappings of the tools functionalities to the sub-processes related to this specific GSBPM phase:

Sub-process	Related Tool	Rationale
3.6) Test statistical business process	SPARQL React	This application consumes and display data and can be used as a baseline for use cases design and development as it allows the data to be used for specific purposes.
3.7) Finalise production systems		This sub-process is not related to the functionality of a specific tool but to all the activities necessary to put the

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

		processes and services into the production environment

Table 16 - Build Phase Mapping

2.4 Collect Phase

This phase deals with the collection of all the data and metadata necessary to analyse the phenomenon, using various methods; this information is then uploaded to the appropriate environment for further processing and analysis. It can include validation of data set formats, but it does not include any transformations of the data.

Below is a brief description of each of its subprocesses:

4.1) Create frame and select sample: This sub-process identifies and specifies the population of interest, defines the sample and determines the most appropriate sampling criteria and methodology.
4.2) Set up collection: The second sub-process ensures that people and technology are ready to collect information, in all the ways foreseen. It takes place over a set period of time, as it includes preparation of collection tools as well as the guarantee of transparency and confidentiality of the data and metadata to be collected.
4.3) Run collection: The third sub-process identifies the tools used to collect the information, includes the request for raw or aggregated data, as well as all associated metadata, includes manual data entry or the management of field survey activities, depending on the origin and method of collection and the monitoring of data collection.
4.4) Finalise collection: This sub-process includes loading the data and metadata into a suitable environment for further processing; it may include the conversion of the formats of files and analysis of the metadata associated with collection to make sure to have met the requirements.

The following table describes mappings of the tools functionalities to the sub-processes related to this specific GSBPM phase:

Sub-process	Related Tool	Rationale
4.2) Set up collection	Idra	This application is concerning the data acquisition and collection: it can federate existing ODMS and standardise the representation of collected open datasets.
4.3) Run collection	Datalift	This application provides the manual (or automatic) data entry at the point of contact: specifically

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

		requires an RDF file as input and it may also include the monitoring of data collection.
	SparQLing	This application provides the manual (or automatic) data entry at the point of contact: specifically requires a Graphol file as input and it may also include the monitoring of data collection.
	Eddy	This application provides the manual (or automatic) data entry at the point of contact: specifically requires a Graphol file as input and it may also include the monitoring or editing of data collection.
4.4) Finalise collection	Excel/CSV to NGSI-LD	This tool provides the functionality of converting the format of an input file, specifically from Excel or CSV formats to NGSI-LD standard.
	Eddy	This tool provides the functionality of converting the format of an input file, specifically from Graphol language to OWL.
	Meta&Data Manager	This application allows the analysis and management of structural metadata associated with data.

Table 17 - Collect Phase Mapping

2.5 Process Phase

This phase deals with manipulating the collected data and is characterized by sub-processes that transform and clean them so that they are ready to be analysed and disseminated.

Below is a brief description of each of its subprocesses:

5.1) Integrate data: This sub-process integrates data from one or more sources through the data combination that comes from multiple sources, integrating geospatial and statistical data or other non-statistical data and performing data fusion. The result is a set of Linked Data.
5.2) Classify and code: This sub-process classifies and codes the input data with automatic coding routines or manual processes.
5.3) Review and validate: This sub-process analyses data to identify potential problems, errors and discrepancies such as outliers and miscoding. Its function is about identifying actual or potential errors, but their correction is done in the next process.

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

5.4) Edit and impute: The purpose of this sub-process is to insert or remove data where is considered incorrect, missing, unreliable or outdated.
5.5) Derive new variables and units: This sub-process derives data for variables that are needed to deliver the required outputs, but which are not explicitly present in the collection.
5.6) Calculate weights: This sub-process is directly connected with 2.5 (Design processing and analysis) in which the system of methods had been outlined to create weights for unit data records; weights can be used for normalization purposes or to make data representative of the target population.
5.7) Calculate aggregates: This sub-process allows to obtain aggregated data starting from microdata or lower-level aggregates. It determines measures of average and dispersion starting from summing data for records sharing certain characteristics.
5.8) Finalise data files: This sub-process brings together the results of the other sub-processes in this phase in a data file, which is used as the input to the "Analyse" phase.
The following table describes mappings of the tools functionalities to the sub-processes related to this specific GSBPM phase:

Sub-process	Related Tool	Rationale
5.1) Integrate data	Eurostat NSI WS	This application provides the integration of data from different sources to create integrated statistics as output.
	Meta&Data Manager	This application provides the integration of data from different sources to manage and public data and structural metadata as output.
	Datalift	This application aims to create, starting from raw data, a set of Linked Data.
	Juma	This application aims to create, starting from raw data, a set of Linked Data.
5.4) Edit and impute	Bauhaus	This tool allows to edit and import the metadata in input in RDF syntax.
	Meta&Data Manager	This tool allows to manage edit and import the data and metadata in input.
5.5) Derive new variables and units	SparQLing	This application allows to derive intermediate data needed to deliver the required outputs: these data represent a SPARQL query useful for querying the triple store.
	Excel/Csv to NGSI-LD	This application allows to derive intermediate data needed to deliver the required outputs: these data

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

		are Entities directly inserted as input of the Context Broker.

Table 18 - Process Phase Mapping

2.6 Analyse Phase

In this process, statistical content is created and analysed in depth. The purpose of this phase is to make sure that at the end of it the data is ready for the dissemination to users; it also includes the sub-processes and activities that enable statistical analysts to understand the data and the statistics produced.

Below is a brief description of each of its subprocesses:

6.1) Prepare draft outputs: It is in this sub-process that the data coming from the previous sub-processes are transformed into statistical outputs such as indexes, trend, accessibility measures, etc. Furthermore, this is where maps and geo-statistical services can also be made to maximize the capacity to analyse the statistical information.
6.2) Validate outputs: This sub-process is where statisticians validate the quality of the outputs produced, in accordance with expectations and with a general quality framework.
6.3) Interpret and explain outputs: In this sub-process an in-depth analysis of the statistical contents is performed: statisticians observe the statistics from all perspectives using different tools and methodologies.
6.4) Apply disclosure control: It verifies that the data and metadata to be disseminated do not violate all the rules on confidentiality according to either organization policies or to the process-specific methodology created in sub-process 2.5 (Design processing and analysis).
6.5) Finalise outputs: In this process statisticians verify that the statistical contents produced reach the required quality level and are thus ready for use by also producing the related supporting documentation.
The following table describes mappings of the tools functionalities to the sub-processes related to this specific GSBPM phase:

Sub-process	Related Tool	Rationale
6.1) Prepare draft outputs	Eurostat NSI WS	This tool allows to obtain statistical outputs in SDMX-ML format.
	Meta & Data Manager	This tool can model and produce statistical multidimensional tables.

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

6.3) Interpret and explain outputs	Bauhaus	This application allows an in-depth management of the linked metadata in input.
	Cube Visualizer	This application allows to view the statistics from all perspectives by creating charts to display data.
	Olap Browser	This application allows to view the statistics from all perspectives by creating tables in two dimensions to display data.

Table 19 - Analyse Phase Mapping

2.7 Disseminate Phase

This phase deals with releasing products and statistical content to users, supporting them in the access and use of these products. Its sub-processes concern all activities associated with assembling and releasing a set of static and dynamic products via a range of channels.

Below is a brief description of each of its subprocesses:

7.1) Update output systems: This sub-process concerns the updating of systems, formatting, loading and linking of metadata activities and also their final check; activities include the management of databases where data and metadata are stored in order to be ready for the dissemination phase.
7.2) Produce dissemination products: This sub-process finalizes the products designed in sub-process 2.1 (Design outputs) to realize the user needs; and they could include printed publications, press releases and websites. The products can take many forms including interactive graphics, tables, maps, public-use microdata, linked open data and downloadable files.
7.3) Manage release of dissemination products: This sub-process manages the release of dissemination products including managing the timing of the release; it also includes the provision of products to subscribers. It also includes the possibility for a statistical organization to retract a product if an error is detected.
7.4) Promote dissemination products: This sub-process is relative to the active promotion of the statistical products produced in a specific statistical business process, to help them reach the widest possible audience.
7.5) Manage user support: This sub-process ensures that user requests in terms of services to be provided and access to data are made correctly within the agreed deadlines. The answers to the most frequently asked questions can be collected on a public page can also be used to populate a knowledge database.

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

This sub-process also includes managing support to any partner organizations involved in disseminating the products.
The following table describes mappings of the tools functionalities to the sub-processes related to this specific GSBPM phase:

Sub-process	Related Tool	Rationale
7.2) Produce dissemination products	Eurostat NSI WS	This tool represents a dissemination environment for distributing data in SDMX-ML format to potential data consumers.
	CEF Context Broker	This tool allows the dissemination of context information to external systems following the ETSI NGSI-LD specification.
	Data Browser	This tool allows data-users to browse, visualize and disseminate datasets based on the SDMX standard.

Table 20 - Disseminate Phase Mapping

2.8 Evaluate Phase

This is the phase of evaluating the results obtained in terms of evaluating the success of the statistical business process and evaluating any improvements and changes. This phase can take place at the end of the instance of the process or on an ongoing basis during the statistical production process.
Below is a brief description of each of its subprocesses:
8.1) Gather evaluation inputs: This sub-process aims to understand the degree of user satisfaction and as part of this phase quality indicators are compiled making them available for the team producing the evaluation. This evaluation process can be automated and take place in a continuous way throughout the whole process.
8.2) Conduct evaluation: This sub-process analyses the evaluation inputs, compares them to the expected results and this analysis is documented within an evaluation report.
8.3) Agree an action plan: In this phase, an action plan based on the report drawn up in the previous sub-process is created. It contains the actions to be taken and it also considerations of a mechanism for monitoring the impact of those actions.
[13]

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

3 INTERSTAT use case ontologies

The following sections describe, for each use case defined in INTERSTAT, the different data sources and related datasets, the exiting ontologies and data models and how they can be integrated and extended to enable cross border interoperability. In this document is reported only the initial phase of this process that will be further refined and completed in the next months before the official start of the project pilots. Moreover, the different ways in which the use cases ontologies are documented and analysed is related to the fact that the different institution and groups that worked on these activities, used, in this initial phase, different tools and methodologies. The different approaches will be harmonised ones the INTERSTAT framework will be ready and will be directly used by the partners.

3.1 The School For You

3.1.1 Overview

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 field and for structural characteristics. A parent who must choose the school for their children needs to know not only the location, but also the educational services that the school offers and the results that the students may achieve in that school. n It is also interesting for a student, who must choose the address of university studies, to know the career opportunities and demands of the job market in a particular area. Finally, the comparison on the dynamics of foreign work or training orientations for example in neighbouring France could provide useful information for the choice.

For policy makers knowing the concentration of students in particular areas can be exploited to define mobility policies and construction of sports facilities or services for young people.

Useful information to meet these needs is available on various portals of statistical institutes and the ministry of education or other public institutions. In addition to being distributed on the WEB on separate sites, the data are often described by non-harmonized classifications and characterized by distinct access methods and the data are made available in different formats. The purpose of this Use Case is therefore:

- Search for sources that meet the training need formulated in this use case
- Outline the original source
- Describe the process of harmonization of the original sources to obtain a complete integration
- Provide the integrated scheme from the conceptual point of view through the ontologies but also the transformations and physical integrations of the data

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

- Provide a technological scheme to achieve the physical integration of data in all their different structures
- Formulate a possible single data access system that resolves the heterogeneity of the diffusion formats of the different sources.

3.1.2 Data and Metadata models

Candidate data sources to be modelled and integrate for this use case are:

1. Education data from Eurostat (EU-EDU).
2. Labour Market data from Eurostat (EU-LFS).
3. Italian Ministry of Education (MIUR) Opendata [24].

Other candidate data sources not yet analysed are:

4. French Ministry of Education [25].
5. Italian Evaluation Institute (INVALSI) [26].
6. OECD PISA evaluation [27].

Almost all data sources contain macrodata, with different aggregation levels, therefore our goal will be to build a generalized model of common macrodata. Dimensional Fact Model as methodology and graphical language was chosen for this purpose [28].
In order to decide which datafiles of sources have to be integrated and at which dimensional level, then macrodata datafiles have to be modelled using Dimensional Fact Model diagram (DFM).
When the datafiles contain microdata, they will be modelled as dimensional attributes associated to the Statistical Unit.
As next step it will be possible to model datafiles to create a generalized Dimensional Fact Model (generalized DFM).
Every structure in the source DFM must be considered as a data source of the generalized Dimensional Fact Model.
The data in every structure of the generalized DFM are provided by many structures from different sources DFM.
For Dimensional attributes, as they are discrete variables, provisioning from source DFM to generalized DFM is operated by a Harmonization function, which transforms values of source domains to generalized domain. This function may be implemented by a decoding table.

Education data from Eurostat

Following the reference framework that describes the INTERSTAT data services, the domains to be integrated into the “School For You” service concern demographic data, the level of education of the

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

population and the training of students, as well as the schools and training institutes present on the national territory of the Italy and France.

Therefore, the following data may be integrated for each territorial level:

1. Population density by age and gender and level of education (e.g., from the 2011 census) in Italy and France
2. Density of schools by address, number of students and results obtained (in Italy available at MIUR and at INVALSI, in France at Ministry of Education, at state level at OECD-PISA)
3. Density of employees by business sector and level of education (ISTAT human capital).

Education-and-training

At the European level, the subject of training, as the level of education and qualification, was comprehensively surveyed on the entire population by the 2011 census and annually is detected by all European NSIs whit the Education and training statistics and the Labor Force survey. Both sources are published on the EUROSTAT website (see "Browse Statistics by Theme"), where data from all NSIs are harmonized with shared classifications.

The Education and training statistics indicators provide information on the participation of individuals in education and training activities, education financing and teaching staff as well as outcomes of education, like in the links [29] and [30].

In this experience, we concentrate on:

- **Participation in Education and Training** [30].

The population classified by ISCED 2011:

Measure of Population (Number of)	Dimensions	Hypercube Code	Time
Pupils enrolled in early childhood education	sex, type of institution and intensity of participation	educ_uoe_enrp01	2012-2019
	sex and age	educ_uoe_enrp02	2012-2019
	sex and NUTS2 regions	educ_uoe_enrp03	2013-2019
Pupils enrolled in primary education	sex, type of institution and intensity of participation	educ_uoe_enrp04	2012-2019
	sex and age	educ_uoe_enrp05	2012-2019
	sex and NUTS2 regions	educ_uoe_enrp06	2013-2019
Pupils enrolled in lower-secondary education by programme orientation	sex, type of institution and intensity of participation	educ_uoe_enrs01	2012-2019
	sex and age	educ_uoe_enrs02	2012-2019

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

Pupils enrolled in upper-secondary education by programme orientation	sex and NUTS2 regions	educ_uoe_enrs03	2013-2019
	sex, type of institution and intensity of participation	educ_uoe_enrs04	2012-2019
	sex and age	educ_uoe_enrs05	2012-2019
	sex and NUTS2 regions	educ_uoe_enrs06	2013-2019
Pupils enrolled in post-secondary non-tertiary education by programme orientation	sex, type of institution and intensity of participation	educ_uoe_enrs07	2012-2019
	sex and age	educ_uoe_enrs08	2012-2019
	sex and NUTS2 regions	educ_uoe_enrs09	2013-2019
Pupils enrolled in vocational upper secondary and post-secondary non-tertiary education by education level	sex and field of education	educ_uoe_enrs10	2013-2019
Students enrolled in tertiary education by education level	programme orientation, sex, type of institution and intensity of participation	educ_uoe_enrt01	2005-2019
	programme orientation, sex, and age	educ_uoe_enrt02	2012-2019
	programme orientation, sex, and field of education	educ_uoe_enrt03	2012-2019
	programme orientation, sex and NUTS2 regions	educ_uoe_enrt06	2012-2019
Ratio of the proportion of tertiary students over the proportion of the population	NUTS1 and NUTS2 regions	educ_uoe_enrt05	

Table 21 - The population classified by ISCED 2011

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

Source Dimensional Fact Model

The hypercubes of **Education and Training** data released by the Eurostat website that were chosen for this experimentation are modeled according to the Dimensional Fact Model, as follows:

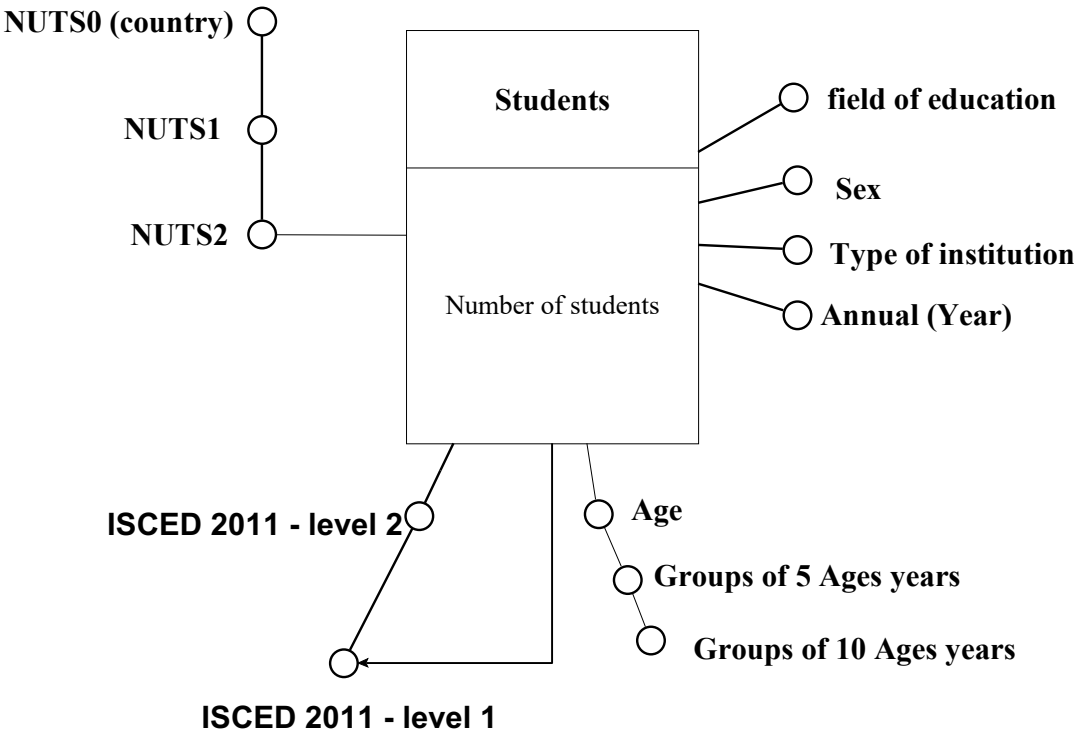


Figure 16 - Hypercube Education and training

Metadata description

In general, the Metadata of the hypercubes published in the Eurostat website are archived in [31].

In particular, the Metadata referred into this work are:

- **Intensity of participation** is annual and referred to the variable **Time**;
- **Field of education and training (ISCED-F 2013)** is defined as "a coherent set or sequence of educational activities designed and organized to achieve pre-determined learning objectives or accomplish a specific set of educational tasks over a sustained period of time" (http://uis.unesco.org/sites/default/files/documents/isced-fields-of-education-and-training-2013-en.pdf);

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

- **ISCED 2011** is the code-list referred to the classification "International Standard Classification of Education" (ISCED), the standard framework used to categorise and report cross-nationally comparable education statistics [32, 33];
- **Type of institution:**

  [PUBL] Public institutions

  [PRIV] Private institutions

  [PRIV_DEP] Private government dependent institutions

  [PRIV_IND] Private government independent institutions

- **Age:** this code-list is characterized by the single year end a group of 5 years, with the exception of the initial and final ages, in particular the classification begins with the list code "less than 2 years" and ends with "65 years or over" (i.e., Age for primary education Or Age for tertiary education)

Labour Market data from Eurostat

The EU Labour Force Survey (EU-LFS) is the largest European household sample survey. Its main statistical objective is to classify the population of working age (15 years and over) into three mutually exclusive and exhaustive groups: **employed persons, unemployed persons,** and the population outside the labour force. The employed population are distinguished in "Self-Employees" and "Employees" by educational qualification, sex and occupation type**,** as reported in [34].

Measure of Population (Number of)	Dimensions	Hypercube Code	Time
Employment	sex, occupation, and educational attainment level	LFSA_EGISED	2011-2020
Self-Employment	sex, occupation and educational attainment level	LFSA_ESGAED	2011-2020
Employees	sex, occupation and educational attainment level	LFSA_EEGAED	2011-2020

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

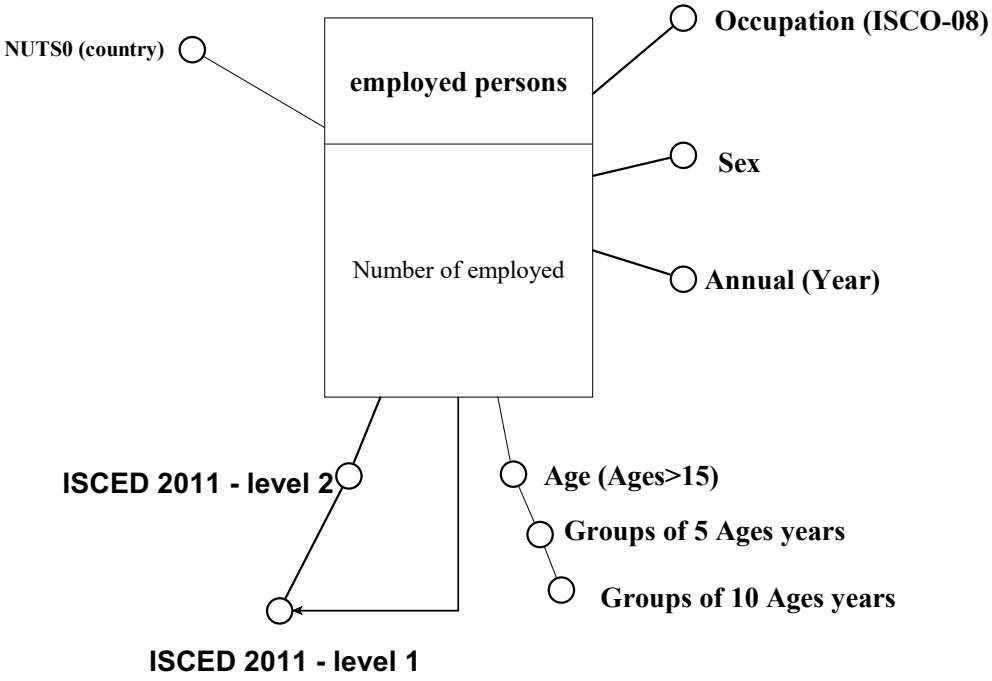


Figure 17 - Hypercube Employment

In these hypercubes we have only:

- **International Standard Classification of Occupations 2008 (ISCO-08)** [35]**:** is an International Labour Organization (ILO) [36] classification structure for organizing information on labour and jobs [37]. It is part of the international family of economic and social classifications of the United Nations [38].

MIUR Data source

Macrodata datafiles are:

Measure of Population (Number of)	Dimensions	Data source (or Source DFM)	Scholastic years covered
Students in Primary (Except nursery school) and Secondary education	• Scholastic year • Sex[1] • Type of institution • Education grade • School plexus code • Class level	ALUCORSOINDCLA	From 2015-2016 To 2019-2020

[1] Female and male measures are provided in different fields of same row

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

	• Scholastic year • Citizenship (ITA/EU/Not EU)[2] • Type of institution • Education grade • School plexus code • Class level	ALUITASTRACIT	From 2017-2018 To 2019-2020
	• Scholastic year • Sex[3] • Type of institution • Education grade • School plexus code • Class level • Kind of high school • Address of high school	ALUSECGRADOIND	From 2015-2016 To 2019-2020
Total square meter of building, **Free square meter of building,** **Volume of building**	• School plexus code • Building code		From 2015-2016 To 2019-2020

Table 22 - Macrodata datafiles

Datafiles are physically provided by one datafile for each scholastic year and different data file for public schools and for private school equivalent to public ones.

Sources DFM schemata are outlined in the following pictures, with original language names (Italian)**:**

[2] provided in different fields of same row
[3] Female and male measures are provided in different fields of same row

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

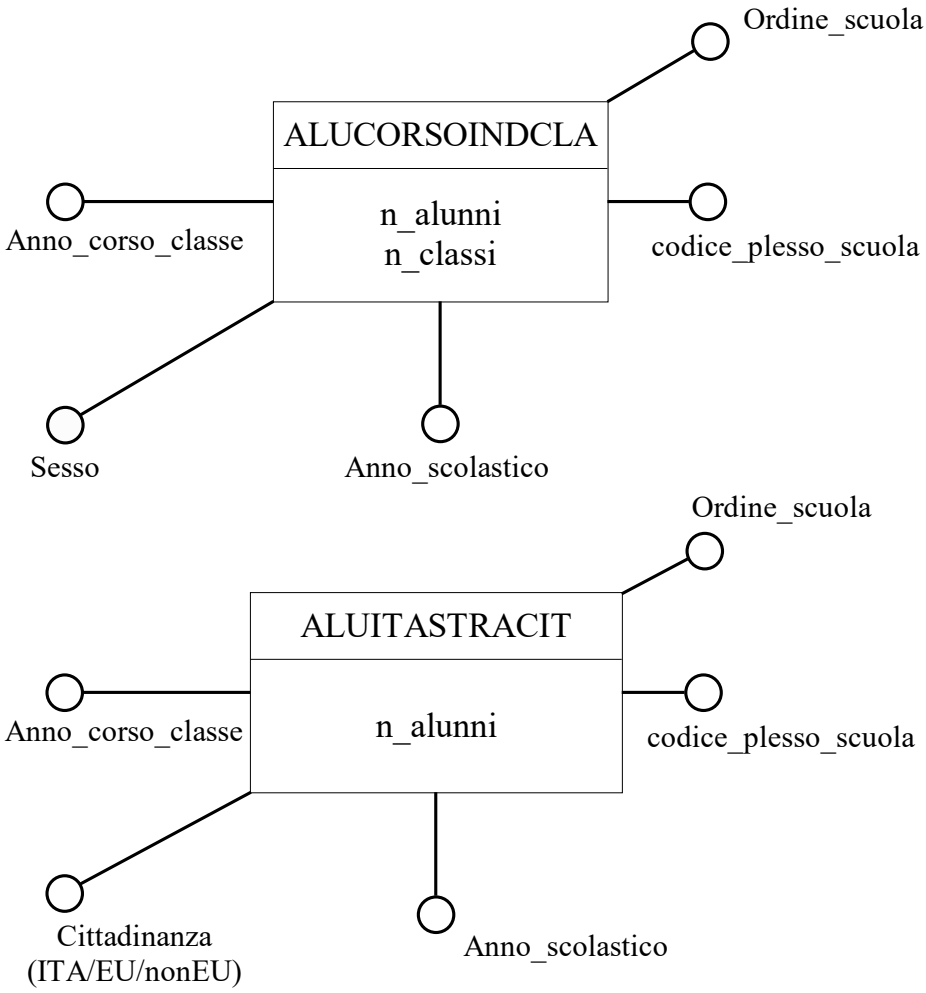


Figure 18 - Hypercube "Alunni"

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

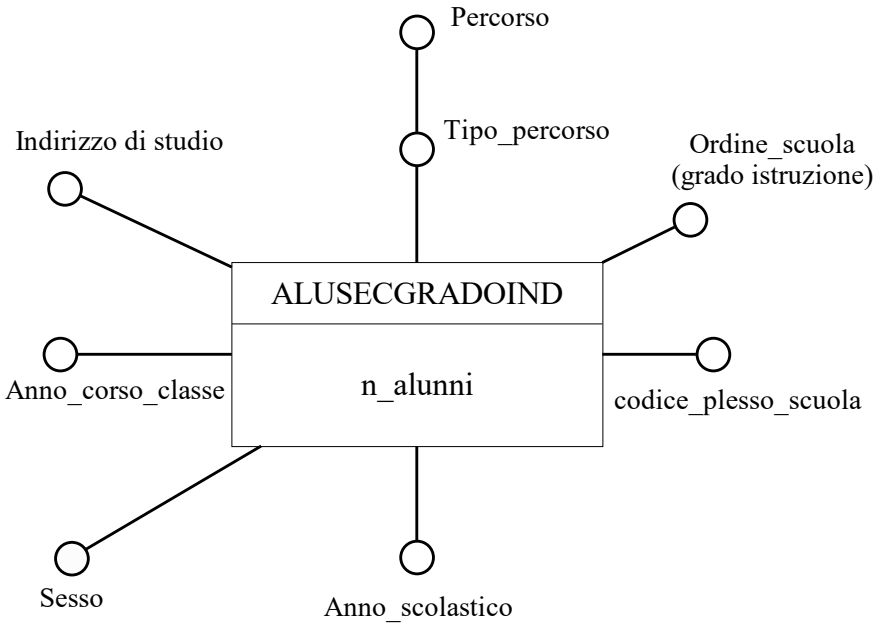


Figure 19 - Hypercube "Alunni 2"

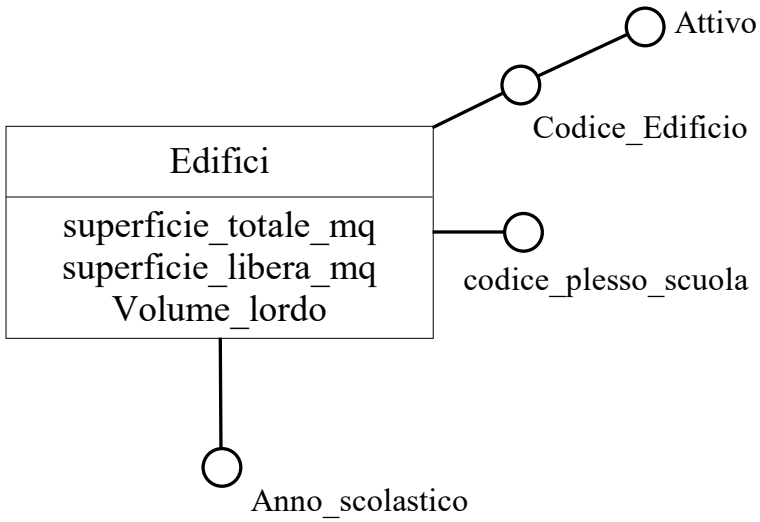


Figure 20 - Hypercube "Edifici"

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

Microdata datafiles are:

- School registry files: registry school data are provided by one file for each scholastic year; files are different (with different structures) for public schools and private schools equivalent to public ones
- School building registry files: registry school building data are provided by one file for each scholastic year. They contain only code of school plexus, code of building and information on the building is active or not

Here registry files schemata outlined as Dimensional attribute tree (with original language names):

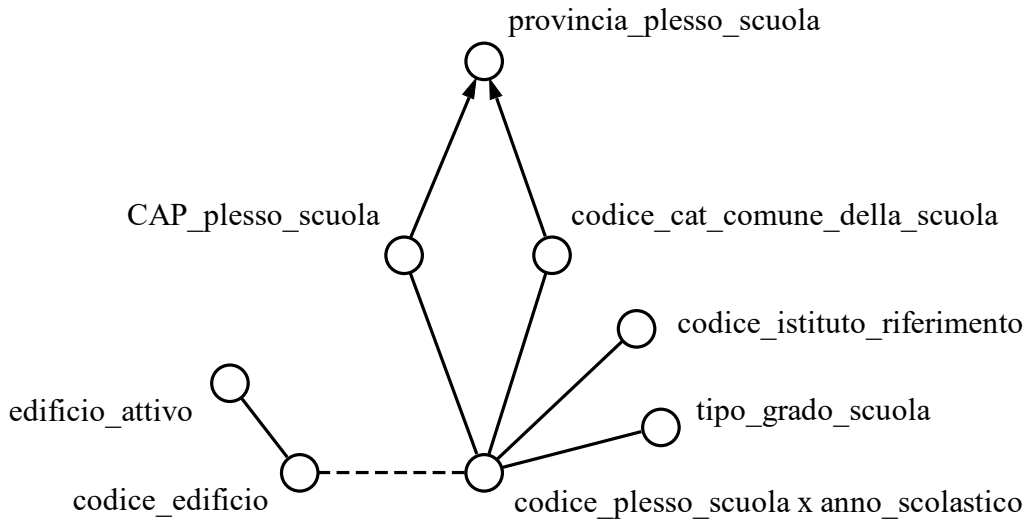


Figure 21 - Hypercube "Plessi"

Every source DFM of MIUR having "codice_plesso_scuola" attribute can be linked with "codice_plesso_scuola x anno_scolastico" in this schema.

Student assessment data in Italy source (INVALSI tests)

INVALSI test is a national-level system of assessment of the quality of learning outcomes of Italian schools. It is carried out yearly, and measures performances in Italian, Math and English with tests that are submitted to all students who attend grades 2, 5, 8, 10 and 13. In additions, forms are to be filled in by teachers and school headmasters.
Results of those tests and other data are available on [39]; registering to the portal and submitting a written and signed request are conditions for downloading individual data.
The data still has to be analysed in depth, however, for the moment we can highlight the following:

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

File	Records	Information provided
School file	1 for every school	School id, province (NUTS-3), municipality, number of classes in school, number of students, official language (IT, DE, SL)
Class file	1 for every class involved in the test	School id, grade, number of students, number of available pcs
Regional files	1 for every region (NUTS-2) and subject	Region (NUTS-2), average scores in math, Italian and English (reading/listening).
Student files	1 for every student involved and subject (Italian, math, English-reading and English-listening, the latter 2 except for grade 2)	Student id, class id, month and year of birth, sex, place of birth, eventually: age at arrival in Italy, Parents' place of birth and profession, school marks in Italian math and English, results for every single question in the test for the given subject, global score in the given subject (different indicators).
Teacher files	1 for every teacher	Teacher id, school id,
Headmaster files	1 for every school	School id, opinions and feedback about the survey, features of school (e.g., IT equipment), social problems (e. g. poor school attendance, lack of discipline), sex/age makeup of teachers, a. s. o.
Teacher files	1 for every teacher and class	School id, subject taught, teacher's curriculum, opinions and feedback about the survey, features of school (e.g., IT equipment), information about interaction with colleagues and headmaster.

Table 23 - Available files

Possible connection to other sources:

Information provided can be linked with other sources via NUTS-3 and/or municipality.
Although information is provided at a school level, linking schools with other sources seems not to be possible, both for technical as well as for legal reasons:

- the identification codes are defined internally, and there is no other secure information which could identify the school, such as the school's name;
- the attempt to identify single schools is not allowed according to the terms of service, and this seems to include not only students and classes, but schools as well (for understandable reasons).

To-do:

Among the tasks that still must be done to include this source in our model, we would like to highlight following ones:

- Choose most interesting and valuable information;
- Finding out whether there is available data already aggregated at municipal or NUTS-3 level;
- Finding out how to calculate average scores, in order to retrieve average scores by varying dimensions;
- Clear any doubts about the impossibility to link single schools with other sources' schools;
- Choosing most significant performance indicators provided, especially with regard to possibility of comparing them with international data;

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

- Finding out how to link INVALSI data with equivalent French data.

3.1.3 Ontologies and mappings

This table define the mapping to link generalized Dimensions and measures with source Dimension and measures.

Generalized structure	Kind of structure	Data source (or Source DFM)	Source field	Harmonization notes
Number of students	measure	ALUCORSOINDCLA	N_alunni	
Number of students	measure	ALUITASTRACIT	N_alunni	
Number of students	measure	ALUSECGRADOIND	N_alunni	
Number of students	measure	ALUITASTRACIT	N_alunni	
Reference_Year	Dimensional Attribute	ALUCORSOINDCLA	Anno_scolastico	Choice the first year of the couple
Reference_Year	Dimensional Attribute	ALUITASTRACIT	Anno_scolastico	Choice the first year of the couple
NUTS2	Dimensional Attribute	ALUITASTRACIT	Provincia	
Relative_citizenship	Dimensional Attribute	ALUITASTRACIT	cittadinanza	ITA->autochthonous EU->EU NoEU->NoEU

Table 24 - Mapping to link generalized and source Dimensions and measures.

The Figure 23 shows the generalized DFM as result of integration step, generalized DFM will be extent as much as new sources are integrated.

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

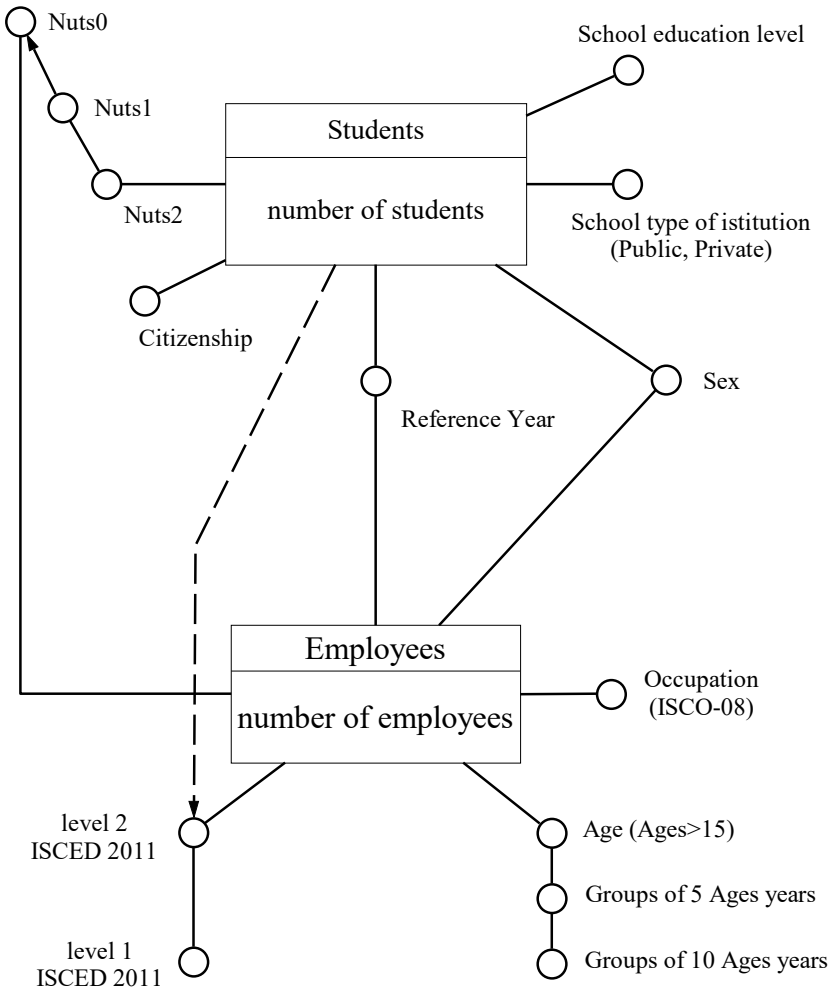


Figure 22 – Generalized Hypercube

An ontology will be defined to semantically describe concept links among data source DFM and generalized DFM.

Following figure shows a draft of this ontology, using the Entity-relationship notation:

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

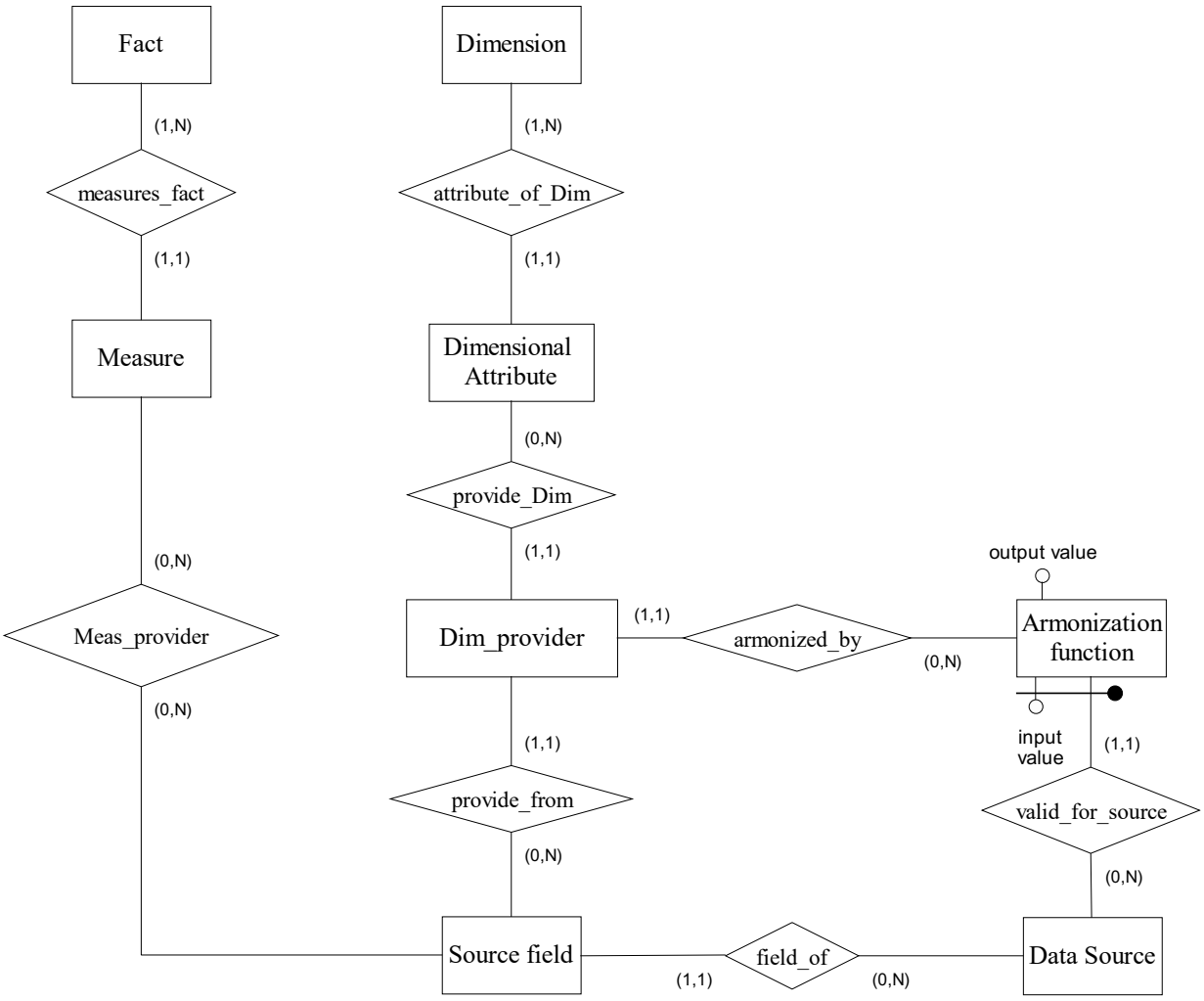


Figure 23- Entity-relationship draft schema of the ontology

3.2 Geolocalized Facilities

3.2.1 Overview

The main objective of the "Geolocalize Facilities" (GF) pilot 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. In the context of the INTERSTAT project, integration and contextualization is provided by the CEF Context Broker, and we also want to maximize the benefits of cross-border interoperability, as explained in deliverable 3.1.

Here, "facilities" (or "equipment", both terms will be used interchangeably) are understood as points of services, merchant or not, which are accessible to the public and operate in domains like education, health, social services, transport, sports, leisure, culture, or tourism.

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. She also would like to know what events are programmed in the nearby stadiums, theatres of cultural venues. From the description of locations or events, it should be simple to navigate on the web for further detail (e.g. on artists or sport teams, 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 neighbourhood, 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.

3.2.2 Data and Metadata models

It results from the previous section than the GF pilot requires mainly data on facilities in France and Italy, enriched with demographic data in both countries. Although not required by the user stories as described above, socio-economic data at fine geographic level would be a good example of source for testing data integration in the "local decider" case. Regarding the "visitor" US, integration is done on the client end through the usual mechanisms of the web.

Data on equipment

The French BPE (Base Permanente des Équipements) source will be the backbone of the GF pilot.

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

The BPE is a statistical base that lists a wide range of equipment and services, whether merchant or not, accessible to the public throughout France on January 1st of each year. It covers more than 180 different types of services and equipment, divided into seven main areas: services to individuals, shops, education, health-social, transport-travel, sports-leisure-culture and tourism.

The BPE is constructed from various administrative sources and merges data on access points to services intended for the population, located at fine geographic levels: municipalities, sub-municipal territories (Iris) and coordinates (x, y) in most cases.
Thanks to the detailed knowledge of the territories that it allows, the BPE constitutes a privileged decision support tool. In particular, it makes it possible to study the structure of the service offer in a territory: volume of equipment, presence or absence, concentration or dispersion, identification of service centres or areas without services, calculation of distances between municipalities equipped and not equipped, calculation of equipment rate by linking equipment and their potential users, constitution of equipment baskets on a particular theme, etc.

The BPE provides a natural integration framework for sources like lists of cultural or sports events, time schedules of educational establishments, or any information related to the facilities listed in the base. A wealth of such sources that can be integrated can be found on open data portals, including the European data portal (category “Education, culture and sports”). Even if ISTAT does not produce any data source as comprehensive as the BPE, comparable datasets exist on some sectors, and thus transnational comparisons can be established for various usages.

The core data model for the BPE is quite simple: the main concepts are i) the equipment itself with its general characteristics (name, address, etc.), ii) the type of equipment, which is a coded property with nearly 200 values, and iii) in the education, sports and leisure domains, specific features that depend on the type of equipment (does a school have a canteen, is a pool heated, etc.). Additionally, geographic coordinates are available for most facilities. Regarding metadata, descriptive information conformant to the SIMS [40] quality reporting standard is available on INSEE’s web site (for the BPE in general [41] and for the 2019 edition of the database [42]), but the same information is also published in RDF form (see for example [43] and [44]), and accessible through INSEE’s metadata API [45]. At the unit level, quality metadata concerning the quality of address geocoding is expressed following a star rating system.

More in detail, the list of variables in the BPE, as well as the possible values for the type of facility, are available from the download page [46].

It should be noted that INSEE also produces from the BPE data in evolution over 5 years (e.g., 2014-2019 for the [46]). This is useful to place the latest figures in a middle-term perspective. An example of such data is the evolution of the amount of equipment by type for various geographic zones (urban areas, departments, etc.).

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

BPE data is available in the venerable dBase format, but also in CSV. Publication as RDF has successfully been tested for the latest version but remains experimental.

The BPE will be associated in the pilot to Italian sources of the same kind. An example of such a source is the data published by the Ministry of Culture [47] about “places of culture” and associated events, available as XML through a REST API. The API is documented in a guide that gives example of queries and returned documents. For each place, details include denomination, description and address, as well as category and type. Type can be for example: Museum, Archaeological Area, Monument, Church, Library, etc., and category refers to a cultural domain and includes: Art, Archaeology, History, Science, Ethnography, etc. Actually, there can be several values for category and type, but the dominant one is specified. It should be noted that the addresses of the places of culture include a “point” element containing latitude, longitude, and altitude (as well as coordinate system), so the Italian cultural facilities are geolocalised, like the facilities from the BPE.

Census data

Concerning the census, the description of data and metadata is provided in the section related to the SEP service and is not repeated here.

3.2.3 Ontologies and mappings

This section describes the main ontologies and vocabularies that will be used to structure the data and metadata associated to the GF pilot. Only the main or specific ontologies will be listed: well-known vocabularies only partially used (Dublin Core, Time, etc.) are not described further.

Data on equipment

For the experimental RDF publication, the core BPE data model was expressed as an OWL ontology with the following overall structure (labels are in French but easily understandable):

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

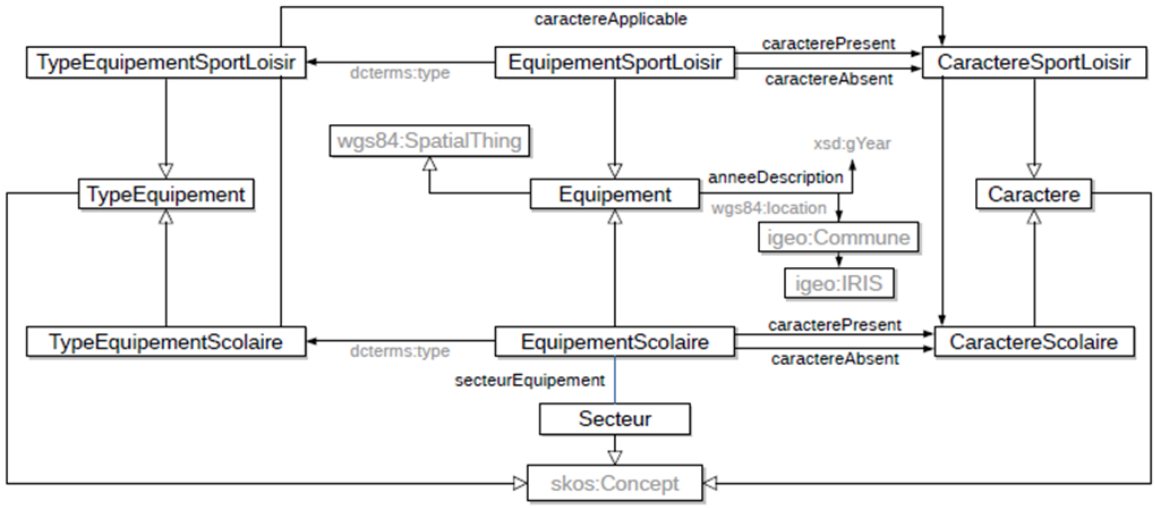


Figure 24 - Ontology on Equipment

This model will be reviewed and adapted to the INTERSTAT pilot context but should not be substantially modified. One of these adaptations can be to connect the BPE ontology to the OGC GeoSPARQL ontology, rather than (or in addition to) the WGS84 Geo Positioning RDF vocabulary (World Geodetic System 1984) in order to associate geometries (here points) to the facilities. This is represented below:

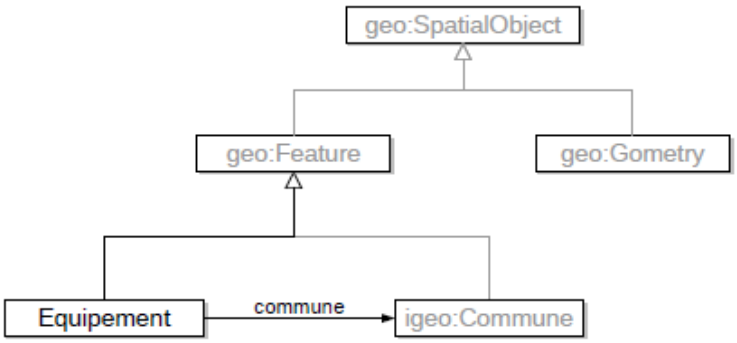


Figure 25 - Connection between ontologies

The geo:Geometry instance will bear coordinates through properties like geo:asWKT or geo:asGML, which allow in particular to specify the coordinate system (the BPE uses several of them). Also, the construction figured above allows the use of GeoSPARQL predicates (contains, covers, touches, etc.) between geometries, for example to check if a given equipment is included in a particular territory.

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

The data model for the Italian data on places of culture is specified by an XML schema (W3C). For example, the “luogo” type structure is (only the mandatory elements are shown):

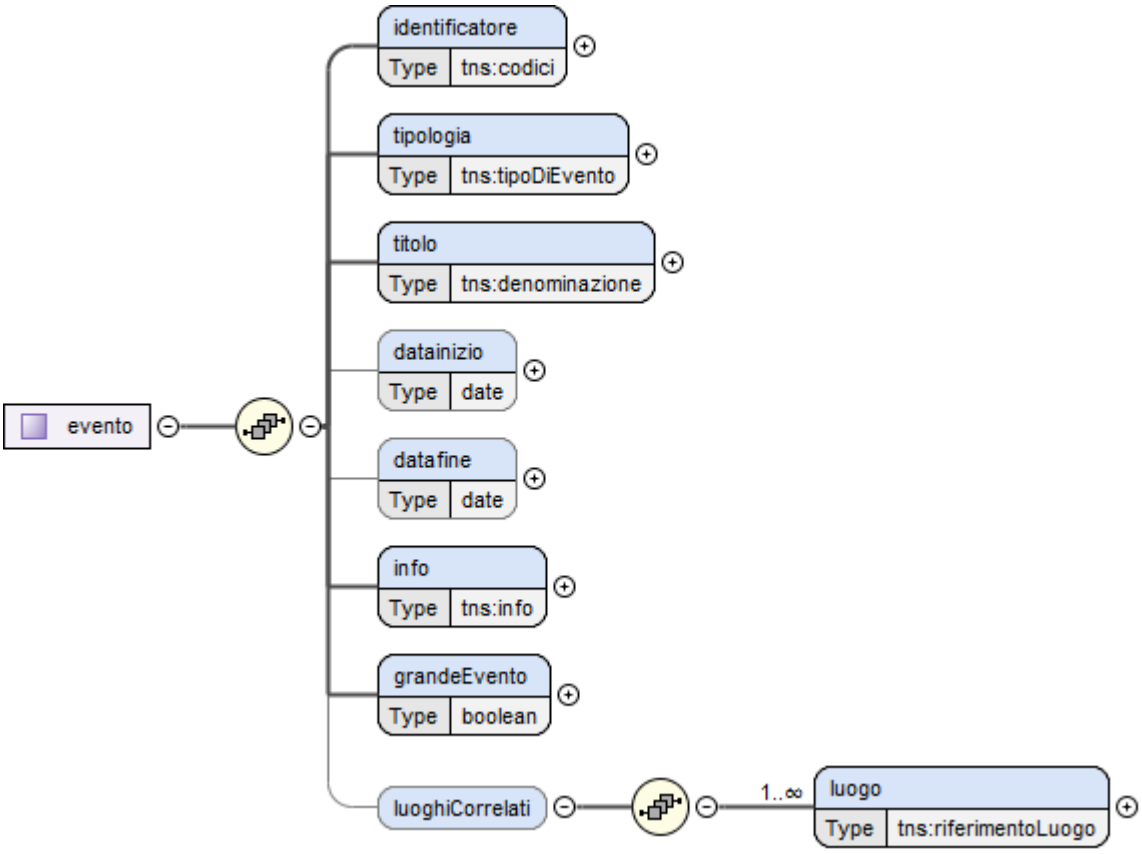


Figure 26 - “Evento” type structure

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

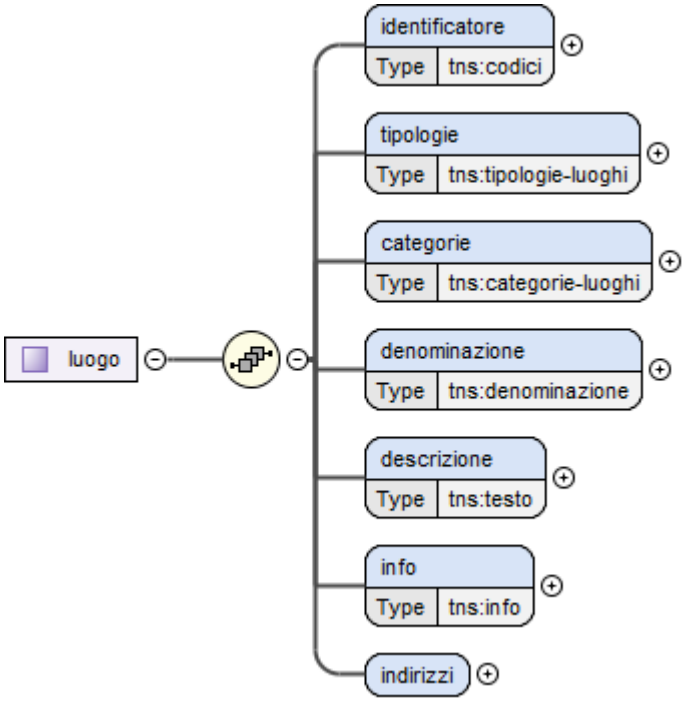


Figure 27 - "Luogo" type structure

This fits in the BPE ontology by defining a third subclass of equipment (EquipmentCulturel) with an associated type that will refer to the category as a specific CaractereCulturel (cultural characteristic).

Event data

Event data are mentioned in the “visitor” use case, but they are not central to the pilot, so a simple model will be used, like the “Event Ontology” or the “Event” construct in schema.org.
In the data from the Italian Ministry of Culture, the events have a quite simple type (some optional elements are omitted):
The type of event is a coded property (Inauguration, Sporting event, Exhibition, Theatrical performance, etc.) whose domain can be expressed as a SKOS concept scheme.

Census data and other data

Census data, as well as data in evolution on equipment and other data from external sources will in general correspond to the “dimensional” (or more specifically “cube”) data model, for which the Dataset part of the Data Cube vocabulary is perfectly adapted and in line with the VTL model that is our reference data model (see deliverable 3.1).

Metadata

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

If we now consider metadata, several vocabularies will be used. For concepts and codes, SKOS will be the reference vocabulary, complemented by XKOS constructs when appropriate. In particular, the code list for the type of facility will be a SKOS concept scheme.

For the specification of cubes, the Data Structure Definition part of Data Cube is easily mapped to the VTL data structure model.

The level of quality of address geocoding for each equipment will be expressed with the help of the Data Quality Vocabulary, and more specifically via instances of dqv:QualityAnnotation targeting the point geometry associated to the equipment as figured below (this mechanism is inherited from the Web Annotation Ontology).

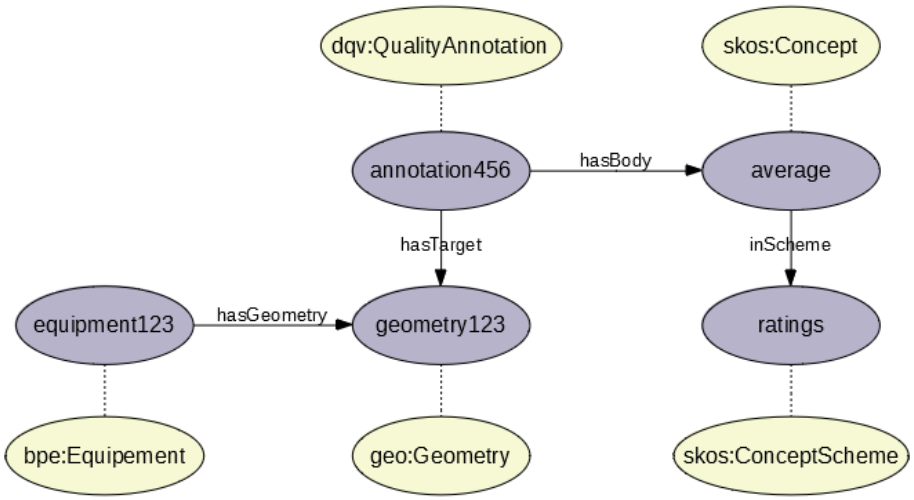


Figure 28 - Geocoding schema

Regarding the descriptive and quality metadata at the source and dataset level, SIMS is not directly expressed as a model but as a SDMX Metadata Structure Definition (see chapter 7 of the SDMX Information Model). The "metadata" part of the SDMX-IM can be adapted in OWL, which allows for more precise data types and improved reusability. Thus, the metadata will take the form of instances of sdmx-mm:MetadataReport grouping metadata attributes corresponding to the different SIMS items. An example of such report has been referenced previously.

Finally, DCAT (or rather DCAT-AP and StatDCAT-AP when possible) and PROV-O will be used as needed for catalogue and process metadata. Also, INSPIRE metadata will be added where appropriate.

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

3.3 Support for Environment Policies

3.3.1 Overview

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: i) sensor data, measuring the concentration of air pollutants and ii) 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: i) the resident population living in areas where air pollutants exceed air quality thresholds; ii) the assessment of the effects of air pollution on vulnerable population groups.

The statistical data service SEP (Support for Environmental Policies) will be implemented to achieve these goals. The SEP service will allow to answer queries combining air pollution and demographic data related to Italian and French territories. In order to test cross-border interoperability, the SEP service could be used to compare air pollution indicators in selected Italian and French areas, such as big cities, particularly Rome and Paris. Concerning cross-domain interoperability, the SEP service could allow to link air quality indicators collected in different stations of each municipality, with the age structure of resident population, to highlight the areas in which the weakest groups are exposed to a higher risk of health issues due to pollution. This analysis could direct the efforts to reduce air pollution in the areas where the concentration of elders is higher. In addition, as one of the main effects of traffic is the increase of NO2 emissions, the proposed data integration could foster the adoption of policies to promote eco-friendly means of transportations.

Before starting any development activity, the following tasks are essential to achieve the semantic harmonization of different domains:

- Exploration of published open data related to the domains of interest
- Inventory of potential data providers to involve
- Selection of specific datasets, description of data models and structures
- Analysis of available ontologies (domain and meta-ontologies), data vocabularies and tools.
-

3.3.2 Data and Metadata models

Starting from the reference framework [48] describing the INTERSTAT data services, the domains to be integrated in the SEP service relate to environmental and demographic data.

Air pollution data

According to current regulations, the main air pollutants monitored to prevent public health issues are:

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

- Nitrogen dioxide (NO2)
- Fine suspended particulate matter (PM2.5)
- Particulate matter with a diameter of less than 10 micro-meters (PM10)
- Ozone (O3).

At national level, the concentration of these particles is gathered through a network of smart devices scattered over the territory. In Italy, regional agencies for environmental protection, collect and publish on their sites daily air quality data. In addition, the Italian Institute for Environmental Protection and Research (ISPRA) collects regional datasets through the InfoAria system. Further, in cooperation with the Ministry of the Environment, ISPRA checks the consistency and completeness of received data, as well as the compliance with recommended formats. Aggregated regional datasets are transmitted to the European Environment Agency (EEA). As reference authority for monitoring and assessing air quality in Italy, ISPRA publishes annual reports, providing downloadable information to the public. The development of national ontologies, as well as the involvement of national stakeholders such as ISPRA are essential to increase semantic interoperability and foster the integration of data collected for different purposes. The following table describes the specific datasets published by ISPRA that could be linked through the SEP service.

Publisher	Dataset	Population	Variables	Unit measure	Classifications and Code-lists	Territorial coverage	Reference time	Data format
ISPRA	QUALITÀ DELL' ARIA AMBIENTE: PARTICOLATO (PM10) – Tabella 1 [49]	Monitoring stations in Italian Municipalities in 2019	PM10 concentration, Annual average value	Micrograms per cubic metre (µg/m3)	Nuts	National	2019	xls
	QUALITÀ DELL'ARIA AMBIENTE: BIOSSIDO DI AZOTO (NO2) – Tabella 1 [50]		NO2 concentration, Annual average value					
	QUALITÀ DELL'ARIA AMBIENTE: OZONO TROPOSFERICO (O3) – Tabella 2 [51]		AOT40 vegetation protection					
	QUALITÀ DELL'ARIA AMBIENTE: PARTICOLATO (PM2,5) – Tabella 1 [52]		PM2,5 concentration, Annual average value					

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

Table 25 - Description of Air pollution datasets published by ISPRA

In France, air quality is monitored through a framework of non-profit licensed associations, Associations Agréées de Surveillance de la Qualité de l'Air ($AASQA_s$) [53]. Each AASQA is managed jointly by national government delegates, regional and local authorities, industries emitting monitored substances, as well as representatives from environmental and consumers advocacy groups and healthcare professions. The activities carried out by the network cover the whole French territory and are coordinated by a central body (Laboratoire Central de Surveillance de la Qualité de l'Air, LCSQA) that is also responsible for scientific and technical support.

The European Environment Agency (EEA) collects air quality information reported by all EU Member States and other reporting countries. Gathered data is published through an air quality database that reports:

- Multi-year time series of air quality data
- Calculated statistics for monitored air pollutants
- Meta-information about the monitoring networks involved (stations and measurements, air quality zones, assessment regimes and compliance attainments reported by data providers).

The main source of environmental data that could be considered for the pilot is: "Air quality statistics calculated by the EEA (AIDE F)" [54]. This dataset contains aggregated concentration or level of air pollutants for both Italian and French territories. The following table describes the specific dataset that could provide French air pollution data to be linked to demographic data through the SEP service.

Publisher	Dataset	Population	Variables	Unit measure	Classifications and Code-lists	Territorial coverage	Reference time	Data format
EEA	Air quality statistics calculated by the EEA (AIDE F)	Air quality measurement stations in reporting countries	Pollutant, AQ values, Sampling Point Local Id, Sampling Point Latitude, Sampling Point Longitude	Micrograms per cubic metre (µg/m3)	Territory, Aggregation Type	EU Member States and other reporting countries	Time series from 2011 to 2019	TSV, CSV

Table 26: Description of Air pollution datasets published by EEA

Census data

The population and housing census provides an overview of the main demographic and social characteristics of persons usually resident in each municipality in a specific reference date.

In Italy, the new permanent census, resulting from the integration of administrative sources and data collected on a representative sample of municipalities and households, allows to publish a set of tables at municipal level. These tables report some of the census variables on an annual basis.

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

In France, the census is currently based on yearly data collection covering all municipalities over a five-year period. Municipalities having less than 10,000 inhabitants conduct a complete enumeration of their population, based on one municipality in five each year. Municipalities with 10,000 residents or more gather information through a sample of addresses, representative of the 8% of their dwellings. The following table reports an overview of the main features of the Italian [55] and French census data, to be linked with air quality indicators.

Publisher	Dataset	Population	Variables	Classifications and Code-lists	Territorial coverage	Reference time	Data format
ISTAT	Demographic characteristics and citizenship	Persons usually resident in Italy in 2017	Age structure, Sex	Sex, Age class, NUTS	National	2018	xls, CSV, PC-axis, SDMX
	Education, work, commuting for studying or working		Current activity status, commuting for studying or working	Current activity status, Location of place of work, school or university, Reason for commuting, NUTS			
Insee	BTX_TD_POP1B_2017 [56]	Persons usually resident in France in 2017	Age structure, Sex	Sex, NUTS		2017	xls
	BTX_TD_POP5_2017		Current activity status, Sex	Current activity status, Sex, NUTS			
	base-excel-flux-mobilite-domicile-lieu-travail-2017 [57]	Active population usually resident in France in 2017	Commuting for working	NUTS			

Table 27 - Description of Italian and French census datasets

SEP Service pipeline

Following the analysis of published data, as well as an inventory of available ontologies, vocabularies and tools, the preliminary step to develop the SEP pipeline is to store the selected datasets to link in a central repository. The main steps to develop the SEP service are:

- Semantic harmonization through the concepts modelled in available ontologies;
- Conceptual mapping between the Air quality ontology and data extracted from the datasets to link;
- Generation of RDF triples.

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

3.3.3 Ontologies and mappings

To harmonize and integrate the information provided by the data sources analysed above, a common representation of the main concepts through the specification of a common ontology is essential. The starting point is an inventory of published ontologies or data models related to the specific domains, to identify the knowledge objects related to the selected use case. The following table reports the ontologies and data models analysed to extract the core concepts for modelling a reference ontology to link the different data sources.

Data domain	Reference ontologies/Data models	Description	Main concepts
Air pollution	AQD data model [58]	It provides some guidelines on how EEA member countries have to report raw data and aggregated statistics to EEA.	Pollutant, Aggregation Type, Media Value, Station, Reporting Year, Time Period, Sampling Point, Concentration
	GeoNames [59] compliant with Inspire Observations/Measures data model	GeoNames ontology describes the domain of the territory in terms of Spatial Object, Geometry, Country, geographic coordinates and so on	Location/Geometry, Longitude and latitude
	SOSA (Sensor, Observation, Sample, and Actuator) Semantic Sensor Network Ontology (SSN) [60]	The ontology describes sensors and their observations, the involved procedures, the studied features of interest, the observed properties. It's included in SSN ontology.	Sensor, Feature of Interest, madeObservation, hasFeatureofInterest
Census data	Census data model	Structure of disseminated data cubes	Total persons usually resident in the analysed area, Age class, Sex, Territory, Current activity status, Commuting

Table 28 - Ontologies and data models of data sources to link

The process design of the Air quality ontology is described in the following ArchiMate[4] [33] diagram: the actor of this process is the ontology designer that analyses the domain and the existing meta-ontologies, vocabularies and data models. The ontology designer produces a list of concepts, roles and attributes that describe the analysed domain. In some cases, the ontology designer reuses existing definitions from available vocabularies to define domain concepts, roles and attributes. The output of this process is an

[4]ArchiMate is an open and independent modelling language, compliant with the Enterprise Architecture standard

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

ontology in OWL2 format. In this case, the application component is implemented by a graphical editor, instead of using a text editor, which facilitates the OWL2 file creation. The ArchiMate objects used in the diagram are described in Annex A.

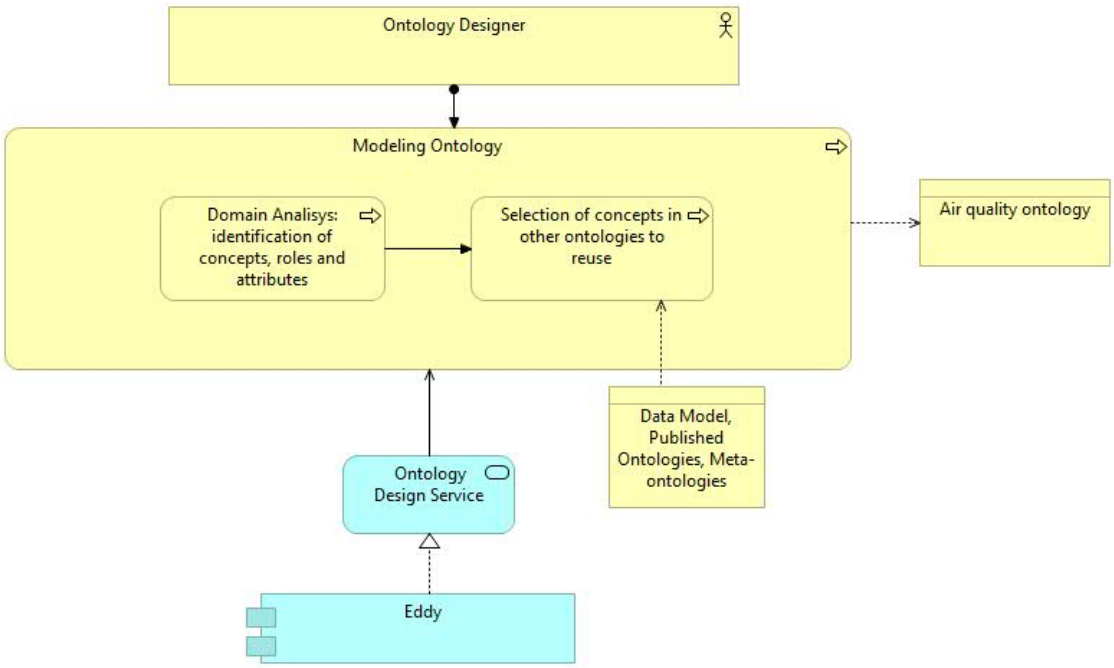


Figure 29 - Process design of the Air quality ontology

The Air quality ontology, modelled in Graphol [61] language and resulting from the combination of relevant concepts listed above is depicted in the following figure. To identify the provenance of each concept extracted from the reference ontologies and vocabularies, and combined in the modelled ontology, the information objects are colour-coded as follows:

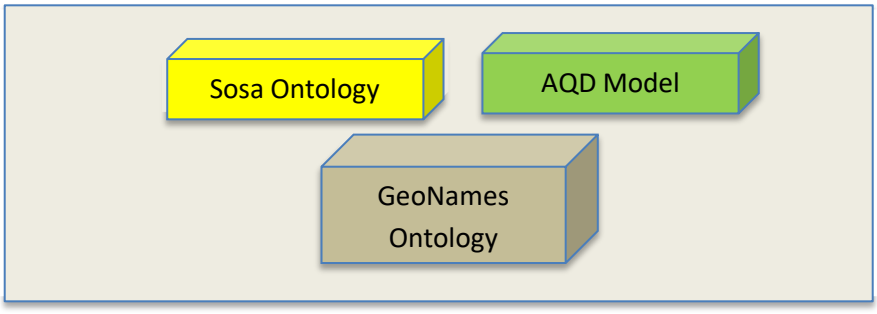


Figure 30 - Ontologies involved in Air quality ontolgy

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

In order to facilitate the visualization, the Air quality ontology is divided in two main parts, according to the analysed domains. The first part refers to air pollution domain while the second part models census data.

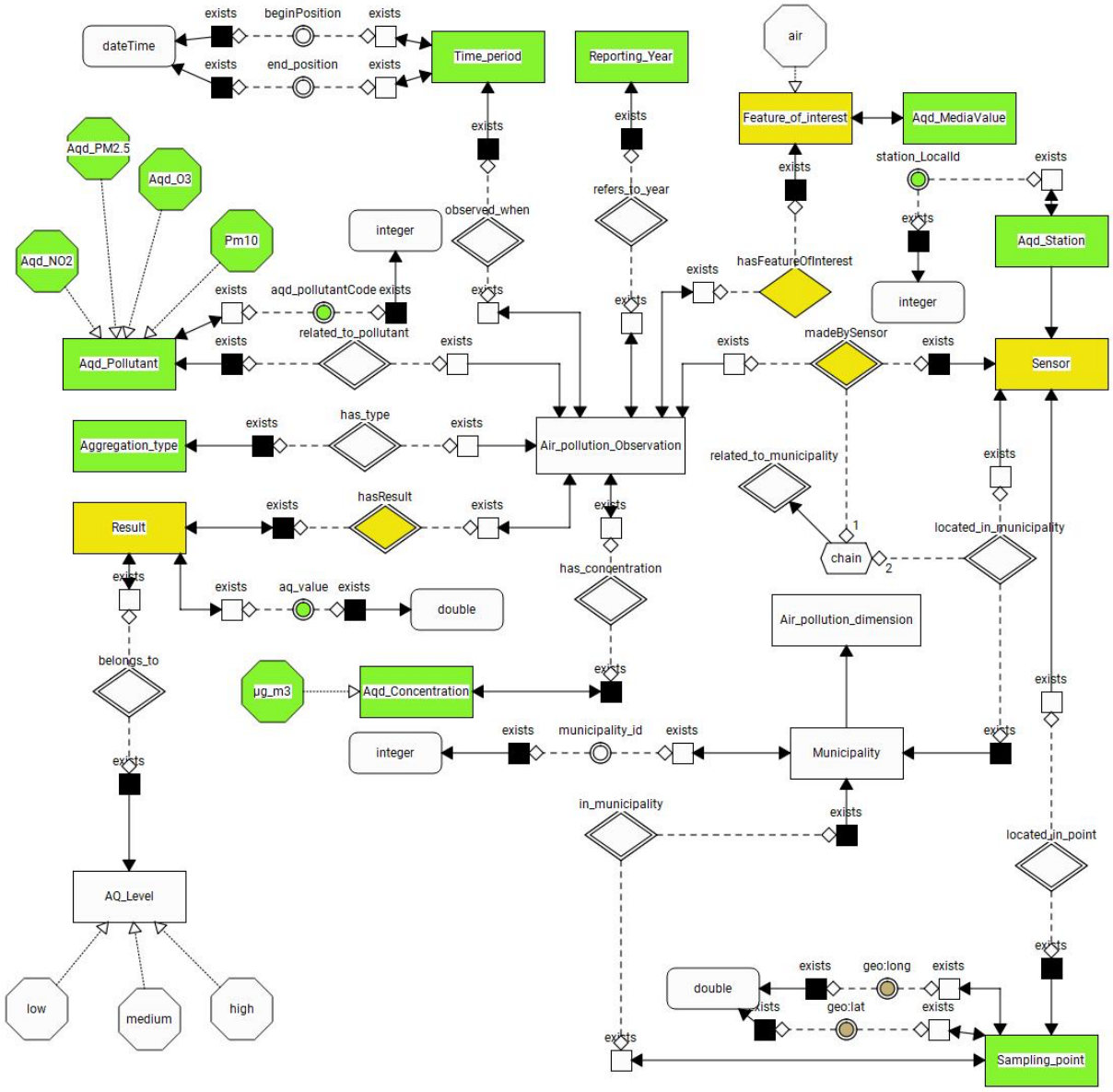


Figure 31 - Air quality ontology

The core concept of the first part is Air Pollution Observation. Each observation is provided by an AQD_ Station, which is a subclass of Sosa Sensor, a device, agent, or software collecting observations which generate some Results. Observation domain is a Feature of Interest of Sosa ontology and corresponds to the air in this use case. The observation refers to one pollutant, for example Pm10. The time plays an

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

important role to describe observations which have a reporting year and a Time_period that represents a specific interval in which observations are gathered.

Observations have a correlated Result object, described by a numeric value (aq_value) that represents the value of observed pollutant. The observation is the result of several measurements gathered in a time interval and aggregated according to a particular aggregation type (for example Monthly average). In order to group the values of observed pollutant in classes, a classification variable (AQ_Level) can be created and linked to each Result. The measurement unit is Aqd_Concentration.

The sensor or station is related to territory in different ways: it can be linked to a sampling point (described by latitude and longitude) or it can be directly linked to its Municipality. This represents an interesting cross-domain concept: Municipality also belongs to census domain. Moreover, through the Nuts (Nomenclature of territorial units for statistics) classification, it is possible to retrieve information about the hierarchical territorial units: Provinces and Regions for Italy, Department and Regions in France. The reference meta-ontology to model the concepts related to Nuts is XKOS.

The following figure shows the concepts that model a subset of census domain to be linked to the Air quality ontology. While in the census ontology, the observation value is directly linked to the census observation, in the Air quality ontology, the observation value is an attribute of Result to link the classification variable (AQ_Level). The concepts refer to census data cubes implemented for dissemination. Census data can be aggregated according to several dimensions. Starting from the selected data cubes, the main dimensions considered for the use case are: Age_class, Current_activity_status, Commuting, Municipality.

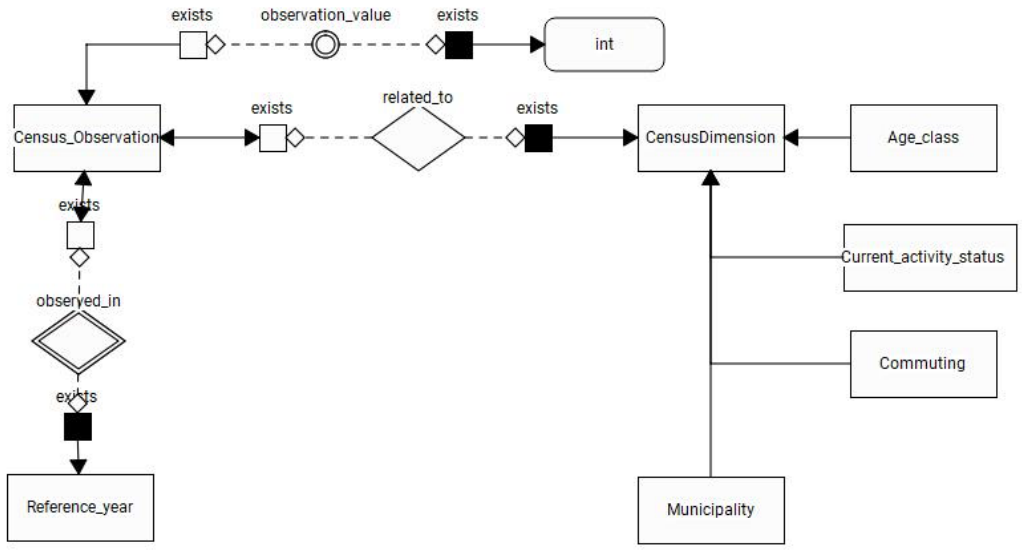


Figure 32 - Excerpt of census domain ontology

As explained above, the concept that links the two domains is the Municipality concept.

Figure 33 - Concept linking air pollution and census domain

The following table provides a brief description of the core concepts of the Air quality ontology, as well as their provenance in terms of ontologies, data models and vocabularies.

Air quality ontology classes or attributes	Description	Source
Aqd_Pollutant	Air polluting substance, level of which is measured and reported to the EEA (http://dd.eionet.europa.eu/vocabulary/aq/pollutant)	**AQD Model**
Reporting Year	Year for which primary data have been reported	**AQD Model**
Aggregation Type	Information about process of data aggregation into annual values (http://dd.eionet.europa.eu/vocabulary/aq/aggregationprocess).	**AQD Model**
AQ Value	Concentration or level of air polluting substance, here given as an aggregation of air pollutant concentration values from primary observation time series.	**AQD Model**
Begin Position	Datetime begin of measurement	**AQD Model**
End Position	Datetime end of measurement	**AQD Model**
Concentration	Measured concentration of air polluting substance.	**AQD Model**
Station	It's the air quality measurement station	**AQD Model**
Sampling Point		**AQD Model**
Sensor	Device, agent (including humans), or software (simulation) involved in, or implementing, a Procedure. Sensors respond to a Stimulus, e.g., a change in the environment, or Input data composed from the Results of prior Observations, and generate a Result	**SOSA**
Feature of Interest	The thing whose property is being estimated or calculated during an Observation to arrive at a Result	**SOSA**
Result	The Result of an Observation	**SOSA**

Table 29 - Main classes of Air quality ontology

Examples of linked data analysis

The following steps could be executed to test cross-border interoperability and provide an example of the additional information provided by linked data related to different countries. The main goal of the analysis

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

is to compare air pollution indicators (AQ values) related to a specific pollutant, measured in Rome and Paris in 2019. Data integration and analysis can be summarized as follows:

- Extract Italian data from the dataset published by ISPRA (reference year 2019);
- Extract French data from the dataset published by the European Environment Agency (reference year 2019) and derive Municipality from Sampling Point Latitude and Longitude, through an external service, or correspondence tables;
- Select AQ values referred to Rome and Paris;
- Create a classification variable (AQ level) to classify AQ values in three main modalities: 'Low', 'Medium', 'High';
- Count the number of Sampling Points for each modality of AQ level;
- Compare the number of Sampling Points having 'Low', 'Medium' or 'High' AQ level in the selected cities.

In order to test cross-domain interoperability, the following is one of the several examples of data analysis, based on the assumptions listed below:

- Extract Italian data from the dataset published by the European Environment Agency (reference year 2018);
- Extract French data from the dataset published by the European Environment Agency (reference year 2017);
- Derive Municipality from Sampling Point Latitude and Longitude through an external service, or correspondence tables;
- Sort AQ values in descending order;
- Select a subset of Municipalities having the highest AQ values;
- Link Italian census data (reference year 2018) to the selected subset of Municipalities belonging to the Italian territory;
- Analyse the resident population characteristics in the most polluted Municipalities. More in detail, analyse: Age structure, Current activity, Commuting for studying or working;
- Repeat the analysis using French census and air pollution data (reference year 2017).

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

4 A generalized pipeline for interoperable services

A generalized service pipeline must be designed as a modular chain of tools for publishing LOSD from heterogeneous sources and several data providers. There is no guarantee that data are provided according to previously agreed standard, so the pipeline must adapt to several contexts. The pipeline must collect data while it standardizes incoming data flow in several steps.

Following the principles from semantic data interoperability, data is decoupled from actual conceptual formalization, and the integration of different data sources is performed through a common ontology. The data sources to link may have several data formats and structures that result in ambiguity or inconsistencies in data formalization. An example of semantic harmonization is shown in the figure reported below. Domain ontology on top layer describes concepts and their relations which are mapped to source schemas, describing the core concepts with respect to a specific context of application. Source schema is described and mapped to actual data files.

A domain ontology is used to integrate multiple source schemas, thus linking distributed data sources. In addition, once the domain ontology is defined, data sources can be dynamically managed in the environment, so dynamic mapping is essential to preserve the relation between domain ontology and source schemas.

Domain ontology allows to integrate the source schemas, by defining a common knowledge model expressed by a shared vocabulary. Source schema describes its own data source and maps knowledge objects by using domain ontology concepts. Domain ontology is mapped to source schemas and in turn source schemas are mapped to data. If a source ontology is available, it will be mapped to the core concepts of the domain ontology.

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

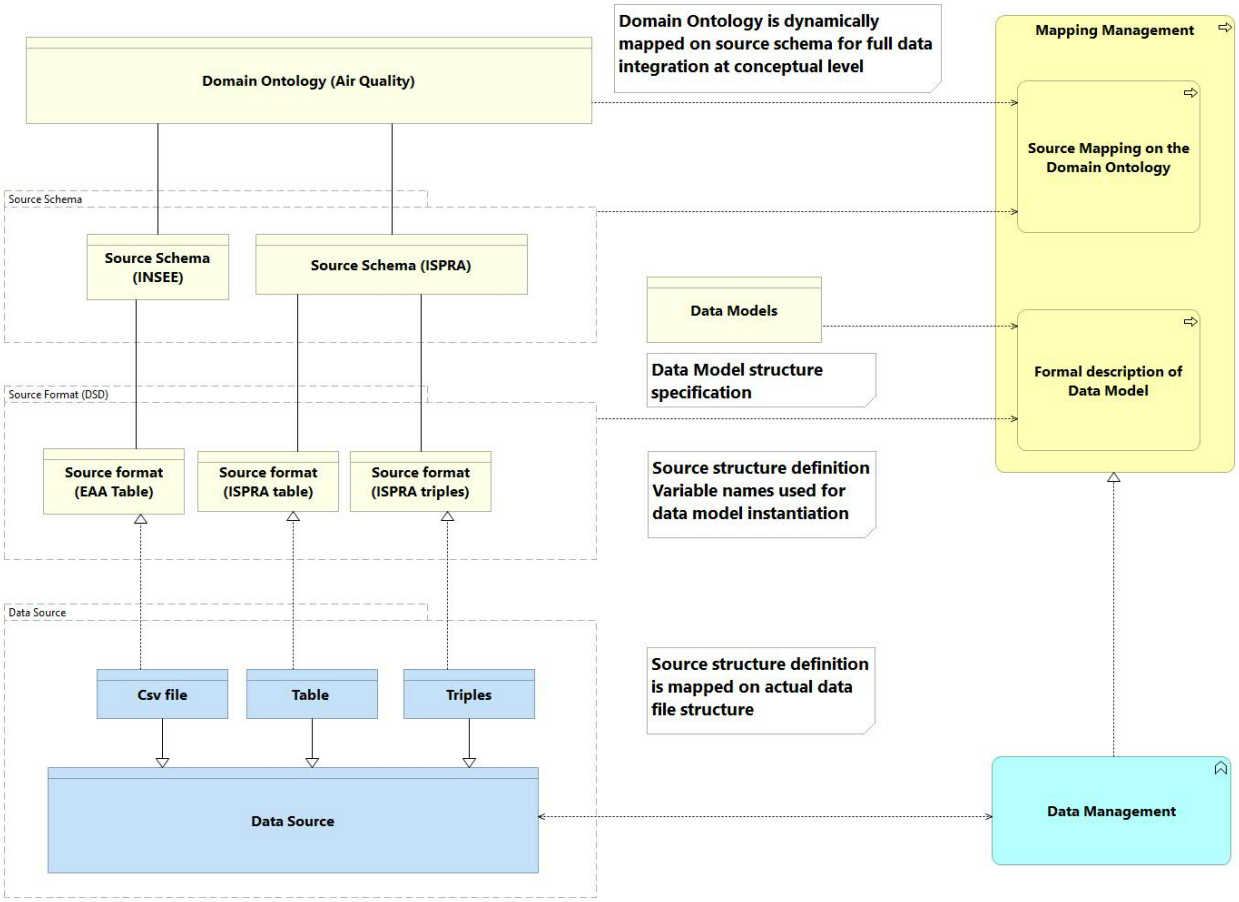


Figure 34 - Data channel outline

Each vertical line in the conceptual schema comprises a data channel. A data channel corresponds to a set of heterogenous data sources, in some cases with an available ontology. Data channels are modular and dynamic. The system must be designed so that data channels can be added or removed dynamically, and data sources can be joined seamlessly. The service pipeline must be also modular to adapt to different data channels that can be sorted with respect to their semantic interoperability level.

Problems about data connectivity and file transfer is not yet addressed here. In a scenario with a low interoperability level, assuming that the data has been acquired using a general format (e.g. csv, xls, etc.), it is possible to analyze the data structure definition, to identify data content and format. Then, if an ontology is available, it can be directly linked to the domain ontology. Any further data file structure must be converted to rdf and then accessed by a SPARQL endpoint to guarantee semantic interoperability.

When a source schema is associated with the dataset, the mapping can be performed smoothly by some tools in the pipeline. Templates can be also put in place to simplify ontology mapping when using a conventional set of variable names and/or formats in the Data Structure Definitions. This is also useful when an ontology is not associated with the data file, but it has a conventional choice of variable names

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

and format that can be easily linked with a template mapping. When this is not possible at all, the service pipeline must provide a tool for manual mapping. The service pipeline must be designed to handle all this different use cases. In the second scenario, having a medium interoperability level, rdf files could be imported to link data, while metadata could be extracted by accessing the data vocabularies provided or through the same federated queries cited above.

The best interoperability level is achieved when data are already available as rdf triples and can be directly accessed at the level of triplestore. They can be hosted on the local triplestore or linked from a remote triplestore and are accessible from a given URI. Data can be directly queried by federated queries in the SPARQL endpoint.

Process details and Application Components

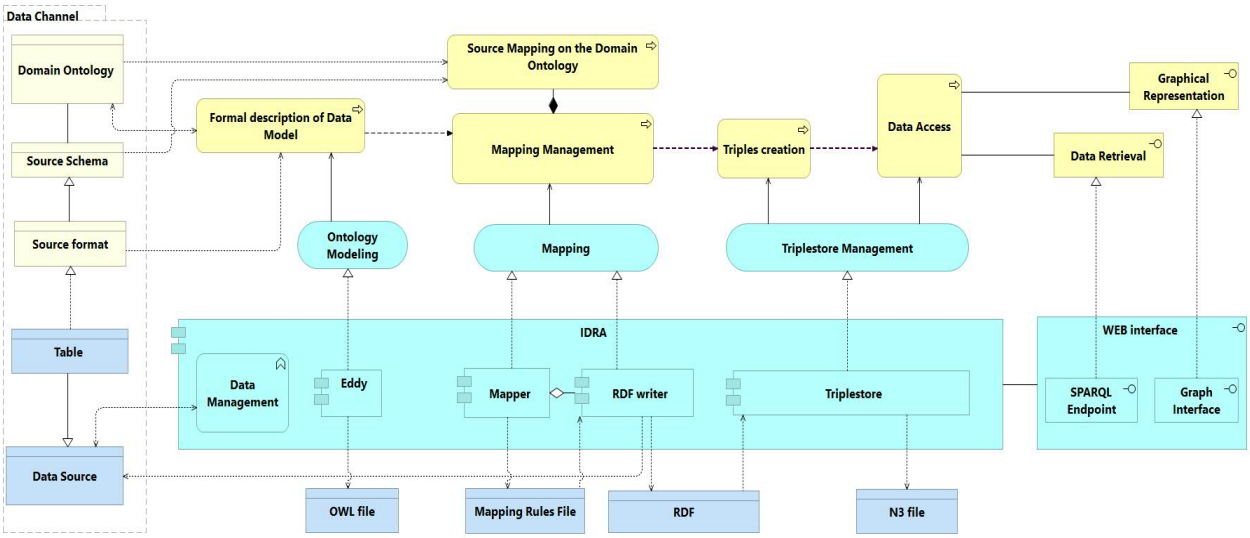


Figure 35 - Process Design and Application Components for data harmonization

In the figure above, the business layer shows the main steps of data harmonization in yellow colour, while the blue colour refers to the components of the application layer. The sketched process can be implemented by several components that can be substituted depending on the chosen software solutions. In addition, the proposed architecture can be integrated in any existing production system.

Such process shows how modular design, and the application components can be adapted to different contexts and can be used to connect to data sources, such as databases or files in several formats. Starting from the provision of a data channel, the process has four stages:

1) Data Modelling stage. Formal description of data models and semantic harmonization are managed here;

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

2) Mapping stage, related to the conceptual mapping between the domain ontology and source schema, as well as definition of mapping rules;
3) Triplestore Management, including Triples creation and storage on a triple store or import of RDF files if available;
4) Data Access. Data are ready for querying by the endpoint. Data are retrieved and used for further analysis.

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

ANNEX A

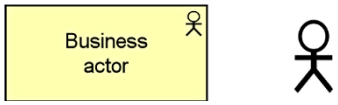
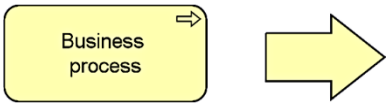
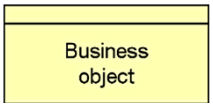
Element	Description	Notation
Business actor	Represents a business entity that is capable of performing behaviour.	
Business process	Represents a sequence of business behaviours that achieves a specific result such as a defined set of products or business services.	
Business object	Represents a concept used within a particular business domain.	

Table 30 - ArchiMate business layer objects [62]

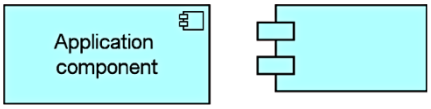
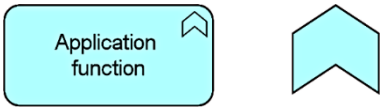
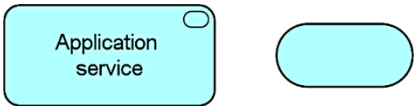
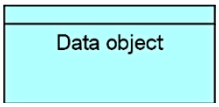
Element	Definition	Notation
Application component	Represents an encapsulation of application functionality aligned to implementation structure, which is modular and replaceable.	
Application function	Represents automated behaviour that can be performed by an application component.	
Application service	Represents an explicitly defined exposed application behaviour.	
Data object	Represents data structured for automated processing.	

Table 31 - ArchiMate application layer objects [62]

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

References

[1] "Eddy," [Online]. Available: https://github.com/obdasystems/eddy.

[2] Obda Systems, "Obda Systems Web Site," [Online]. Available: https://www.obdasystems.com/.

[3] "Olap Browser," [Online]. Available: https://github.com/LOSD-Data/qb-olap-browser.

[4] F. Cotton, "ESSnet Linked Open Statistical Data (LOSD) project," Insee, [Online]. Available: https://ec.europa.eu/eurostat/cros/system/files/los_-_d1-1-1_-_lod_platform_-_draft.pdf.

[5] Derilinx, "Derilinx Web Site," [Online]. Available: https://derilinx.com/.

[6] "Cube Visualizer," [Online]. Available: https://github.com/LOSD-Data/CubeVisualizer.

[7] "SparQLing," [Online]. Available: https://github.com/picorana/sparqling#documentation-index.

[8] "Bauhaus," [Online]. Available: https://github.com/InseeFr/Bauhaus.

[9] "SPARQL React," [Online]. Available: https://github.com/LOS-ESSnet/SPARQL-Inside.

[10] "Juma," [Online]. Available: https://github.com/Al-Meehan/juma-losd.

[11] Eurostat, "SDMX Reference Infrastructure," [Online]. Available: https://ec.europa.eu/eurostat/web/sdmx-infospace/sdmx-it-tools/sdmx-ri.

[12] Eurostat, "Eurostat Web Site," [Online]. Available: https://ec.europa.eu/eurostat.

[13] "Excel/CSV to NGSI-LD," [Online]. Available: https://github.com/jason-fox/csv-to-json.

[14] Eurostat, "STATISTICAL ATLAS project," [Online]. Available: https://ec.europa.eu/eurostat/web/gisco/gisco-activities/statistical-atlas.

[15] "Excel2csv," [Online]. Available: https://sdmxistattoolkit.github.io/mydoc_Excel2csv_Software.html.

[16] "Meta & Data Manager," [Online]. Available: https://sdmxistattoolkit.github.io/mydoc_MDM_Software.html .

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

[17] ISTAT, “SDMX Istat Toolkit,” [Online]. Available: http://statistics.caricom.org/Files/Meetings/HLF3/Brochures/SDMX_Istat_Toolkit.pdf.

[18] “Data Browser,” [Online]. Available: https://sdmxistattoolkit.github.io/mydoc_DB_Software.html.

[19] “CEF Context Broker,” [Online]. Available: https://ec.europa.eu/cefdigital/wiki/display/CEFDIGITAL/Context+Broker.

[20] “Idra - Open Data Federation Platform,” [Online]. Available: https://github.com/OPSILab/Idra.

[21] “FESTIVAL project,” [Online]. Available: https://www.festival-project.eu/.

[22] “Datalift,” [Online]. Available: https://gforge.inria.fr/scm/?group_id=2935.

[23] S. Vale, “The Generic Statistical Business Process Model,” [Online]. Available: https://statswiki.unece.org/display/GSBPM/Generic+Statistical+Business+Process+Model.

[24] MIUR, “Italian Ministry of Education (MIUR) Opendata,” [Online]. Available: https://dati.istruzione.it/opendata/opendata/.

[25] French Ministry of Education, “French Ministry of Education Web Site,” [Online]. Available: https://data.education.gouv.fr/.

[26] INVALSI, “Italian Evaluation Institute (INVALSI) Web Site,” [Online]. Available: https://www.invalsiopen.it/.

[27] OECD PISA, “OECD PISA evaluation,” [Online]. Available: https://www.oecd.org/pisa/data/.

[28] Wikipedia, “Dimensional fact model,” [Online]. Available: https://en.wikipedia.org/wiki/Dimensional_fact_model.

[29] Eurostat, “Education and training – Overview,” [Online]. Available: https://ec.europa.eu/eurostat/web/education-and-training/overview.

[30] Eurostat, “Education and Training - Database,” [Online]. Available: https://ec.europa.eu/eurostat/web/education-and-training/data/database.

[31] Eurostat, “Education administrative data from 2013 onwards (ISCED 2011),” [Online]. Available: https://ec.europa.eu/eurostat/cache/metadata/en/educ_uoe_enr_esms.htm.

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

[32] UNESCO Uis, “International Standard Classification of Education,” [Online]. Available: http://uis.unesco.org/sites/default/files/documents/international-standard-classification-of-education-isced-2011-en.pdf.

[33] “ArchiMate Tool,” [Online]. Available: https://www.archimatetool.com/.

[34] Eurostat, “Labour Market, including Labour Force Survey (LFS) — Overview,” [Online]. Available: https://ec.europa.eu/eurostat/web/labour-market/overview.

[35] International Labour Office, “International Standard Classification of Occupations,” [Online]. Available: https://www.ilo.org/wcmsp5/groups/public/@dgreports/@dcomm/@publ/documents/publication/wcms_172572.pdf.

[36] Wikipedia, “International Labour Organization,” [Online]. Available: https://en.wikipedia.org/wiki/International_Labour_Organization.

[37] Wikipedia, “Labour economics,” [Online]. Available: https://en.wikipedia.org/wiki/Labour_economics.

[38] Wikipedia, “United Nations,” [Online]. Available: https://en.wikipedia.org/wiki/United_Nations.

[39] INVALSI, “Results of INVALSI test,” [Online]. Available: https://www.invalsiopen.it/.

[40] Eurostat, “Single Integrated Metadata Structure V 2.0 (SIMS V2.0),” [Online]. Available: https://ec.europa.eu/eurostat/documents/64157/4373903/SIMS-2-0-Revised-standards-November-2015-ESSC-final.pdf/47c0b80d-0e19-4777-8f9e-28f89f82ce18.

[41] Insee, “Permanent database of facilities,” [Online]. Available: https://www.insee.fr/en/metadonnees/source/serie/s1161.

[42] Insee, “Permanent database of facilities 2019,” [Online]. Available: https://www.insee.fr/en/metadonnees/source/operation/s1524.

[43] Insee, “Results of node <http://id.insee.fr/operations/serie/s1161> published in RDF form,” [Online]. Available: http://id.insee.fr/operations/serie/s1161.

[44] Insee, “Results of node <http://id.insee.fr/qualite/rapport/1967> published in RDF form,” [Online]. Available: http://id.insee.fr/qualite/rapport/1967.

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

[45] Insee, “INSEE's metadata API,” [Online]. Available: https://api.insee.fr/catalogue/site/themes/wso2/subthemes/insee/pages/item-info.jag?name=M%C3%A9tadonn%C3%A9es&version=V1&provider=insee.

[46] Insee, “List of variables in the BPE,” [Online]. Available: https://www.insee.fr/fr/statistiques/3568638?sommaire=3568656.

[47] Ministry of Culture, “Ministry of Culture catalogues,” [Online]. Available: https://www.beniculturali.it/catalogo-di-dati-metadati-e-banche-dati.

[48] INTERSTAT project, “Deliverable 3.1 - Report of the use cases to demonstrate the cross-border benefits of the proposed solution,” [Online]. Available: https://cef-interstat.eu/resources/.

[49] ISPRA, “Environmental Data Yearbook,” [Online]. Available: https://annuario.isprambiente.it/sys_ind/448.

[50] ISPRA, “AMBIENT AIR QUALITY: NITROGEN DIOXIDE (NO2),” [Online]. Available: https://annuario.isprambiente.it/sys_ind/450.

[51] ISPRA, “AMBIENT AIR QUALITY: TROPOSPHERIC OZONE (O3),” [Online]. Available: https://annuario.isprambiente.it/sys_ind/451.

[52] ISPRA, “AMBIENT AIR QUALITY: PARTICULATE (PM2.5),” [Online]. Available: https://annuario.isprambiente.it/sys_ind/452.

[53] Library Of Congress, “Regulation of Air Pollution: France,” [Online]. Available: https://www.loc.gov/law/help/air-pollution/france.php.

[54] European Environment Agency, “Air quality statistics calculated by the EEA (AIDE F),” [Online]. Available: http://aidef.apps.eea.europa.eu/?source=%7B%22query%22%3A%7B%22match_all%22%3A%7B%7D%7D%2C%22display_type%22%3A%22tabular%22%7D.

[55] ISTAT, “Italian census data,” [Online]. Available: http://dati-censimentipermanenti.istat.it/?lang=en.

[56] Insee, “French census datasets BTX_TD_POP1B_2017 and BTX_TD_POP5_2017,” [Online]. Available: https://www.insee.fr/fr/statistiques/4515539?sommaire=4516122.

[57] Insee, “Professional mobility in 2017: travel home - workplace,” [Online]. Available: https://www.insee.fr/fr/statistiques/4509353.

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

[58] EEA, “AQD data model,” [Online]. Available: https://ftp.eea.europa.eu/www/aqereporting-3/AQeReporting_products_2018_v1.pdf.

[59] Geonames, “GeoNames ontology,” [Online]. Available: https://www.geonames.org/ontology/ontology_v3.2.rdf.

[60] W3C, “Semantic Sensor Network Ontology (SSN),” [Online]. Available: https://www.w3.org/TR/vocab-ssn/.

[61] Obda Systems, “Graphol, A visual language for ontologies,” [Online]. Available: http://obdasystems.com/graphol.

[62] The Open Group, “ArchiMate® 3.1 Specification,” [Online]. Available: https://pubs.opengroup.org/architecture/archimate3-doc/.

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