INTERSTAT

Open Statistical Data Interoperability Framework

www.cef-interstat.eu

D3.1 - Report of the use cases to demonstrate the cross-border benefits of the proposed solution

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

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

Project full title
INTERSTAT - Open Statistical Data Interoperability Framework

Grant Agreement No.
INEA/CEF/ICT/A2019/2063524

Project Document Number
Deliverable 3.1 (Activity 3)

Project Document Delivery Date
31.12.2020

Deliverable Type and Security
Report – Public

Author
Franck Cotton (Insee)

Contributors
Monica Scannapieco (Istat), Carlo Vaccari (Istat), Martino Maggio (ENG)

Reviewers
Marina Gandolfo (Istat)
Stéphane Dufour and Davide Taibi (External Expert Advisory Board)

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

Table of Contents

List of figures

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

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

Executive Summary

The scope of this document is to analyse the “cross-border” benefits of the use cases defined in the INTERSTAT project.

In the chapter 1 are defined different criteria for assessing the cross-border value of INTERSTAT services and clients, and in particular the pilot use cases. More than a uniformly-structured evaluation benchmark, we will describe a list of topics that will allow us to characterise the cross-border value of the upstream and downstream data services and of the clients from different points of view.

The chapter 0 will focus on the cross-border added value brought by the INTERSTAT technical framework itself highlighting in particular the interoperability capabilities.

The last chapter will describe the INTERSTAT pilot services in the light of the criteria previously defined. For each use case it will be provided a description, the proposed datasets to be used for the service implementation and the cross-border and cross domain features.

Introduction

In this deliverable are discussed the "cross-border" benefits of the use cases defined in the INTERSTAT project proposal. Beforehand, we are going to define more precisely this notion of cross-border value or benefit, and list some aspects that can be used to evaluate it.

The following figure presents an overview of the solution that will be delivered by the INTERSTAT project. The pilot use cases are pictured in the upper left corner. They feed on data exposed through an API based on the INTERSTAT framework. The framework itself merges data from different statistical offices, which can be heterogeneous in terms of representations but are conceptually aligned on shared ontologies in order to allow for their interoperability.

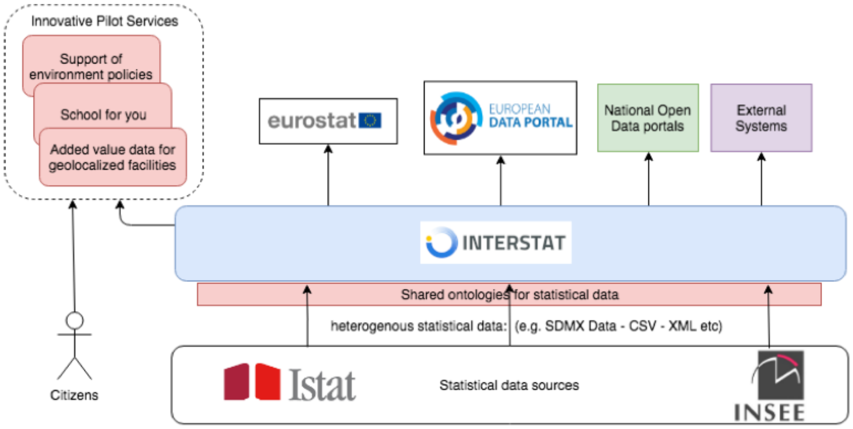
Innovative Pilot Services
Support of environment policies
School for you
Added value data for geolocalized facilities
eurostat
EUROPEAN DATA PORTAL
National Open Data portals
External Systems
INTERSTAT
Shared ontologies for statistical data
heterogenous statistical data: (e.g. SDMX Data - CSV - XML etc)
Citizens
Istat
Statistical data sources
INSEE

Figure 1 - INTERSTAT framework high-level overview

We can see that different layers of data services are present in the solution, mainly the upstream data services produced by the statistical institutes and consumed by the INTERSTAT framework, and downstream services produced by the INTERSTAT framework and consumed by client applications, and in particular by pilot use cases.

In the INTERSTAT installation, upstream data services are provided by National Statistical Institutes, so the notion of "border" appears at this stage. One of the main objectives of INTERSTAT is to ensure that national data are made interoperable so they can be used in transnational client applications valuable for their users. So the benefit of transnational client applications for their users relies, not only on their intrinsic business or technical value, but also on the value and level of interoperability of the downstream services they rely on, which in turn are determined by the value and level of interoperability of upstream data services, but also by the value added by the INTERSTAT framework, which can be measured along various dimensions of data and metadata quality. These different components are reflected in the structure of the paper. In a first part, we will try to define more clearly the notion of cross-border value for statistical data services, and the factors that it involves. The second part focuses on the value added by the INTERSTAT platform. Then we will review each of the INTERSTAT use cases in the light of this notion. Finally, we will analyse how it is possible to adjust the use case specifications in order to improve their potential benefits as cross-border services.

1 Cross-border services

In this section, we define cross-border services and how they can be categorized with respect to various characteristics.

1.1 Definition

In coherence with INTERSTAT’s objectives, we will focus on services whose main function is to provide statistical data and related metadata. In the rest of the document, this kind of services will be referred to as SDSs (Statistical Data Services).

By definition, an SDS will be characterised as a cross-border service if it is available in two or more countries. In its narrowest sense, “cross-border” would require the countries to be bordering, but we will not restrict ourselves to this specific case. On the other hand, we will not require a strict identity of the services features, and indeed we will establish different degrees of similarity between equivalent services across the border.

“Border” can be understood in the physical sense (the line that separates two countries) or more generally as a limit between two legislative, linguistic, cultural, etc., systems. The first meaning is particularly applicable in the case of SDSs with a geographic dimension, which is what we will mostly consider in the case of INTERSTAT since the project aims at delivering geolocalized context information, but other meanings can also be relevant. For example, the linguistic dimension is important for services that involve a textual aspect.

1.2 Features

In this part, we list characteristics of SDSs that can be used to define and assess the cross-border benefits of these services. In order to be able to express a clear description of some of these characteristics, we first have to define some vocabularies and concepts related to the model governing the data served by the SDSs. This will also be useful for the specification of the pilot services later on.

1.2.1 Data model

In the statistical community, it is customary to differentiate between unit data and dimensional data. Those are for example clearly distinguished in the “Structure” group of the Generic Statistical Information Model, or GSIM [1],which is the international reference information model for Official Statistics developed

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

by the High-Level Group for the Modernisation of Official Statistics (HLG-MOS) within the Unece Conference of European Statisticians[1].

A unit dataset is a collection of data that conforms to a known structure and describes aspects of one or more units: an individual, household, enterprise, point of interest, etc. This type of data is also known as microdata and can be found in surveys, registers, administrative files, scientific and public-use files[2] and so on. The Data Documentation Initiative (DDI [2]) standard provides for very rich and precise descriptions of unit data.

The use case on geolocalized facilities includes an example of unit data: the French "Permanent database of facilities" lists individual characteristics of the facilities, for example their type, their geographic coordinates, their name, etc. In contrast, results published from the Census are for example population counts by age, sex, type of activity, etc. [3]

Indeed, microdata are opposed to macrodata or aggregated data, which provide a summarised information in the form of counts, ratios, frequencies or other statistics. GSIM uses the notion of dimensional data, where each value corresponds to a measure and a combination of identifying dimensions. For example, in a table counting the number of citizens by country and gender, the country and gender would be identifier components and the number of citizens a measure component. This vocabulary is largely inspired by the SDMX [4] data model, which remains the reference model for statistical dimensional data. SDMX is typically used for the dissemination of hypercubes of statistical data. The model has been adapted for the publication of multidimensional linked data in the "RDF Data Cube Vocabulary" W3C Recommendation [3].

The comparison between GSIM, DDI and SDMX standards has been a subject of discussions between experts for years, but in the context of INTERSTAT we can safely simplify the question and consider a basic convergence model established for the specification of the SDMX Validation and Transformation Language [5].

The Validation and Transformation Languages (VTL) model defines logical datasets containing data points (or rows) and conforming to a data structure made of (column) components: identifiers, measures or attributes. Atomic data items are characterized in terms of variables, each of which has a representation: a data type and a value domain. The value domain can be enumerated, which means that the possible values belong to a code list.

[1] See https://unece.org/ar/node/4425.

[2] See for example https://ec.europa.eu/eurostat/web/microdata and https://ec.europa.eu/eurostat/web/microdata/public-microdata.

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

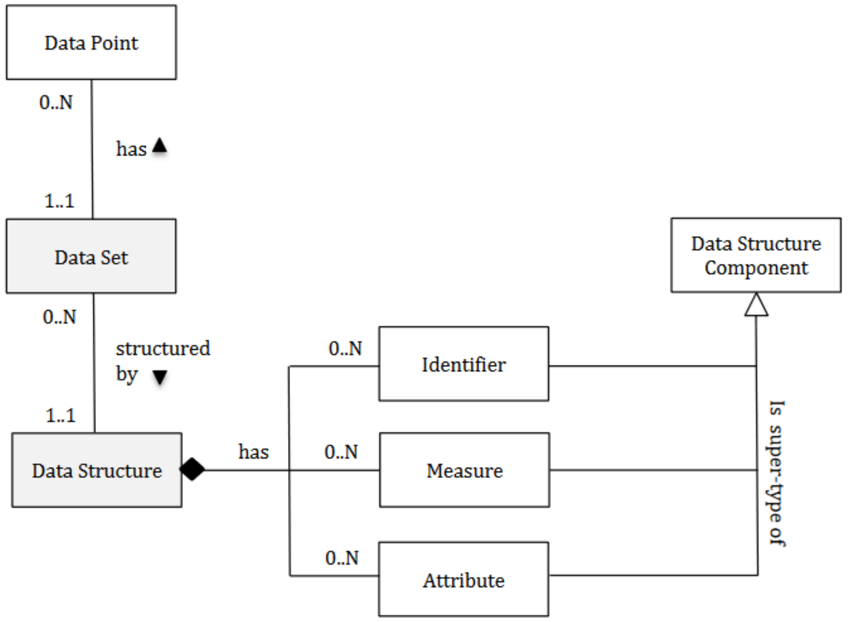

Figure 2 - Base VTL Data Model

One of the advantages of using the VTL model for the description of the INTERSTAT SDSs is that it will easily allow us to express validation expressions or even transformations to perform on the data, for example recodifications, aggregations, imputations, etc. This will be very useful to assess or improve comparability between analogous data services. The VTL language remains at the logical (specification) level, but different tools exist that can translate it at implementation level (e.g. to Java, JavaScript, etc.).

We now have a basic version of a formal but simple framework that we can use to describe the INTERSTAT data services, and in particular those used in our first use cases. This allows us to define criteria that will help us to define cross-border benefits.

1.2.2 Multi-geographic

This feature refers to a dataset that involves two or more geographic identifier components (or dimensions, in SDMX terms). More specifically, since we are focused on cross-border services, the geographic components will be related to different countries.

Examples of such data sets abound in the statistical domain. We can mention statistics on tourism, international trade, etc. In the case of tourism, typical statistics count the visitors to a destination country by country of origin and other dimensions (age, sex, type of accommodation...). In some cases, the geographic breakdown (e.g. NUTS 1 or 2[3]) is finer for the destination, or even for the origin. For international trade, goods and services are similarly distributed according to origin and destination.

[3] See https://ec.europa.eu/eurostat/web/nuts/background

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

Services based on multi-geographic datasets have by nature a high value from a cross-border perspective. They are more than the simple juxtaposition of country-level services: they really describe a transnational phenomenon.

1.2.3 Interoperability

When several SDSs are used together in constructing a business solution, the first question that arises is how interoperable they are. Can a data element from service A be related to and used together with another data element from service B? And thus, is it possible to create additional value from the use of the services together?

Regarding interoperability, a well-established reference framework is the European Interoperability Framework (EIF [6]). The EIF defines different models, 12 relevant principles[4] and 49 related recommendations. The best-known model is the layered interoperability model, which defines four levels: technical, semantic, organisational and legal.

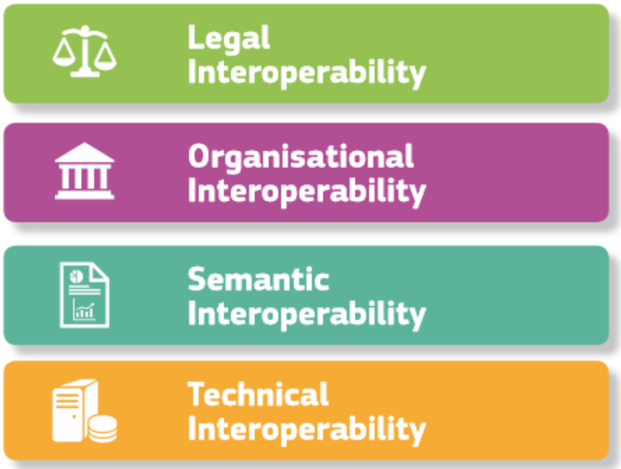
Legal Interoperability
Organisational Interoperability
Semantic Interoperability
Technical Interoperability

Figure 3 - EIF Levels of Interoperability

Technical interoperability is covered by the INTERSTAT platform itself. The platform can also partially address semantic interoperability through ontology alignments.

In the context of INTERSTAT, the providers of upstream data are NSIs, and more precisely members of the European Statistical System (ESS[5]). There are a number of transformative processes going on in the ESS and in international Official Statistics in general, that aim at improving NSI's business processes through standardization, sharing, industrialization, etc. These are fuelled in particular by the "Modernisation of Official Statistics (MOS)" initiative led by the HLG-MOS under the umbrella of the Unece CES [7].

[4] Subsidiarity and proportionality, openness, transparency, reusability, technological neutrality and data portability, user-centricity, inclusion and accessibility, security and privacy, multilingualism, administrative simplification, preservation of information, assessment of effectiveness and efficiency

[5] The members of the ESS are NSIs, Eurostat and the ONAs, or Other National Authorities, see details in https://ec.europa.eu/eurostat/web/european-statistical-system.

Specifically, the MOS vision produced high-level business and activity models (GSBPM [8] and GAMSO [9]) that are powerful tools of business process alignment, and thus organisational interoperability, for statistical organizations at the global level.

The level of organisational interoperability is of course even much higher within the European Statistical System, where intensive coordination between NSIs is organised by Eurostat regarding processes (e.g. for some surveys) and their outputs ("standardization by the outputs" policy).

In most cases, European statistics, and therefore the statistical products developed, produced and disseminated in the ESS, are specified in great detail by legal acts like EU Regulations on European statistics in several domains. A good example is the Population Census [10], whose statistical outputs are described very precisely in the Regulation (EC) No 763/2008 on population and housing censuses[6] and the related implementing acts as regards the technical specifications of the topics and of their breakdowns[7], establishing the reference year and the programme of the statistical data and metadata[8] and as regards the modalities and structure of the quality reports and the technical format for data transmission[9].

One can also mention the EBS[10] (European Business Statistics), which represent a Framework Regulation Integrating Business Statistics[11] (FRIBS[12]) and the common framework for European statistics relating to persons and households, based on data at individual level collected from samples, which represent the Integrated European Social Statistics (IESS[13]). The standardization of concepts and variables implemented by such frameworks is obviously an essential tool for improving the comparability of data between countries.

The case of key semantic assets like EU official classifications is also worth mentioning. Most of the time, they are also provided within EU legislation acts (Regulation, Directive or Decision). The example of the Nomenclature of territorial units for statistics (NUTS) has already been cited, but we can add the European classification of economic activities (NACE[14]), statistical classification of economic activities, or the CPA[15], statistical classification of products by activity. In some cases, EU legislation simply enforces the sectorial use of classifications published by external organisations, for example the International Standard Classification of Occupations (ISCO) maintained by the International Labour Organisation (ILO).

Of course, when statistical products and services are specified by EU legislation acts, or when they use international official classifications, we can expect them to be more stable, comparable, documented, and thus to bring much more potential cross-border value. Having an EU legal base will thus be an important criterium for SDSs from a cross-border value perspective.

When mixing NSI data with external data, legal interoperability can still happen through the use of official classifications, but otherwise semantic interoperability becomes very important. In many cases, it is

[6] See https://eur-lex.europa.eu/eli/reg/2008/763/oj and related implementing acts.
[7] See https://eur-lex.europa.eu/eli/reg_impl/2017/543/oj.
[8] See https://eur-lex.europa.eu/eli/reg/2017/712/oj.
[9] See https://eur-lex.europa.eu/eli/reg_impl/2017/881/oj.
[10] See https://eur-lex.europa.eu/eli/reg/2019/2152/oj.
[11] See https://www.insee.fr/fr/information/4254231 (in French).
[12] See https://www.insee.fr/fr/information/4254231 (in French).
[13] See https://www.insee.fr/fr/information/4254233 (in French) and https://eur-lex.europa.eu/eli/reg/2019/1700/oj.
[14] See https://eur-lex.europa.eu/eli/reg/2006/1893/oj
[15] See https://eur-lex.europa.eu/eli/reg/2008/451/oj

difficult to understand the concepts used in the external sources, so mapping with NSI data can be hard. This is why it is crucial that NSIs publish and document the statistical concepts that they use so that third-party data providers can use that reference.

In conclusion, we see that the main subject for INTERSTAT is semantic interoperability, and we can even distinguish between identity or alignment of concepts, and interoperability of concept representation. This is detailed in the next section.

1.2.4 Integration

We define in this section the notion of integration between cross-border services, which is actually a specific view on service interoperability, assuming technical interoperability and focusing on semantic interoperability. We also assume that the data structures of the services align on the VTL model, and levels of integration are defined on the common structure components. We specify three levels of service integration:

- **Comparable**: the services use the same concepts for measures and dimensions, and textual elements are available in a common language, but notable differences in concept representations or implementation. The following examples illustrate service integration at the "comparable" level:
  - use of a dimension on economic activity, but measured with Ateco for the Italian service and NAF for French service[16];
  - use of geographic dimension, but represented with different NUTS levels or different NUTS versions;
  - use of an "age" dimension, but with different code lists for age groups, of with an ordinal age in one case and a numerical age in the other.

- **Integrated:** the services use the same data structures and a common language, the underlying concepts are identical, but minor differences or discrepancies exist in concept representations. An example of this level could be services conforming to the same data structure definition with a geographic dimension involving a representation as grid data, but the grid contours are defined at the country level and do not connect at the border.
- **Seamless**: in seamless integration, we have integrated services where concepts and concept representations are identical and connected. In the previous example, if grids are defined at the European level, and thus connect at the borders, the services are seamlessly integrated. For example, the results of the 2021 round of censuses should be disseminated using a common 1 km^2 grid mapping defined in the INSPIRE framework on European Spatial Data Infrastructure. This implies that population counts can be obtained even on grid cells that cover two or more countries[17]. This is obviously a huge improvement in potential value for cross-border use of the services. In the NAF/Ateco example above, if the activity is recoded at the NACE level, the integration becomes seamless. Also, for services using comparable age dimensions, it is often possible to derive seamless integration through coding or recoding of the age. More generally,

[16] The Ateco and the NAF are respectively the Italian and French refinements of the European classification of economic activities (NACE). They are identical down to the "class" level, but they differ at the "sub-class" level.

[17] See https://ec.europa.eu/eurostat/statistics-explained/index.php/Population_grids for details

seamless integration of comparable services can often be achieved at the price of a loss of detail or precision.

1.2.5 Temporality

All data services use implicitly or explicitly a time dimension, so temporality is an important aspect of interoperability. Also, time is often used with different meanings for a given data set: date of observation, of reference, of publication, etc. It is rare that this information is available or clearly documented, or even understood, and even when they are there can be very tricky cases in some circumstances[18]. Limiting ourselves to an explicit use of the time dimension, we can distinguish several options:

- **Dated**: data published without any dates are practically worthless, so we will require from SDSs used in INTERSTAT pilots to be at least dated, for example indicating a reference date for the observed phenomenon. Comparability between dated data sets or services depends largely on the concept measured: for example, for a slowly varying phenomenon, comparison can still make sense if reference dates are close.
- **Periodic**: for data published periodically (e.g. every month, year...), an explicit (non-degenerate) time dimension is introduced. As for other dimensions, comparability between periodic dated data sets or services depends essentially on the compatibility of the time representations used on both sides. As an example, Insee publishes every year the COG, which is the official list of French administrative territories (regions, departments, municipalities…) as of the first of January of the year. Actually, the COG is more of a metadata set that would be used for the representation of a geographic identifier or dimension in an actual data set or service[19].
- **Historic**: this type of data also implies an explicit time dimension, but past observations are kept in the data set or can be recreated. Consequently, the data can be recalculated at any given point in the past, which of course improves hugely comparability between services, and specifically user value for cross-border services. Elaborating on the previous example, Insee also publishes every year the list of dated events concerning administrative territories (mergings, splittings, etc.), so past COG lists can be recreated, not only on the first of January but at any date.
- **Continuous**: it is a specific case where a service produces data continuously (it includes for instance real-time data). In theory, it is possible to reconstruct dated, periodic of historic data from a continuous service, but this can raise practicality issues for fast or voluminous data.

We see that the time dimension raises very specific and complex issues, so it will be important to study and document precisely this aspect in the data services consumed or published by INTERSTAT. From the point of view of cross-border value, previous considerations on interoperability apply, but we can also add that data has no value when it does not come with a time reference.

[18] For example, the French official population counts for 2017, published in December 2019, came into force on January 1, 2020. They are based on geographic zones as they were on January 1, 2019.

[19] We observe that using a geographic dimension introduces a new “hidden” time dimension, as is the case for every classification evolving over time (and most of them do).

1.2.6 Quality

Quality is a very large concept, which applies to a variety of things: data, metadata, processes, services, etc. Regarding Official Statistics, a number of reference frameworks and documents dealing with quality exist [11]. At EU level, the main one is the European Statistics Code of Practice (ESCoP) [12], which is “the cornerstone of the common quality framework of the European Statistical System”. The ESCoP defines 16 principles covering the institutional environment, statistical processes and statistical outputs. It also specifies indicators for each of the principles.

In line with the ESCoP, the European Statistical System Quality Assurance Framework (ESSQAF) [13] constitutes the common quality framework for the European Statistical System. It complements and breaks further down the ESCoP and identifies possible methods, tools and good practices that can provide guidance for its implementation. Based on these frameworks, high-quality European Statistics are developed, produced and disseminated.

Other ESS reference frameworks have been defined at a more operational level, in particular for quality reporting. The ESS has adopted in 2015 the version 2 of the Single Integrated Metadata Structure (SIMS) [14] as a convergence model based on previous standards defined for quality reporting targeted at different publics (users, producers, etc.). The SIMS contains about 80 topics organized hierarchically in 19 sections[20]. It also includes quality and performance indicators (QPI), which are very precisely defined, for example non-response rate, imputation rate, etc.

Technically, the SIMS is formalized as an SDMX Metadata Structure Definition, which allows to disseminate quality information in a standardized way along with the data service itself. The SDMX metadata model can be adapted to a linked data context (in the same fashion that the Data Cube specification translates the SDMX data model), so quality meta-information can be attached in a standard way to the data services.

In the INTERSTAT context, some quality dimensions are particularly relevant from the point of view of cross-border value: accessibility (including multilanguage features) and clarity, quality management, update policy, comparability and coherence.

Other quality frameworks exist that apply to data or data services in general, not only data from Official Statistics. One of the most popular is described by the so-called FAIR principles[21] (Findable, Accessible, Interoperable and Reusable), endorsed by the G20[22] and mentioned in the Communication from the Commission on the European strategy for data (COM/2020/66[23]). More information on the FAIR principles and their interest for European data is available in [15].

[20] Contact, Metadata update, Statistical presentation, Unit of measure, Reference period, Institutional mandate, Confidentiality, Release policy, Frequency of dissemination, Accessibility and clarity, Quality management, Relevance, Accuracy and reliability, Timeliness and punctuality, Comparability and Coherence, Cost and burden, Data revision, Statistical processing, Comment

[21] See for example https://www.go-fair.org/fair-principles/.

[22] https://ec.europa.eu/commission/presscorner/detail/en/STATEMENT_16_2967.

[23] See https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52020DC0066.

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

Another well-known quality framework is the 5-star scale[24] for open data proposed by Tim Berners-Lee. At the highest level, data (and metadata) is made available in a non-proprietary structured open format (RDF), identified by URIs and linked to other data. Thus, the 5-star scheme is more specifically targeted at the publication of linked open data, and will consequently be of particular interest in the INTERSTAT context.

Of course, there are overlaps and correspondences between the different quality frameworks described in this section, in particular between the FAIR principles and the 5-star scheme for open data, but each of them is more relevant in given contexts: ESCoP for statistical data, 5-star for linked data and FAIR for third-party data services. So we will be called to use all of them in the INTERSTAT context.

We mostly focused in this section on aspects of quality that apply to data, but there are also other relevant dimensions, in particular regarding data services, either upstream or downstream. Service availability is an important example, and so are quality of support, standardisation of protocols, process reproducibility, etc. For service-oriented dimensions, the ITIL framework[25] is the most widely recognized reference.

[24] See https://5stardata.info/.

[25] See https://www.axelos.com/best-practice-solutions/itil.

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

2 Cross-border interoperability in the INTERSTAT technical framework

INTERSTAT aims at building an interoperability framework to enable and provide cross-border services in the statistics' domain. Figure 4 depicts the proposed framework architecture.

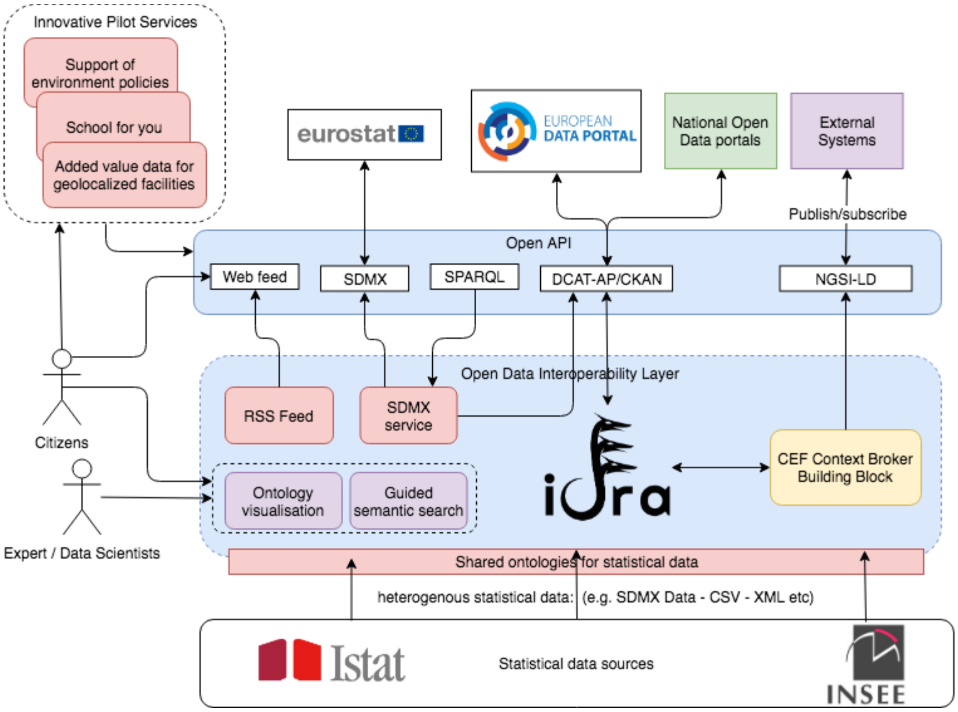
Innovative Pilot Services
Support of environment policies
School for you
Added value data for geolocalized facilities
eurostat
EUROPEAN DATA PORTAL
National Open Data portals
External Systems
Publish/subscribe
Open API
Web feed
SDMX
SPARQL
DCAT-AP/CKAN
NGSI-LD
Open Data Interoperability Layer
RSS Feed
SDMX service
Citizens
Ontology visualisation
Guided semantic search
idra
CEF Context Broker Building Block
Expert / Data Scientists
Shared ontologies for statistical data
heterogenous statistical data: (e.g. SDMX Data - CSV - XML etc)
Istat
Statistical data sources
INSEE

Figure 4 - INTERSTAT Framework Architecture

The INTERSTAT framework has been defined to support from technical point of view cross-border scenario through a set of features and capabilities:

- **International standards**: the adoption of international standards, such as ETSI NGSI-LD, DCAT-AP, StatDCAT-AP or SDMX or the facto ones such as CKAN API, guarantees the interoperability with different systems, fostering and simplifying the implementation of cross-border service that for definition should rely on common and well know international technologies and (de-facto) standards. Moreover, these adopted standards ease the dissemination of the statistical datasets produced within the project to the European Data portal.

- **Internationalization**: language barriers can be a critical aspect to consider when building a cross-border service. The tools included in the INTERSTAT framework are designed to support internationalisation and cross-language interoperability, for instance by the use of the EuroVoc Thesaurus[26].

- **Open APIs**: INTERSTAT framework will be designed to facilitate the interaction with external third-party applications through the provisioning of a set of public and open APIs in order to simplify the development of innovative services. These APIs will be also used as a single point of access for heterogenous data coming from different cross-border data sources.
- **Open Source Tools**: the usage of open source components to build the INTERSTAT framework avoids vendor lock-in and, thus, foster the adoption of the framework also in the context of Public Administrations of different Countries it simplifies the convergence to common technologies.
- **Common Ontologies**: the adoption common ontology models, such as the ones provided (e.g. SDMX or DCAT-AP) to harmonize statistical data and metadata, enables the open data interoperability among the national statistical open-data portals simplifying the development of cross-border services.
- **End user multi-modal access**: INTERSTAT technological solutions are also aimed to enable the provisioning of the information to end-users through different channels and devices. The complete decoupling of the framework back-end and the application layer, through the paradigm of micro-services, allows the easy development, for instance, of mobile applications to support the cross-border use case scenarios

Among the several technological components that the INTERSTAT framework is going to exploit, **Idra**[27] and the **CEF Context Broker**[28] are going to provide the most of the interoperability capabilities. Idra, among the other functionalities, harmonizes the metadata coming from heterogenous Open Data portals to DCAT-AP format, provides multilingual search functions through the EuroVoc thesaurus and expose a set of CKAN API that can be used to interact with the European Data Portal. The CEF Context Broker, implementing the ETSI NGSI-LD API, acts as the bridge with external system exposing the statistical data and metadata though its NGSI-LD APIs.

[26] https://data.europa.eu/euodp/en/data/dataset/eurovoc
[27] https://github.com/OPSILab/Idra
[28] https://ec.europa.eu/cefdigital/wiki/display/CEFDIGITAL/Context+Broker

3 INTERSTAT use cases

In this section, we list the pilot services that will be developed by the INTERSTAT project and describe each of them in relation with the framework established in the previous sections.

Some of the available services will be cross-domain, i.e. integrating the statistical domain with other domains like, e.g. air pollution data. For the statistical domain, EU-level content harmonization is mostly performed. For other domains it can be the case that content harmonization is there (e.g. mostly for air pollution data) or not (e.g. school-related data), and hence a preliminary step of intra-domain (cross-border) content harmonization should be done. However, once overcome these harmonization issues, the possibility of having services "cross-domain" really increases the value of the services themselves by simplifying ways to data access and hence fostering the knowledge derivable from them.

3.1 The School For You

3.1.1 Description

The "S4Y – The School for You" service allows families, who have to choose the most suitable school for their children, to access integrated information related to data coming from different schools' web portal of a chosen city's area and data coming from the Census of Population and Dwellings (e.g. as resident population, age, sex, level of education, marital status, the construction period and the state of preservation of the dwelling) related to the same geographical area. In fact, families, in addition to the location of the structure, information easily obtained on Google Maps, are interested in obtaining information about the quality and type of services provided from the school.

At the end of each school year, in the school portal are published information about: the total number of students for each year of the course of study; statistics on the progress of students (grades, results of end-of-course exams); choices made by students for the following academic path; comparison between the results of school students in the national tests, compared to the average of other schools at regional level; organization of school schedules; taught disciplines; laboratories present in the school, as well as the supplementary activities offered. This information is published in PDF file or integrated in the web page.

In Italy, the Ministry of University Education and Research (MIUR) publishes data for each Italian school, on grades, results of end-of-course exams, results of school students in the national tests in CSV, JSON, RDF and XML formats. Also, in Italy is available data on the status of the buildings of each school, and also the services offered by each school[29]. In France, equivalent data are available from the Ministry of National Education, Youth and Sports[30].

Another important source of information is the European data portal, currently over 9000 educational datasets are published. The service "S4Y – The School for You" allows users to compare information about

[29] See https://dati.istruzione.it/opendata/opendata/catalogo/elements1/?area=Edilizia%20Scolastica.

[30] See for example https://data.education.gouv.fr/explore/dataset/fr-en-indicateurs-de-resultat-des-lycees-denseignement-general-et-technologique/table/, and related descriptive information at https://www.education.gouv.fr/les-indicateurs-de-resultats-des-lycees-1118.

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

different schools in the area, integrating them with information obtained from the Population and Dwelling census. This comparison allows deducing information on the school's catchment area (attending families, average age of the population in the territory) as well as age and maintenance status of the school building. It could be also interesting to integrate data related to schools with data related to public transportation.

A specific SDS service, named S4Y (School for you) will be developed in order to reach the above described objectives. The S4Y service will be able to be invoked to answer queries combining both school's data and demographic data for both Italian and French territories.

3.1.2 Data used

The proposed service will use as data sources following datasets:

- Italian Census data: Data of the Census of Population and Housing aggregated at different territorial levels coming from ISTAT (http://dati-censimentopopolazione.istat.it/Index.aspx)
- French Census data: Data of the Census of Population and Housing aggregated at different territorial levels coming from Insee (latest data at https://www.insee.fr/fr/information/2008354)
- Schools data: data published in schools web portals in Italy and France, in Italy also available at Education Ministry with the portal "Scuola in chiaro" at https://cercalatuascuola.istruzione.it/cercalatuascuola/. In France, open data on education is available as a DCAT catalogue at https://data.education.gouv.fr/api/v2/catalog/exports/rdf
- Schools national self-assessment outcomes: data on results of schools students in national tests (in Italy available at INVALSI, in France at Ministry of Education, at state level at OECD-PISA) https://www.invalsiopen.it/ , https://www.oecd.org/pisa/data/
- Datasets on education from European Data Portal https://www.europeandataportal.eu/en/highlights/open-education-data-european-data-portal

3.1.3 Evaluation

The S4Y SDS will be designed either to provide a "seamless" integration level or at least an "integrated" one. The service will solve content harmonization issues between nations, between domains and between contents coming from public sector and contents coming from private portals performing schools evaluation like https://www.eduscopio.it/.

The most important business benefit of the SDS service is to show how realizing a **cross-border and cross-domain statistical** data service. Indeed:

- S4Y will be cross-border, involving at least data from France and Italy.
- S4Y will be cross-domain, having the relevant and interesting feature of integrating data from the statistical domain (population census) with data from the education domain (school results

evaluation). While dedicated efforts have been paid for EU level data harmonization within a domain, S4Y will show how these efforts can be exploited also for cross-domain services.

The service will provide cross-border and cross-domain integrated information available thought Italian and French NSIs Data Portal and other sources of information, integrating at semantic level data coming from different web portals and providing users a single access point to this information giving him a support to his decision-making activities.

Given that the service can be used by families, the opportunity to develop an app to be used on smartphones will be evaluated.

3.2 Geolocalized Facilities

3.2.1 Description

The BPE[31] (Base permanente des équipements, or Permanent database of facilities) is a database published annually by Insee which contains the localisation and characteristics of more than 2.5 million facilities covering a wide range of sectors (health, sports, culture, education, etc.) all over the French territory. The usability of the BPE has been considerably improved recently as Insee switched from ancient proprietary formats to CSV and released data in evolution.

The BPE is a very rich source of information, especially at the local level ; in particular, it can be combined with fine population estimates from the census[32] or at small grid level[33] to offer insight on where to invest for given types of facilities, develop market studies, etc.

The BPE also provides a natural integration framework for sources like lists of cultural or sports events, time schedules of educational establishments, or actually 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 (e. g. OpenAgenda[34]), including the European data portal, where more than ten thousand datasets are listed in 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. It will be important to describe limitations, constraints and level of comparability in each case: a high degree of transparency will augment trust in the information made available.

The BPE can support a large variety of use cases, out of which two are proposed: the visitor use case and the local decider one. The user stories will be refined during the starting phase of the project, but a broad idea is presented below.

[31] https://www.insee.fr/en/metadonnees/source/serie/s1161
[32] https://www.insee.fr/fr/information/2008354
[33] https://www.insee.fr/fr/statistiques/4176290?sommaire=4176305&q=carreau
[34] https://public.opendatasoft.com/explore/dataset/evenements-publics-cibul/

In the first case, we will 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. This information should be made available on the user's smartphone via an easy-to-use interface. 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. The decider 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 combine this information with other sources more specifically relevant for his particular problem, so it would be interesting to provide the data through a simple web application based on web services.

3.2.2 Data used

The proposed service will use as data sources following datasets:

- BPE, Permanent database of facilities, published by: Insee (latest data are published at https://www.insee.fr/fr/statistiques/3568656)
- French Census: yearly data of the population census aggregated at different territorial levels, published by: Insee
- Italian Census: data of the Census of Population and Housing aggregated at different territorial levels, published by: Istat http://dati-censimentopopolazione.istat.it/Index.aspx
- Italian cultural facilities database: database published by Italian Ministry of Cultural Heritage and Activities (MIBACT), containing cultural events (exhibitions, conferences, conventions, seminars, and so on), published by: MIBACT https://www.beniculturali.it/open-data-e-linked-data
- French grid data: socio-demographic data at fine grid (200m) level, published by: Insee
- Open Agenda: database on public events, published by: Opendatasoft https://public.opendatasoft.com/explore/dataset/evenements-publics-cibul/table/?disjunctive.tags&disjunctive.placename&disjunctive.city
- European Data Portal datasets: a selection of datasets available on the European Data Portal will be made, based on interest and reliability, published by: Publications Office of the European Union https://www.europeandataportal.eu/en/highlights/cultural-institutions-and-cultural-open-data

3.2.3 Evaluation

This service will provide integrated cross-border and cross-domain integrated information available thought Italian and French NSIs Data Portal and other sources of information. The BPE data is seamlessly integrated with the French Census data through the use of common geographical classifications and identical concepts for facilities. Italian and French Census Data are integrated at the legal interoperability level. BPE data and French Census data are published every year, so the evolution of facilities can be compared to the changes in population.

3.3 Support for Environment Policies

3.3.1 Description

This use case has the objective to support local policy makers who have to take decisions about environmental policies to be applied in a city. In particular, local policy makers can benefit from integrated datasets deriving from: (i) sensor data concerning air pollution and (ii) statistical data regarding demographic characterization of the city's areas. On the basis of such data, they are able to make several analyses to help planning and governing their interventions.

As an example, possible analyses could be related to set priorities between the city' areas in order to plan in which intervene first, namely: (i) if there are areas in which the air pollution levels exceed the permitted threshold levels, among them, those that are densely populated could have a highest priority of intervention; (ii) it could be interesting to make analyses related to the weaker sections of the population and assess the impact of the air pollution levels on them, etc.

3.3.2 Data used

A specific SDS service, named SEP (Support for environmental policies) will be developed in order to reach the above described objectives. The SEP service will be able to be invoked to answer queries combining both air pollution data and demographic data for both Italian and French territories.

The proposed service could use as data sources:

- air pollution data published by Environmental Protection Agencies e.g in Italy at https://www.isprambiente.gov.it/it/banche-dati
- statistical demographic data coming from the Population Census and data published by the European Environment Agency.
- In Italy, for example, the ARPA (Regional Agency for Environmental Protection) publishes daily data on air quality on its site. ARPA gathers pollution data through its smart devices spread over the territory and publish them on grid areas of one km^2 https://qa.arpalazio.net/exportData.php
- In France, the IMREDD (Mediterranean Institute for Risks, Environment and Sustainable Development) publishes on its site the same data on air pollution on grid areas of 25 m^2.
- In addition, the European Environment Agency publishes several datasets related to air pollution for each EU country https://www.eea.europa.eu/themes/air

The SEP SDS will be designed either to provide a "seamless" integration level or at least an "integrated" one. Same data models will be used, but the territorial dimensions can pose some content harmonization issues, as explained below.

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

3.3.3 Evaluation

The most important business benefit of the SDS service is to show how realizing a cross-border and cross-domain statistical data service. Indeed:

- SEP will be cross-border, involving at least data from France and Italy.
- SEP will be cross-domain, having the relevant and interesting feature of integrating data from the statistical domain (population census) with data from the environmental domain (air pollution). While dedicated efforts have been paid for EU level data harmonization *within* a domain, SEP will show how these efforts can be exploited also for cross-domain services.

As mentioned, the SEP SDS will be designed to provide “seamless” and “integrated” functions. Discrepancies with respects to the geographical grids of air pollution data will be taken into account and where possible overcome, for instance in all the cases where policy makers are interested to city-level data, results can be related to Italian and French cities (i.e. the cities will be taken as reference geographical entities instead of the grid cells). Instead, for all cases in which grid discrepancies cannot be overcome, an “integration” level of the service will be anyway ensured.

Conclusion

In this document, we have proposed a framework for the analysis of the services provided through the INTERSTAT platform according to a number of criteria based on well-known references. We have reviewed the pilot use cases proposed by the project in the light of this framework in order to assess their value from a cross-border point of view.

The present document can be considered as a first shot at a service evaluation framework and how the pilot use cases can be used to in a continuous improvement approach to improve both the pilot services and the framework itself.

We believe that these analysis and assessments confirm the relevance of the INTERSTAT project as a means to improve the value and facilitate the efficient use of statistical data for users across Europe.

References

[1] United Nations Economic Commission for Europe , “Generic Statistical Information Model (GSIM) - version 1.2,” [Online]. Available: https://statswiki.unece.org/display/gsim/.

[2] The Data Documentation Initiative Alliance, “DDI Lifecycle – version 3.3,” [Online]. Available: https://ddialliance.org/Specification/DDI-Lifecycle/3.3/.

[3] W3C Recommendation, “RDF Data Cube Vocabulary,” [Online]. Available: https://www.w3.org/TR/vocab-data-cube/.

[4] “SDMX information model – version 2.1,” [Online]. Available: https://sdmx.org/wp-content/uploads/SDMX_2-1_SECTION_2_InformationModel_2020-07.pdf.

[5] “VTL (Validation & Transformation Language) – version 2.0,” [Online]. Available: https://sdmx.org/wp-content/uploads/VTL-2.0-User-Manual-20180416-final.pdf.

[6] European Commission, “The European Interoperability Framework,” [Online]. Available: https://ec.europa.eu/isa2/eif_en.

[7] United Nations Economic Commission for Europe, “High-Level Group for the Modernisation of Official Statistics,” [Online]. Available: https://statswiki.unece.org/display/hlgbas.

[8] United Nations Economic Commission for Europe, “Generic Statistical Business Process Model (GSBPM) – version 5.1,” [Online]. Available: https://statswiki.unece.org/display/GSBPM.

[9] United Nations Economic Commission for Europe, “Generic Activity Model for Statistical Organisations (GAMSO) – version 1.2,” [Online]. Available: https://statswiki.unece.org/display/GAMSO .

[10] “EU legislation on the 2021 population and housing censuses – Explanatory notes,” 2019. [Online]. Available: https://ec.europa.eu/eurostat/documents/3859598/9670557/KS-GQ-18-010-EN-N.pdf.

[11] “Overview on quality for European statistics,” [Online]. Available: https://ec.europa.eu/eurostat/web/quality.

[12] “The European Statistics Code of Practice,” [Online]. Available: https://ec.europa.eu/eurostat/web/quality/european-statistics-code-of-practice.

[13] “Quality Assurance Framework of the European Statistical System,” [Online]. Available: https://ec.europa.eu/eurostat/documents/64157/4392716/ESS-QAF-V1-2final.pdf/bbf5970c-1adf-46c8-afc3-58ce177a0646.

[14] “Single Integrated Metadata Structure (SIMS),” [Online]. Available: https://ec.europa.eu/eurostat/documents/64157/4373903/03-Single-Integrated-Metadata-Structure-and-its-Technical-Manual.pdf.

[15] European Commission, “Turning FAIR into reality – Final report and action plan from the European Commission expert group on FAIR data,” [Online]. Available: https://op.europa.eu/s/oArl.