

WP4: Methodology and Quality

*Deliverable D4.7 BREAL - Big Data REference Architecture and Layers for web
scraped data*

Final version 1.0
September 2024

Prepared by:

Olav ten Bosch (CBS)

Romain Lesur, Antoine Palazzolo (INSEE)

Sonia Quaresma (INE)

Francesca Inglese, Annalisa Lucarelli, Renato Magistro, Giulio Massacci (ISTAT)

Task coordination:

Giuseppina Ruocco (ISTAT)

This document was funded by the European Union.

The content of this deliverable represents the views of the author only and is his/her sole responsibility. The European Commission does not accept any responsibility for use that may be made of the information it contains.



Web Intelligence
Network



**Funded by
the European Union**

Table of Contents

1. Introduction	3
2. Web Intelligence Hub: main objectives.....	4
2.1. Business requirements	5
2.2. Functional requirements	8
2.3. Enhancing User Experience	10
3. Big data REference Architecture and Web Intelligence Hub implementation.....	12
3.1. Big Data REference Architecture and use cases lifecycle.....	12
3.2. Big Data REference Architecture and use cases workflows	15
3.3. The Web Intelligence Hub and open source.....	17
4. From experimental to production model	19
4.1. Key elements of the statistical production model.....	19
4.2. Web Intelligence Hub user stories.....	23
4.3. Combining actors, roles, responsibilities.....	26
4.4. Interaction between the Web Intelligence Hub and the National Statistical Institutes	28
5. Final results	34
5.1. Big Data REference Architecture enhancement for web data	34
5.2. Big Data REference Architecture extension for web data.....	39
5.3. Conclusions and lessons learnt.....	40
References	41

1. Introduction

The development of a Web Intelligence Hub (WIH) providing services for collecting and processing web data to produce official statistics is one of the main goals of the project “Trusted Smart Statistics - Web Intelligence Network (WIN)”. Within the project, Work package 4 was conceived to extend and enhance the findings of the previous ESSnet Big Data II, concerning the quality assessment, the methodological aspects and the architectural framework. More in detail, the focus of the architectural task is the “Extension and Enhancement (E&E)” of the BREAL framework (BREAL – Big Data REference Architecture and Layers¹), through the insights gained from the WIH use cases during the project.

In this report, starting from an overview of the WIN project and a focus of the architectural task, the second chapter analyses the business and functional requirements of the WIH, in terms of BREAL Business Functions (BBFs). Special attention is paid to the user experience, to highlight the impact of technical choices on the services usability analysed from the users’ perspective.

The third chapter analyses the implemented BBFs according to the three main stages of a use case, from the feasibility and exploration phases to the deployment in the production environment. The lifecycle and the maturity of the WIH use cases are key elements to consider for an E&E of the BREAL framework. The E&E of the BREAL framework based on the use cases experience aims at accelerating the WIH implementation. In order to strengthen the web scraping community within the European Statistical System (ESS), a final section highlights the relevance and the benefits of open source software for the development of sharable building blocks.

The fourth chapter explores a generalised workflow, to detail the several interconnected dimensions to consider for the definition of a statistical production model. These dimensions concern mainly the workflow execution, the workflow assessment, the several actors involved and their roles. A series of user stories provide examples of the transition from a fully centralised process to a set of tasks performed in a shared infrastructure. From the organizational perspective, in relation to the WIH functionalities and use cases requirements, the specification of “*Who can do What*” and “*Who is the owner of each task*” is essential to manage several actors and roles. The analysis of the steps that can be centralised and the tasks that must be performed locally in the national environments enables the integration of the national production systems and the WIH, as well as process transparency.

The last chapter focuses on the specialization of BBFs for web data, based on the experience gained during the implementation activities, concerning: i) the balance between process standardisation, the specific features of each use case and the associated statistical domain; ii) national regulations, country-specific practices and infrastructures. In order to face the unexpected issues that prevent the integration of a use case into production, and to manage technical or organisational aspects, planned activities and related outputs, a new BBF “Strategy and Process Management” is conceived. This additional BBF supports the definition of a statistical production model for each use case and its transition to production.

Overall, the architectural task has realised the E&E of the BREAL framework to turn the challenges encountered during the use case development into opportunities. The lessons learnt during the project point to possible ways of building a common vision and implementing a holistic approach for the integration of web data into official statistics at EU level.

¹ In the following also referred to as Big data REference Architecture.

2. Web Intelligence Hub: main objectives

The main goal of the Web Intelligence Network (WIN) project is to foster collaboration within the European Statistical System to:

- Support National Statistical Institutes (NSIs) in the development and use of tools and technologies for collecting and processing web data (starting from Web data scraping, including Natural Language Processing techniques, Machine Learning algorithms)
- Create a network, and implement a centralized infrastructure: the Web Intelligence Hub (WIH), to develop a set of use cases exploiting web data sources for specific domains
- Promote the integration between tools and services provided by the WIH and the national statistical production systems
- Explore potential extensions of the WIH through new data sources and applications
- Develop sound methodologies and a quality framework for producing statistics within the WIH.

The WIH is a common infrastructure to develop, test, share and document reusable tools for collecting and processing web data. The WIH performs several functions, serving as:

- Service provider for acquisition of web data
- Web data repository
- Environment for processing web data.

Although these functions are interconnected, the development activities depend and have an impact on the priorities and the maturity level of each WIH use case, influencing its integration in production. In this context, the term ‘Service’ refers to any code, script or procedure enabling the execution of a task, regardless of the language, or the software used.

The technical activities are grouped in the following strands of work:

- Work package 2 (WP2), responsible for moving Online Job Advertisements (OJA) and Online Based Enterprise Characteristics (OBEC) use cases into the production environment
- Work package 3 (WP3), assessing the potential of web data sources for the development of six new use cases (UC1 – Characteristics of the real estate market, UC2 – Construction activities, UC3 – Online prices of household appliances and audio-visual, photographic and information processing equipment, UC4 – Experimental indices in tourism statistics, UC5 – Business Register (BR) quality enhancement and UC6 – Faster Economic Indicators using new data sources)
- Work package 4 (WP4), conceived to extend and enhance the findings of the previous ESSnet Big Data II, concerning the quality assessment, the methodological aspects and the architectural framework (BREAL – Big Data REference Architecture and Layers).

The strategy adopted to achieve the main goals of the project is based on the combination of a top-down and bottom-up approaches. As stated in the Grant Agreement, WP2 is in charge of the development activities. The architectural task within WP4 was conceived to deal with the validation and the eventual “Extension and Enhancement (E&E) of the BREAL framework”, based on WP2 and WP3 experience and domain peculiarities. BREAL is a reference architecture resulting from the ESSnet Big Data II project, conceived to support NSIs in planning Big Data investments. BREAL provides a set of artifacts concerning the business objectives, the application components and the related data models to build a statistical process based on Big Data.

2.1. Business requirements

The starting point for this top-down approach, based on the above assumptions, is a focus on the business layer (WHAT), built on the exploration of the application layer (HOW). The business layer concerns the description of processes, abilities, actors and roles. The application layer refers to the software solutions and application services that are implemented according to the business requirements (expressed in terms of business functions). In relation to the objectives of the architectural task, within WP4, the analysis and modelling of the technical layer is out of scope. In this context, the application layer is considered only for the identification of functional requirements of the services to be provided by the WIH.

The strategy to bridge the high-level and top-down approaches is to start from scratch the analysis of the business and application layers and to identify a set of minimum requirements for the WIH platform. Adopting a learning by doing approach, the analysis of the WIH use cases workflow is performed while developing WIH services. Based on these premises, the architectural task has focused on: BREAL Business Functions (BBFs), WIH High-Level Requirements, BREAL and WIH use cases.

BREAL Business Functions

As highlighted in the first report delivered by WP4 (Deliverable 4.1: Minimal guidelines and recommendations for implementation), BBFs describe behaviors in order to organize resources, skills, or knowledge and are divided into two main subsets: “Development, Production and deployment” and “Support”. The first subset groups the core abilities related to data ingestion and processing, while the second one includes auxiliary abilities to enable and monitor their correct functioning.

The following figure shows the BBFs grouped according to these subclasses. The different colors relate to the official statistical standards (Generic Statistical Business Process Model - GSBPM, Generic Statistical Information Model - GSIM and others) and to the architectural frameworks (Enterprise Architecture Reference Framework - EARF) combined in BREAL.

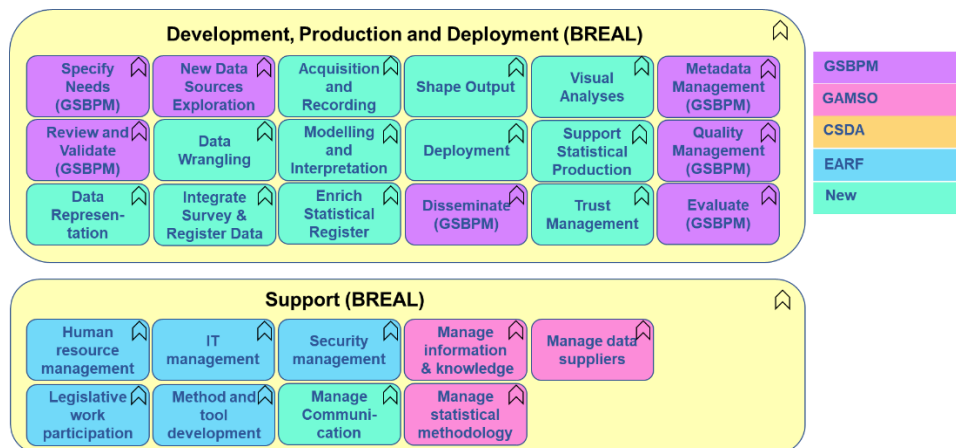


Figure 1: BREAL Business Functions²

The following analysis takes into account a subset of the BBFs, explored with respect to the life cycle of the WIH use cases. Each BBF can be associated with one or more steps of a generalized workflow

² Source: Scannapieco M., Bogdanovits F., Gallois F.; Fischer B. et al. (2019): BREAL. Big Data Reference Architecture and Layers. Version 2019-12-09. Edited by EUROSTAT

based on the WIH use cases. The figure below gives an overview and a summary of the BBFs belonging to the first subset, derived from the original description.

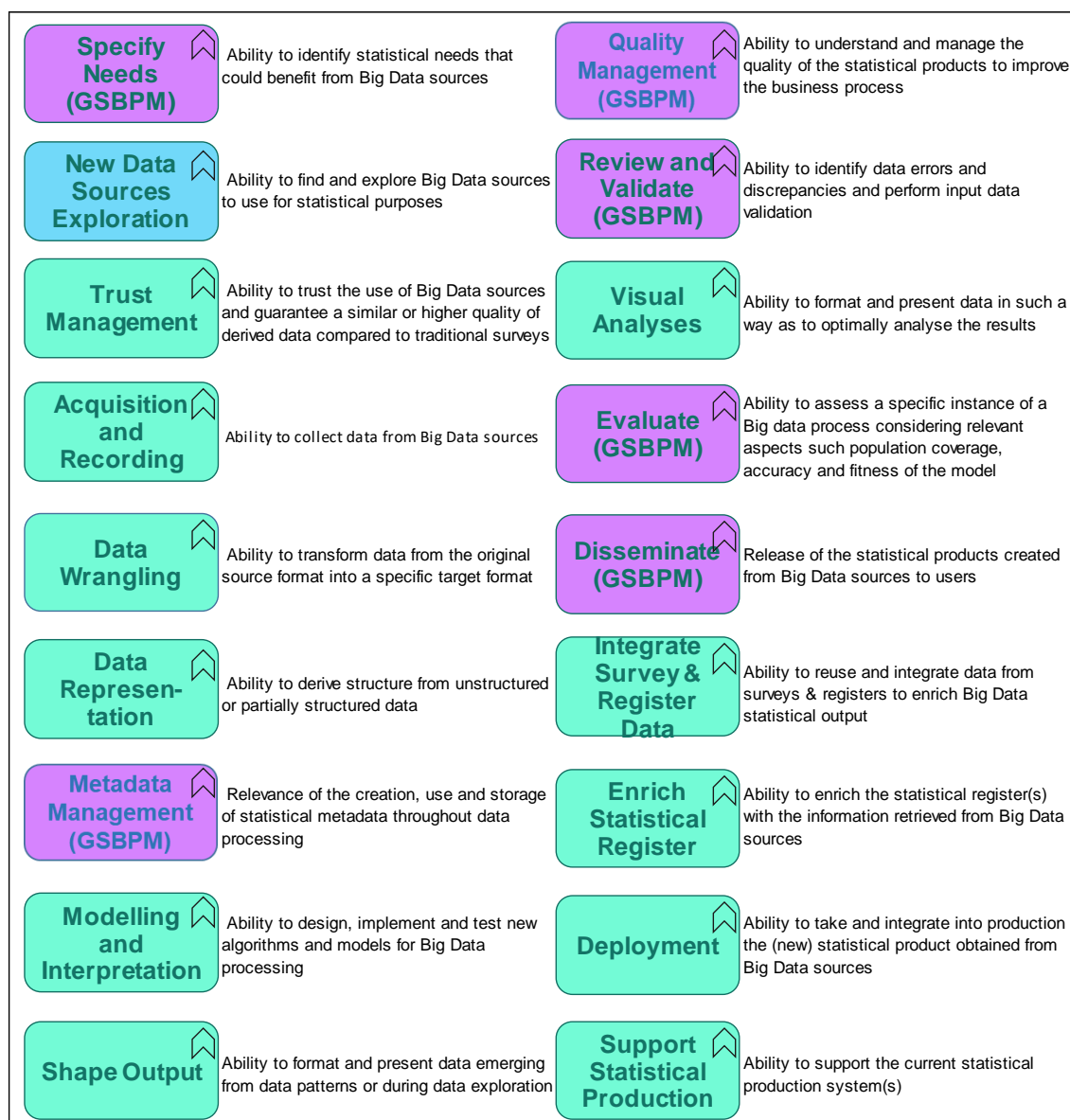


Figure 2: BREAL – Overview of Development, Production and Deployment business functions

The following figure reports the list and a summary to recall the key concepts of the Support BFFs. As above, the definitions correspond to the short descriptions reported in the first WP4 deliverable³.

³ Kowarik A., Daas P. et al. (2021) Deliverable 4.1: Minimal guidelines and recommendations for implementation. Version 2021-07-30. Edited by EUROSTAT.

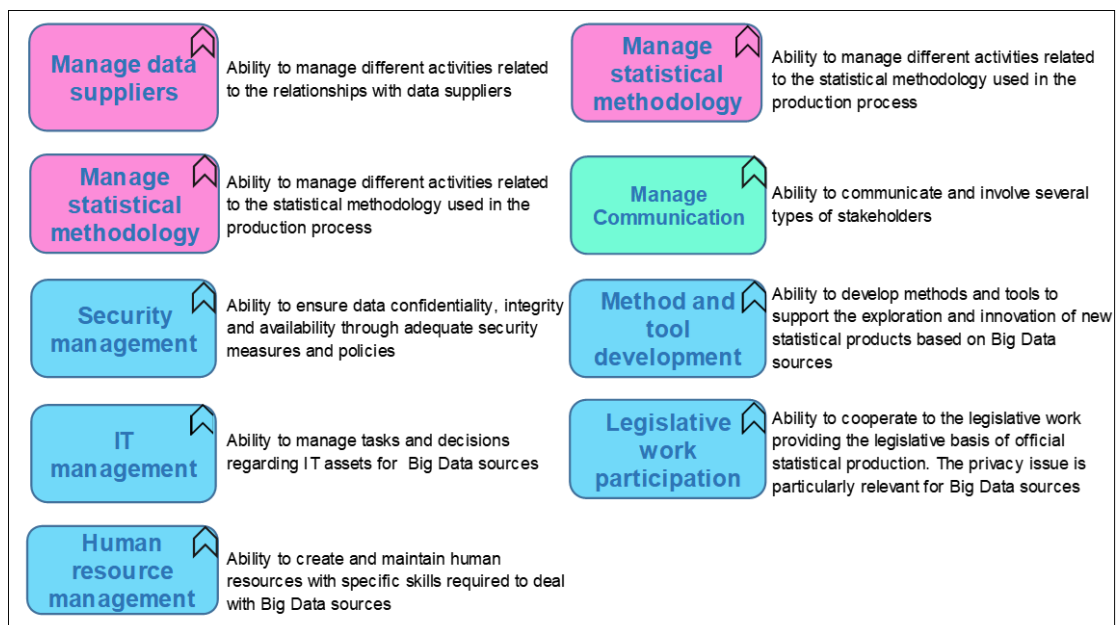


Figure 3: BREAL – Overview of Support business functions

WIH High-Level Requirements

In relation to the general objectives, as mentioned above, the high-level requirements of the WIH can be expressed in terms of BBFs. The aim of this approach is to explore how the BREAL framework can support the implementation of the WIH use cases, and how this framework can be enriched to be specialized for web data. Indeed, the BBFs group the main abilities and activities to develop, aimed at exploiting big data sources for statistical purposes. Therefore, the WIH was conceived to enable the achievement of the following general purposes:

- Providing a set of tools to support NSIs in the acquisition and processing of web data
- Managing data and metadata storage and updates
- Disseminating microdata derived from web data content to enrich official statistics
- Promoting the collaborative work among NSIs through shareable and reusable solutions regardless of the specific use cases
- Ensuring incremental improvement of the developed solutions according to methodological and quality enhancements.

The high-level purposes listed above are interdependent and can be associated to several sub-processes, for example: URLs landscaping and selection, URLs scraping, Analysis of scraped content, Monitoring URLs stability over time. Each sub-process can be associated to one or more BBFs and performed by one or more WIH services, depending on the developers' choices and users' needs.

Linking the definition of BBFs and the workflow of each use case under development, the WIH aims at supporting the following tasks:

- Management of the main issues related to the acquisition of web data
- Versioning of the web data stored in the WIH repository
- Tracking and monitoring data format transformation
- Testing and implementing algorithms, methods and accuracy indicators

- Making available to WIH users a set of process metadata (e.g., parameters set for the crawling, Machine learning metrics and parameters) for process tracking and assessment
- Standardize the statistical output to disseminate and the related quality indicators.

2.2. Functional requirements

The description of high-level requirements enables the definition of the end functions to be implemented through the WIH services. An overview of the functional requirements of each service should reduce the gap between the top-down approach, based on BBFs, and the bottom-up approach related to the implementation activities.

Despite the definition of general requirements for process and service functionalities, the project experience has revealed several challenges related to the extent to which a service can be shared, reused and standardized. In other words, the implementation activities have enhanced that service sharability is influenced by the national context and the statistical domain. This insight underlines the need to assess if some functionalities can be standardized for all countries and domains or need to be customized to meet specific users' requirements. Some issues highlighted in the development of all use cases, regardless their maturity stage, are listed below:

- Lack of data standardization: each provider represents the data in different ways, inducing difficulties to harmonize them
- Identification of the statistical unit: aspects related to data deduplication, especially when the same unit is represented between two or more providers
- Difficulty of a generalized data scraping tool: provider sites have different structures which do not allow the development of a single software
- Problems related to legal aspects: the terms under which data can be downloaded are not always clear.

Considering a set of services realized for the acquisition and processing of web data, the following table reports an example of functional requirements. Regardless of the technical implementation of the services, the main aim of this list is to highlight the relevance of gathering user needs in terms of the tasks to be performed.

Functional Requirements (FR)	Description	Tasks
FR1: Source evaluation	Evaluation of the reliability of the data source	Assess the source of the data to be acquired, comparing it with data from other sources if possible
FR2: Agreed API connector	Read data from API's	API description, protocol and endpoint
FR3: Data provenance & terms of use	Provide documentation including legal aspects and usage rules	Compile provenance metadata to track the process from the beginning, also for data provided or preprocessed by third parties
FR4: Service configuration	Configure, monitor and save the parameters of the service used	Standard or customized set-up of services parameters
FR5: Data ingestion	Ingest data considering specific acquisition criteria	List of URLs from landscaping and/or specified keywords to be included for URLs finding
FR6: Data transformation	Perform data cleaning, transformation and reduction	Standardization and harmonization of data format and rules applied for data deduplication and cleaning

Functional Requirements (FR)	Description	Tasks
FR7: Data processing	Execute algorithms and models	Compile process metadata to assess and improve the models to be used
FR8: Data analysis & visualization	Access statistical output for data analysis	Summarize the statistical output
FR9: Results download	Download of the results for output analysis	Standardize output data structure

Table 1: Examples of functional requirements for acquisition and processing of web data

The requirements listed above refer to running services deployed through the WIH. A valid alternative to shared WIH services is to make available and executable the source code locally, in the NSI's infrastructure or in the WIH environment. This can be possible thanks to collaboration tools, which also allow users to manage the source code, as well as its improvements and versioning.

Based on the project experience, a benefit for the WIH is the integration of different types of APIs for the users, which can be categorised as follows:

1. **Data APIs:** these APIs provide access to external data sources like databases, data warehouses, or data lakes. They allow retrieval of data from external sources for potential integration into the WIH. This enables the bulk uploads/downloads of large datasets, comparisons and studies using nationally or centrally collected data.
2. **Machine Learning APIs:** these APIs offer machine learning capabilities like natural language processing (NLP), image recognition, or predictive analytics (e.g., TensorFlow, IBM Watson). They enable the integration of advanced analytics and predictive models into the WIH.
3. **Analytics APIs:** these APIs provide access to analytics platforms or services, allowing users to retrieve data analytics insights or perform data analysis within the WIH. Examples include Google Analytics API or Microsoft Power BI API.
4. **Third-Party APIs:** these APIs, provided by third-party services and platforms, offer additional functionalities. Examples include social media APIs (e.g., Twitter, Facebook), weather APIs (e.g., OpenWeatherMap), news APIs (e.g., NewsAPI, New York Times API), or mapping APIs (e.g., Google Maps, Mapbox).

The second category, Machine Learning APIs, deserves special attention as it is an important part of the work to be carried out through the WIH. Currently, the most prominent APIs in this area for text prediction and classification are:

1. **OpenAI API** that provides an API offering different natural language processing capabilities, including text classification. It allows developers to build custom text classification models using their GPT (Generative Pre-trained Transformer) models
2. **Google Cloud Natural Language API** that offers powerful text analysis capabilities, including text classification. It can categorize documents into predefined categories or user-defined custom categories. The API also provides sentiment analysis, entity recognition, and syntax analysis features
3. **Amazon Comprehend** that provides natural language processing capabilities, including text classification. It can categorize text documents into predefined categories or custom categories specified by the user. It also supports sentiment analysis, entity recognition, and key phrase extraction
4. **Microsoft Azure Text Analytics** that provides text classification capabilities, allowing developers to categorize text documents into predefined categories or custom categories. The service also offers sentiment analysis, entity recognition, and key phrase extraction

5. **IBM Watson Natural Language Understanding** that offers a natural language understanding service to analyze text and extract insights like sentiment, entities, keywords, and categories. It supports text classification by categorizing content into user-defined or predefined categories
6. **Aylien Text Analysis API** that offers a text analysis API that includes text classification capabilities. It can categorize text documents into predefined or custom categories, as well as extract entities, sentiment, and other insights from text data.

2.3. Enhancing User Experience

Though the analysis of the technological layer is out of scope, technical choices have an impact on the service usability from the users' perspective. This section aims to highlight the challenges associated with the technical architecture of the WIH Data Lab, in order to provide insights for enhancing user experience in the future. The analysis focused mainly on the function of the WIH as a data repository, to facilitate access by statisticians and domain experts. For these actors, interaction with the WIH Data Lab may be more challenging than for IT experts who are used to accessing different platforms, portals and environments. In addition, the data acquisition service was not available until later in the project, so it could not be thoroughly evaluated.

The WIH Data Lab underwent a transition to a new platform, based on the Big Data Test Infrastructure (BDTI) project created by DIGIT in 2019. This transition occurred in February 2024 and offers a cloud-based analytics test environment free of charge for Public Administrations in the European Member States.

Regarding the user experience (UX), the initial interaction with the Data Lab platform involves the connection process. In the previous version of the Data Lab, SSH tunneling was predominantly promoted for connection, which, although secure, complicated user access and deviated from best practices in cloud-native application development. With the new platform, HTTP(S) access is promoted, aligning more closely with current best practices. Furthermore, the management of user access identities transitioned from Microsoft services to EU Login, addressing concerns about sovereignty expected from an EU service and eliminating reliance on third-party proprietary systems. However, at present the Sharepoint documentation of the platform remains managed by Microsoft, although alternatives are being explored.

Once logged in, users gain access to the main portal, as shown in the figure below.

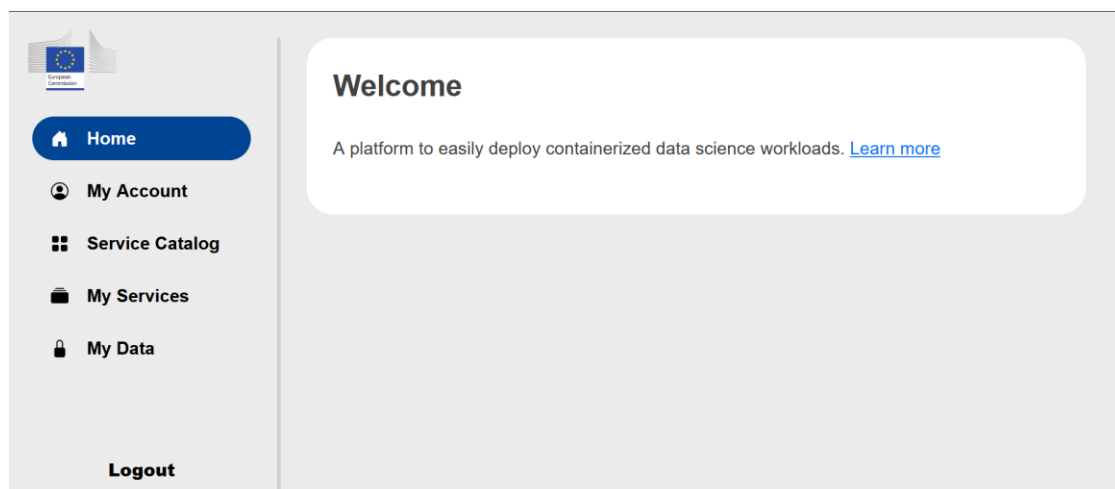


Figure 4: New Data Lab Landing page

Onboarding sessions and hands-on demos are regularly organized to facilitate user familiarity with the Data Lab. However, the need for clear guidelines about the platform has been raised, particularly as the interaction between services launched by different users can be challenging, and not all users are data scientists. Nevertheless, improvements have been made compared to the previous Data Lab, for instance with regard to the design of the pages, or the availability of a centralized overview of all the accessible services. The new service catalog offers an access to all the services below:

- Data Processing and Workflow Management
 - Apache Airflow v2.7.1: Workflow automation and scheduling
 - Apache Spark v3.4.1: Big data processing and analytics
- Data Visualization and Business Intelligence
 - Apache Superset v2.1: Data exploration and visualization
 - Kibana v8.5.1: Data visualization and exploration for Elasticsearch
 - Metabase v0.47.1: Business intelligence and data visualization
- Data Storage and Management
 - Elasticsearch v8.5.1: Search and analytics engine
 - MongoDB v6.0.9: NoSQL database
 - MinIO v2023.07.07: Object storage
 - PostgreSQL v15.4.0: Relational database management system
 - Virtuoso v07.2.10: Multi-model database.
- Data Science and Machine Learning
 - H2o v42.0.4: Machine learning platform
 - Knime v5.1.0: Data analytics, reporting, and integration platform
 - RStudio v4.3.1: Integrated development environment (IDE) for R
- Development and Analysis Tools
 - JupyterLab v4.0.4: Interactive development environment for notebooks, code, and data
 - JupyterLab (Spark) v4.0.4: JupyterLab with integrated Apache Spark support
 - PgAdmin4 v7.6: PostgreSQL database management tool.

In contrast to the previous platform, opened services are now independent for each user. This means that it is no longer a single virtual machine that is shared by all the users. This not only enhances security by preventing access to other users' work, but also facilitates horizontal scalability and reproducibility. Additionally, users can customize the resources required for each created service.

Overall, the transition to the new platform appears to be positive, provided that users can easily adapt. However, the BDTI does have limitations, some of which are outlined in the Data Lab documentation. Firstly, the infrastructure in itself is not yet open source and the release of the source code is planned to happen later in 2025. An example of an open source alternative is [Onyxia](#), the reference infrastructure for the One-Stop Shop, a new project started in April 2024⁴. An earlier version of the BDTI was created in 2019, which had a monolithic approach that was very different from the current Data Lab or Onyxia.

Leveraging Onyxia paves the way for future improvements in the development of shared infrastructures. Regarding user experience, gathering feedback from users on the new platform will be crucial for drawing conclusions and further enhancing the overall experience.

⁴ One-stop-shop for Artificial Intelligence and Machine Learning for Official Statistics (AIML4OS) <https://cros.ec.europa.eu/dashboard/aiml4os>

3. Big data REference Architecture and Web Intelligence Hub implementation

Bridging the top-down and bottom-up approaches reinforced the need to analyze the WIH use cases workflow implementing BBFs, to identify:

- The minimum set of tasks to be developed in relation to the maturity of a use case for accelerating the transition to the production environment
- The degree of standardization of each task
- Prerequisites and dependencies for task execution
- Centralized or in-house task management
- Technical issues related to the environment and/or procedures
- Degree of scalability, reuse and sharability of WIH services.

The following sections focus on increasing the maturity of WIH use cases with respect to a generic workflow, modelled for each stage of a use case lifecycle and associated with BBFs. This workflow is a result of the experience and the insights gained from the development of the OJA and OBEC use cases.

3.1. Big Data REference Architecture and use cases lifecycle

The key elements to consider for an Extension & Enhancement (E&E) of the BREAL framework for web data relate to the lifecycle and the maturity of a use case. A use case starts with an exploratory phase that is a feasibility analysis, generally performed through the development of Proofs of Concept (PoCs). In the following experimental stage, the core activities are aimed at implementing building blocks, to realize the main “Development, Production and Deployment” BBFs. In the last phase, the development of each building block is completed and the interaction and the coordination with the services realizing the support BBFs are finalized, so that the use case can be integrated in production.

Throughout the different stages, additional activities are carried out on a larger scale. The figure below provides a general example of the combination of these three stages and the solutions implemented in relation to BBFs.

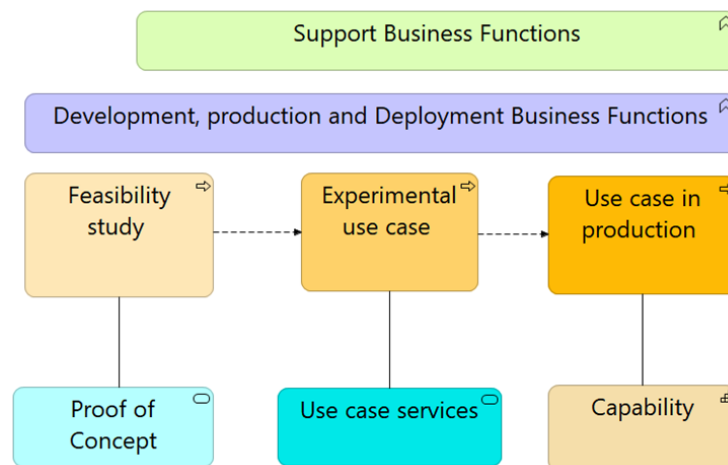


Figure 5: Example of combining a use case lifecycle and BREAL Business Functions

Beyond the iterative nature of the statistical process, the analysis of BBFs with respect to the WIH use cases supports the modelling of the application layer in terms of functional requirements of the solutions to be developed. The BBFs selected for a generic use case in the different phases relate to the main abilities needed to produce the statistical output from web data sources. The following analysis focuses on the business layer, to highlight the tasks to execute in order to transform web data content in statistical output.

Starting from the first stage, the relevant objectives of the feasibility study concern the exploration of URLs, the assessment of their relevance for official statistics and the inventory of available tools in order to avoid developing from scratch. Management of web data providers and an initial assessment of the statistical output are essential to evaluate the new data source in relation to the information needs. As shown in the following figure, the focus of the feasibility study is to select a set of websites and/or portals for assessing their relevance with respect to specific statistical needs within a given domain. To produce and evaluate the quality of the statistical output, in this stage the following activities are performed:

- Definition of criteria for the selection of URLs (Six et al., 2023)
- Set up of crawler/scrapper tool and parameters
- Data storage
- Data deduplication and standardization
- Definition of an initial set of ontologies and validation rules.

In relation to the application level, this stage produces PoCs, realised to test the initial assumptions and executed either on local infrastructure, or centrally, in a common infrastructure. In the WIN project, the new use cases explored through WP3 tasks have started from the feasibility phase. In this stage, the implementation of the BBFs, in the figure below modelled through ArchiMate⁵ language, was the focus of the development activities in the WP3 use cases⁶.

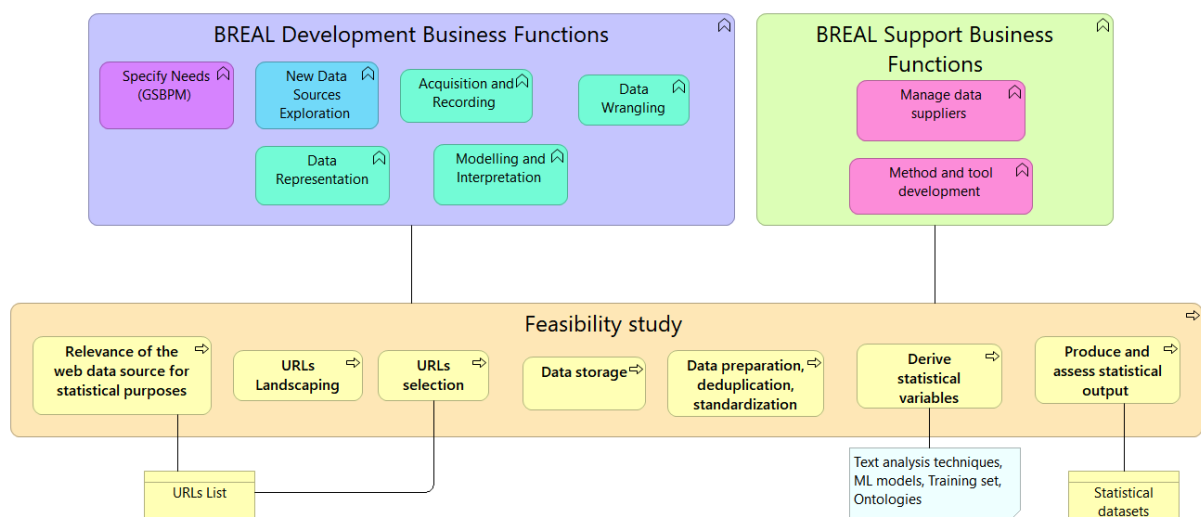


Figure 6: Feasibility study – Business layer

⁵ ArchiMate is an open and independent language for architectural modelling according to the Enterprise Architecture standard, available from: <https://www.archimatetool.com/>

⁶ For a more in depth description of the implementation activities, see: Deliverable D3.3 “WP3 3rd Interim technical report”.

In the experimental phase, the previously implemented tasks are further developed and adapted, with particular attention to process monitoring and quality assessment. As an example, for the data acquisition, the list of URLs explored in the previous stage is checked to assess the relevance and stability of the selected websites. The figure below shows the process steps for the realisation of the remaining BBFs in order to upgrade the feasibility phase.

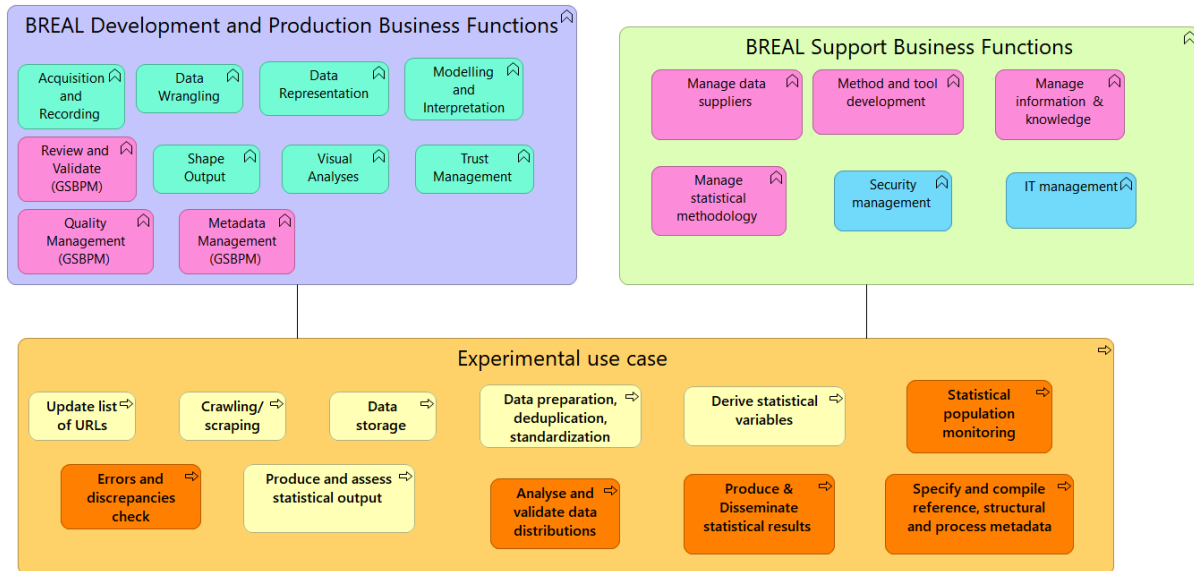


Figure 7: Experimental use case – Business layer

The subtasks highlighted in dark orange show the following additional activities:

- Statistical population monitoring
- Errors and discrepancies check
- Analysis and validation of data distributions
- Production and dissemination of statistical results
- Specification and description of reference, structural and process metadata.

The main goal of these activities is to improve the accuracy of the statistical output, and perform data validation, both at micro and macro level, in order to produce and disseminate experimental statistics. To this end, the creation of reference, structural and process metadata allows the output to be disseminated with the auxiliary information explaining the benefits and the limitations of the statistics produced. Concerning the environments and the actors involved, this phase is essential to establish the existing tools which can be reused and shared, the activities centrally and locally managed, as well as the skills needed to perform the tasks. With the aim of developing a common infrastructure and web data network, this phase is also important to identify specific requirements of a particular use case or domain.

In order to facilitate the transition to the production phase, the tasks developed and enhanced during the experimental phase will be further improved to realise the whole set of BBFs. In this case, the focus is mainly on activities that can be managed or harmonised by a common infrastructure. However, the addition of the supporting BBFs "Integrate Survey & Register Data" and "Enrich Statistical Register" highlights the tasks that can be carried out within the infrastructure of an NSI to avoid privacy issues.

In the figure below, the additional activities developed with respect to the experimental stage mainly concern:

- Dissemination of statistical results and related metadata (reference, structural and process)
- Overall assessment of the process instance
- Integration with survey data and BRs
- Enrichment of BRs.

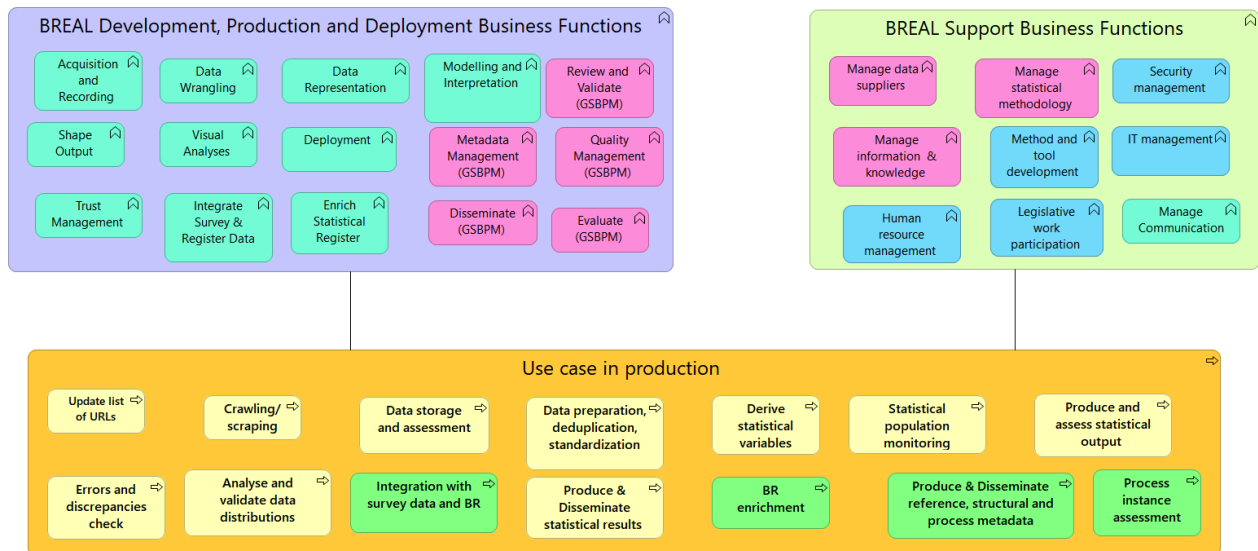


Figure 8: Architectural layers – Use case in production – Business layer

When a use case is ready for production, based on the developments of the previous phase, the process is clearly planned and the tasks, centrally or locally managed, are well defined. The organisations and the key stakeholders involved in the tasks are also identified, as is the task owner that is the entity responsible for executing and monitoring each task.

3.2. Big Data REference Architecture and use cases workflows

Modelling a generalized workflow in relation to the different stages of a use case life cycle accelerates the transition of an experimental use case in the production environment. Mapping this standardised workflow to BBFs enables the confluence between the top-down and the bottom-up perspectives. More specifically, modelling the process steps associated with BBFs is essential in order to have a fairly detailed template against which to assess the maturity of a use case. By abstracting from a particular domain, this template supports all use cases, focusing on the main aspects that need to be orchestrated to accelerate the deployment to production.

Based on the WIH experience, in the transition from the experimental to the production phase, the most challenging activities are not related to the implementation of the workflow, but to its execution and assessment, as well as to the management of its components and actors. These aspects, detailed in the figure below are grouped in three main subsets: workflow execution, workflow assessment and BREAL actors and roles.

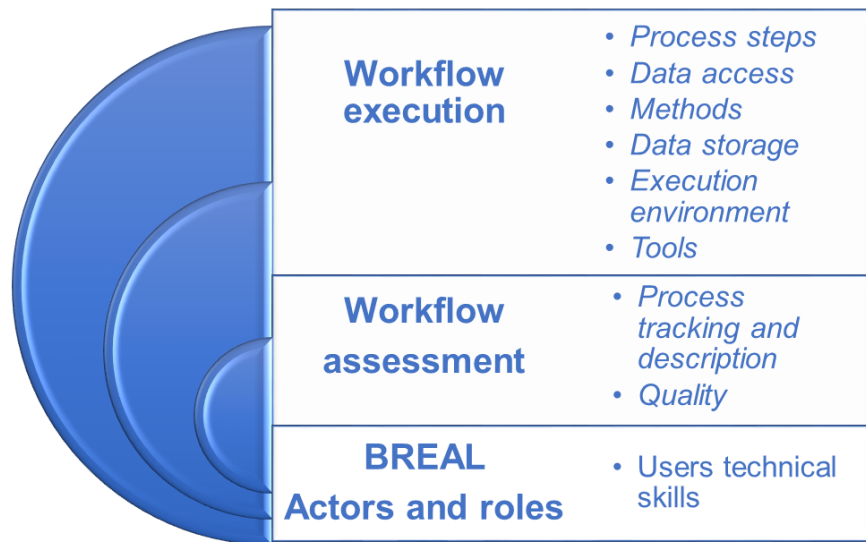


Figure 9: Workflow components and actors

Each subset relates to one or more BBFs and has an impact on the others. Indeed, all these elements must be in place to perform a process step in production: the environment, the access to input data, collected or stored in a repository, the tools performing the methods to apply, the actors involved in the task execution.

Once each task is in place, the evaluation and reporting of the results may lead to a step iteration. Aiming at the adoption of a common production model, this analysis is useful to decide which elements need to be centralised, whether it is only the environment or also the methods, the tools and the quality assessment. In other words, the steps which can be centrally executed in a common infrastructure and the steps which must be standardised and performed locally in NSI's infrastructure. The combination of several tasks, managed either centrally or locally, requires the coordination of different actors belonging to different organisations. In terms of skills, these actors can be grouped into three subsets: data scientists, methodologists and domain experts.

Considering the OJA use case, mapping its workflow to the generalised workflow modelled for the production stage, highlights the steps to be locally or centrally managed. In fact, a task can be fully or partially centralised. It is fully centralised when it is performed by one organisation in a common infrastructure. Otherwise, it is partly centralised if it is carried out in a common environment, in most cases using standardised or shared tools. Conversely, a task is managed locally if it is executed in-house, even if it is performed through standardised or shared tools. The combination of centralised and localised tasks will differ depending on the roles and interactions between the organisations working together in a particular use case.

As an example, in the figure below, the OJA use case is analysed in relation to these categories. Particularly, the yellow coloured tasks are fully centralized, executed for all countries by Eurostat. The tasks highlighted in green are partially centralised. They are carried out at central level by Eurostat with the contribution of the NSIs participating in the WIN project, which evaluate and analyse the statistical output using the common infrastructure. The orange sub-processes correspond to the activities locally performed by NSIs.

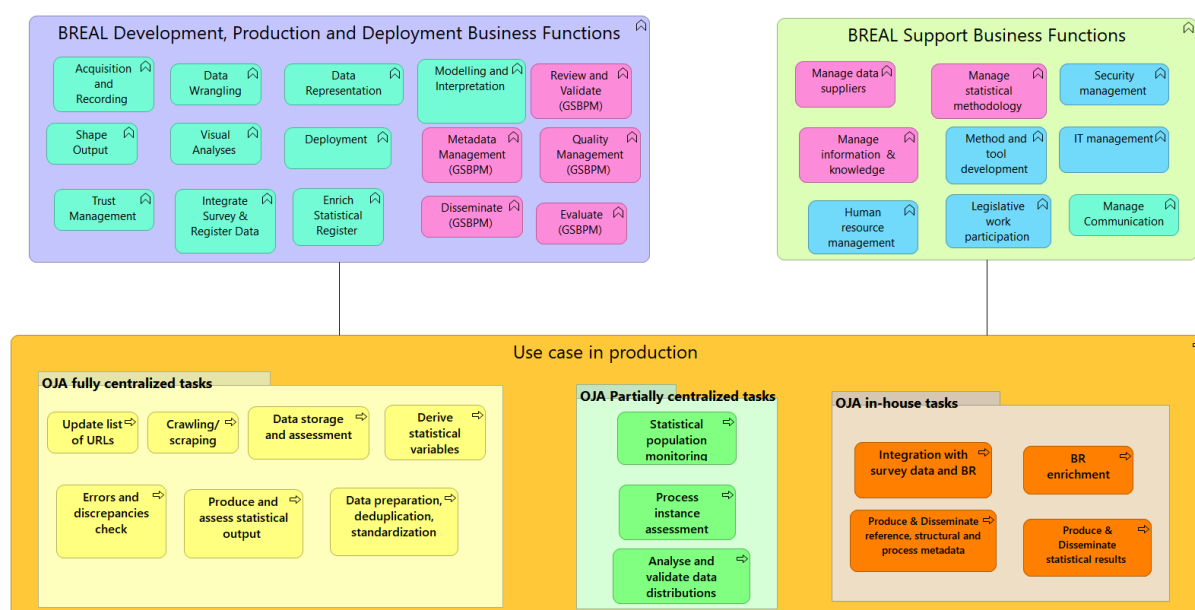


Figure 10: Centralized and local tasks in the OJA use case

3.3. The Web Intelligence Hub and open source

In the current implementation of the WIH, the majority of its source code is closed, and some sub-processes rely on vendor services. In this section we motivate why the WIH software should be open source and adopt open source software building blocks, where possible and feasible, so that it can open up possibilities for strengthening the ESS web scraping community as a whole.

The motivation for opening up software in governmental organizations as open source and in official statistics has been formulated already in many documents. We briefly repeat the most important:

- The EC has defined an EU-Open source software strategy in 2020⁷, “Think Open”, which “sets out the vision for encouraging and leveraging the transformative, innovative and collaborative power of open source, its principles and development practices. It promotes the sharing and reuse of software solutions, knowledge and expertise, to deliver better European services that benefit society and lower costs to that society. The Commission commits to increasing its use of open source not only in practical areas, for example IT, but also in areas where it can be strategic”. On 8 December 2021 the Commission adopted new rules aimed at facilitating and accelerating the publication of the software source code it owns⁸. We think the ESS is an excellent example of such strategic area.
- The Conference of European Statisticians (CES) conference⁹, 71st plenary session, 2023, held a session about adopting open source in the ESS, with papers from six NSIs. In the discussion it was pointed out that “Adopting open-source technologies could provide NSOs with an opportunity to: Redesign and standardize statistical production processes, or orientate them around workflows; Embrace practices of working in collaboration with others in an open, transparent and efficient way”.

⁷ https://commission.europa.eu/about-european-commission/departments-and-executive-agencies/digital-services/open-source-software-strategy_en

⁸ https://commission.europa.eu/news/commission-adopts-new-rules-open-source-software-distribution-2021-12-08_en

⁹ <https://unece.org/statistics/events/CES2023>

- Eurostat and member States created a group on open source, called OS4OS, in 2022-2023, with representatives of eighteen NSIs and two international organizations (Eurostat and OECD). Results¹⁰ of the group were a set of principles, bottlenecks and best practices, all adopted by all group members.
- The UNECE performed a project on adoption of open source in 2023 and is continuing this work in a new project¹¹ in 2024.
- As part of the ESSnet I3S, a proposal for an ESS Open Source strategy implementation roadmap was formulated, see¹².

This list shows that the reuse of software assets across organizations in the official statistics community via open source solutions has been stressed many times, that many successful examples¹³ have shown to be effective, and that there is support on management level. Writing down the intentions is still different from reaching open source maturity in reality. The WIH is no exception to this. The reuse of costly scraping can only be gained if a true open source approach is also adopted as a common principle within the WIH community. From an architectural perspective, the actual scrapers might be shared in the official statistics community. This means that WIH OJA portal scrapers (frontend, API or headless) for portals that contain national content, can be reused in a national context by the corresponding National Statistical Offices (NSOs). This prevents duplication of scraper development on the same portals. The same holds for generic text processing and coding software for OJA variables or skills detection. If open sourced, they can be reused in national OJA and skills projects. It also holds for the data processing steps behind text coding (e.g., the classification models), and the dashboard dissemination software. If open source, it could be reused and collectively improved by national job vacancy/skills projects, making the result stronger. If not open source, there is no choice for other OJA/skills projects to implement their own, which is totally in contradiction with the intentions in the groups mentioned above. For OBEC software, ideally the e-commerce detection and social media and possibly other variables extraction from website content should be released as open source to enable maximum reuse and transparency in national ICT statistics. The use cases in WP3 are still in early development, but even here – or maybe especially here (see principle “work in the open”), shared development via open source will pay off in the end result.

To enable them to be reused, WIH developments must also rely on open source software and libraries, without which the codes could not be executed in other environments. The WIH, in its current implementation, uses some cloud vendor services whereas there are open source alternatives¹⁴. At the moment, the use of Athena is for managing the data access and provide easier access to the parquet files in S3. The main advantage of this solution is that it can be easily replaced by other alternatives for data management and access.

The very nature of web scraping requires the utmost transparency. Many web scraping activities on the internet are carried out maliciously, with the aim of stealing the content of websites, even illegally. This is why websites owners seek to protect themselves from web scraping by putting in place defence mechanisms, for instance blocking IP addresses or honeypots that send false data to scrapers. Ethical and legitimate scrapers therefore need to distinguish themselves, through a formal agreement with data providers, so that they are not considered as malicious agents by default. In order to promote transparency, the WIH is using the following user agent and description: `http.agent.name "Web Intelligence Hub"`, `http.agent.description: "The WIH is run by Eurostat, the official statistics office of`

¹⁰ <https://os4os.pages.code.europa.eu/pbbp>

¹¹ [https://unece.org/sites/default/files/2023-](https://unece.org/sites/default/files/2023-11/HLG2023%20ProjectProposal2024%20Statistical%20Open%20Source%20Software.pdf)

[11/HLG2023%20ProjectProposal2024%20Statistical%20Open%20Source%20Software.pdf](https://unece.org/sites/default/files/2023-11/HLG2023%20ProjectProposal2024%20Statistical%20Open%20Source%20Software.pdf)

¹² <https://i3s-essnet.github.io/Documents/2022/oslo/osos/os4os-oslo-document.html>

¹³ <https://github.com/SNStatComp/awesome-official-statistics-software>

¹⁴ For example, AWS Athena is used in the WIH whereas DuckDB is a cost-effective open source alternative.

the European Union". All the data retrieved by the WIH is strictly used only for statistical purposes, in accordance with Regulation (EC) No 223/2009 of the European Parliament and of the Council on European statistics."

The ESS Web Content Retrieval Guidelines¹⁵ adopted by the DIME-ITDG in June 2022 are part of ethical scraping. These guidelines include "be[ing] open about the tools, methods and processes used for web data retrieval". In order for third parties, for example website owners, to be able to audit by themselves that the activities carried out by the WIH comply effectively with these guidelines, it is important to go beyond their publication and prove that these principles are indeed implemented in practice by opening up the WIH source code.

The statements above should not be taken as an effort to do extra where it is not profitable. Instead, as has been proven in the throughput domain by many generic statistical packages for validation, error localization, imputation and statistical disclosure control, the open source way of working is in the end profitable though multiple organizations. The principles from the OS4OS group underlying open source in official statistics can be taken as guidance. Reducing duplicated efforts through co-investments increases efficiency. Moreover, sharing software as open source software (OSS) contributes to the principle of transparency as contained in the Statistical Code of Practice and is in line with EU and national open source strategies.

4. From experimental to production model

In order to identify a possible statistical production model to be adopted during and after the implementation of the WIH, this chapter explores some key dimensions involved in the interaction between NSIs and the WIH. These dimensions are all interconnected and concern all the elements developing capacity building, from skills, processes and behaviours to a mindset change.

4.1. Key elements of the statistical production model

Based on the experience gained so far, the development of a statistical production model depends on several dimensions affecting the architectural choices, the statistical output, the management of actors and roles. The following figure shows how, starting from BREAL and the associated roles and actors, the statistics to be produced from web data sources impact on the WIH functionalities and requirements. The latter are also highlighted and complemented by some user stories. The main aim of these user stories is to highlight the skills required to implement and interact with the WIH environment from both an individual user and the NSI perspective.

In this context, the relationship between statistical domains and the information provided by web data is a cornerstone to identify and manage the key concepts related to a specific statistical area. WIH use cases enable the analysis of common features and specific issues related to each domain. This avoids the duplication of information from different sources, creates synergies and promotes the reuse of implemented solutions.

¹⁵ https://cros-legacy.ec.europa.eu/content/web-content-retrieval-guidelines-0_en

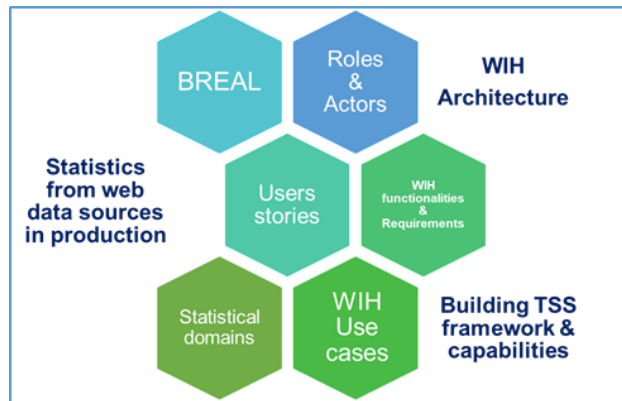


Figure 11: Elements impacting the statistical production model

Another aspect to consider is process scheduling, specifying not only the sequence, but also the duration and artefacts of each task. Standardising the information objects involved in the process also helps to manage the specific features of each domain and use case. The adoption of the BREAL framework also allows to identify the information objects used, modified or created by each task, thus facilitating the transition to the production environment and supporting the description and monitoring of the entire process.

Starting from BBFs, modelling the main tasks to execute also enables the inventory and the set-up of the core and auxiliary datasets to be processed in the central infrastructure, or locally. This inventory is essential for the tracking of data transformation throughout the process, and ensuring that each element is in place for the production phase. As an example, the two tables below provide a list of information objects associated to the tasks described in section 3.2 and based on the OJA use case. More in detail, the first table groups the BBFs mainly related to the web data source assessment, ingestion and editing.

BREAL Business Functions	Production stage tasks	Information Objects
Specify Needs (GSBPM)	Exploration of the potentials of a web data source for a specific statistical domain	• Specification of information needs satisfied by web data sources
		• Reference statistical sources
		• Population frame
		• Lists of reference/target units
		• Mapping between statistical concepts and web information
Acquisition and Recording	URLs update and monitoring	• Lists of reference/target units
		• List of not accessible URLs
		• Validation of URLs list
		• List of key words for web searching
	Crawling/scraping	• Collected data
		• Scraping and crawling errors
		• Scraper/crawling parameters
	Data storage and assessment	• Report on connection errors
		• Website structure changes
Data Wrangling	Data preparation, deduplication, standardization	• Transform HTML content in structured format
		• Count of empty HTML pages,
		• Records concerning duplicated units,
		• Deduplication rules
Data Representation	Derive statistical variables	• Structured data description
		• Ontologies for NLP techniques
		• Auxiliary information
Review and Validate (GSBPM)	Statistical population monitoring	• Model parameters
		• Population frame
	Errors and discrepancies check	• Auxiliary data sources for coverage assessment
		• Validation rules
		• Validation report
		• Quality indicators of model accuracy and degradation over time

Table 2: Subset of BBFS, tasks and related information objects

The second table displays the main tasks and information objects related to the subset of BBFs, mainly concerning quality assessment, modelling of statistical output, dissemination and enrichment of surveys and registers.

BREAL Business Functions	Production stage tasks	Information Objects
Modelling and Interpretation	Produce and assess statistical output	<ul style="list-style-type: none"> • Confusion matrix • Other assessment metrics • ML training set • ML validation set • ML test data
Shape Output		<ul style="list-style-type: none"> • Emerging data patterns
Visual Analyses	Analyse and validate data distributions	<ul style="list-style-type: none"> • Aggregated data tables
Quality management (GSBPM)	Produce reference, structural and process metadata	<ul style="list-style-type: none"> • Quality indicators • Reference, structural, process metadata • Official classifications • Code lists
	Benchmark with external sources, accuracy assessment through statistical analysis and/or manual revision of a sample of records	<ul style="list-style-type: none"> • Ground truth samples
Disseminate (GSBPM)	Produce reference, structural and process metadata	<ul style="list-style-type: none"> • Explanation of disseminated results
	Produce & Disseminate statistical results	<ul style="list-style-type: none"> • Experimental statistics and indicators
Integrate Survey & Register Data	Integration with survey data and BR	<ul style="list-style-type: none"> • National statistical data sources
Enrich Statistical Register	BR enrichment	<ul style="list-style-type: none"> • National Business Register

Table 3: Subset of BBFS, tasks and related information objects

4.2. Web Intelligence Hub user stories

The following user stories were conceived to provide examples of roles and actors accessing the WIH, focusing mainly on one of the most advanced use cases, the OJA workflow. The aim was to analyse users' needs and perspective, starting from BREAL roles and actors, to improve their experience of accessing the WIH. Although the focus is mainly on the staff of the NSIs, improving the user experience increases the benefits of a common infrastructure also for potential external stakeholders. The latter, although not directly involved in the WIH, can access the platform and analyse the statistical output for research purposes. In addition, the final user story describes WP3's approach to sharing and standardising the use of a web data collection service.

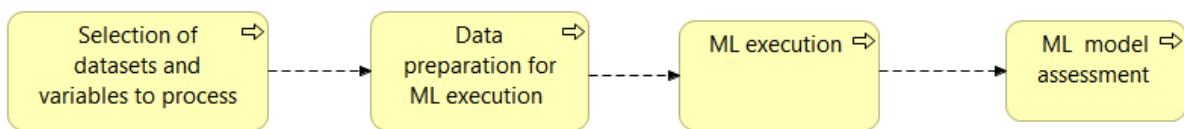
Using big data capabilities

Description: A user, having a training dataset, accesses the WIH platform to train a ML algorithm using data available in the platform. The cross-validation step then makes it possible to assess the accuracy of the model.

Actors: Data scientist

Key pre-conditions and assumptions: Training dataset, ML algorithm

Tasks to execute



Data and metadata management:

- Data structure description to select the datasets and variables of interest
- Tracking of the main process steps for process auditability and reproducibility
- ML quality indicators.

Harmonizing traditional and big data sources

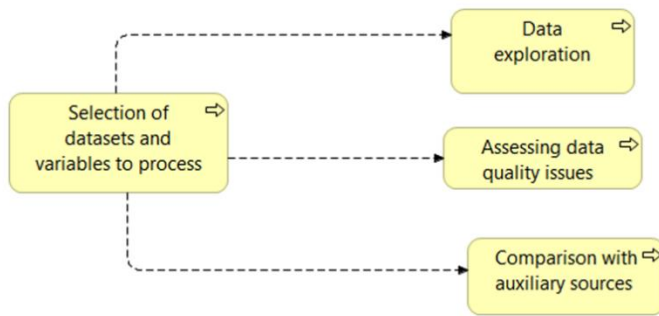
Description: A user accesses the WIH platform to run statistical methods for analysing web data to:

- Enrich the information collected through traditional survey modes and reduce the respondents' burden
- Test different methods to combine survey and web data sources
- Provide an assessment of web data sources in terms of representativeness of the statistical population
- Highlight coverage issues affecting specific subsets of units

Actors: Methodologist

Key pre-conditions and assumptions: Auxiliary data sources

Tasks to execute



Data and metadata management:

- Description of data structures to select the datasets and variables of interest
- Tracking of the main process steps for process assessment and reproducibility
- Indicators for assessing the output of applied methods.

Analyzing statistical output from web data

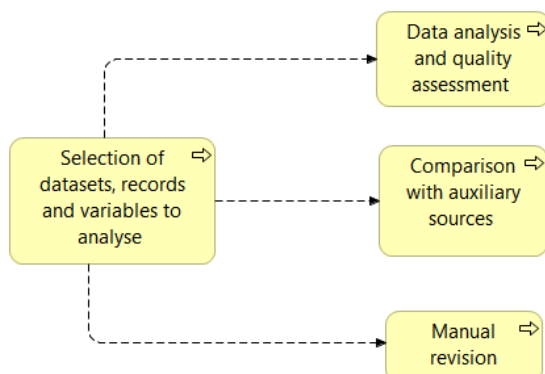
Description: a user, involved in the statistical production accesses the WIH platform to contribute to the data validation process through:

- A benchmark of aggregated statistical output extracted from web data with auxiliary data sources and official statistics
- Assessment of data accuracy in terms of coherence and comparability
- Manual revision of statistical output to validate and improve the WIH data workflow

Actors: Domain specialist

Key pre-conditions and assumptions: Selection of subset of units for manual revision, auxiliary data sources, data validation methods

Tasks to execute



Data and metadata management:

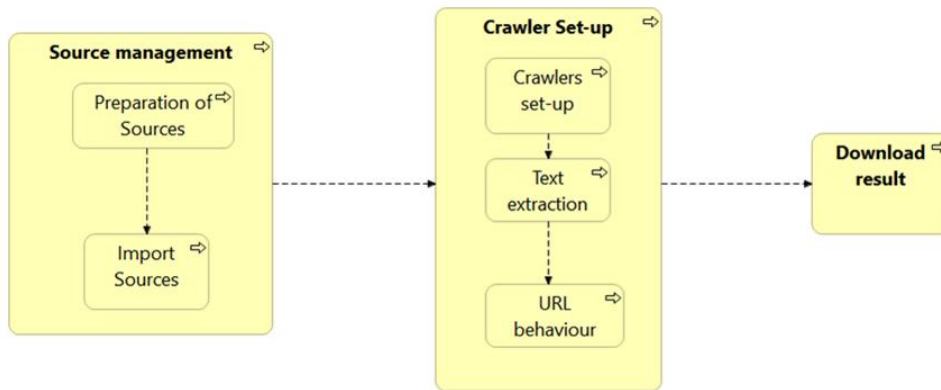
- Description of data structures to select the datasets and variables of interest
- Tracking of the main process steps for process auditability and reproducibility
- Indicators for assessing the output of applied methods.

Web crawling template

Description: In order to facilitate the use of the web data collection service, the NSIs involved in the use of web data for real estate and construction statistics have prepared a template. The main goal of this model ('URL to WIH Data acquisition Platform') is to document and share the main steps to execute web data crawling. The figure below provides a general overview of the service execution.

Actors: NSIs' staff involved in UC1 and UC2 use cases

Tasks to execute



The adoption of this template for URLs landscaping in UC1 (Characteristics of the real estate market) and UC2 (Construction activities) has proved that modelling the main steps of a service helps to:

- Test, share, reuse and standardize code scripts and configuration parameters
- Highlight technical issues
- Realize an iterative improvement of the process steps.

4.3. Combining actors, roles, responsibilities

In defining a statistical production model, the high-level description of BREAL actors and roles should be applied to the actual tasks and organisational issues related to the implementation of the WIH. For each use case, in addition to the specific skills of the NSI staff accessing the WIH, the project experience has highlighted the need for management skills to facilitate the process changes. This is not only due to the specific nature of web data sources, but also to moving from a fully centralised process to a set of tasks performed in a shared infrastructure. From an organisational perspective, in terms of WIH functionalities and use cases, the specification of “*Who can do What*” and “*Who is the owner of each task*” is essential to manage several actors and roles. For each WIH use case, the analysis of the steps that can be centralised and the tasks to be performed in the NSI’s infrastructure is essential to realise the integration between the NSI’s production system and the WIH, as well as to establish accountability criteria.

One way to simplify the project management issues is to adopt a RACI matrix to identify the assignments and responsibilities associated to each actor that has access to the WIH, including external stakeholders, for example web data providers. To illustrate this, the table below provides an example of facilitating cooperation between actors, taking into account the main tasks analysed in section 3.2 for a general mature use case in the production phase. This analysis concerns not only NSI’s staff, but also the actors dealing with non-functional requirements, related for example to security and privacy issues.

In the following example, the actor responsible for the overall management of the WIH platform, including the technological infrastructure and maintenance, is assumed to be the WIH Owner (WO). As one of the main goals of the WIN project is to involve also external organizations (e.g. Academia, public institutions other than NSIs) in the production process, they are taken into account in the example below. Another general assumption is that there are the same internal roles for each type of actor involved in the process, namely:

- Data provider, for use cases in which web data are collected and pre-processed outside the WIH, either by the NSIs or by other organisations
- Domain expert, to involve specialists in identifying, assessing and monitoring web data sources that meet statistical needs and the statistical output produced
- Methodologist and Data scientist, to apply, share and harmonise sound methodologies, tools and best practices for using web data sources for statistical purposes
- IT staff, providing support in technical and interoperability issues concerning the interaction between the WIH and the in-house environment.

In the RACI matrix, each actor can have at least one of the following roles:

- **Responsible**: actors involved in the work implementation, or in the decision making process. Actors belonging to different organizations can jointly cover this role
- **Accountable**: actor who is the “owner” of the work, responsible for assigning and authorising tasks, and monitoring the activities performed. Only one person/organisation can be responsible for each task and/or deliverable.
- **Consulted**: actors providing input, usually at the beginning of the process, thus actively participating to the inception and the evaluation of the activities
- **Informed**: actors receiving updates and decisions, although not directly involved in the process.

BREAL Business Functions	Production stage tasks	WO	NSI		External stake-holder	Final user
			Management	Technical staff		
Specify Needs (GSBPM)	Exploration of the potentials of a web data source for a specific statistical domain	A	R		C	I
Acquisition and Recording	URLs update and monitoring	A	R		I	I
	Crawling/scraping	A	R		I	I
	Data storage and assessment	A	R		I	I
Data Wrangling	Data preparation, deduplication, standardization	A	R		I	I
Data Representation	Derive statistical variables	C	A	R	C	I
Review and Validate (GSBPM)	Statistical population monitoring	R	A	R	I	I
	Errors and discrepancies check	R	A	R	I	I
Modelling and Interpretation	Produce and assess statistical output	C	A	R	I	I
Shape Output						
Visual Analyses	Analyse and validate data distributions	C	A	R	I	I
Quality management (GSBPM)	Produce reference, structural and process metadata	C	A	R	I	I
	Benchmark with external sources, accuracy assessment through statistical analysis and/or manual revision of a sample of records	C	A	R	I	I




BREAL Business Functions	Production stage tasks	WO	NSI		External stake-holder	Final user
			Management	Technical staff		
	Produce reference, structural and process metadata	R	A	R	I	I
	Produce & Disseminate statistical results	R	A	R	I	I
	Integration with survey data and BR	C	A	R	I	I
	BR enrichment	C	A	R	I	I

Table 4: Example of RACI matrix for a use case in production

4.4. Interaction between the Web Intelligence Hub and the National Statistical Institutes

Adopting a ‘learning by doing approach’, the main objectives of the following user stories are to:

- Imagine the potential uses of the WIH capabilities in the near future
- Provide an overview of the main challenges, issues and solutions related to the use of the WIH services
- Highlight how the centralised infrastructure of the WIH communicates and interacts with the infrastructure and components of the NSIs from the NSI perspective
- Summarise and put the lessons learnt into practice, drawing practical examples from the WIN experience.

To this aim, the main tasks of the production use case modelled in Section 3.2 and the insights gained from the project experience are taken into account. For simplicity, these tasks are expressed in terms of BBFs, focusing mainly on the Development, Production and Deployment subset to describe the business layer, and emphasise the interaction between the WIH and the NSIs. In addition, the following user stories concern different domains and are designed to demonstrate that, despite the maturity of its solutions, the WIH is also an inventory of use cases, as experimental statistics from web data sources continue to increase.

Experimental statistics from web data when the WIH is in place

Description: A preliminary analysis of statistical needs has identified a number of international data portals providing information on tourism accommodations supply. Based on the data centrally collected on the web, a set of experimental indicators is produced through the WIH, resulting from the cooperation between the WIH (WIH Owner, WO) and the Member States. In this case, the whole process is performed, either using the WIH services, or running local services in the national infrastructure.

Actors: WO, NSIs

Key pre-conditions and assumptions:

- List of international tourism portals resulting from the collaboration of domain experts, data scientists and methodologists
- Tested models for information extraction and classification
- Involvement of external stakeholders for the assessment of the achieved results
- Compliance with regulations concerning webscraping/crawling and websites terms of use.

Tasks executed in the WIH

The following tasks are executed centrally by the WO, under the monitoring of each NSI, in summary: Acquisition of web data, Data storage, Data deduplication and pre-processing, Extraction of statistical content, Data modelling, Data validation, Computation of quality indicators. These tasks correspond to the BBFs grouped in the violet box of the figure below.

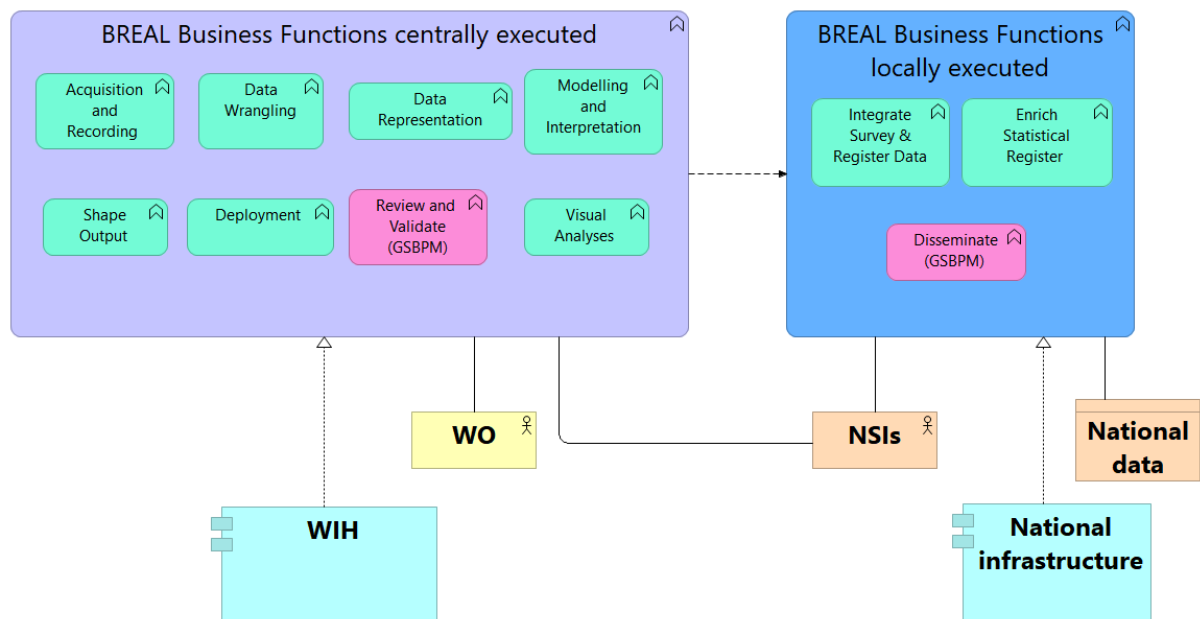


Figure 12: First user story - BREAL Business functions

NSIs are responsible for quality assessment and cooperate with each other to realize each of the BBFs in the blue box above. They agree on the methods to apply, the rules for data validation, as well as the statistical output to produce. NSIs can raise an issue if there are discrepancies with the expected results. In this case, any deviation will be analysed with the input of both the WO and the NSIs, who are actively involved in the decision-making process.

The tasks carried out by the NSIs in the WIH, which focus on the analysis of national data, are mainly intended to assess the accuracy of the centrally managed and executed steps:

- Monitoring and evaluation of each task
- Assessment of the relevance and stability of web portals over time
- Comparison of raw and processed data at micro level
- Analysis of variable distributions for accuracy assessment at macro level.

Tasks executed in-house

NSIs can download national data from the WIH repository at any stage of the process, to explore and/or evaluate the data in more detail. The following activities, relating to national statistical sources, are carried out at local level:

- Benchmark of statistical output with external or auxiliary national data sources
- Integration with and/or enrichment of national survey data and statistical registers
- Dissemination of national statistical results.

Data and metadata management:

Tasks executed in the WIH

- Upload of the list of tourism portals
- Description of unstructured data tables
- Description of structured data tables
- Mapping with official classifications and code lists management
- Applying methods for textual data processing
- Storing training and test datasets for applying ML algorithms
- Creation of metadata for quality reporting.

Tasks executed in-house

Local management of:

- Data structure of national statistical sources used as benchmark
- Official classifications, Code lists
- Experimental statistics to disseminate
- Explanation of disseminated statistics.

Mapping with the WIH use cases

The user story described above is inspired by the current situation of the OJA use case where many of the tasks are carried out centrally. Although most of the tasks are executed by third parties, the active participation of NSIs in each step and their involvement in decision making improves the monitoring and the accuracy of the whole process. In terms of RACI matrix, NSIs are responsible not only for the process implementation, coordinated by the WO, but also for the monitoring and management of national data throughout the process.

Collection of web data using the WIH

Description: The information published on the web can be used to measure the impact of local policies aimed at increasing the number of residents in specific areas inhabited only in the summer. For this purpose, an NSI prepares a list of keywords and uses a service offered by the WIH, to identify a set of websites providing information about the companies operating in the area under study and to collect data.

Actors: NSIs

Key pre-conditions and assumptions:

- List of keywords to use for the URLs search.
- Tested models for information extraction and classification
- Compliance with regulations concerning webscraping/crawling and websites terms of use.

Tasks executed in the WIH by the NSIs

Using the WIH infrastructure, the NSI performs URLs finding, sets the required parameters and performs the data crawling. The data collected from the web is stored in the WIH repository.

Tasks executed in-house

The NSI downloads the collected data in the local infrastructure and executes all the steps to transform the scraped content in statistical output. The figure below shows the combination of tasks performed by the NSI, both in the WIH and in the local infrastructure.

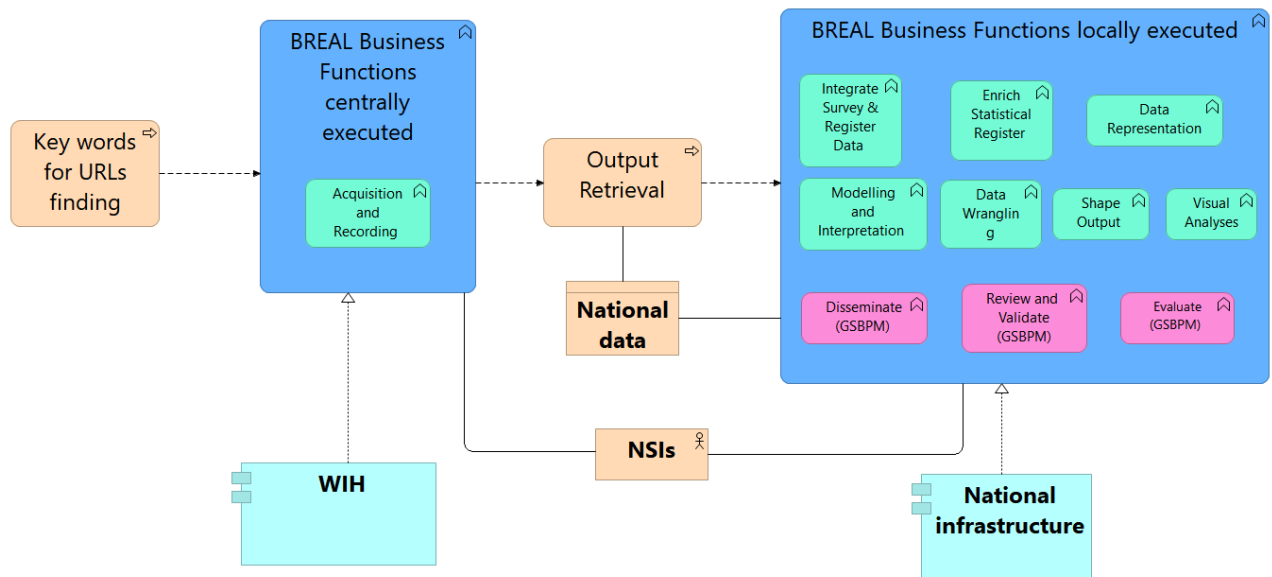


Figure 13: Second user story - BREAL Business functions

Data and metadata management:

Tasks executed in the WIH by the NSIs

- Upload of the list of keywords for URLs finding
- Description of unstructured data tables

Tasks executed in-house

- Description of structured data tables
- Applying methods for textual data processing
- Training and test datasets for applying ML algorithms
- Creation of metadata for quality reporting
- Linkage with national statistical sources used as benchmark
- Mapping with official classifications and code lists management
- Dissemination of the statistical output
- Explanation of disseminated statistics.

Mapping with the WIH use cases

This user story is based on OBEC and WP3 use cases, in which NSIs are responsible for the list of URLs and/or for URLs finding. In order to collect data, NSIs use a service offered by the WIH and executed in the common infrastructure by NSIs. At this stage, the output is downloaded from the WIH and processed in-house. The information objects involved in this user story concern mainly the national level.

Using the WIH services for web data processing

Description: An NSI aims to enrich business statistics by measuring the degree of social responsibility in terms of gender equality. For this purpose, the NSI performs locally web data scraping and processing. Instead of starting from scratch to calculate some indicators, the NSI adapts a service developed and shared by the WIH community to enrich the BR statistics. As a common infrastructure, the WIH also promotes the reuse and sharing of methods and tools developed for collecting and processing web data.

Actors: NSIs

Key pre-conditions and assumptions:

- Sharing of code scripts within the WIH community
- Inventory of available WIH services with instructions on how to access them.

Tasks executed in the WIH

The NSI accesses the WIH to consult the inventory of the implemented solutions shared by the WIH community and gets the link to a code script for the calculation of a set of indicators for the enrichment of the BR.

Tasks executed in-house

The NSI performs locally the collection and processing of national web data. These tasks are expressed in terms of BBFs and modelled in the figure below. To harmonise methods and tools beyond cross-border and domain peculiarities, the NSI aligns internal activities with the best practices emerging from the WIH experience. This approach promotes the sharing of tools and methods, both within the NSI and the WIH community.

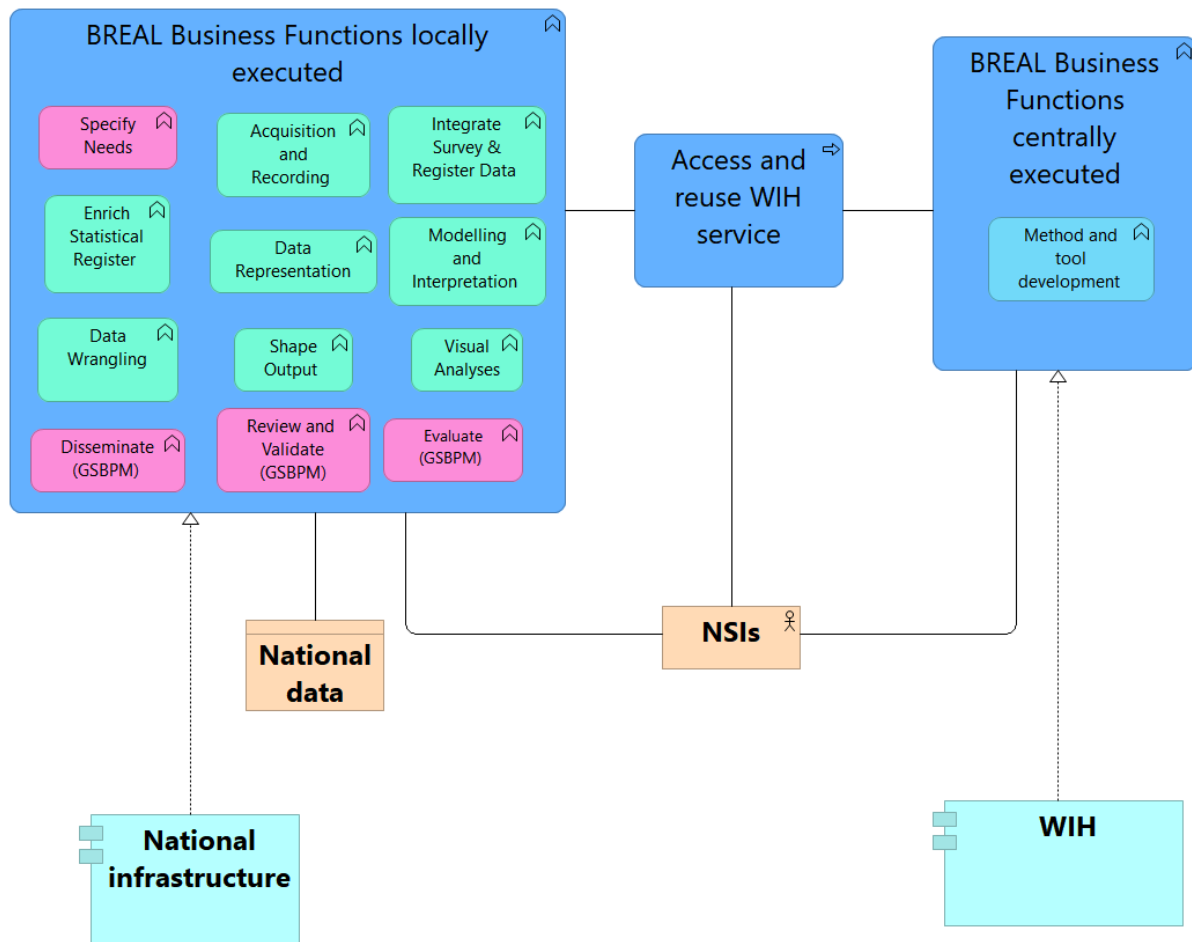


Figure 14: Third user story - BREAL Business functions

Data and metadata management:

Tasks executed in the WIH by the NSI

Inventory and explanation of methods and tools developed and shared by the WIH community.

Tasks executed in-house

- Creation of the list of keywords for URLs finding
- Description of unstructured data tables
- Description of structured data tables
- Applying methods for textual data processing
- Training and test datasets for applying ML algorithms
- Creation of metadata for quality reporting
- Linkage with national statistical sources used as benchmark
- Mapping with official classifications and code lists management

- Dissemination of the statistical output
- Explanation of disseminated statistics.

Mapping with the WIH use cases

The objective of this user story is to provide an example of sharing and reuse of tools. Integrating and enriching survey data and statistical registers is a common goal for several WIH use cases, enabling the creation of a common infrastructure and providing adaptable and reusable code scripts. As an example, in the OJA use case, Italy and Bulgaria have shared and compared the code implemented in R and Python for computing a set of statistical indicators. This approach lays the foundations for harmonising methods and tools in the long term.

5. Final results

One of the main goals of the architectural task is the ‘Enhancement and Extension (E&E) of BREAL’. The enhancement of the BREAL business layer is achieved by specialising the main BBFs, based on the activities performed within each WIH use case. The E&E of the application and information layers can only be realised at a very high level, as these layers are closely linked to the implementation tools and the specific requirements of each use case. For this reason, the E&E of BREAL has mainly concerned the BBFs implemented through the WIH development activities.

More in detail, in relation to the BREAL extension, the analysis of the architectural challenges has highlighted the completeness of BREAL in terms of “Development, Production and Deployment” BBFs, and the need to integrate the subset of BBFs related to the supporting activities and the management of roles and actors. These conclusions are drawn from the project experience and relate mainly to operational issues and organisational challenges. On the basis of these premises, the following sections describe the enrichment of the BREAL framework with regard to the following aspects:

- Specialisation of some BBFs belonging to the “Development, Production and Deployment” subset
- Addition of a new BBF in the “Support” subset, to foster the transition of a use case from the experimental to the production phase.

5.1. Big Data REference Architecture enhancement for web data

The activities for the development of the WIH have highlighted the relevance of BBFs for the inception, implementation and monitoring of a statistical process from web data sources. However, the BREAL version released at the end of the Essnet Big data II was conceived to provide a general overview of the main abilities and tasks for all types of Big data, grouping several types of data sources. The insights resulting from the use cases implementation, regardless their maturity level, have underlined specific issues and challenges, thus suggesting the specialisation of BREAL for web data.

In order to integrate the top-down and the bottom-up approaches, BREAL enhancement has concerned the BBFs subset “Development, Production and Deployment”. The analysis has excluded the BBFs concerning output shaping, data visualisation and dissemination, data revision and validation which are described in detail in BREAL and are applied to web data sources as such.

The table below reports the list of BBFs, directly related to data collection and processing, specialised for web data on the basis of the evidence provided by the project experience. The BBFs listed in the

table are ordered according to the sequence of a hypothetical workflow. The enhancement is not intended to replace the original description, but to enrich it.

BREAL Business Functions	
Original Description	Enhancement for web scraped data based on the project experience
Specify needs	
<p><i>Big Data specific</i></p> <p>When using Big Data sources the needs are derived in an iterative manner. At the start, the scope of the need can be very broad. During the exploration of the source (see business function “New data sources exploration (EARF)”) the need becomes more detailed based on the possibilities of the source</p>	<p>Identify web data sources meeting statistical needs, involving researchers, domain experts and other stakeholders supporting national and European statistical systems</p>
New data sources exploration	
<p><i>Big Data specific</i></p> <p>Besides the exploration of the new data sources the ability to find Big Data sources and to make these sources available for statistical research and development becomes important. The latter is part of the business function “Manage data suppliers”</p>	<ul style="list-style-type: none"> • Define and share a set of criteria to assess and select potential web data sources based on a ranking approach • Explore web data sources to compare key statistical concepts and the information provided by web data for a specific domain of interest • Adopt an iterative approach to identify and compare web data portals and variables, population frames, lists of reference/target units
Acquisition and recording	
<p>The ability to collect data from a given Big Data source, e.g. through API access, web scraping, etc. In addition, this function includes the ability to store and make data accessible within the NSI</p>	<p>The ability to: identify and list relevant URLs; collect and store data from the web e.g. through API access, web scraping or crawling.</p> <ul style="list-style-type: none"> • After an initial phase of URL selection and landscaping, also through a list of keywords, monitoring of stability and relevance of URLs over time, as well as URLs accessibility issues. • Identifying and defining the reference/target units to enable the creation of population frames. Early validation of scraped data to prevent storing inconsistent information
Data Wrangling	
<p>The ability to transform data from the original source format into a desired target format, which is better suited for further analysis and processing. Data Wrangling consists of Extraction (retrieving the data), Cleaning (detecting and correcting errors in the data) and Annotation (enriching with metadata). It can be mapped to the GSBPM steps 5.1. Integrate data, 5.2. Classify and code, and 5.4. Edit and impute</p>	<p>The ability to transform web content into a target format and extract the relevant information from the website. This ability also involves:</p> <ul style="list-style-type: none"> • The performance of a first round of data cleaning to drop empty and duplicated records • The integration of the derived features with statistical sources at macro or micro level, whether web reference units correspond to statistical units

BREAL Business Functions	
Original Description	Enhancement for web scraped data based on the project experience
<i>Data Representation</i>	
The ability to derive structure from unstructured data (e.g. from text) or partially structured data (e.g. data in CSV files or XML files). This includes data modelling, i.e. establishing a data structure to represent the data	The ability to derive structure from unstructured scraped information, using explorative techniques e.g. text mining techniques and lists of reference keywords (annotations)
<i>Modelling and Interpretation</i>	
The ability to design, develop and test new algorithms and models to process Big Data sources. This includes approaches like machine learning and predictive modelling (model to predict outcome)	The ability to design, develop and test algorithms and models to process and transform web texts and derive statistical variables. In the application of ML techniques, the adoption of measures to prevent and monitor concept and data drift is particularly relevant
<i>Enrich Statistical Register</i>	
The ability to enrich the statistical register(s) with the information retrieved from the Big Data source	Mapping of the variables derived from the web source to the information needed to enrich the statistical registers
<i>Integrate Survey & Register Data</i>	
The ability to reuse and integrate other data like survey and register data in order to enrich the results derived so far	<p>In relation to a specific domain, mapping key statistical concepts and units with web data concepts and units enables:</p> <ul style="list-style-type: none"> • The integration at micro/macro level between data from traditional sources and web data • Quality assessment of the statistical output derived from web data, using survey and registers data as benchmark • Reuse and enrichment of the information provided by data from traditional sources and web data
<i>Support Statistical Production</i>	
The ability to support the statistical production system(s) already in place	The ability to assess the benefits of the web data in terms of saved resources, reduced respondent burden, timely analysis of emerging phenomena in relation to a specific domain

Table 5: Specialisation of BBFs for web data

The following table shows the specialisation of the main overarching BBFs, concerning the auxiliary information required to assess, monitor and improve the tasks and the information objects related to the BBFs analysed above.

BREAL Business Functions	
Original Description	Extension and Enhancement for Web scraped data
Metadata Management (GSBPM)	
For the original description, see GSBPM, paragraph 118 through 1211: Good metadata management is essential for the efficient operation of statistical business processes. Metadata are present in every phase, either created or carried forward from a previous phase. The emphasis is on the creation, use and archiving of statistical metadata. The key challenge is to ensure that these metadata are captured as early as possible, and stored and transferred from phase to phase alongside the data they refer to	In order to ensure transparency, in addition to reference and structural metadata, a set of process metadata is essential to document web data workflow, particularly if a process is managed by several actors. As an example: number and characteristics of scraped/crawled websites, description of input/output information objects, methods applied and actors involved in the main tasks
Quality Management (GSBPM)	
For the original description, see GSBPM, paragraph 106 through 114: The main goal of quality management within the statistical business process is to understand and manage the quality of the statistical products. In order to improve the product quality, quality management should be present throughout the statistical business process model. All evaluations result in feedback, which should be used to improve the relevant process, phase or sub-process, creating a quality loop	Considering the specific features of web data sources, a set of auxiliary information is required to assess, monitor and document the quality of web data, the process and the results. In particular, a set of quality indicators is required to evaluate a web data source: <ul style="list-style-type: none"> • Websites selection criteria and accessibility • Accuracy of collected information and applied methods • Representativeness with respect to population frames • Comparability with other statistical sources • Timeliness and relevance of the final results
Evaluate (GSBPM)	
<i>Big Data specific</i> When Big Data sources are used, evaluation plays an important role. Most of the specificities of Big Data are related to its quick pace of change, both in terms of the population covered and of their behaviour. Thus issues like coverage, accuracy and fitness of the model must be constantly assessed and monitored	Starting from the auxiliary information related to the previous BBFs (Metadata Management and Quality Management), the ability to assess and monitor the main issues (methodological, technical, operational, organizational) affecting the workflow developed for a specific use case. As an example: <ul style="list-style-type: none"> • Unexpected websites changes • Agreements with websites owners to ensure data accessibility over time • Models decay due to data or concept drifts • Quality improvements through manual revisions of a sample of units • Staff trainings
Trust Management	
The ability to gain reliability. Trust to use Big Data sources in a secure and rightful manner is needed to be able to gain access to these sources. Trust	The ability to derive consistent statistical output from web data sources, relying on:

BREAL Business Functions	
Original Description	Extension and Enhancement for Web scraped data
to be able to derive the same or even higher quality of data using Big Data sources in comparison to the more traditional way of making statistics is needed to create new or to replace existing statistical products	<ul style="list-style-type: none"> • Transparency of the workflows developed for specific use cases achieved through process metadata documenting the tasks performed • Harmonization of methods and tools enabling the enhancement and reuse of developed solutions • Cooperation between the different actors involved in the process • Active involvement of the research community and of the final users • Continuous improvements according to the Plan-Do-Check-Act cycle
Deployment	
The ability to take the (new) statistical product using Big Data sources and process it into production. This is to ensure that the statistical product is created and supported for a longer period of time	The ability to deploy statistical output in the production chain, depending on the status of each use case. From the exploratory phase of a use case, test each stage of the workflow in the production environment to identify issues that prevent it from being improved and maintained, taking into account national environments

Table 6: Specialisation of overarching BBFs for web data

In relation to the BBFs belonging to the ‘Support’ subset, the experience gained during the development activities showed discrepancies due to the different legislative and organisational contexts within the ESS. Although the BBFs conceived to harmonize the legislative framework and achieve common European policies, additional efforts are needed to foster:

- The endorsement of official statistics from web data sources at national level
- Acknowledging web scraping/crawling as a method of collecting data for statistical purposes
- Maintenance of common infrastructures for the centralized execution of specific tasks
- Relationship between NSIs and website owners.

The specialisation of BBFs results from the challenges reported in the WP3 use cases and also observed in the OBEC and OJA use cases, summarised as follows:

- Partnership with web data providers
- URLs finding through APIs
- Assessment and monitoring of the source stability
- Deduplication between portals
- Choices of technologies and set up of the working environment, software development and its testing
- Definition of common guidelines to develop and test reusable and sharable software (tools).

5.2. Big Data REference Architecture extension for web data

In relation to the BREAL extension, to overcome the initial challenge and balancing the top-down and the bottom-up approaches, the project experience has underlined the need for an additional overarching BBF, to accelerate and monitor the transition of each use case from the experimental to the production phase.

In order to promote process management and orchestration, the addition of a new BBF to BREAL in the ‘Support’ subset allows to deal with several implementation issues, most of which are unexpected and closely related to each other. These issues concern the integration of a use case into production and can affect technical or organisational aspects, planned activities and associated outputs. Furthermore, depending on the specific use case, coordination becomes essential to manage: i) the interaction of several actors belonging to different organisations; ii) the combination of tasks managed in a common infrastructure and/or in a local environment; iii) the constraints imposed by national policies.

The BBF “Strategy and Process management”, for the alignment and coordination of the dimensions showed in the figure below, enables the definition of a statistical production model and accelerates the transition in production of the WIH use cases.

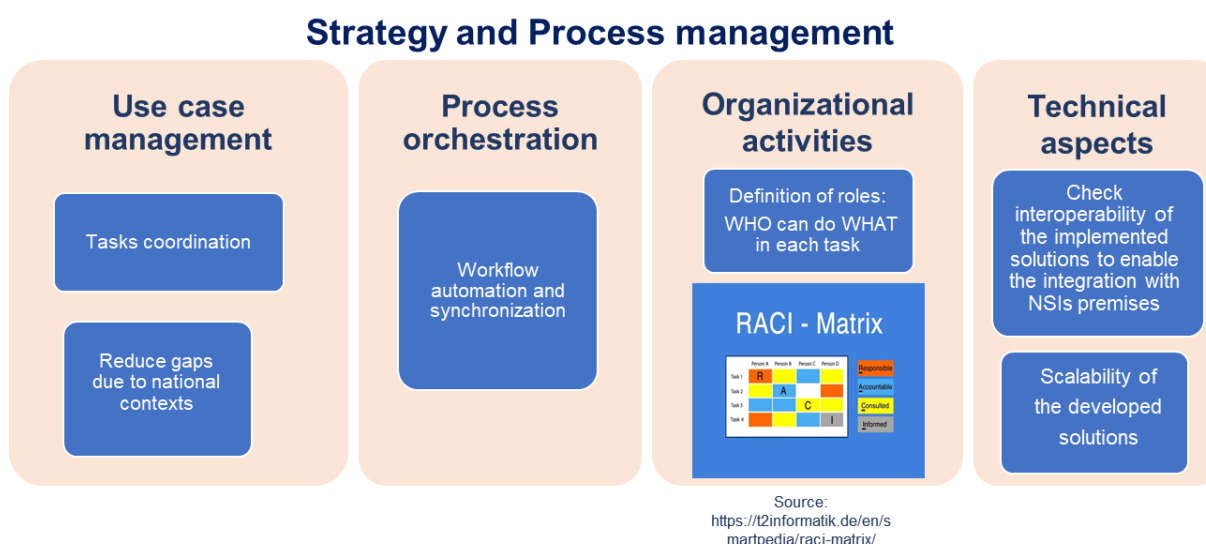


Figure 15: Elements of the new overarching Business Function

As a point of confluence between operational and organizational tasks, this new BBF focuses on:

- Checking the execution of each task, connecting the implemented solutions with the related tasks and input/output
- Enabling the workflow execution by different users and coordinating the management of centralized and in-house tasks
- Management of the alignment between the common infrastructure and the national environment
- Monitoring and assessing the scalability of the shared solutions in the production environment
- Preventing delays due to national gaps and/or unpredictable issues during the design of a use case
- Identifying country-specific issues to promote methods and tools harmonization.

5.3. Conclusions and lessons learnt

The WIN experience has demonstrated several advantages and challenges deriving from the implementation of a common infrastructure to produce trust statistics based on web data. The effort to reuse, share and harmonise the solutions developed can appear challenging at first, but in the long run this strategy results in a win-win approach, enabling cost reductions and efficiency gains.

The WIH is a capability under development, for the collection and processing of web data, serving several use cases as:

- Service provider for acquisition of web data
- Web data repository
- Environment for processing web data.

Due to the wide range of functions offered by the WIH, there is no "one size fits all" production model. Depending on the combination of tasks, actors and execution environment (local or shared infrastructure), each use case falls into a specific production model which can be standardised.

Starting from the project experience, the main goal of the architectural task was the Enhancement and Enrichment (E&E) of the BREAL framework. The analysis of the WIH use cases led to the revision of the BBFs (BREAL Business Functions) as follows:

- Specialisation of some BBFs belonging to the "Development, Production and Deployment" subset
- Addition of a new BBF, "Strategy and Process management", in the "Support" subset, to help with the transition of a use case from the experimental to the production phase.

The WIH use cases have also highlighted the relevance of harmonizing methods and tools to promote the reuse and sharability of available solutions, overcoming country-specific issues.

The E&E of the BREAL framework aimed to highlight the interconnection between methods, tools, data transformations and use case management, thus building a common vision of the WIH and a holistic approach for the integration of web data into official statistics at EU level.

References

- Scannapieco M., Bogdanovits F., Gallois F.; Fischer , Kostadin G., Paulussen R., Quaresma S. et al. (2019): (Deliverable F1) BREAL. Big Data REference Architecture and Layers. Business Layer. Version 2019-12-09. Edited by EUROSTAT
- Cedefop (2019). Online job vacancies and skills analysis: a CEDEFOP pan-european approach. https://www.cedefop.europa.eu/files/4172_en.pdf
- Generic Statistical Business Process Model – GSBPM, <https://statswiki.unece.org/display/GSBPM>
- Generic Statistical Information Model – GSIM, <https://statswiki.unece.org/display/gsim>
- ESS Enterprise Architecture Reference Framework – EARF, https://joinup.ec.europa.eu/sites/default/files/document/2018-10/ESS_Enterprise%20Architecture%20Reference%20Framework_Version%201.1_A1_Introductory%20document.docx
- Scannapieco M., Bogdanovits F., Gallois F.; Fischer, Kostadin G., Paulussen R., Quaresma S. et al. (2021): (Deliverable F2) BREAL. Big Data Reference Architecture and Layers. Application layer and Information layer. Version 2021-03-31. Edited by EUROSTAT
- Kowarik A., Daas P. et al. (2021) Deliverable 4.1: Minimal guidelines and recommendations for implementation. Version 2021-07-30. Edited by EUROSTAT
- Reis F. (2022). The Web Intelligence Hub – A tool for integrating web data in Official Statistics. IAOS 2022 conference. Available at: https://www.iaos2022.pl/presentations/?drawer=Session%2008*Reis%20Fernando
- Stateva G., Dabrowski D., Lasslop G. et al. (2022). (Deliverable 3.1) WP3 1st Interim technical report. Final version of first interim technical report for WP3 of the ESSnet WIN. March 2022
- Six, M., Kowarik, A., Gussenbauer, J. (2023). Landscaping of Websites for Webscraping with Focus on Selection Modes. Draft version of the report for WP4 of the ESSnet WIN. Oct. 2023
- Six, M., & Kowarik, A. (2023). Issue 10 - Quality aspects of web scraped data - Focus on landscaping and selection of sources. Web Intelligence Network Blog. <https://cros.ec.europa.eu/book-page/issue-10-quality-aspects-web-scraped-data-focus-landscaping-and-selection-sources>
- Six, M., Kowarik, A., Daas P. et al. (2023) Deliverable 4.6: WP4 Methodology report on using web scraped data. Draft version of the WIN Deliverable 4.6. Nov. 2023
- Six, M., Kowarik, A. et al. (2023). Deliverable 4.5: Quality Guidelines for acquiring and using web scraped data. Draft version of the WIN Deliverable 4.5. Nov. 2023
- Inglese F., Lucarelli A. et al. (2024). Experimental OJA based indicators on labour demand changes: opportunities and challenges. European Conference on Quality in Official Statistics. Available at: <https://airdrive.eventsair.com/eventsairwesteuprod/production-leading-public/84caa32b6ac74d63a1c466c02ef18622>