



Open Statistical Data Interoperability Framework

www.cef-interstat.eu

D3.3 – Impact assessment report



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 3.3 (Activity 3)

Project Document Delivery Date

31.08.2023

Deliverable Type and Security

Report – Public

This document is licensed under a [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/)

Authors: Raffaella Maria Aracri (ISTAT), Francesca D'Agresti (ENG), Romina Filippini, (ISTAT), Rosa Maria Lipsi (ISTAT), Paolo Francescangeli (ISTAT), Renato Magistro (ISTAT), Giulio Massacci (ISTAT), Francesco Ortame (ISTAT), Giuseppina Ruocco (ISTAT)

Contributors

Giovanna Brancato (ISTAT), Franck Cotton (Insee), Romain Tailhurat (Insee), Dubois Thomas (Insee), Fernando Lòpez (FIWARE)

Reviewers

Martino Maggio (ENG)

Table of Contents

| | | |
|----------|---|-----------|
| 1 | Introduction..... | 7 |
| 1.1 | Scope and objectives of the document..... | 7 |
| 1.2 | Overview of project outcomes..... | 8 |
| 1.2.1 | Technical website enhancements..... | 8 |
| 1.2.2 | Pilot applications improvements..... | 11 |
| 2 | Methodology for the impact assessment | 13 |
| 2.1 | Assessment dimensions | 13 |
| 2.2 | KPIs for the impact assessment..... | 14 |
| 2.3 | Stakeholders' engagement: the Assessment Survey | 16 |
| 3 | Assessment of INTERSTAT impact..... | 19 |
| 3.1 | Analysis of respondents..... | 19 |
| 3.2 | Impact of the INTERSTAT framework..... | 24 |
| 3.2.1 | Use cases and added value of INTERSTAT tools..... | 33 |
| 3.3 | Pilot applications and users experience..... | 44 |
| 3.4 | Summary of survey results | 48 |
| 3.5 | MQA compliance | 51 |
| 4 | Lessons learned and guidelines | 64 |
| 5 | Conclusions | 66 |
| | References..... | 67 |
| | Annex: Assessment survey Questionnaire..... | 69 |

List of figures

| | |
|--|----|
| Figure 1: Relevant official standards fostering data and metadata interoperability | 9 |
| Figure 2: Example of a LOSD workflow | 11 |
| Figure 3: Respondents by main field of activity | 20 |
| Figure 4: Respondents by type of organization..... | 20 |
| Figure 5: Respondents background in relation to the INTERSTAT framework | 21 |
| Figure 6: Respondents by LOSD experience | 22 |
| Figure 7: Respondents by LOSD experience and background..... | 23 |
| Figure 8: Respondents by LOSD experience and main field of activity | 23 |
| Figure 9: Respondents by navigation purposes | 25 |
| Figure 10: Respondents by LOSD experience and navigation purposes..... | 25 |
| Figure 11: Respondents evaluation of the technical website | 26 |
| Figure 12: Respondents by LOSD experience and evaluation of the technical website | 27 |
| Figure 13: Respondents evaluation of the INTERSTAT tools | 28 |
| Figure 14: Respondents evaluation of the INTERSTAT tools and technical website | 29 |
| Figure 15: Respondents evaluation of open tools to integrate in the INTERSTAT framework..... | 30 |
| Figure 16: Respondents evaluation of cross-border and cross-domain analysis | 45 |
| Figure 17: Respondents assessment of pilots' navigation | 45 |
| Figure 18: KPIs overview assessing the project impact | 48 |
| Figure 19: Overall rating of the INTERSTAT project | 50 |
| Figure 20: INTERSTAT open data catalogue published on Idra platform..... | 52 |
| Figure 21: List of datasets of the INTERSTAT catalogue on Idra platform..... | 53 |
| Figure 22: Schematic image of harvesting catalogue using Idra | 54 |
| Figure 23: INTERSTAT catalogue published in EDP | 54 |
| Figure 24: Results of the MQA related to the GF pilot Dataset [15] | 59 |
| Figure 25: INTERSTAT Catalogue MQA | 60 |
| Figure 26: INTERSTAT Catalogue MQA - Rating evolution | 61 |
| Figure 27: INTERSTAT Catalogue MQA - Findability | 61 |
| Figure 28: INTERSTAT Catalogue MQA - Accessibility..... | 62 |
| Figure 29: INTERSTAT Catalogue MQA - Interoperability | 62 |
| Figure 30: INTERSTAT Catalogue MQA - Reusability | 63 |
| Figure 31: INTERSTAT Catalogue MQA – Contextuality | 63 |

List of tables

| | |
|---|----|
| Table 1: List of tools provided by the INTERSTAT framework | 8 |
| Table 2 - Initial list of KPIs and assessment dimensions | 15 |
| Table 3: KPIs by type of assessment and stakeholders to engage | 18 |
| Table 4: First section questions | 19 |
| Table 5: Second section questions | 24 |
| Table 6: Scores assigned to Section 2 questions | 31 |
| Table 7: KPIs assessing the INTERSTAT framework | 32 |
| Table 8: Third section questions | 44 |
| Table 9: Scores assigned to Section 3 questions | 46 |
| Table 10: KPIs assessing the Pilot services..... | 47 |
| Table 11 - Profiles for Synthetic Index decoding | 49 |
| Table 12: Interstat open datasets harvested by the EDP | 56 |
| Table 13: Subset of metadata provided for each Dataset harvested by the EDP | 58 |

Executive Summary

At the end of the implementation tasks of INTERSTAT project, this deliverable describes the activities related to Milestone 7 – “Impact assessment report” and complete Activity 3 of the INTERSTAT initiative (“Pilot services execution and assessment”). The evaluation of the project output has concerned both the technical framework and the pilot applications developed to test the efficiency of the solutions provided by the framework.

The first chapter briefly describes the main outputs of the project, the enhancements of the technical framework and the services applications, which continued evolving after the completion of Milestone 6 - Pilot services deployed and working in real environments.

The second chapter focuses on the methodology adopted to assess the project impact, mainly based on the feedback collected through external stakeholders and the launch of an Assessment Survey (AS).

The third chapter reports the main results of the AS, and the KPIs computed for each dimension to evaluate. In addition, the assessment analysis has considered the compliance with the Metadata Quality Assurance (MQA) framework, validating the metadata related to the datasets published in the INTERSTAT open catalogue harvested by the European Data Portal (data.europa.eu).

The report ends with an overview of the main lessons learned during the development activities and some final considerations about the impact of the project, fully compliant with the initial expectations to foster the reuse and the dissemination of open statistical data.

1 Introduction

1.1 Scope and objectives of the document

At the end of the implementation activities of the INTERSTAT project, one of the main tasks performed has concerned the evaluation of the project impact. More in detail, the effort was concentrated on assessing the relevance of the outcome of the action, considering not only the pilot applications implemented, or the benefits for the main groups of stakeholders, but also the increase of technical and semantic interoperability due to the use of the tools integrated in the INTERSTAT framework.

The measurement of the impact of the action thus included several dimensions and engaged internal and external stakeholders. These dimensions are related to the key milestones of the project, such as:

- The services applications designed and implemented to test the INTERSTAT tools
- The compliance degree of the INTERSTAT data catalogue published in the European Data Portal (data.europa.eu, hereinafter referred to as EDP) with the metadata quality framework, conceived to help data providers and data portals to check their metadata against the FAIR (Findable, Accessible, Interoperable, Reusable) principles.

This report provides an overview of the methodology adopted to define a set of KPIs (Key Performance Indicators) to measure the technical achievements in terms of efficiency, efficacy, reusability and shareability of implemented solutions. For this purpose, both the enhancements of the INTERSTAT framework and the pilots improvements have continued after the deadlines scheduled for the deliverables. The KPIs described in the next paragraphs are the result of a survey launched to collect basic information related to the different validation dimensions. After data processing, a final rating of KPIs related to each dimension has been summarized to have an overall assessment of the project relevance.

In the project plan, the "Impact assessment report" corresponds to Milestone 7 (M7), which complements Activity 3 ("Pilot services execution and assessment"), extending until the end of the project the effort dedicated to the improvement, monitoring and assessment of the pilots.

1.2 Overview of project outcomes

The next paragraphs provide an overview of the most relevant enhancements of both the INTERSTAT framework and the pilots, to highlight and make the achieved results more user-friendly, mainly for non-technical end-users. Actually, the steps to produce and publish Linked Open Statistical Data (LOSD) are quite difficult to describe to other domain experts. Consequently, after the first release of the implemented solutions, the main efforts were directed to the improvement of the user experience for not-expert users. After the pilots' deployment, the technical activities have also concerned the development of a converter from SDMX format to NGSI-LD specification, thus supporting the dissemination of LOSD in the EDP through the FIWARE Context broker (formerly CEF Building Block).

1.2.1 Technical website enhancements

The INTERSTAT framework provides a set of tools to enable and increase semantic and technical interoperability. In order to improve the navigation experience of the technical website, the tools were grouped according to the main steps of a generic data pipeline for LOSD production and dissemination, as reported in Table 1.

| | | |
|---|--|--|
| Tools for Data Publication <ul style="list-style-type: none"> • Idra Open data federation platform • GraphDB | Tools for Data Dissemination <ul style="list-style-type: none"> • CEF Orion Context broker • Data browser • Eurostat NSI Web service | Tools for Data or Metadata Management <ul style="list-style-type: none"> • Adminer MySQL • Meta and Data Manager • SparQling • Monolith |
| Tools for Data Visualization <ul style="list-style-type: none"> • Cube Visualizer • Olap Browser | | Tools for Data Transformation <ul style="list-style-type: none"> • Excel or CSV to NGSI-LD • SDMX/NGSI-LD Parser • Eddy • Juma Editor |

Table 1: List of tools provided by the INTERSTAT framework

In addition to the tools, catalogue and manuals, the framework was enhanced to offer an entire section concerning the main official standards to improve technical and semantic interoperability

and the data pipelines. Currently, in this section of the technical website [1], the side navigation menu provides the link related to each standard shown in Figure 1, complemented by the following descriptions:

- The Generic Statistical Business Process Model (GSBPM) is the reference framework to describe and design the statistical process. GSBPM allows identifying the several steps of the statistical business process and the connection between them.
- SDMX: Statistical Data and Metadata eXchange is a standard for the exchange of statistical data and metadata among international organisations.
- DCAT-AP/statDCAT-AP: DCAT Application profile for data portals in Europe is a specification based on W3C's Data Catalogue vocabulary (DCAT) for describing public sector datasets in Europe. StatDCAT-AP is a DCAT-AP extension for the exchange of metadata for statistical datasets.
- ETSI NGSI-LD: The Context Information Management API allows users to provide, consume and subscribe to context information in multiple scenarios and involving multiple stakeholders.



Figure 1: Relevant official standards fostering data and metadata interoperability

Following the official standards, the side navigation menu allows to access a brief description of the data pipelines implemented for the pilot applications. The two approaches converged at the end of the data workflow, in the dissemination step are:

- The ETL approach, based on a generalised ETL (Extract, transform, load) pattern that generates RDF triples from a CSV Dataset through Python procedures. The main advantages of this approach are:
 - ✓ Openness: the code, developed using open tools, is available in the INTERSTAT GitHub repository
 - ✓ Maximal automation, to avoid manual treatments, save time and improve traceability
 - ✓ Reproducibility, resulting from automation and code documentation
 - ✓ Efficiency, increased by the execution of the pipeline in a distributed environment.
- The Domain Knowledge approach, based on the:
 - ✓ Description of the domain of interest through an ontology, representing the core concepts
 - ✓ Definition of a logical Common Data Model to link heterogeneous data sources with ontology concepts.

The main steps of this data pipeline are:

- ✓ Data Acquisition: the datasets to link are downloaded from sources and uploaded in a DBMS
- ✓ Data Processing: Data from different sources are loaded in a DBMS with the Common Data Model. In this step data are not integrated, but are harmonized and federated
- ✓ Conceptual integration, to link ontology concepts to Data in the Common Data Model, using Monolith tool and query data through ontology concepts.

Scrolling the menu, a user can also analyze in detail an example of a data pipeline for LOSD production and dissemination, complemented by the tools executing each step, as depicted in the following figure.

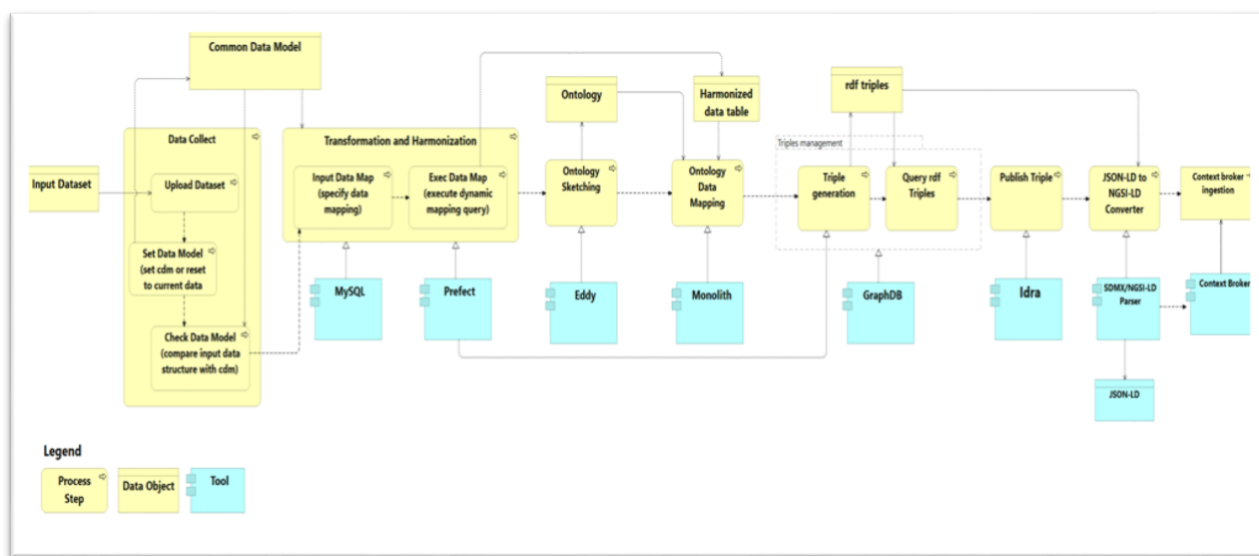


Figure 2: Example of a LOSD workflow

1.2.2 Pilot applications improvements

Once achieved Milestone 6 (M6) of the project, concerning the "Pilot services deployed and working in real environments", the work has continued with additional activities, dedicated to the monitoring and assessment of the pilots. Therefore, the applications services, properly documented in "Deliverable 3.2 - Report on pilots' development and deployment" [2], have evolved by continuous improvements until the end of the project. The revision of the pilots has concerned mainly the front-end, to highlight the cross-border and cross-domain analysis, and emphasize that the main goal of the service applications was to test the technical solutions provided by the INTERSTAT framework.

The following descriptions provide an overview of the use cases resulting from the integration of open statistical data sources published by different data providers and related to several domains.

Support for Environment Policies (SEP)

The SEP [3] is a pilot application developed using the tools provided by the INTERSTAT framework and linking air pollution and demographic data related to Italian and French regions. To test cross-border interoperability, the SEP pilot allows comparing air pollution indicators in selected Italian and French areas, such as big cities, particularly Rome and Paris. Concerning cross-domain

interoperability, the SEP service combines air quality indicators collected in different stations of each Municipality, with the age structure of resident population.

Geolocalize Facilities (GF)

The GF client application [4] is an example of cross-border and cross-domain interoperability, linking and geolocalising data collected in France and Italy concerning public infrastructures. As a result, information about facilities, equipments and events, is contextualized in space and integrated with other data sources, such as the population census.

The School For You (S4Y)

The S4Y [5] integrates Italian and French data concerning the resident population and educational characteristics, such as school attendance and the distribution of public and private schools.

2 Methodology for the impact assessment

This section focuses on the several aspects and the elements combined to develop the methodology for the impact assessment. The key elements, analysed in the next paragraphs, are:

- The **dimensions** to evaluate, corresponding to the areas to define for assessing project efficacy and effectiveness
- The **KPIs** to compute for measuring the relevance of the action
- The **stakeholders** to engage and the tasks to execute for the impact assessment.

2.1 Assessment dimensions

The methodology applied to assess the impact of the project aims to answer the following questions:

1. **WHAT** are the main dimensions to consider for the assessment? In relation to each dimension, which are the most relevant KPIs to evaluate the action?
2. **HOW** to measure the project performance?
3. **WHO** are the stakeholders to involve in the assessment?

Concerning the first issue, after several consultations within the Consortium, the areas identified for the impact measurement, related to the main outcomes of the project are:

- The **INTERSTAT framework** to assess the benefits of its usage, focusing on tools shareability, service reuse, increase of technical interoperability
- The **Data pipelines** - only for more experienced users - to assess the pilots architecture, through the evaluation of the data pipelines developed for the use cases
- The **Client applications**, to evaluate the implemented use cases with a specific focus on the front-end and the user experience
- Compliance with **Metadata Quality Assurance (MQA)**, to measure the accuracy of the metadata descriptions of the INTERSTAT data catalogue harvested by the European Data Portal.

2.2 KPIs for the impact assessment

The guiding principle adopted for the definition of the KPIs to assess each dimension is the "SMART criteria". According to this principle, the KPIs should have the following characteristics¹:

- The measure relates to a **Specific** area to assess or improve
- The assessment is **Measurable**, to assign a value of the KPI, either quantitative or qualitative
- The defined goals have to be **Achievable**, although challenging
- The improvement has to be **Relevant** with respect to the organization objectives, and **Time** phased, that is related to a specific period.

The table below reports the list of KPIs defined for each dimension to validate.

| What | | | |
|------------------------|--|---|--------------|
| Assessment dimension | KPI | Rationale behind the KPI | KPI type |
| 1. INTERSTAT Framework | 1.1 - Number of use cases covered by the tools integrated in the framework | Description of the different use cases covered by the framework to demonstrate that tools can be combined in different pipelines. Efficiency assessment of the framework in terms of open tools and degree of technical interoperability | Quantitative |
| | 1.2 - Percentage of Open Tools | Measuring the degree of openness and shareability of INTERSTAT tools | |
| | 1.3 -Main benefits of using the INTERSTAT framework | Overview of the expectations and requirements of different type of users met by the INTERSTAT framework | Qualitative |
| | 1.4 - Clarity of key concepts (communication efficacy) | Efficacy of the INTERSTAT framework in promoting principles and tools to increase interoperability | Qualitative |

¹ https://en.wikipedia.org/wiki/SMART_criteria

| What | | | |
|---|--|--|-----------------|
| Assessment dimension | KPI | Rationale behind the KPI | KPI type |
| | 1.5 - Effectiveness of INTERSTAT tools | Relevance of the tools provided by INTERSTAT framework | Qualitative |
| | 1.6 - Degree of completeness | Evaluate the framework efficacy in terms of exhaustiveness of the tools provided to execute a data pipeline | Qualitative |
| 2. Data pipeline assessment (only for Technical User - IT Staff) | 2.1 - Degree of replicability | Evaluate the degree of resilience of data pipelines in terms of replicability and scalability | Qualitative |
| | 2.2 - Degree of scalability | | |
| | 2.3 - Degree of adaptability: - difficulty to add a new source - difficulty to translate into another programming language | Evaluate the degree of code adaptivity to manage data sources or programming languages changes | Qualitative |
| 3. Client applications | 3.1 - Relevance of cross-border analysis | Efficacy of the pilots to demonstrate the feasibility of cross-border services | Qualitative |
| | 3.2 - Relevance of cross-domain analysis | Efficacy of the pilots to demonstrate the feasibility of cross-domain services | Qualitative |
| | 3.3 - Users experiences assessment | Assess whether and to each extent the pilots interface is user-friendly. Measuring users satisfaction during pilot navigation | Qualitative |
| 4. MQA compliance | 4.1 - MQA score for each Dataset harvested by the European Data portal | Assessment of each pilot through the score resulting from the MQA assessment service | Quantitative |

Table 2 - Initial list of KPIs and assessment dimensions

2.3 Stakeholders' engagement: the Assessment Survey

After the selection of the assessment dimensions and the related KPIs, the next steps have concerned the definition of the best method to proceed and involve external stakeholders. Finally, to collect feedbacks related to the different dimensions from potential users of the INTERSTAT framework, the Consortium decided to launch an Assessment Survey (AS).

As a starting point, external stakeholders were grouped in the following subsets, according to their technical skills and requirements:

- *Technical User / IT staff* dealing with statistical data. They work with all the technologies that concern data acquisition and processing, as well as metadata. They may take advantage of the tools and technologies that the INTERSTAT framework offers to increase data interoperability. From a technical point of view, they have greater IT skills than statistical domain experts and data analysts. Specifically, they can manage tools and IT infrastructures to execute and monitor the main steps of a LOSD data pipeline.
- *Statistical Expert or equivalent (Data Analyst, Data Scientist)*, using open data for a statistical scope. They are users who have subject matter knowledge, and the ability to extract new knowledge from the data previously created or processed by the technical users. They often interact and support IT staff by suggesting additional data analysis. It is not supposed that they have advanced technical skills: they are experts in understanding statistical data, and can benefit from the interoperability between different standards and data models provided by the INTERSTAT framework, thus enabling new value-added analysis across different domains.
- *Other Domain Expert - General User*: citizens, decision makers or other type of users who intend to approach to the INTERSTAT framework without having particular technical skills or knowledge of languages for extracting information. They may want to read open data, access open data portals or may use very simple tools to explore data; they can check a web portal with some information, tables, and maps. They will mainly be able to explore open data through the various types of visualization that the INTERSTAT framework

offers, or use functions previously implemented within the framework by more experienced users.

The following table summarizes the KPIs identified in the exploratory analysis, and the solutions to the last two questions guiding the assessment investigation: How to measure the project performance and what type of stakeholders to engage in the assessment? The self-assessment has involved the members of the INTERSTAT Consortium (INTERSTAT partners).

| What | | How | Who |
|--|--|---------------------------------|-------------------------------|
| Assessment dimension | KPI | Performance Measurement | Stakeholders to involve |
| 1. INTERSTAT Framework | 1.1 - Number of use cases covered by the tools integrated in the framework | Self-assessment | INTERSTAT partners |
| | 1.2 - Percentage of Open Tools | | |
| | 1.3 - Main benefits of using the INTERSTAT framework | Assessment Survey Questionnaire | External & INTERSTAT partners |
| | 1.4 - Clarity of key concepts (communication efficacy) | | |
| | 1.5 - Effectiveness of INTERSTAT tools | | |
| | 1.6 - Degree of completeness | | |
| Data pipeline assessment (only for Technical User - IT Staff) | 2.1 - Degree of replicability | Assessment Survey Questionnaire | External & INTERSTAT partners |
| | 2.2 - Degree of scalability | | |
| | 2.3 - Degree of adaptability: - difficulty to add a new source - difficulty to translate in another programming language | | |
| Client applications | 3.1 - Relevance of cross-border analysis | Assessment Survey Questionnaire | External & INTERSTAT partners |
| | 3.2 - Relevance of cross-domain analysis | | |
| | 3.3 - Users experiences assessment | | |

| What | | How | Who |
|-----------------------|---|---|---------------------------|
| Assessment dimension | KPI | Performance Measurement | Stakeholders to involve |
| MQA compliance | 4.1 - MQA score of the Datasets harvested by the European Data portal | Self-assessment: for each pilot, report the score resulting from the MQA assessment service | INTERSTAT partners |

Table 3: KPIs by type of assessment and stakeholders to engage

After the definition of the target groups, a questionnaire was designed to collect basic information for each assessment dimension and then compute the related KPIs. The questionnaire, reported in the Annex, is composed by the following sections:

- **General information**, for profiling the respondents interested or curious about the project achievements and tasks
- **Assessing the INTERSTAT Framework**, to gain external insights about the technical website and the tools offered for producing and disseminating LOSD
- **Assessing the Data Pipeline** (only for Technical Users - IT Staff), to ask a benchmark about the data workflow implemented for the pilot applications. Due to the very low number of Technical Users participating to the Assessment Survey, this dimension was evaluated through the Client applications analysis
- **Assessing the Client Applications**, to gather the respondents viewpoint about the front-end, and the relevance of cross-border and cross-domain analyses
- **Final Comments and Feedback**, a text box allowing the respondent to add some remarks.

3 Assessment of INTERSTAT impact

The evaluation of the impact of the action has involved both, the Consortium members and external stakeholders. The Assessment Survey was launched to collect feedback from potential users of the project outcome and verify whether their needs were met. The data collection started on July 10, for two weeks, gathering feedback on the developed solutions from 30 respondents. This chapter reports the main results of the survey, used to derive the KPIs for each assessment dimension, and summarized through a final score.

3.1 Analysis of respondents

The first section of the questionnaire was conceived to profile the several types of stakeholders, potentially dealing with interoperability issues, and accessing the INTERSTAT framework to explore available solutions. The information collected does not include personal data, but refers mainly to the respondent's knowledge and experience in LOSD. More in detail, this preliminary section is composed of four questions, reported in the table below.

| Section1 - General information | | | |
|---|--|--|---|
| Question 1.1 - What is your main field of activity ? | Question 1.2 - What type of organization do you work for? | Question 1.3 - What type of user are you with respect to the INTERSTAT framework? | Question 1.4 - Rate your experience in Linked Open Statistical Data (LOSD) |

Table 4: First section questions

The following figure shows the distribution of the respondents according to their main field of activity. Most of them come mainly from the Education, Information technology and Statistics sectors (about 53%), while the remaining operates in other specific sectors, such as Transportation & Logistics and Economics.

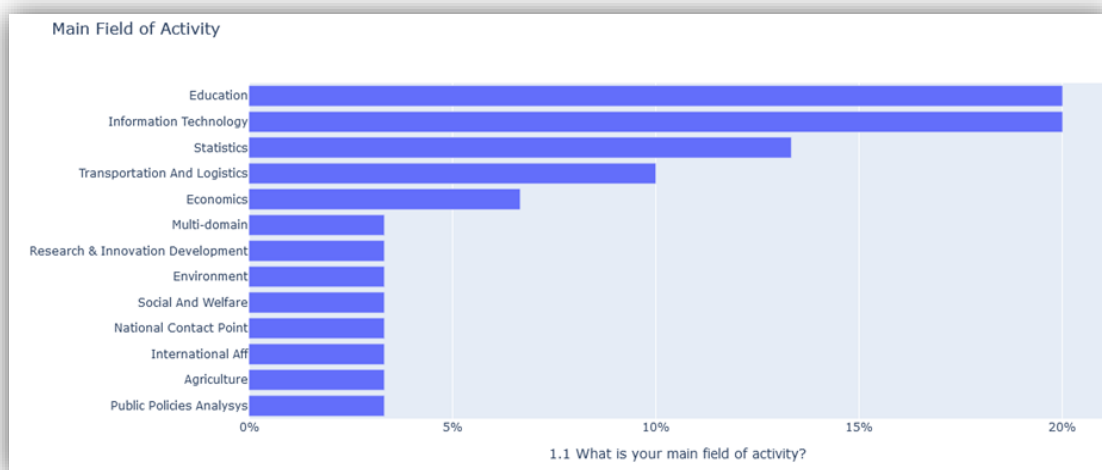


Figure 3: Respondents by main field of activity

Concerning the type of organization, almost all respondents come from the public sector (more than 90%), as illustrated in the following chart.



Figure 4: Respondents by type of organization

According to the three main subsets (Technical User - IT Staff, Other domain expert - General user, Statistical expert or equivalent) described in the previous chapter, the distribution of stakeholders reveals that 60% of respondents have a background purely as Data Analyst or a Data Scientist. The remaining is composed of Other Domain Experts, or General Users, while only one Technical User participated to the survey.

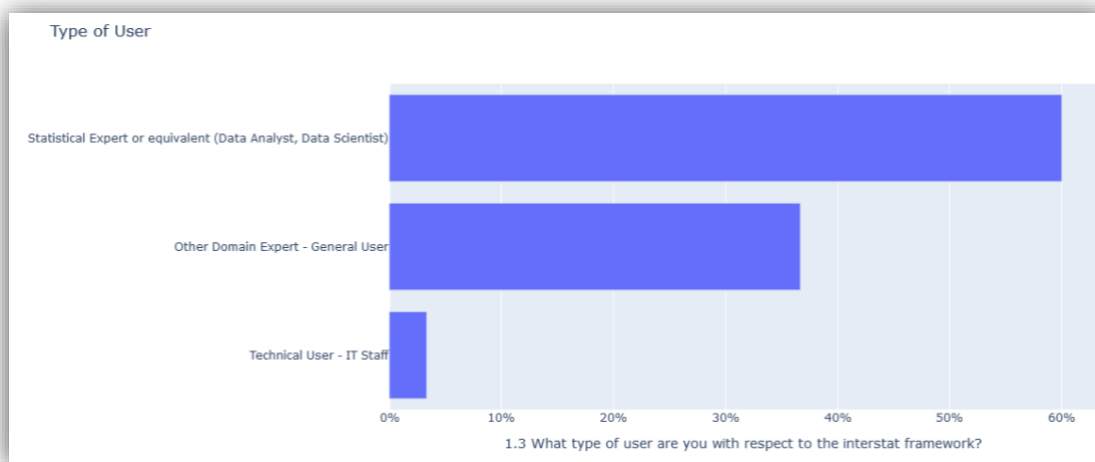


Figure 5: Respondents background in relation to the INTERSTAT framework

Considering the level of experience in LOSD, the distribution of the respondents is mainly divided into "Beginner" users (60%,) and "Intermediate" users (33%). Only a small portion of them consider themselves experts (about 7%).

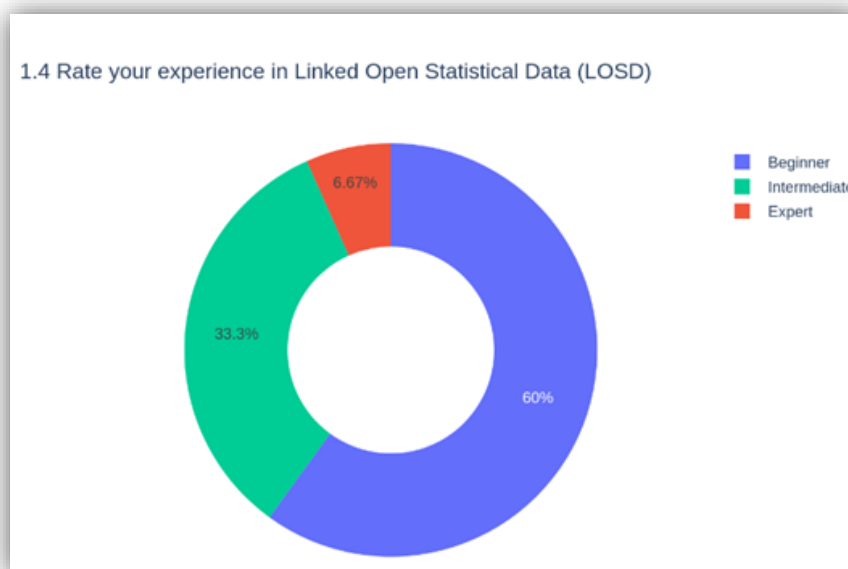


Figure 6: Respondents by LOSD experience

The joint analysis of the respondent experience in LOSD and the user type with respect to the INTERSTAT framework reveals that Beginner users are mainly General users, while Intermediate users and those who have a great deal of knowledge in open data are mostly concentrated in the subset of Statistical experts or equivalent (Data analyst/Data scientist).

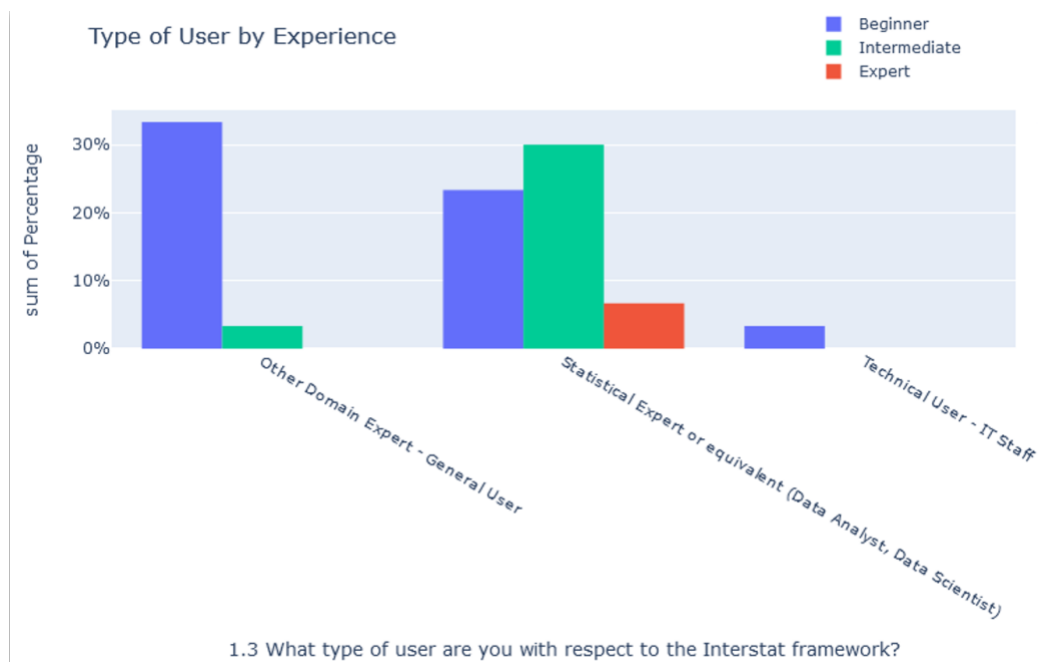


Figure 7: Respondents by LOSD experience and background

In order to complete the initial exploration and respondent profiling, the intersection of the level of experience in LOSD and the main field of activity highlights that Beginners work mainly in IT and Education sectors (26%). People with an Intermediate level of experience belong mostly to the Statistical sector (10%), while Experts are distributed uniformly in IT sector and other domains.

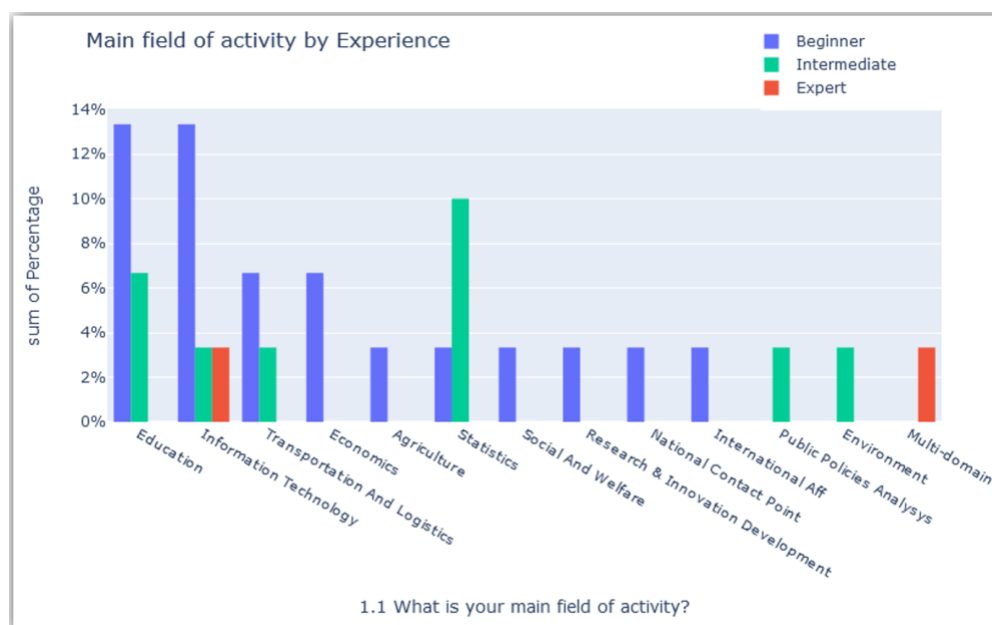


Figure 8: Respondents by LOSD experience and main field of activity

3.2 Impact of the INTERSTAT framework

The second section of the questionnaire: “Assessing the INTERSTAT Framework” was designed to collect an external feedback, with respect to the technical website realized for reusing existing open source tools tested for the pilots development. The list of KPIs and the related questions are reported in the table below.

| KPIs | | | |
|---|--|--|--|
| 1.3 - Main benefits of using the INTERSTAT framework | 1.4 - Clarity of key concepts (communication efficacy) | 1.5 - Effectiveness of INTERSTAT tools | 1.6 - Degree of completeness |
| Section 2: Assessing the INTERSTAT Framework | | | |
| Question 2.1 - For what purposes would you navigate the INTERSTAT framework? | Question 2.2 - About the content of the INTERSTAT framework, how do you rate the different sections of the website? | Question 2.3 - According to your experience and needs, what are the most relevant tools in the INTERSTAT framework? | Question 2.4 - Are there any additional open tools to be integrated in the framework? |

Table 5: Second section questions

Analysis of survey responses

According to the answers, the main reason for using and/or browsing the INTERSTAT framework, is to get documentation/information regarding LOSD (86%), confirming that the technical website covers the main aspects of LOSD production and dissemination.

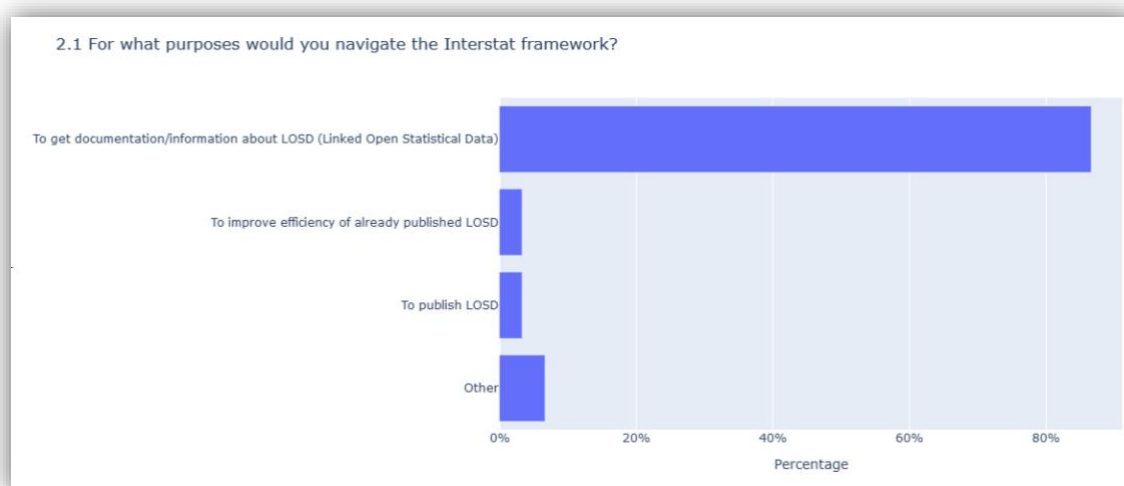


Figure 9: Respondents by navigation purposes

The intersection of the level of experience in LOSD and the purposes to navigate the INTERSTAT framework points out that also LOSD experts would consult the technical website to improve their knowledge and skills.

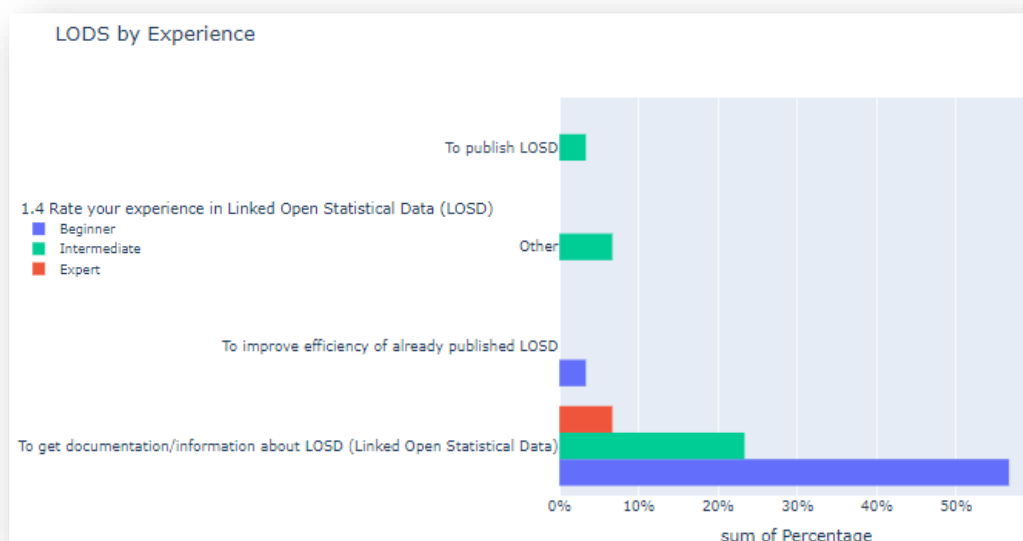


Figure 10: Respondents by LOSD experience and navigation purposes

Regarding the evaluation of communication efficacy, respondents are divided in those who find the INTERSTAT framework clear and easy to navigate (50%) and those who find it slightly confusing (46%). The minority of respondents consider it hard to navigate or have not answered (about 4%).

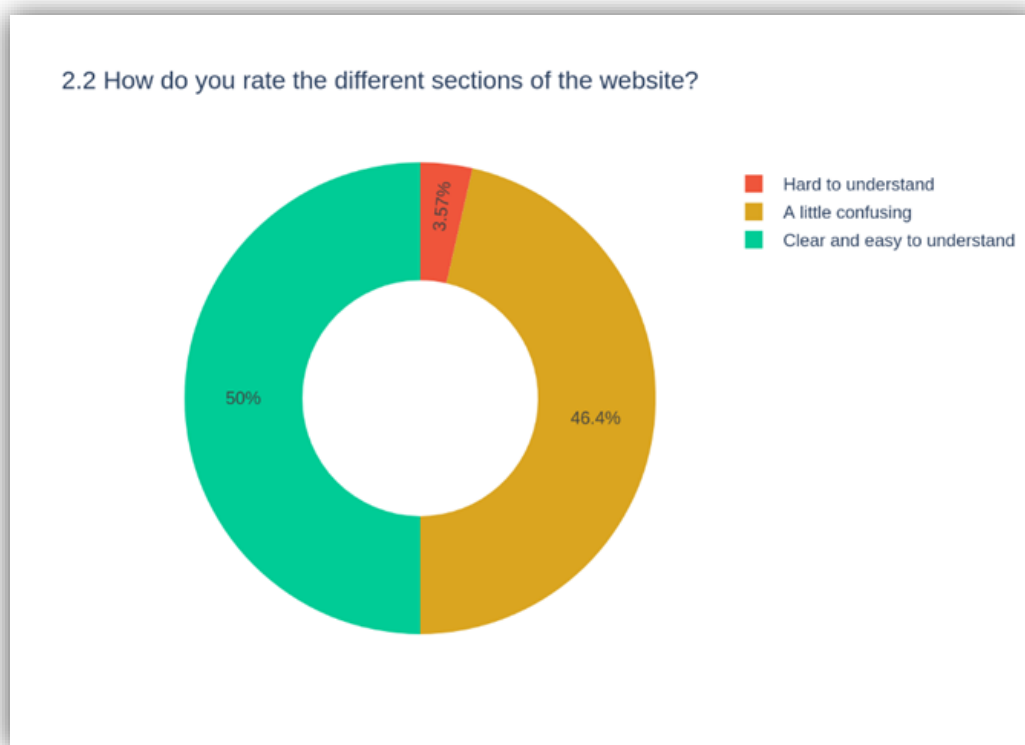


Figure 11: Respondents evaluation of the technical website

Crossing the clarity of the technical website with the level of experience highlights that most of those who consider the INTERSTAT framework a little confusing are Beginners (52% of Beginners, 30% of total respondents), while those who find it clear belong to an Intermediate level (66% of Intermediate respondents, 20% of total respondents). The analysis indicates that the framework is built mainly to be used by a person with a higher level of experience concerning open data.

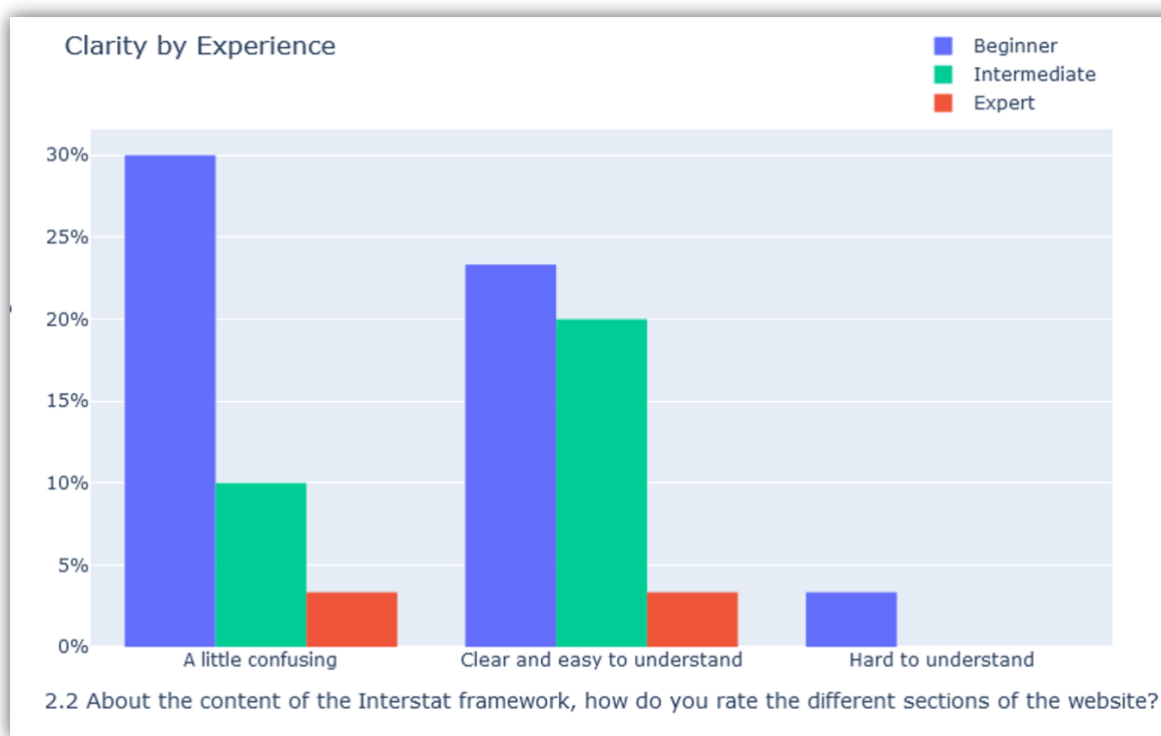


Figure 12: Respondents by LOSD experience and evaluation of the technical website

In relation to the relevance of the tools provided by the INTERSTAT framework, the ones for data analysis are considered the most effective (43%), followed by the ones for data publishing (26%). This result could be explained by the incidence of statistical experts highlighted in the previous section.

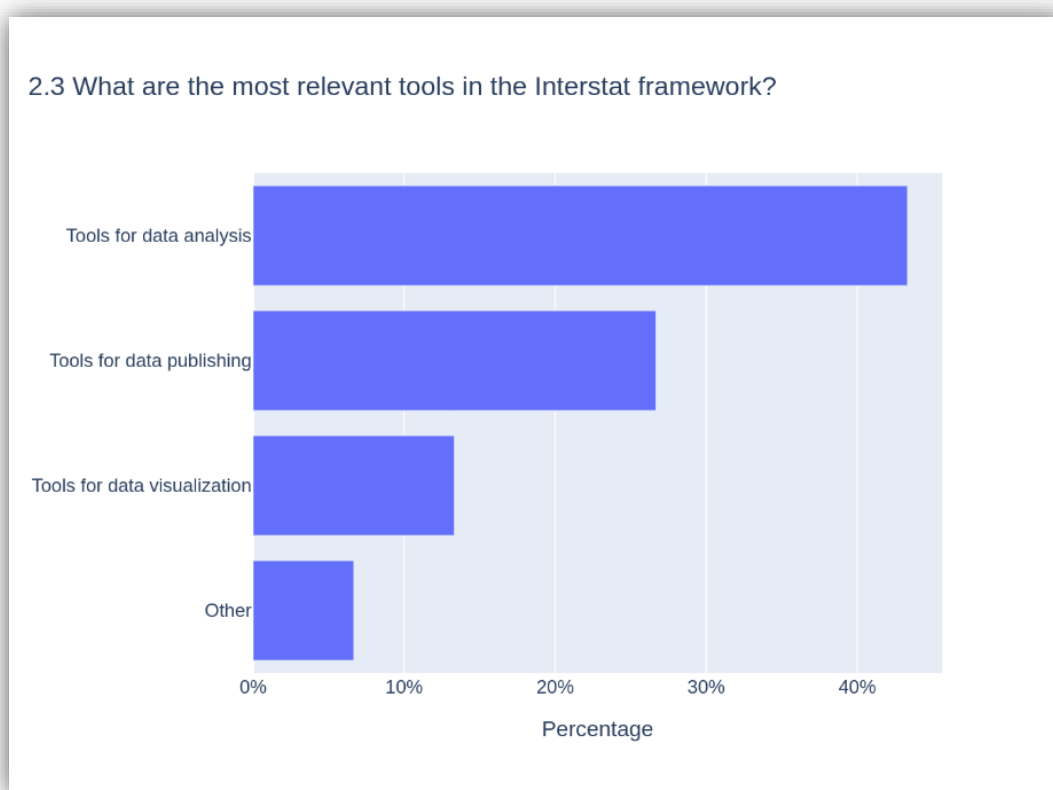


Figure 13: Respondents evaluation of the INTERSTAT tools

Considering each subset of tools with respect to the user experience in navigating the INTERSTAT framework, although their relevance, the ones for data analysis (in the framework grouped as Tools for Data or Metadata management) are the most difficult to explain to non-technical users. This result may depend on the core features of these tools, which are hard to describe without using technical concepts.

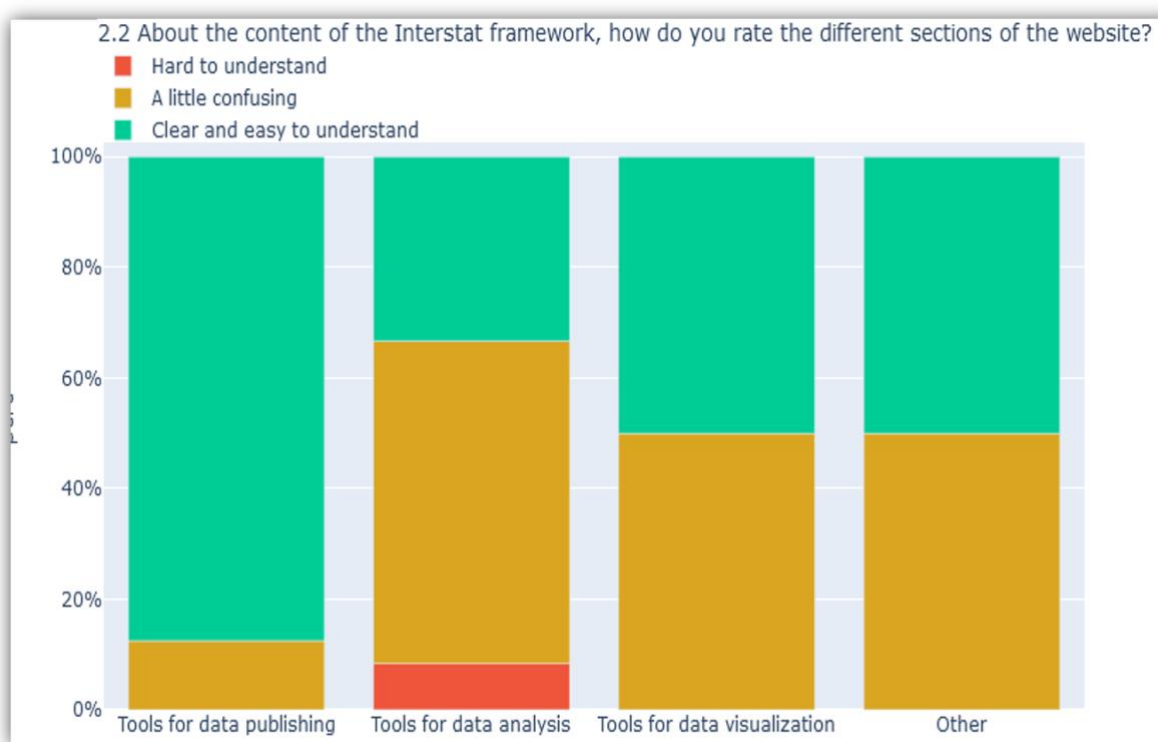


Figure 14: Respondents evaluation of the INTERSTAT tools and technical website

Regarding the framework completeness, for 25% of respondents there is no need to add other applications to the INTERSTAT framework, while most of them do not know (about 71%). This outcome confirms the relevance of the project, promoting and collecting tools, standards and workflows to increase knowledge and skills for LOSD production.

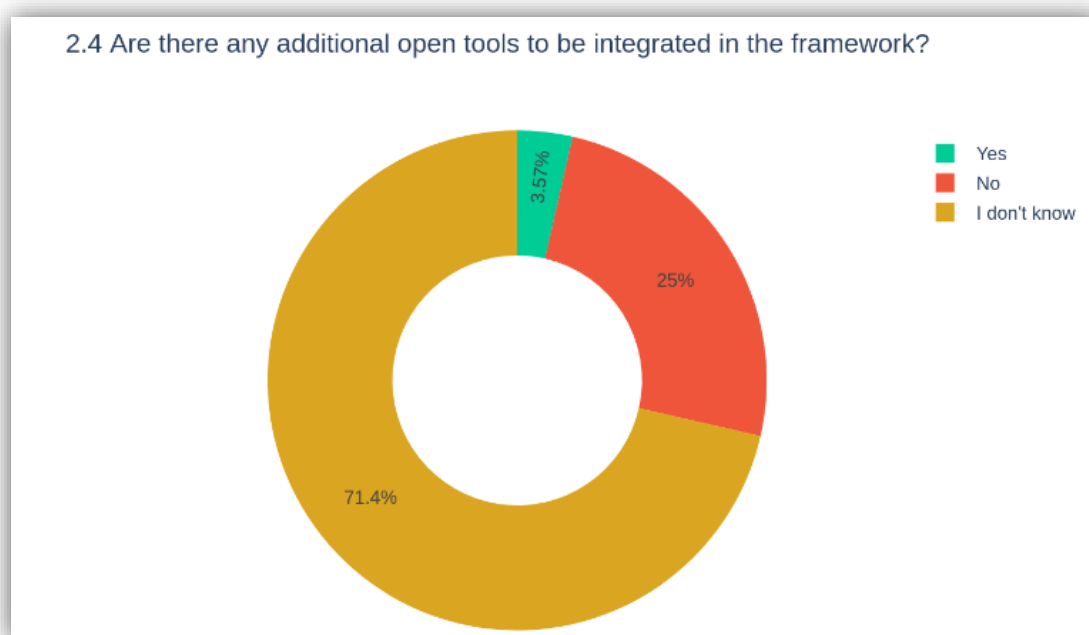


Figure 15: Respondents evaluation of open tools to integrate in the INTERSTAT framework

KPIs assessing the INTERSTAT Framework

The computation of KPIs is based on the score assigned to the answer choices, to rank and summarize the answers provided for a specific assessment dimension. The answer choices directly related to LOSD production and dissemination have the highest score. Then, for each KPI, the scores assigned to each answer were rescaled in order to obtain a value between 0 and 100, according to the main objectives of the project. Once all the qualitative items were translated into numerical values for each respondent, the average score was calculated for a single aspect applying the following formula:

$$KPI = \sum_{i=1}^P \frac{n_i * score_i}{N}$$

Where N is the total number of respondents

and then for all the dimensions to assess. The following table reports the score attributed to each question of Section 2.

| KPI | Question | Answer Choices (p) | Rating scores | 0-100 Scale |
|---|--|---|---------------|-------------|
| 1.3 - Main benefits of using the INTERSTAT framework | Question 2.1 - For what purposes would you navigate the INTERSTAT framework? | To get documentation/information about LOSD | 1 | 33.3 |
| | | To publish LOSD | 3 | 100.0 |
| | | To improve efficiency of already published LOSD | 2 | 66.7 |
| | | Other | 0 | 0.0 |
| | | | | |
| 1.4 - Clarity of key concepts (communication efficacy) | Question 2.2 - About the content of the INTERSTAT framework, how do you rate the different sections of the website? | Clear and easy to understand | 3 | 100.0 |
| | | A little confusing | 1,5 | 50.0 |
| | | Hard to understand | 0 | 0.0 |
| | | | | |
| 1.5 - Effectiveness of INTERSTAT tools | Question 2.3 - According to your experience and needs, what are the most relevant tools in the INTERSTAT framework? | Tools for data publishing | 3 | 100.0 |
| | | Tools for data visualization | 2 | 66.7 |
| | | Tools for data analysis | 1 | 33.3 |
| | | | | |
| 1.6 - Degree of completeness | Question 2.4 - Are there any additional open tools to be integrated in the framework? | Yes | 2 | 66.7 |
| | | No | 1 | 33.3 |
| | | I don't know | 1,5 | 50.0 |
| | | | | |

Table 6: Scores assigned to Section 2 questions

Adopting this approach, the values of KPIs concerning the INTERSTAT Framework are reported in the next table. Regarding the score calculated for each KPI, it is possible to see how the clarity had the major score (73.2), instead the completeness, the effectiveness and main benefits had the worst scores (46.4, 60, 34.4 respectively).

| What | | | |
|-------------------------------|--|---|------------------------------------|
| Assessment dimension | KPI | Rationale behind the KPI | KPI Value |
| 1. INTERSTAT Framework | <i>Self-assessment</i> | | |
| | 1.1 - Number of use cases covered by the tools integrated in the framework | Description of the different use cases covered by the framework to demonstrate that tools can be combined in different pipelines. Efficiency assessment of the framework in terms of open tools and degree of technical interoperability | 5 |
| | 1.2 - Percentage of Open Tools | Measuring the degree of openness and shareability of INTERSTAT tools | 13 out of 14 tools are open-source |
| | <i>Assessment Survey</i> | | |
| | 1.3 -Main benefits of using the INTERSTAT framework | Overview of the expectations and requirements of different type of users met by the INTERSTAT framework | 34.4 |
| | 1.4 - Clarity of key concepts (communication efficacy) | Efficacy of the INTERSTAT framework in promoting principles and tools to increase interoperability | 73.2 |
| | 1.5 - Effectiveness of INTERSTAT tools | Relevance of the tools provided by INTERSTAT framework | 60.0 |
| | 1.6 - Degree of completeness | Evaluate the framework efficacy in terms of exhaustiveness of the tools provided to execute a data pipeline | 46.4 |

Table 7: KPIs assessing the INTERSTAT framework

3.2.1 Use cases and added value of INTERSTAT tools

Based on the self-assessment of the tools integrated in the technical website, the following use cases complement the analysis of the relevance of the INTERSTAT framework, highlighting how tools can be combined for:

- Statistical data visualization and browsing
- Data pre-processing to answer users' requests
- Creating SPARQL queries templates for data browsing
- Mapping and Publishing Statistical Data
- Publishing data in the European Data Portal and obtaining published data in real-time.

The description of each use case shows how INTERSTAT tools can meet the needs of the three categories of stakeholders previously described (Technical User / IT staff, Statistical Expert or equivalent, Other Domain Expert - General User). In addition, optimizing the cooperation between different actors makes it easier to obtain new content starting from available statistical data, beyond a simple navigation of published data. In addition, this analysis underlines how the interoperability, between national and international web catalogues and statistical portals, allows to exploit the great potential of open statistical data.

Use case 1 - Statistical data visualisation and browsing

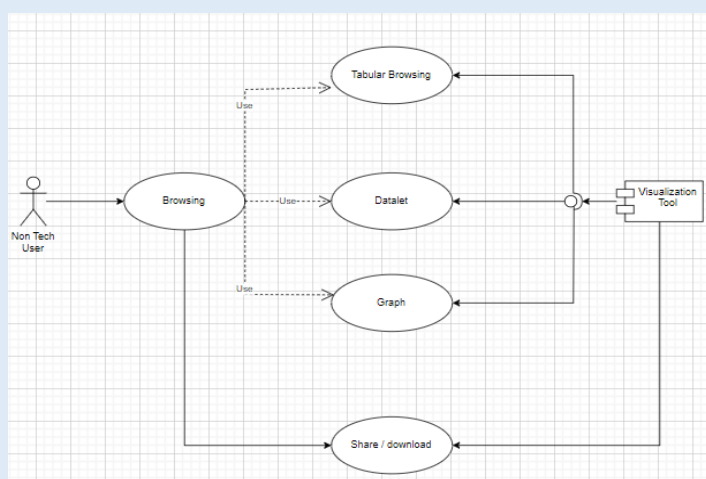
| Actors | Tasks | Tools and Technologies Involved |
|----------------------|---|--|
| General users | <p>Pre-conditions</p> <ul style="list-style-type: none"> i. Linked Open Statistical Data has been previously created, stored and published. ii. The user does not have any knowledge of RDF or Linked Data. <p>Description and Solution</p> <p>A non-technical end-user may want to browse through statistical data, make graphical analyses, even comparing data from different countries and/or domains. In order to make statistical data understanding easier, the framework proposes methods and tools which, cooperating together, can allow any non-technical end-user to be able to query such data in a simplified and facilitated way.</p> <ol style="list-style-type: none"> 1. A general user would like to explore relevant statistical data through browsing a dataset of a catalogue (e.g., CKAN) published in the Idra Portal. 2. If the selected dataset is in RDF format, the user will be able to navigate data in tabular form through Olap Browser, or in graphical format through Cube Visualiser, based on the type of view he wants to obtain. <ul style="list-style-type: none"> a. The data of interest will be visually displayed to the user in order to explore them. b. The user can share the data he discovered via the page URL, or download them. 3. If the dataset to browse is not in RDF format, the user will be able to visually analyse the data using the Deep tool integrated in Idra which will allow to create a graphical visualization (<i>Datalet</i>). <ul style="list-style-type: none"> a. The user can easily create a chart based on the federated open dataset. b. The user can save, download or share the newly created resource. | <p>Idra Portal</p> <p>Deep Tool (Idra)</p> <p>Olap Browser</p> <p>Cube Visualiser</p> <p>Data Browser</p> |

4. Another type of dataset visualization is provided to the user through the **Data Browser** tool: it allows data-users to browse and visualize Census datasets.

It is the front end that has to be used for browsing all the disseminated data (including open data and LOSD). The main functions for a user allow to switch between the available dashboards, switch between different distributed databases (web services), browse one or more tree-themes and select the dataset of interest and create graphs and thematic maps.

Post-conditions

- i. A Linked Open statistical data has been visualized, shared via a browsable URL or downloaded in the chosen data format.
- ii. The user is able to interact and explore similar data across different countries with the ability to discover and compare the data according to several combinations of variables (i.e., by region, age, gender etc.).



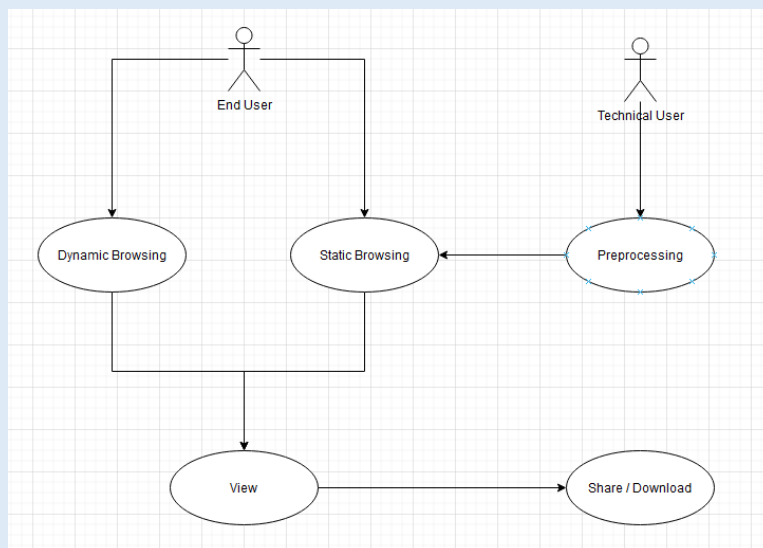
Use case 2 - Data pre-processing to answer users' requests

| Actors | Tasks | Tools and Technologies Involved |
|--|--|--|
| General users Statistical domain experts/Data Analyst | <p>Pre-conditions</p> <ul style="list-style-type: none"> i. The SPARQL queries underlying the user requests, must have already been properly implemented by technical end-users. ii. The user does not have any knowledge of SPARQL language or Linked Data. iii. In order to use SparQLing, the user must be familiar with ontologies and with the specific visual language Graphol. <p>Description and Solution</p> <p>A general user can extrapolate data of interest regardless of technical knowledge in two ways:</p> <p>1. By accessing the main page of the GraphDB SPARQL endpoint exposed by the central Triple store, in which each request is associated with a specific button of the interface and executing the request of interest. For example, the requests underlying these interface buttons could be predefined SPARQL or SDMX queries, to allow the user to access the underlying statistical data without any further instructions.</p> <p>The pre-processed requests can relate to any generic data query that could be useful for a user using the framework from simpler requests concerning for example geographic data (e.g., particular regions, departments, districts and municipalities or demographic data) to queries that exploit cross-border interoperability (e.g. relating French and Italian data).</p> <p>A more experienced user, who is at least able to interpret a selected ontology, can use the SparQLing tool to write queries in a simpler and faster way. Some queries made previously by technical users could be saved in the SPARQL endpoint of the framework, and made available to less experienced user.</p> <p>2. The pilots developed using the INTERSTAT tools are an example of specific dynamic SPARQL queries prepared by technical users,</p> | <p>GraphDB</p> <p>SPARQL endpoint</p> <p>SparQLing</p> <p>INTERSTAT pilot applications</p> |

querying the RDF underlying statistical data and showing the results. The requests are found in different sections of the web application and the user has the possibility, in this case, to choose the cities of interest to navigate Italian and French data, view geographic maps and so on. Further, the user can choose to download the data obtained in a specific data format.

Post-conditions

Thanks to the common queries implemented to run across similar data for different countries, the user can analyse data of interest in the same domain. Data are selected and accessed without any specific knowledge in the field of statistics or Linked Data.

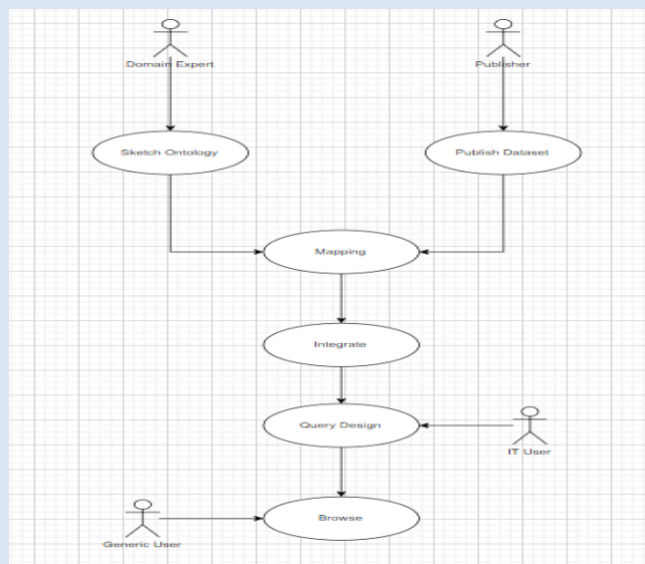


Use case 3 - Creating SPARQL queries templates for data browsing

| Actors | Tasks | Tools and Technologies Involved |
|----------------------------|---|---|
| Technical experts/IT staff | <p>Pre-conditions</p> <ul style="list-style-type: none"> I. Advanced technical knowledge. II. Interaction between subject matter experts creating domain ontologies and IT staff responsible for data publishing and data visualization. <p>Description and Solution</p> <p>In this use case, a group of technical users wants to enable browsing by creating pre-defined queries to aid less experienced users (citizens or non-technical users), starting from ontologies created by domain experts.</p> <p>1. A subject matter expert can model a domain ontology using the functionalities provided by Eddy tool, a desktop application implementing the OBDA approach (Ontology-Based Data Access). The reference ontology, saved in <i>Graphol</i> format, can be exported into OWL format. The mapping process (matching data source to the domain ontology) takes in input Owl ontology and data and produces in output the materialized triples (if the OBDA architecture is composed by triple stores) or the on-the-fly triples in the case of a relational database underlying.</p> <p>In this step, the user can perform a syntactic check of the entire ontology: the tool allows to edit or identify errors in a <i>Graphol</i> ontology selected by the user.</p> <p>2. A Technical User can use the domain ontology as input in SparQLing, to create queries in a faster and simpler way, by browsing and selecting the appropriate resources and properties.</p> <p>At this stage, the user has the possibility to decide to save the query created, so that it can be used by other portal users through the GraphDB SPARQL end-point. Finally, the user can choose the format to save the response, such as JSON or XML.</p> | <p>Eddy (Desktop App)</p> <p>SparQLing</p> <p>GraphDB SPARQL end-point</p> |

Post-conditions

- i. A new SPARQL query, based on a domain ontology has been created and can be shared via **GraphDB SPARQL end-point**, to help users without technical knowledge to browse data (via **use case 2**).
- ii. Linked Statistical data, obtained in two different formats, can be shared and used for future analyses.
- iii. The Statistical Data Analyst has the possibility to implement cross-border and cross-domain queries to extract further knowledge from available statistical data.



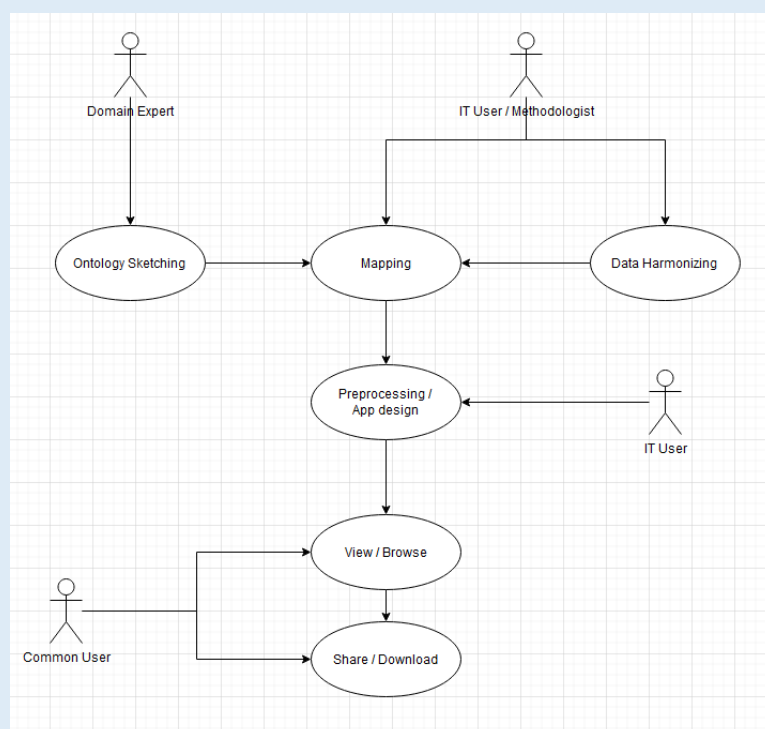
Use case 4 - Mapping and Publishing Statistical Data

| Actors | Tasks | Tools and Technologies Involved |
|-----------------------------------|--|--|
| Technical experts/IT staff | <p>Pre-conditions</p> <p>Availability of statistical data to be converted into LOD format for dissemination or uploading in the Idra Portal.</p> <p>Description and Solution</p> <p>A National Statistical Institute (NSI) aims at publishing statistical data in a Linked Open Statistical Data format in the local platform and can do it as follows:</p> <ol style="list-style-type: none"> 1. Using the Adminer tool, the data must be loaded into a staging MySQL database, which operates as data source for the Monolith tool. 2. In relation to the data layer, to harmonize sources having different data structure, the data must be converted into a conventional structure, the Common Data Model. 3. Common Data Models are implemented through mapping files within Monolith tool, specifying the correspondence between the domain concepts and the actual data. The use of the same mapping for similarly structured data will allow to link data belonging to the same domain in different countries, and therefore to create common interoperable queries (through use case 3) and then compare data from different countries. 4. Once data are mapped, it is possible to create queries through Monolith tool, using SPARQL language. It is also possible to export the query result in the desired format, including RDF itself. 6. Once the RDF file has been obtained, if the publisher decides to save the created Linked Open Statistical Data in a catalogue, the user can publish it in Idra. Once published in Idra, data will be harvested by the European data portal, becoming discoverable and usable by third parties. | <p>Adminer tool</p> <p>Monolith tool</p> <p>Idra</p> <p>Context Broker</p> |

8. The data published in the **Idra Portal** can be automatically loaded into the **Context broker**, through the DCAT-AP/NGSI-LD data model which exports the DCAT-AP metadata from Idra, publishing them as Entities in the Context Broker. The Context Broker can thus notify changes on the datasets to the external systems and enables to query published data through APIs.

Post-conditions

- i. Reusing available **mappings** allows to expand the system's data management capabilities incrementally. It is also possible to integrate new data sources, as well as different datasets, if compliant with the stored ontology concepts.
- ii. The newly converted Linked Open Statistical Data is now **stored and published on the Idra Portal and Context Broker**, so the data can also be reached by third-party's systems (e.g: External Systems and National Open Data portals).



Use case 5 - Publishing data in the European Data Portal and obtaining published data in real-time

| Actors | Tasks | Tools and Technologies Involved |
|--|--|--|
| <p>Technical experts/IT staff</p> <p>Statistical domain experts/Data Analyst</p> | <p>Pre-conditions</p> <p>The data must have been published in Idra Portal using use case 4.</p> <p>Description and Solution</p> <p>The following use case enables interoperability among different national statistical portals and the European Data Portal through the adoption of technical standards (DCAT-AP and NGSI-LD) and the interactions between tools. Technical interoperability allows to reuse harmonized statistical data in combination with open datasets (e.g., city-related data) from European Data Portal and other national open data portals.</p> <p>1. Publishing data in Idra (via the use case 4) allows to automatically publish the datasets also in the Context Broker, reachable from external systems thanks to the integration of the DCAT-AP data model in Idra and the CEF Context Broker.</p> <p>Using DCAT-AP metadata and de-facto standard APIs, Idra realizes the connection to the European Data Portal.</p> <p>2. The Context Broker, therefore, using NGSI specifications, enables external applications to provide real-time updates and access to information that contextualises data being displayed.</p> <p>The Context Broker is conceived for the management and querying of the context information in a structured manner, based on linked data standards following the ETSI NGSI-LD specification.</p> <p>Precisely, it allows external systems to create, modify and delete Entities and their attributes, and enables batch operations to create/update a set of Entities in a single request, to query/retrieve Entities, with a rich set of filters and a powerful query language.</p> <p>External applications, through the creation of Subscriptions, can get notifications on changes in Entities, instead of actively polling the broker, and through Registrations, can extend the broker with entities living inside external context sources or brokers. The core platform also allows to test the use of real-time data.</p> | <p>Idra</p> <p>Context Broker</p> <p>SDMX/JSON-LD Gateway</p> |

| | | |
|--|---|--|
| | <p>3. The publication of open data in Idra and in the Context Broker allows to connect the data to the European Data Portal, thus linking harmonized statistical data with other open datasets from this portal and other national open data portals.</p> <p>4. Interoperability is guaranteed also thanks to the SDMX/JSON-LD gateway: The SDMX/JSON-LD gateway translates the SDMX Data Model (RDF Turtle format) into JSON-LD and upload it into the Context Broker using ETSI NGSI-LD API allowing external applications accessing the statistical data from Context Broker.</p> <p>5. Data publishers have the possibility to analyse, share, manage and use data simultaneously.</p> <p>Post-conditions</p> <p>The interoperability between different standards and data models such as NGSI-LD, JSON-LD, DCAT-AP, and SDMX was performed.</p> | |
|--|---|--|

3.3 Pilot applications and users experience

The assessment of the pilot applications developed using the tools of the INTERSTAT framework has the main goal to prove the added value of reusing open statistical data through cross-border and cross-domain analysis. This dimension focuses on the front-end implemented to enable a general user to navigate the open statistical sources integrated in each pilot. The evaluation of this dimension was totally based on the external feedback, through the questions and KPIs reported in the next table.

| KPIs | | |
|--|--|---|
| 3.1 - Relevance of cross-border analysis | 3.2 - Relevance of cross-domain analysis | 3.3 - Users experiences assessment |
| Section 3: Assessing the Client Applications | | |
| Question 3.1 - Considering the topics selected for the pilots, the relevance of cross-border analysis is: Low, Medium, High | Question 3.2 - Considering the topics selected for the pilots, the relevance of cross-domain analysis is: Low, Medium, High | Question 3.3 - How do you rate the pilot interface ? |

Table 8: Third section questions

Analysis of survey responses

The following figure reports the distributions of respondents with respect to the relevance of cross-border and cross-domain analysis provided by the application services developed. The chart underlines that, according to the respondents, the cross-border analysis seems to be more relevant compared to the cross-domain data linkage.

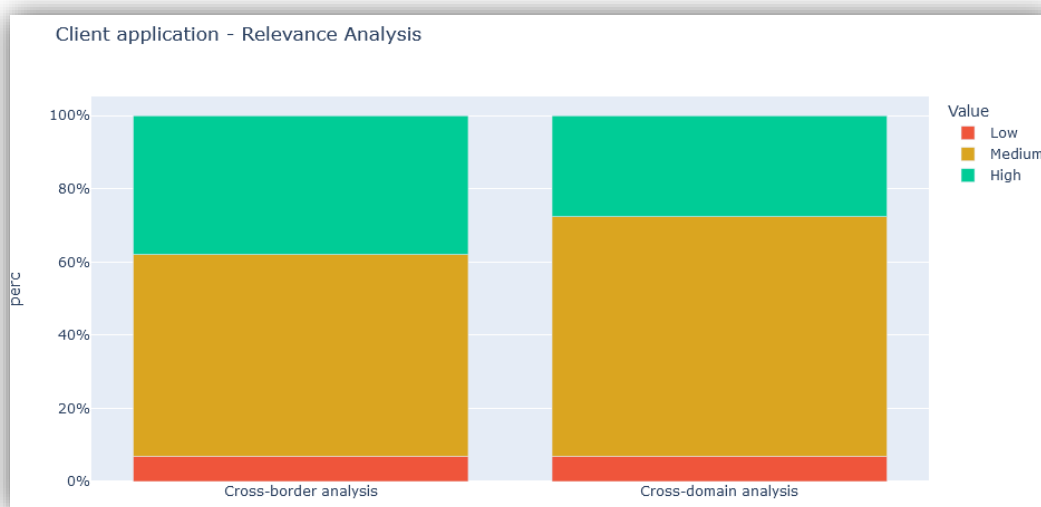


Figure 16: Respondents evaluation of cross-border and cross-domain analysis

Considering the client applications, the assessment of the pilot interfaces is similar to the navigation and user experience related to the INTERSTAT framework. In this case, end-users are divided in those who answered clear and easy to navigate (42%) and those who found the navigation a little bit confusing (53%).

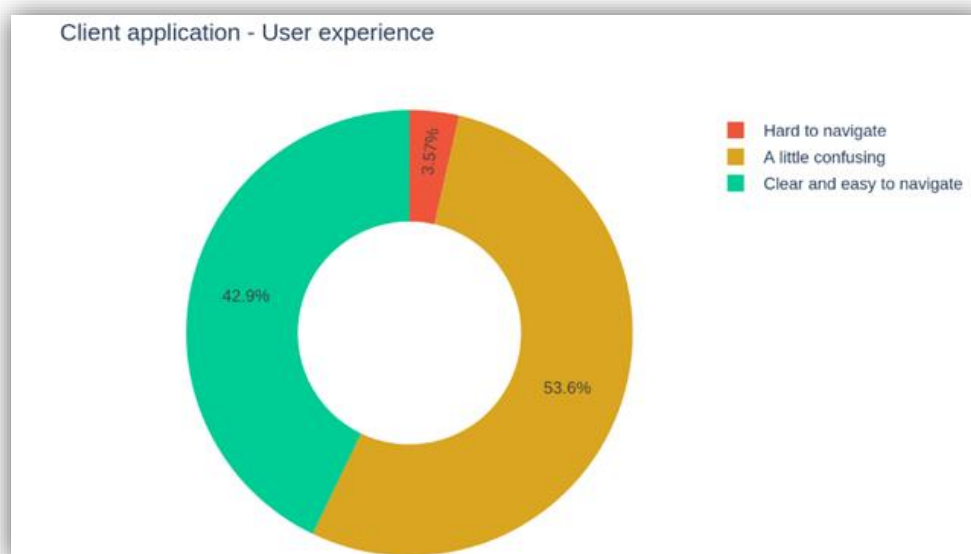


Figure 17: Respondents assessment of pilots' navigation

KPIs assessing the Client applications

In order to compute the KPIs, based on the collected answers, the score assigned to each answer choice is reported in the table below.

| KPI | Question | Answer Choices (p) | Rating scores | 0-100 Scale |
|---|--|----------------------------|---------------|-------------|
| 3.1 - Relevance of cross-border analysis | Question 3.1 - Considering the topics selected for the pilots, the relevance of cross-border analysis is: Low, Medium, High | Low | 1 | 33.3 |
| | | Medium | 2 | 66.7 |
| | | High | 3 | 100.0 |
| | | | | |
| 3.2 - Relevance of cross-domain analysis | Question 3.2 - Considering the topics selected for the pilots, the relevance of cross-domain analysis is: Low, Medium, High | Low | 1 | 33.3 |
| | | Medium | 2 | 66.7 |
| | | High | 3 | 100.0 |
| | | | | |
| 3.3 - Users experiences assessment | Question 3.3 - How do you rate the pilot interface ? | Clear and easy to navigate | 3 | 100.0 |
| | | A little confusing | 2 | 66.7 |
| | | Hard to navigate | 1 | 33.3 |

Table 9: Scores assigned to Section 3 questions

The following table shows the values of the KPIs assessing the client applications. The three indicators confirm the relevance of cross-border and cross-domain analysis (respectively, 77.0 and 73.6), as well as a good user experience navigating the pilots (79.8).

| What | | | |
|---------------------------------|--|---|------------------|
| Assessment dimension | KPI | Rationale behind the KPI | KPI Value |
| <i>Assessment survey</i> | | | |
| 3. Client applications | 3.1 - Relevance of cross-border analysis | Efficacy of the pilots to demonstrate the feasibility of cross-border services | 77.0 |
| | 3.2 - Relevance of cross-domain analysis | Efficacy of the pilots to demonstrate the feasibility of cross-domain services | 73.6 |
| | 3.3 - Users experiences assessment | Assess whether and to each extent the pilots interface is user-friendly. Measuring users satisfaction during pilot navigation | 79.8 |

Table 10: KPIs assessing the Pilot services

3.4 Summary of survey results

Starting from the KPIs computed for each assessment dimension, the following figure provides an overview of the main performance indicators and their relationships.

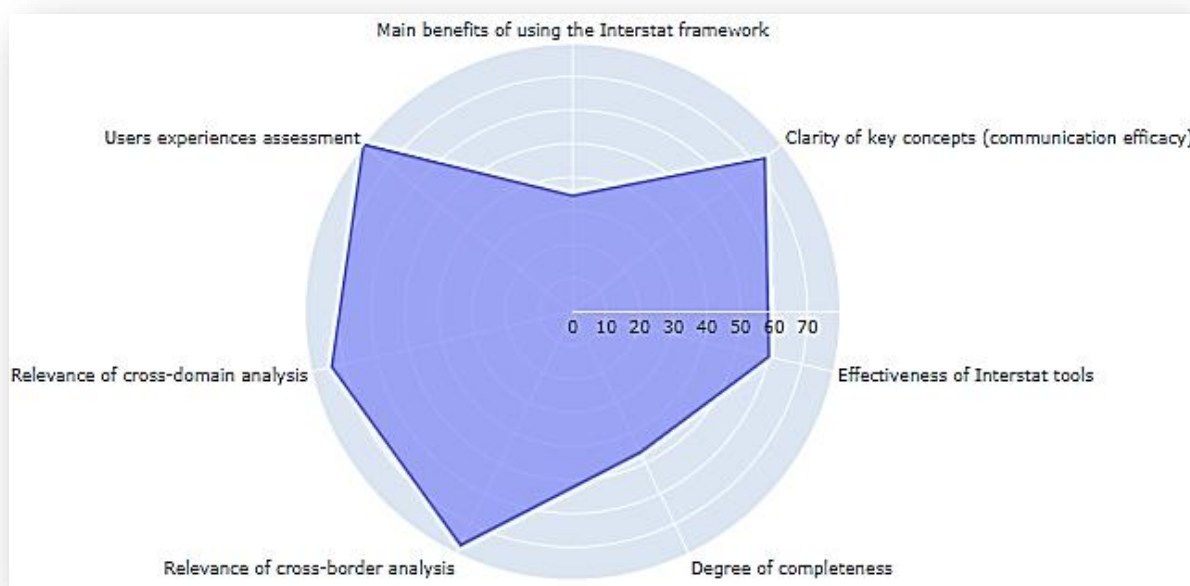


Figure 18: KPIs overview assessing the project impact

As highlighted above, the chart shows that the dimensions related to User experience, Relevance of cross-border analysis and Clarity of Key concepts have the highest rating compared to the other dimensions. This result underlines that the INTERSTAT framework has succeeded in implementing a technical environment dealing with the core aspects related to LOSD production, from open data conversion in rdf triples, to data harvesting by open data portals. Assessing the effectiveness of the INTERSTAT tools was hard, due to the low number of LOSD experts able to evaluate the technical aspects, the issues faced and the solutions adopted during the development of the framework and the pilot services.

In order to achieve an overall assessment of the project impact provided by external stakeholders, the scores assigned to each item and dimension were summarized to obtain a total indicator. A Synthetic Index was computed by processing the average scores of each respondent and

standardizing the output to the 0 -100 scale, divided in three main categories described in the following table. In this case, the applied formula is the following:

$$Synthetic\ Index = \sum_{i=1}^{DIM} KPI_i / DIM$$

Where DIM is the total number of assessment sub-dimensions (a KPI is associated to each sub-dimension)

| Synthetic Index Value | Synthetic Index Decoding |
|-----------------------|---|
| 0 to 40 | Essentially open to experts/skilled users: the project outcome is clear mainly to open data experts, who are used to deal with interoperability issues through the tools and the technical standards integrated in the INTERSTAT framework |
| 40 to 70 | Fit for purpose: the project can impact several types of stakeholders and domains, even those who are not LOSD experts, covering several aspects of LOSD and addressing the most relevant issues preventing open data interoperability |
| 70 to 100 | Open by design: the project meets the needs of all types of stakeholders, and fosters capacity building to prevent and solve technical and semantic interoperability issues and barriers, regardless of prior skills and knowledge |

Table 11 - Profiles for Synthetic Index decoding

Applying this method, the overall score is 62.8 (Figure 19), placing the INTERSTAT project in the 'Fit for purpose' category.

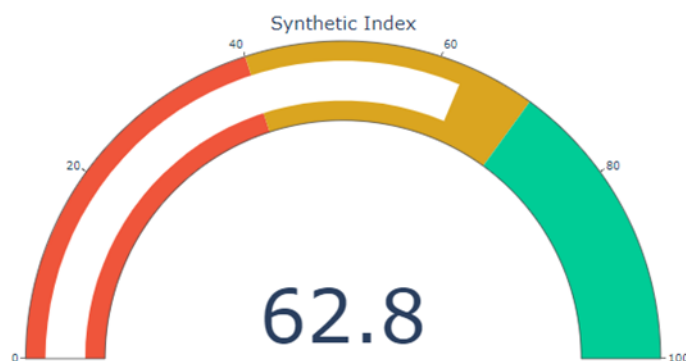


Figure 19: Overall rating of the INTERSTAT project

This synthetic score highlights that the project has delivered solutions with high potential that could be improved to increase usability for a real use by non-technical experts. Despite the prevalence of highly technical tasks, respondents have acknowledged the efforts made to clearly document and describe the project deliverables.

3.5 MQA compliance

One of the project milestones, considered as a remarkable dimension for the action assessment, is the compliance of the pilots metadata with the Metadata Quality Assurance [6]. The dissemination of the outcome through the European Data Portal (EDP) [7] was one of the core requirements for the project, launched within the Connecting Europe Facility (CEF) programme. The EDP enables the reuse of public sector data by providing a single access point to explore and download public datasets, grouped in a catalogue. A common format for metadata descriptions, the DCAT Application Profile (DCAT-AP) [8], guarantees cross-border interoperability. A human-readable website or a machine-readable API can access the EDP, which supports metadata search and management. The Metadata Quality Assessment (MQA) is a tool offering several functionalities to evaluate the quality of metadata from data publishers and data portals, based on W3C Data Quality Vocabulary (DQV) [9]. The dimensions validated by the MQA correspond to the FAIR principle², namely:

- **Findability**, measuring to which extent data can be reused and accessed by humans and machines - data and metadata should be easy to find
- **Accessibility**, to check the access to the data distributions
- **Interoperability**, to assess whether the format and media type of distributions are machine readable for data analysis, storage and processing
- **Reusability**, to check the adequacy of data and metadata descriptions to enable data replicability and integration.

The MQA provides a score for each dimension derived from the FAIR principle and an overall score for the whole catalogue, based on the evolution of each assessment dimension. Each harvest is complemented by the metadata check performed by the MQA.

In order to reach this milestone, the INTERSTAT catalogue was created, grouping the main datasets and the related metadata produced during the implementation of the pilots. This catalogue was published in Idra.

Idra is a web application 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 uniform representation of collected open datasets,

² <https://www.go-fair.org/fair-principles/>

thanks to the adoption of international standards (DCAT-AP [8]) and provides a set of RESTful APIs to be used by third party applications. Idra is an open-source software developed by Engineering Ingegneria Informatica SpA [10].

Figure 20 represents the homepage of Idra platform, which shows the main tags relating to the INTERSTAT catalogue federated and published on the platform [11].

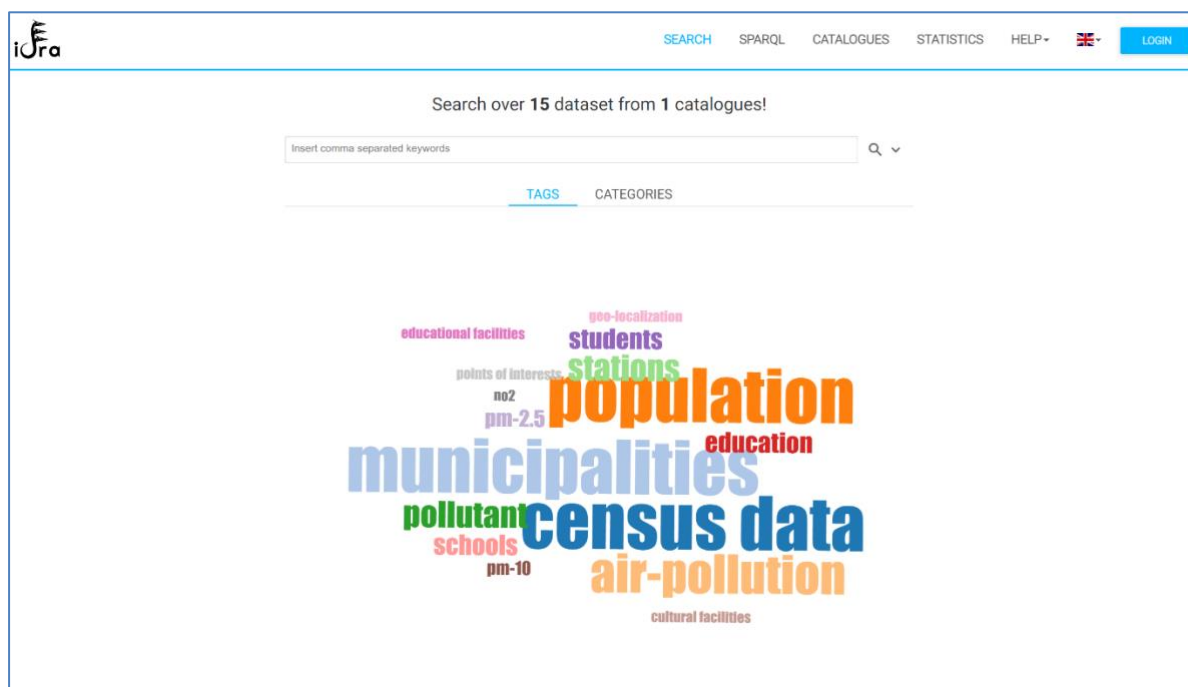


Figure 20: INTERSTAT open data catalogue published on Idra platform

Starting from the homepage shown in Figure 20, it is possible to select a tag from the tag-cloud to filter the search using the selected tag or search dataset by Categories. By clicking on the search icon, it is possible to perform the search on all of the federated datasets. The result of the search is a list of the dataset that match with the requested filter.

Figure 21 illustrates the result of a search operation. In this case, it shows the INTERSTAT catalogue content, reporting the list of related Datasets federated in it. In this page it is possible to navigate results, to change the order and the number of the results per page; moreover, It is possible to filter data using a facet approach. Different facets are available, in particular: Tags, File Formats, File Licenses, Catalogues or Categories. Each Dataset may contain one or more distributions, depending on the format adopted for data representation.

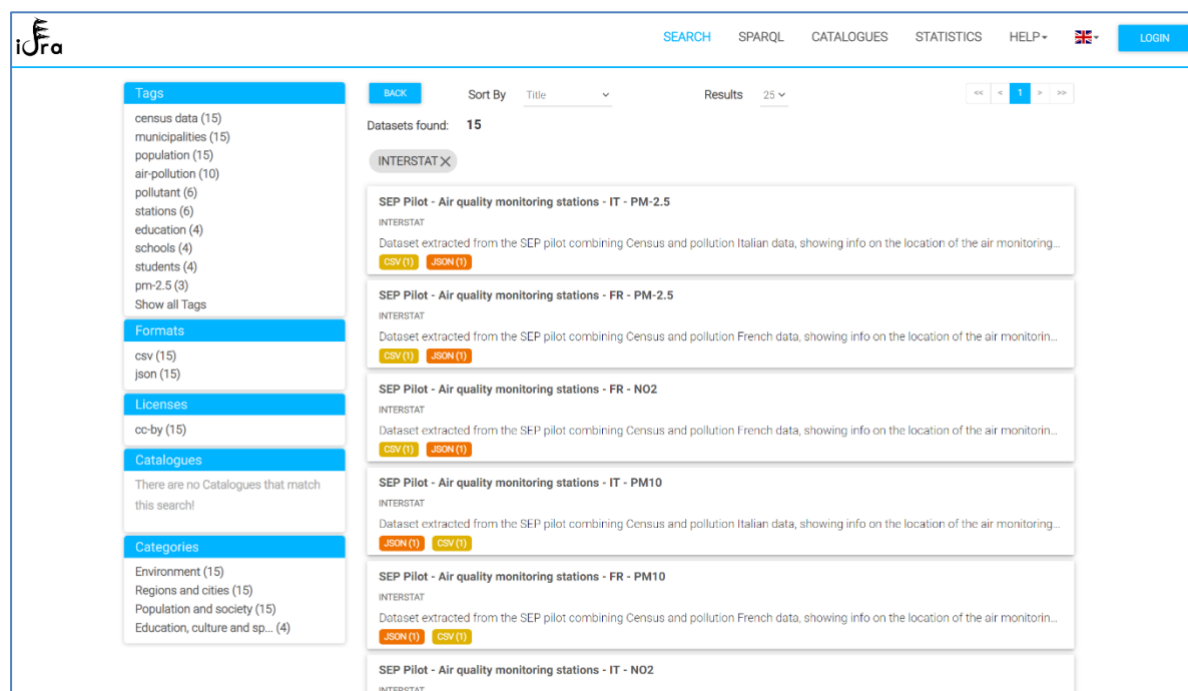


Figure 21: List of datasets of the INTERSTAT catalogue on Idra platform

Idra therefore, as shown schematically in Figure 22, allows a unique and standard access to the open data to foster and simplify the access to data held by national governments. This platform provides compliance to DCAT-AP standard and European Data Portal specifications and enables the possibility to be harvested by European or national open data catalogues. In the particular case of the INTERSTAT catalogue, it was created in Idra using the *SPARQL* [12] connector and extracting the Linked Open Data produced during the project, from the official repository and SPARQL endpoint adopted to collect them, GraphDB [13].

Idra, then, allowed the INTERSTAT catalogue to be collected by the European Data Portal, under which the MQA can be executed on a regular schedule.

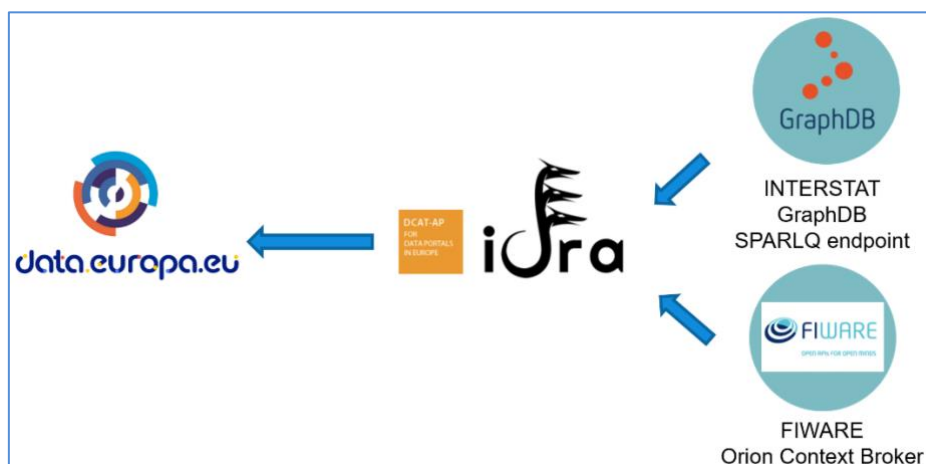
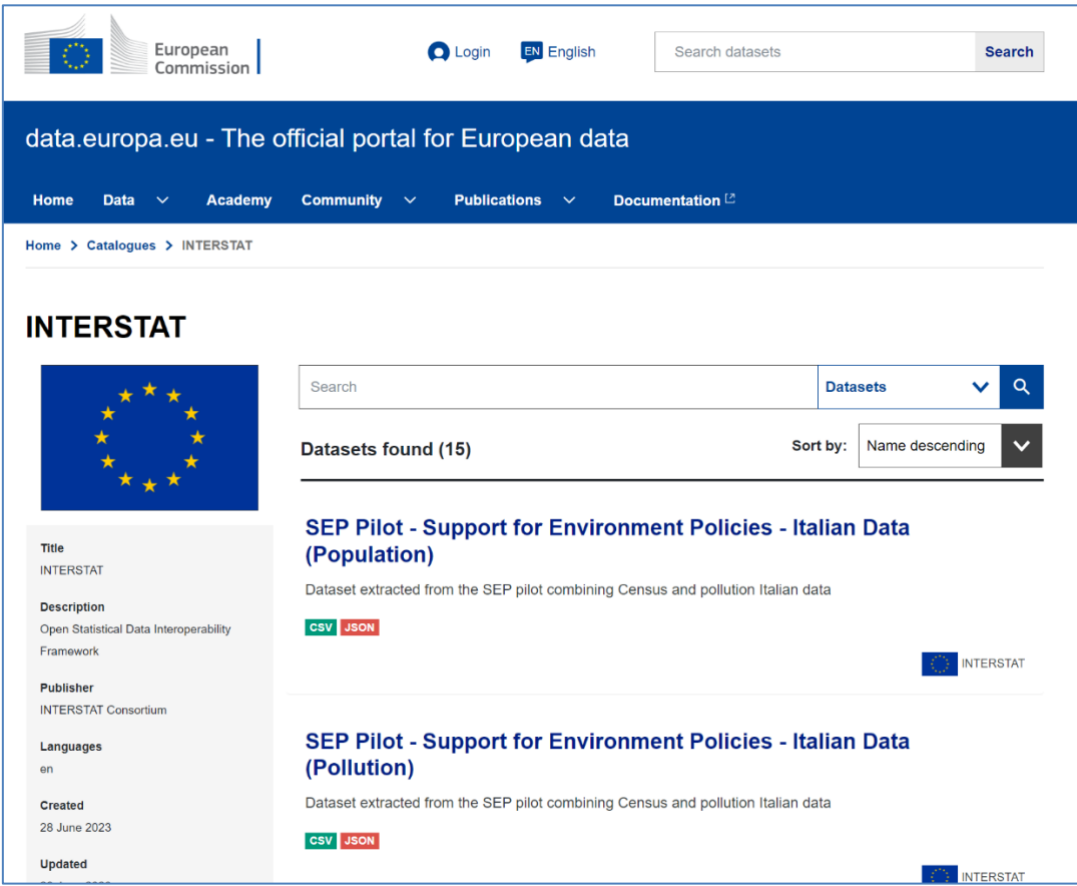


Figure 22: Schematic image of harvesting catalogue using Idra

The publication of the INTERSTAT catalogue in EDP was therefore achieved [14], as shown in the next figure



The screenshot shows the official portal for European data. The main content area displays the INTERSTAT catalogue with a search bar and a list of datasets. The first dataset is 'SEP Pilot - Support for Environment Policies - Italian Data (Population)', described as 'Dataset extracted from the SEP pilot combining Census and pollution Italian data'. It is available in CSV and JSON formats. The second dataset is 'SEP Pilot - Support for Environment Policies - Italian Data (Pollution)', also described as 'Dataset extracted from the SEP pilot combining Census and pollution Italian data', available in CSV and JSON formats. The left sidebar provides metadata for the INTERSTAT dataset, including its title, description, publisher (INTERSTAT Consortium), language (en), creation date (28 June 2023), and update date.

Figure 23: INTERSTAT catalogue published in EDP

Table 12 reports the list of the pilot Datasets stored in Idra platform and harvested by the EDP.

| Pilot applications Datasets | | |
|---|---|--|
| Name | Description | Distribution |
| SEP Pilot - Air quality monitoring stations - IT - NO2 | Dataset extracted from the SEP pilot combining Census and pollution Italian data, showing info on the location of the air monitoring stations and the detected values regarding the NO2 pollutant | Monitoring stations in Italy |
| SEP Pilot - Air quality monitoring stations - IT - PM10 | Dataset extracted from the SEP pilot combining Census and pollution Italian data, showing info on the location of the air monitoring stations and the detected values regarding the PM-10 pollutant | Monitoring stations in Italy |
| SEP Pilot - Air quality monitoring stations - FR - PM-2.5 | Dataset extracted from the SEP pilot combining Census and pollution French data, showing info on the location of the air monitoring stations and the detected values regarding the PM-2.5 pollutant | Monitoring stations in French |
| SEP Pilot - Air quality monitoring stations - FR - PM10 | Dataset extracted from the SEP pilot combining Census and pollution French data, showing info on the location of the air monitoring stations and the detected values regarding the PM-10 pollutant | Monitoring stations in French |
| SEP Pilot - Support for Environment Policies - French Data (Population) | Dataset extracted from the SEP pilot combining Census and pollution French data | Pollution levels in France Municipalities with highest population |
| SEP Pilot - Support for Environment Policies - Italian Data (Population) | Dataset extracted from the SEP pilot combining Census and pollution Italian data | Pollution levels in Italian Municipalities with highest population |
| SEP Pilot - Support for Environment Policies - French Data (Pollution) | Dataset extracted from the SEP pilot combining Census and pollution French data | Resident Population in France Municipalities with highest pollution |
| SEP Pilot - Support for Environment Policies - Italian Data (Pollution) | Dataset extracted from the SEP pilot combining Census and pollution Italian data | Resident Population in Italian Municipalities with highest pollution |

| Pilot applications Datasets | | |
|--|--|--|
| Name | Description | Distribution |
| SEP Pilot - Air quality monitoring stations - FR - NO2 | Dataset extracted from the SEP pilot combining Census and pollution French data, showing info on the location of the air monitoring stations and the detected values regarding the NO2 pollutant | Monitoring stations in French areas |
| SEP Pilot - Air quality monitoring stations - IT - PM-2.5 | Dataset extracted from the SEP pilot combining Census and pollution Italian data, showing info on the location of the air monitoring stations and the detected values regarding the PM-2.5 pollutant | Monitoring stations in Italy |
| GF Pilot - Geolocalized Facilities | Dataset extracted from the GF pilot combining Census data and information about cultural facilities | Cultural facilities in Rome Municipality |
| S4Y Pilot - The Schools for You - FR Schools | Dataset extracted from the S4Y pilot combining Census data and information about scholar services in France | Number of Schools and resident population in Paris Municipality |
| S4Y Pilot - The Schools for You - IT Schools | Dataset extracted from the S4Y pilot combining Census data and information about scholar services | Number of Schools and resident population in Rome Municipality |
| S4Y Pilot - The Schools for You - FR Students | Dataset extracted from the S4Y pilot combining Census data and information about scholar services | Number of Students and resident population in Paris Municipality |
| S4Y Pilot - The Schools for You - IT Students | Dataset extracted from the S4Y pilot combining Census data and information about scholar services | Number of Students and resident population in Rome Municipality |

Table 12: Interstat open datasets harvested by the EDP

Concerning metadata validation, the DCAT-AP specification is based on terms from several common data vocabularies and standards, identifying mandatory, recommended and optional classes and properties. The initial subset of metadata and the related properties provided for each Dataset (mandatory class) and distribution (optional class) of the INTERSTAT catalogue is reported in Table 13.

| Property | URI | Range | Description | Card |
|--|--|--|--|------|
| Mandatory properties for Dataset | | | | |
| description | dct:description | rdfs:Literal | Free-text describing the Dataset | 1..n |
| title | dct:title | rdfs:Literal | Dataset name | 1..n |
| Recommended properties for Dataset | | | | |
| contact point | dcat:contactPoint | vcard:Kind | Contact information to provide comments about the Dataset | 0..n |
| dataset distribution | dcat:distribution | dcat:Distribution | Links between the Dataset and an available Distribution. | 0..n |
| keyword/tag | dcat:keyword | rdfs:Literal | Keyword or tag describing the Dataset | 0..n |
| publisher | dct:publisher | foaf:Agent | Entity (organisation) responsible for making the Dataset available | 0..1 |
| theme/category | dcat:theme, subproperty of dct:subject | skos:Concept | Category of the Dataset | 0..n |
| Optional properties for Dataset | | | | |
| landing page | dcat:landingPage | foaf:Document | landing page of the data provider | 0..n |
| spatial/geographical coverage | dct:spatial | dct:Location | Geographic region covered by the Dataset | 0..n |
| temporal coverage | dct:temporal | dct:PeriodOfTime | Temporal period covered by the Dataset | 0..n |
| access rights | dct:accessRights | dct:RightsStatement | Property specifying whether the Dataset is open data, has access restrictions or is not public | 0..1 |
| release date | dct:issued | rdfs:Literal typed as xsd:date or xsd:dateTime | Date of formal issuance (e.g., publication) of the Dataset | 0..1 |
| update/modification date | dct:modified | rdfs:Literal typed as xsd:date or xsd:dateTime | Last date on which the Dataset was changed or modified | 0..1 |
| Mandatory properties for Distribution | | | | |
| access URL | dcat:accessURL | rdfs:Resource | URL that gives access to a Distribution of the Dataset | 1..n |
| Recommended properties for Distribution | | | | |
| description | dct:description | rdfs:Literal | Free-text describing the Distribution | 0..n |

| Property | URI | Range | Description | Card |
|---|---|---|---|------|
| format | dct:format | dct:MediaTypeOrExtent | File format of the Distribution | 0..1 |
| licence | dct:license | dct:LicenseDocument | Licence under which the Distribution is made available. | 0..1 |
| Optional properties for Distribution | | | | |
| media type | dcat:mediaType, subproperty of dct:format | dct:MediaTypeOrExtent | Media type of the Distribution | 0..1 |
| download URL | dcat:downloadURL | rdfs:Resource | URL that is a direct link to a downloadable file | 0..n |
| rights | dct:rights | dct:RightsStatement | Specification of the rights associated with the Distribution. | 0..1 |
| byte size | dcat:byteSize | rdfs:Literal typed as xsd:decimal | Distribution size in bytes | 0..1 |
| release date | dct:issued | rdfs:Literal typed as xsd:date or xsd:dateTime | Date of formal issuance (e.g., publication) of the Distribution | 0..1 |
| update/ modification date | dct:modified | rdfs:Literal typed as xsd:date or xsd:dateTime | Last date on which the Distribution was changed or modified | 0..1 |

Table 13: Subset of metadata provided for each Dataset harvested by the EDP

Figure 24 shows the score assigned by the EDP to each Dataset belonging to the INTERSTAT catalogue for every MQA assessment dimension.

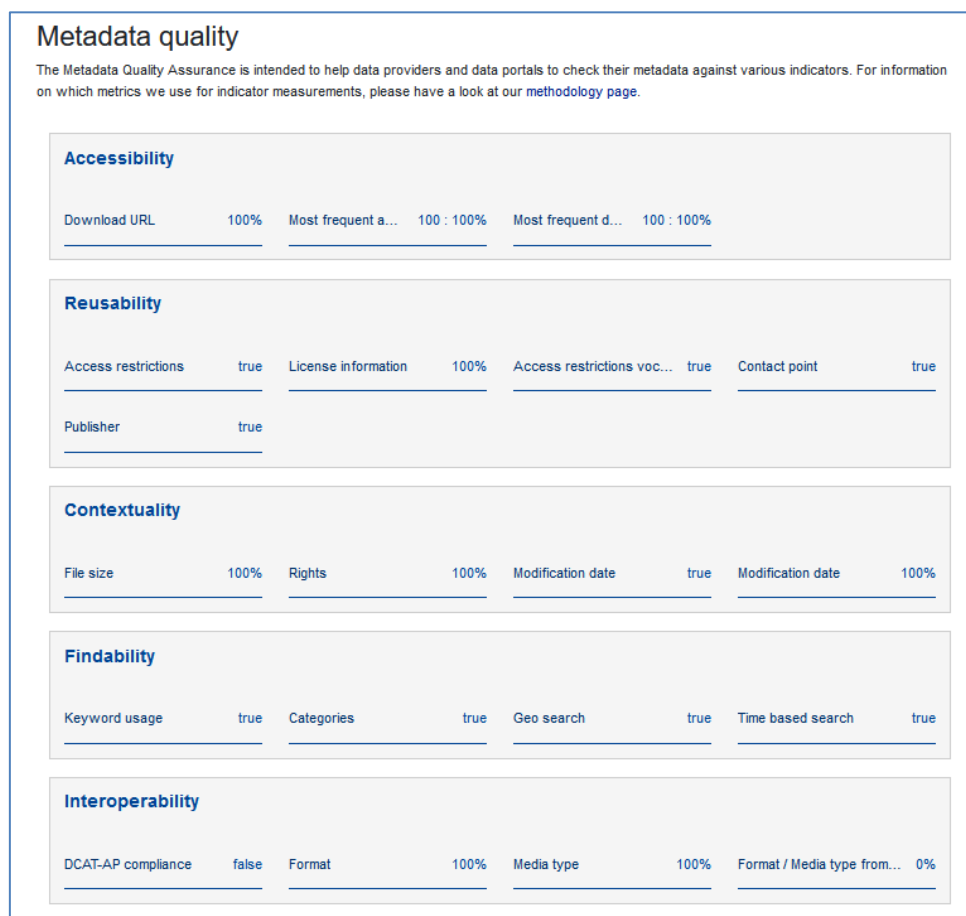


Figure 24: Results of the MQA related to the GF pilot Dataset [15]

The EDP also provides an overall rating for the whole INTERSTAT Catalogue accessed at the current date (end of August 2023), showed in the figures below.

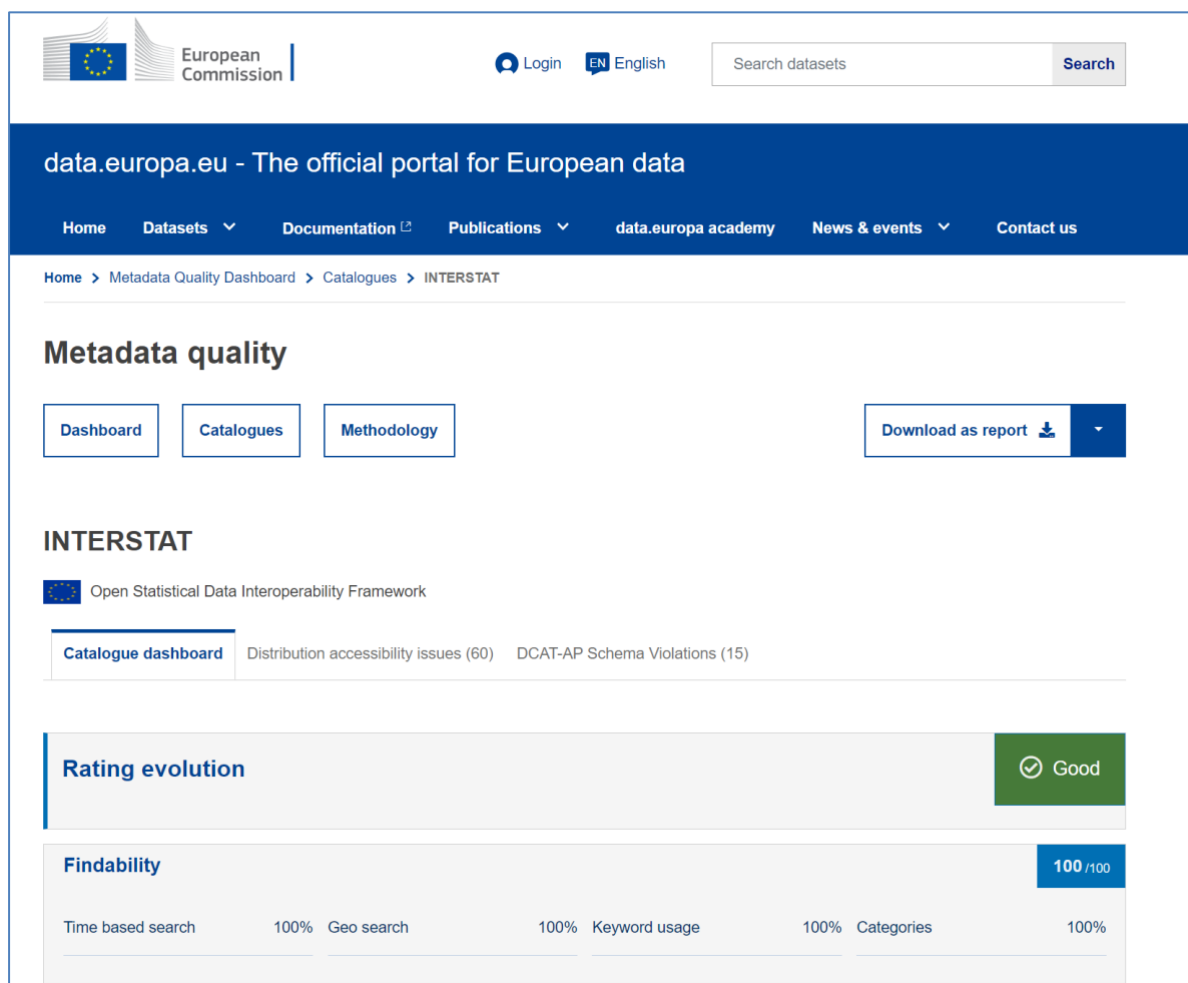


Figure 25: INTERSTAT Catalogue MQA

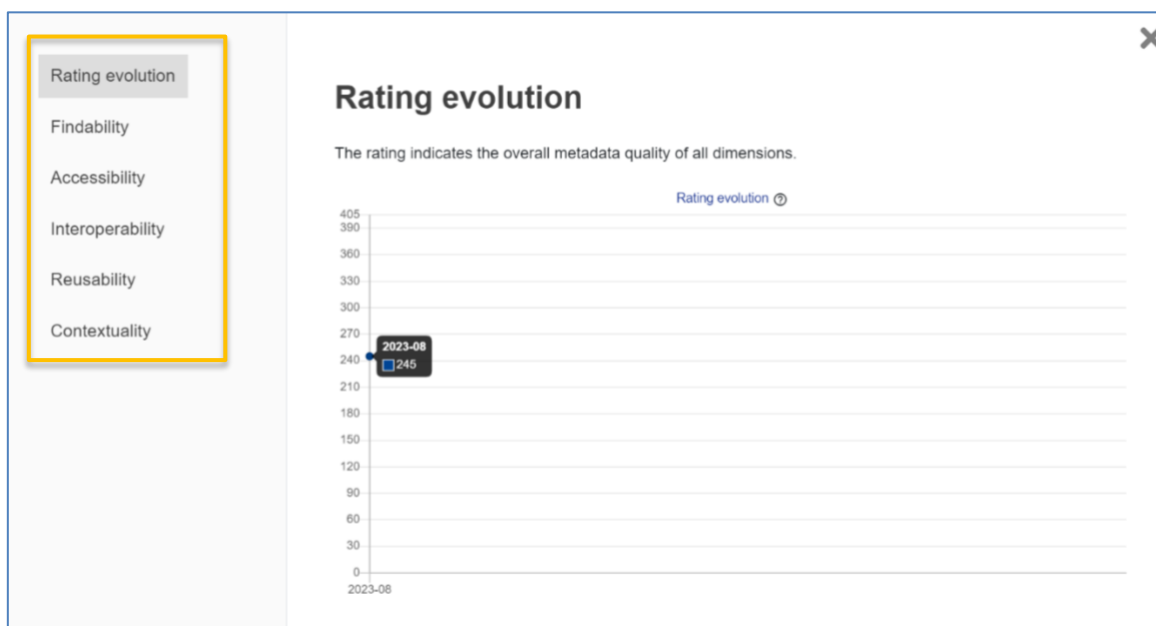


Figure 26: INTERSTAT Catalogue MQA - Rating evolution

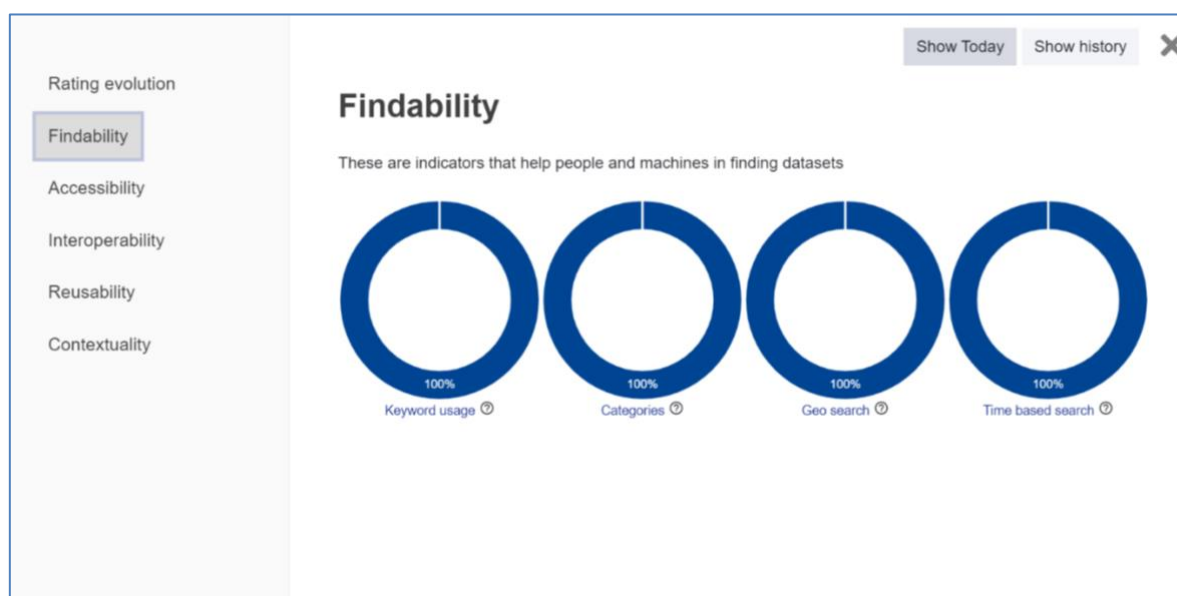


Figure 27: INTERSTAT Catalogue MQA - Findability

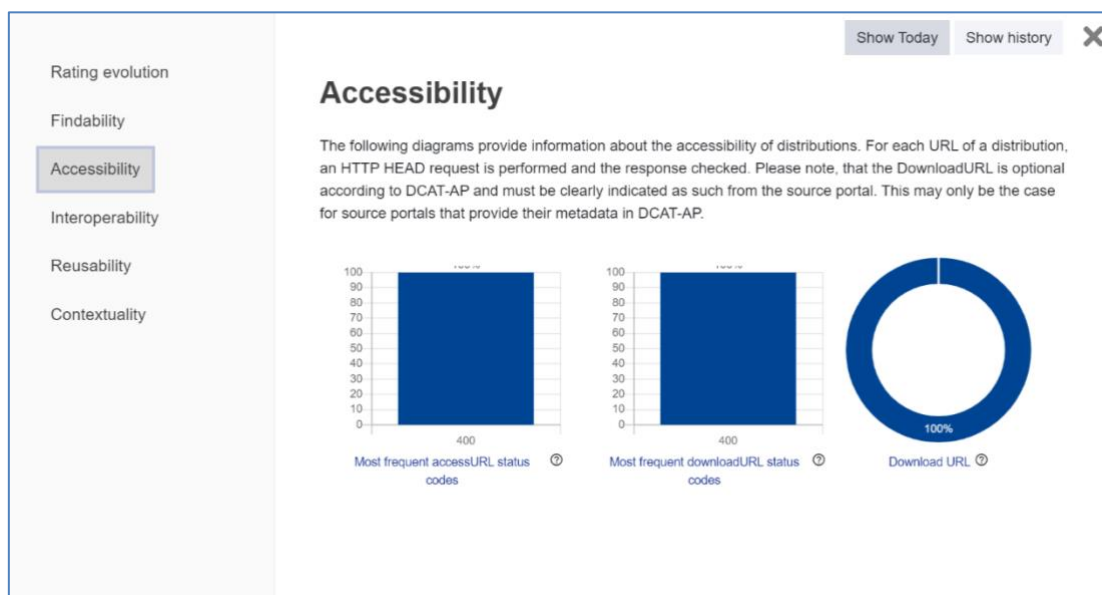


Figure 28: INTERSTAT Catalogue MQA - Accessibility

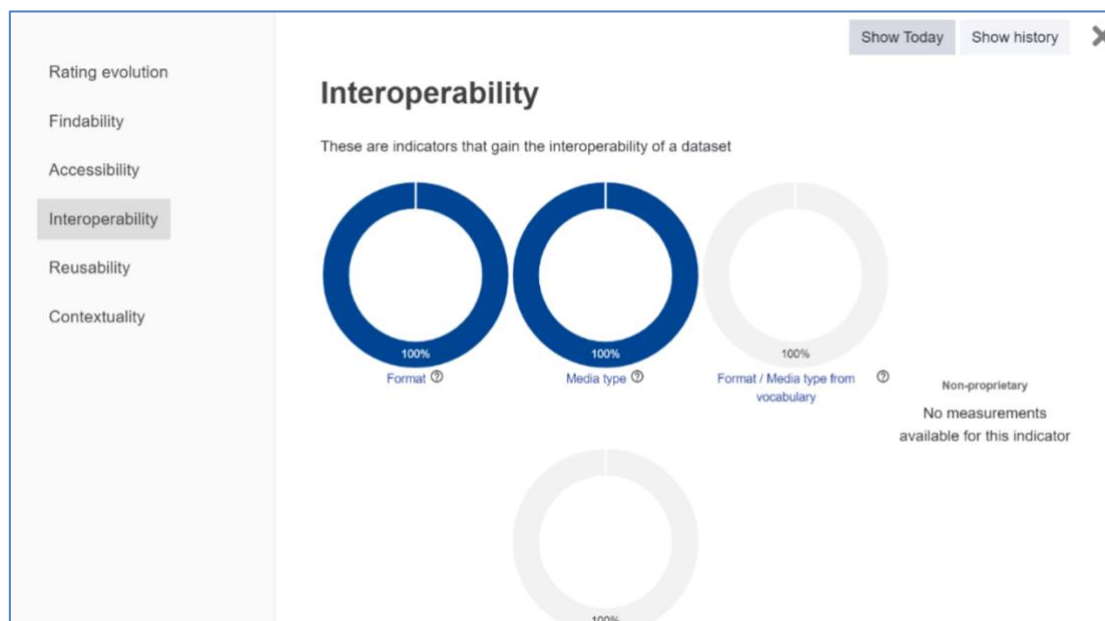


Figure 29: INTERSTAT Catalogue MQA - Interoperability

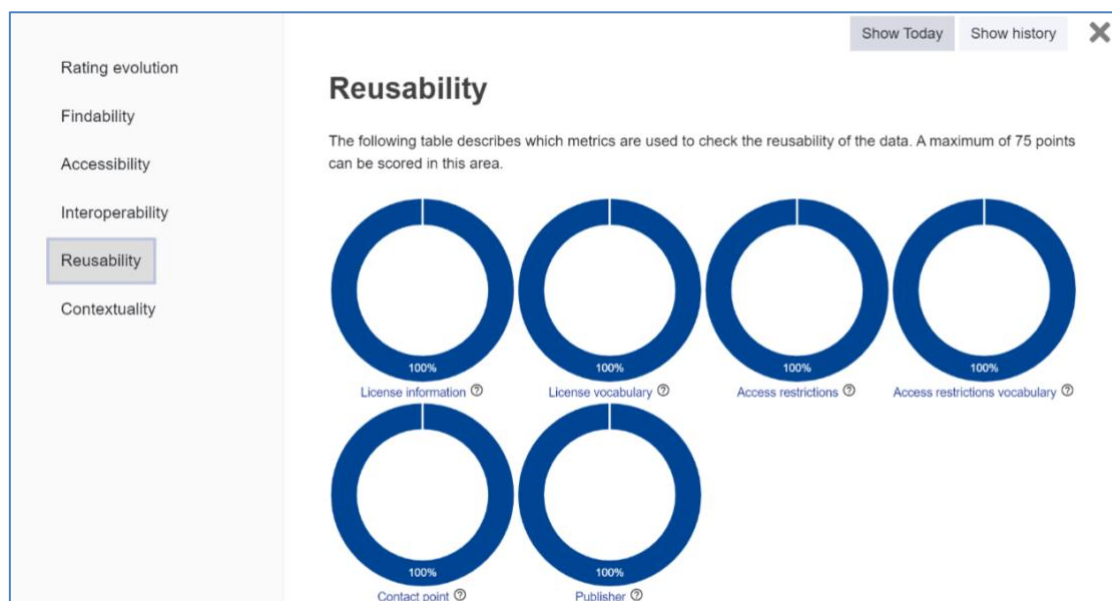


Figure 30: INTERSTAT Catalogue MQA - Reusability

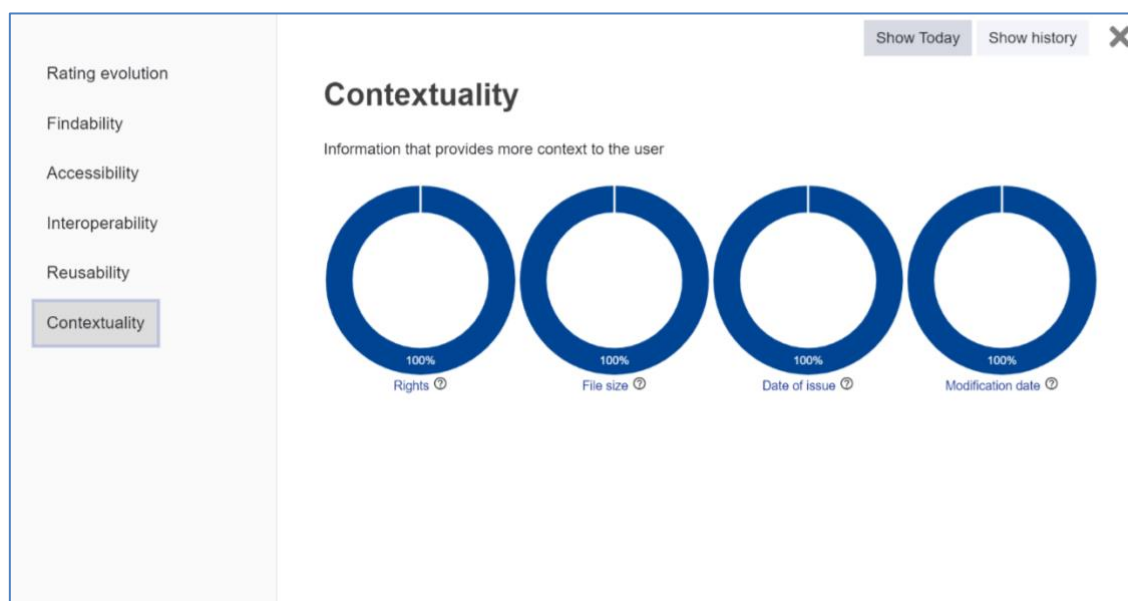


Figure 31: INTERSTAT Catalogue MQA – Contextuality

The metadata of the INTERSTAT Catalogue will be further revised to improve the rating concerning the Accessibility and Interoperability dimensions and also to fix the highlighted DCAT-AP Schema violations. In addition, another Dataset will be added concerning the data published into the FIWARE Orion Context Broker [16].

4 Lessons learned and guidelines

The main lessons learned during the project concern all layers of the interoperability-by-design paradigm promoted by the New European Interoperability Framework [17]: legal, organisational, semantic and technical. Although these aspects are intertwined, they are analyzed separately in this context in order to focus on the main issues and to draw the following insights:

- Technical tools and standards are essential to foster the reuse and integration of open data, but must be complemented by semantic descriptions. The availability of domain ontologies, meta-ontologies and common data models made the pilot applications development easier
- Accessing to well documented data and metadata impacted the selection of input data sources during the design of the use cases, highlighting the added value of combining statistical sources from different countries and domains
- The development activities for implementing the INTERSTAT framework and the pilot services involved several skills, mainly technical and domain expertise
- Relevance and need of user-friendly solutions to transform the format of open statistical data and obtain machine-readable datasets.

Based on the insights listed above, the best strategy for producing and publishing LOSD is to start upstream and concentrate the efforts on implementing common capabilities to improve interoperability. Starting small, the proposed approach that could be applied either within a single statistical organization or on a community level is composed by the following elements:

- Process revision, to detect and remove the barriers preventing the release of statistical data in open formats, ready to be reused by end-users or external portals
- Alignment of technical and semantic assets, to reuse the developed data pipelines and gain process resilience beyond data changes. Technical and semantic interoperability are two sides of the same coin, and should be developed at the same pace
- Knowledge sharing and training sessions for all the teams dealing with statistical processes, to increase their awareness of the main issues and solutions concerning data interoperability
- Cooperation between subject-matter experts and IT staff to improve data interoperability, starting from the core concepts of the statistical domains particularly relevant for the implementation of High Value Datasets

- Creation of common repositories to share the implemented solutions, domain ontologies, lists of common data models, meta-ontologies and common data vocabularies
- Revision of data and metadata formats, to make them readable by humans and machines.

In the medium-long term, the expected outcome of these guiding principles is to promote data interoperability by addressing and harmonizing the different dimensions and stakeholders involved in the statistical production.

5 Conclusions

The assessment task has confirmed the relevance of the INTERSTAT project in developing a consistent framework for the production and the publication of LOSD. Although some peculiarities requiring a technical background, the outcome of the project is significant for several types of stakeholders dealing with open statistical data. The INTERSTAT framework has integrated open tools and technical standards to meet the requirements of LOSD experts, as well as to support non-technical end users querying and exploring linked open statistical data.

The pilot applications, regardless of the specific domain investigated, have demonstrated the need for cross-border and cross-domain analysis. The dissemination of the datasets produced by the pilots through an open data catalogue harvested by the European Data Portal proves the efficacy of the INTERSTAT framework in increasing data interoperability. In addition, the compliance of the metadata documenting the published datasets with the Metadata Quality Assessment is a further result confirming the impact of the action.

Beyond the maturity of the developed solutions, the project has effectively addressed the key issues related to LOSD, and created a unique environment combining tools, technical standards and semantics. The results achieved are very encouraging and can be a starting point for building capabilities to share tools and to re-use statistical open data, thus realizing interoperable data systems to increase the value of official statistics.

References

- [1] INTERSTAT, «INTERSTAT technical website,» 2023. [Online]. Available: <https://framework.cef-interstat.eu/>.
- [2] INTERSTAT, «D3.2 - Report on pilots' development and deployment,» 2022. [Online]. Available: <https://cef-interstat.eu/resources/>.
- [3] INTERSTAT, «INTERSTAT SEP Client Application,» 2023. [Online]. Available: <https://framework.cef-interstat.eu/SEP-pilot-client/>.
- [4] INTERSTAT, «INTERSTAT, GF Client Application,» 2023. [Online]. Available: <https://framework.cef-interstat.eu/GF-pilot-client/>.
- [5] INTERSTAT, «INTERSTAT S4Y Client Application,» 2023. [Online]. Available: <https://framework.cef-interstat.eu/S4Y-pilot-client/>.
- [6] data.europa.eu, «Metadata Quality Assurance Methodology,» 2023. [Online]. Available: <https://data.europa.eu/mqa/methodology?locale=en..>
- [7] data.europa.eu, «Official portal for European data,» 2023. [Online]. Available: <https://data.europa.eu/en>.
- [8] data.europa.eu, «DCAT-AP standard release 2.0.1,» 2023. [Online]. Available: <https://joinup.ec.europa.eu/collection/semantic-interoperability-community-semic/news/dcat-ap-release-201>.
- [9] W3C, «Data on the Web Best Practices: Data Quality Vocabulary,» 2023. [Online]. Available: <https://www.w3.org/TR/vocab-dqv/>.
- [10] Engineering Ingegneria Informatica SpA, «Idra - Open Data Federation Platform,» 2023. [Online]. Available: <https://idra.readthedocs.io/en/latest/>.
- [11] Engineering Ingegneria Informatica SpA, «Idra Portal,» 2023. [Online]. Available: <https://framework.cef-interstat.eu/IdraPortal/>.

- [12] W3C, «SPARQL Query Language for RDF,» 2023. [Online]. Available: <https://www.w3.org/TR/rdf-sparql-query/>.
- [13] Ontotext, «GraphDB,» 2023. [Online]. Available: <https://www.ontotext.com/products/graphdb/>.
- [14] data.europa.eu, «INTERSTAT catalogue,» 2023. [Online]. Available: <https://data.europa.eu/data/catalogues/interstat?locale=en>.
- [15] data.europa.eu, «INTERSTAT Catalogue - GF pilot Dataset,» 2023. [Online]. Available: <https://data.europa.eu/data/datasets/de99ff51-499f-4ffe-a316-bf70404d1767/quality?locale=en>.
- [16] FIWARE, «Orion Context Broker,» 2023. [Online]. Available: <https://fiware-orion.readthedocs.io/en/master/>.
- [17] data.europa.eu, «The New European Interoperability Framework,» 2023. [Online]. Available: https://ec.europa.eu/isa2/eif_en/.

Annex: Assessment survey Questionnaire

Disclaimer

The European Commission is not responsible for the content of questionnaires created using the EUSurvey service - it remains the sole responsibility of the form creator and manager. The use of EUSurvey service does not imply a recommendation or endorsement, by the European Commission, of the views expressed within them.



1 General information

Question 1.1 - What is your main field of activity?

- ☐ Health
- ☐ Economics
- ☐ Education
- ☐ Social and Welfare
- ☐ Information Technology
- ☐ Environment
- ☐ Transportation and Logistics
- ☐ Tourism



- ☐ Agriculture
- ☐ Industry
- ☐ Other

Question 1.2 - What type of organization do you work for?

- ☐ Public Administration
- ☐ Private Company
- ☐ Other

Question 1.3 - What type of user are you with respect to the INTERSTAT framework?

- ☐ Technical User - IT Staff
- ☐ Statistical Expert or equivalent (Data Analyst, Data Scientist)
- ☐ Other Domain Expert - General User

Question 1.4 - Rate your experience in Linked Open Statistical Data (LOSD)

- ☐ Beginner
- ☐ Intermediate
- ☐ Expert

2 - Assessing the [INTERSTAT Framework](#)**Question 2.1 - For what purposes would you navigate the INTERSTAT framework?**

- ☐ To get documentation/information about LOSD (Linked Open Statistical Data)
- ☐ To publish LOSD
- ☐ To improve efficiency of already published LOSD

☐ Other

Question 2.2 - About the content of the INTERSTAT framework, how do you rate the different sections of the website?

☐ Clear and easy to understand

☐ A little confusing

☐ Hard to understand

Question 2.3 - According to your experience and needs, what are the most relevant tools in the INTERSTAT framework?

☐ Tools for data publishing

☐ Tools for data visualization

☐ Tools for data analysis

☐ Other

Question 2.4 - Are there any additional open tools to be integrated in the framework?

☐ Yes

☐ No

☐ I don't know

3 - Assessing the Client Applications

Pilot Client Overview

[SEP - Air Quality](#)

[School For You](#)

[Geo-localized Facilities](#)

Question 3.1 - Considering the topics selected for the pilots, the relevance of cross-border analysis is:



- ☐ Low
- ☐ Medium
- ☐ High

Question 3.2 - Considering the topics selected for the pilots, the relevance of cross-domain analysis is:

- ☐ Low
- ☐ Medium
- ☐ High

Question 3.3 - How do you rate the pilot interface?

- ☐ Clear and easy to navigate
- ☐ A little confusing
- ☐ Hard to navigate

Question 4 - Final Comments and Feedback

