WEB DATA SOURCES AND OFFICIAL STATISTICAL STANDARDS: THE WIN EXPERIENCE[1]

Giuseppina Ruocco, Renato Magistro, Giulio Massacci

Abstract. In order to promote the integration of web data sources for official statistics, Eurostat has launched a four-year initiative, the Web Intelligence Network project. One of the main objectives of the project is to create a network within the European Statistical System (ESS) and to develop a common infrastructure, the Web Intelligence Hub (WIH), providing services and tools for web data collection and management. The WIH is designed to support National Statistical Institutes (NSIs) in all stages of web data collection and processing. In the long term, the WIH can be further developed to explore the potential of additional innovative data sources for official statistics. During the WIN project, the architectural task has dealt with the Enhancement and Enrichment (E&E) of the architectural standard Big Data REference Architecture and Layers (BREAL). BREAL is an architectural framework developed in the Big Data II project to support NSIs in planning their investments in big data. It provides a set of tools for defining the business objectives, application components and data models needed to develop statistical processes based on big data. One of the main outcomes achieved by the architectural task is the specialization of BREAL for web data. The adoption and E&E of the BREAL framework has enabled the harmonization between the requirements of the project use cases and the services provided by the WIH. BREAL specialization results from the combination of the project experience and it is intended to promote the deployment of web data workflows in the production environment. This enhancement of BREAL enables a NSI to assess the maturity level of a use case. In addition, the adoption of this approach increases process standardization and the development of shareable tools that can be reused by other statistical organizations, and/or adapted to other use cases and/or statistical domains.

1. Introduction

The availability of new data sources resulting from technological developments represents a significant opportunity for official statistics. The new data sources increase the efficiency of statistical processes, reduce the burden on respondents, and improve the quality of the output produced. This improvement is evident in terms of both timeliness and the enrichment of statistical products, resulting in an enhancement of disseminated results.

[1] The individual contributions are as follows: Giuseppina Ruocco: Paragraphs 3,6; Renato Magistro: Abstract, paragraphs 4,5; Giulio Massacci: Paragraphs 1,2.

Although traditional data sources have been thoroughly investigated and standardised, new data sources are characterised by heterogeneity in terms of both their origin and data format. Examples include satellite images and information gathered using sensors in various contexts.

In all these cases, the issue of heterogeneity affects the entire process, not just the input and output phases. Particular focus is given to quality assessment. This assessment should consider potential new sources of error and factors that have not yet been fully investigated by the statistical frameworks adopted for traditional data sources. In order to collect, process and validate these new types of data, the methodology, tools and technological infrastructures must first be designed and tested in relation to this heterogeneity. The main challenge will then be to adapt or create new reference standards.

Starting from the experience of the Web Intelligence Network project, this article provides an overview of the Enhancement and Enrichment (E&E) of the architectural standard Big Data REference Architecture and Layers (BREAL) for web data. This enhancement of BREAL enables a National Statistical Institute (NSI) to assess the maturity level of a use case. In addition, the adoption of this approach increases process standardization and the development of shareable tools that can be reused by other statistical organizations, and/or adapted for other use cases and/or statistical domains.

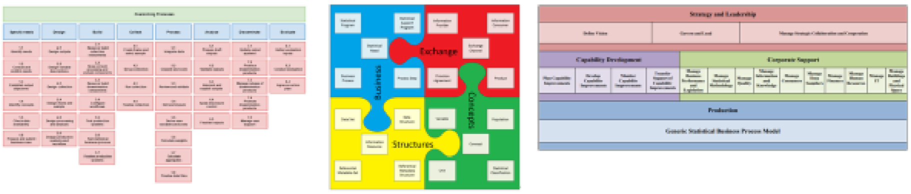
2. Official standards and new data sources

The main official statistical standards reported in Figure 1 have so far played a significant role in improving the efficiency and quality of statistical processes based on traditional data sources. The following standards have improved the efficiency of NSIs with regard to the management of processes, data and metadata. In particular:

- The GSBPM (Generic Statistical Business Process Model) provides a standardised model to describe all phases of the statistical process with harmonised and internationally shared terminology, from the definition of information needs to the dissemination of data, including the final evaluation. It is useful within the NSIs for mapping, analysing and improving statistical processes, but it is not a “rigid” framework in which all steps must be followed in a strict order.
- The GSIM (Generic Statistical Information Model) is a reference framework that describes different types of information (information objects), their attributes, and their relationships. Its purpose is to model the information used to standardise the management and use of data and metadata throughout the entire statistical production process.

- The GAMSO (Generic Activity Model for Statistical Organisations) is a standard that focuses on organisational management, aiming to describe all the activities of a statistical organisation, not just those related to production. Indeed, GAMSO builds upon the GSBPM by incorporating strategic, institutional, and support activities such as resource management, governance, planning, innovation, and external relations.

Figure 1 – *GSBPM, GSIM and GAMSO models.*



GSBPM: Process standardization

GSIM: Data and metadata modelling

GAMSO: Generic Activity Model for Statistical Organizations

Source: Generic Statistical Business Process Model – GSBPM, Generic Statistical Information Model (GSIM) – User guide, Generic Activity Model for Statistical Organisations – GAMSO

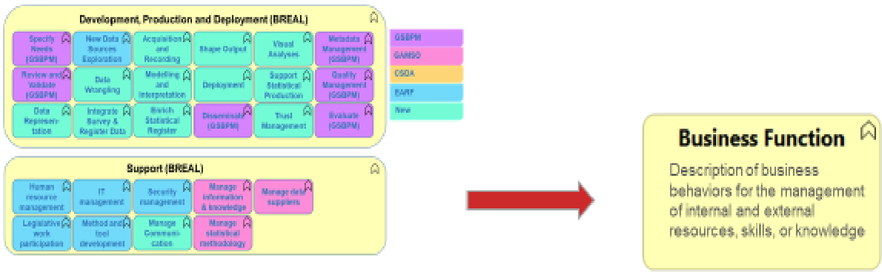
3. BREAL: an architectural framework for big data sources

At the European level, several projects integrating new data sources into statistical processes have been launched in order to address these new challenges and respond to increasing complexity. In this context, one of the main outputs of Big Data II was the BREAL framework. The goal of BREAL is to provide guidance for investments in big data and facilitate the transition from using standards in well-known environments (traditional sources) to adapting and using them in innovative contexts.

Specifically, the BREAL framework consists of several elements, each of which focuses on a particular stage of the data processing lifecycle. Together, these elements provide a roadmap that considers all aspects when dealing with new data sources, known as big data, at the beginning of their use.

During the project, the BREAL framework was applied to several use cases, the most important of which were OBEC (Online Based Enterprise Characteristics) and OJA (Online Job Advertisements). OBEC relates to the use of information on company websites to enrich the information contained in the Business Register, while OJA relates to the use of job advertisements published on job portals to develop new labour demand indicators. The framework has been designed to help statistical organisations deal with the complexities of using big data, while ensuring consistency with existing standards. Among the various components of the BREAL framework, the Business Layer is particularly relevant for modelling and standardising statistical processes based on new sources. It shows the main Business Functions aimed at acquiring and subsequent processing. From an architectural point of view, the business layer in the BREAL framework defines “WHAT” is to be realised in terms of behaviour for managing internal and external resources, competencies or knowledge. As shown in Figure 2, the BREAL business layer includes functions from various official statistics standards, including those relating to the statistical process, such as GSBPM and GAMSO, as well as architectural elements derived from CSDA and EARF, and new Business Functions related to new data sources. The Business Functions are generally divided into two main categories: the first is aimed at implementing and putting big data processes into production (Development, Production and Deployment subset), while the second relates to all the functions that support the statistical process at various levels (Support subset).

Figure 2 – *The BREAL business layer and Business function definition*.


Development, Production and Deployment (BREAL)
Specify Needs (GSBPM)
New Data Sources Exploration
Acquisition and Recording
Shape Output
Visual Analyses
Metadata Management (GSBPM)
Review and Validate (GSBPM)
Data Wrangling
Modelling and Interpretation
Deployment
Support Statistical Production
Quality Management (GSBPM)
Data Representation
Integrate Survey & Register Data
Enrich Statistical Register
Disseminate (GSBPM)
Trust Management
Evaluate (GSBPM)
GSBPM
GAMSO
CSDA
EARF
New
Support (BREAL)
Human resource management
IT management
Security management
Manage information & knowledge
Manage data suppliers
Legislative work participation
Method and tool development
Manage Communication
Manage statistical methodology
Business Function
Description of business behaviors for the management of internal and external resources, skills, or knowledge

Source: For BREAL business layer - Scannapieco M. et al. (2019): BREAL. Big Data Reference Architecture and Layers. Version 2019-12-09. Edited by EUROSTAT

In order to identify the methodological implications of implementing such architecture across the various countries, an inventory of the different use cases should be made to achieve:

- A classification of each kind of implementation by type of source (input)
- A list of the methods adopted for the pre-processing and data processing steps (throughput activities)
- The outputs produced and their integration within the related statistical domains.

Methods and, above all, quality indicators must be associated with each business function for each source type, so that the tools and procedures that implement these methods can be shared among different NSIs.

The WIN project’s experience demonstrates how the BREAL framework achieves a balance between generalisation at the European level and the necessary customisation at the national level with regard to web data.

Although one of the goals of the WIN is to produce a standardised architecture and reusable tools, its effectiveness ultimately hinges on NSIs' ability to adapt it to their institutional, regulatory, and operational contexts.

BREAL defines a set of Business Functions that are universally applicable to NSIs at a 'high' level. The WIN facilitates the sharing of open source tools for data acquisition and pre-processing (e.g. deduplication and structuring), as well as data processing. In addition, it explored some quality indicators to evaluate the web data sources. This approach is consistent with the GSBPM/GSIM/GAMSO models and ensures comparability and interoperability between countries. However, while based on a shared core, each NSI should adapt the framework to its own specific requirements. The main areas of customisation include:

- Legal and ethical compliance. Development of data acquisition strategies to national context with respect to data protection and use legislation (e.g. specific GDPR interpretations, local data provider agreements).
- Integration with national IT infrastructure. It involves, for example, the inclusion of WIN tools within existing IT architectures, focusing on compatibility with cloud infrastructures.
- Statistical domains and national priorities, focusing on use cases that are most relevant to each country's needs (e.g. tourism indicators for some NSIs and labour market indicators for others).

The proposed framework addresses these challenges by offering sufficient flexibility for national adaptations, all the while maintaining a robust shared methodological framework. In this sense, the framework developed by the WIN project should not be understood as a fixed solution, but rather as an adaptable architecture designed to promote both harmonisation and innovation. By combining shared services with opportunities for national customisation, the framework provides a practical approach to integrating web data into official statistics within the ESS.

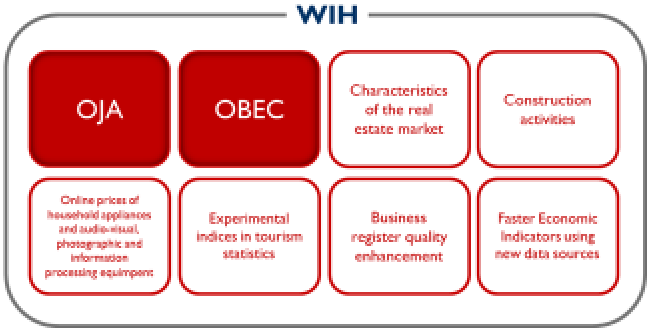
4. The WIN project and the BREAL specialization for web data

The Web Intelligence Network (WIN) project was launched in 2021 with the aim to bring the most mature use cases into production and develop the Web Intelligence Hub (WIH). The WIH is a shared European infrastructure for the efficient management of the acquisition and processing of web data for statistical purposes related to specific statistical domains. The WIH is designed to provide services for supporting NSIs at every stage of web data collection and processing.

During the WIN project, several use cases examined how web data could be utilised in various statistical fields, including the real estate market, tourism, electronics retail, labour demand trends, and business characteristics. Regardless of the statistical domain, the following characteristics were common to all of these use cases:

- The selection of URLs relating to portals and business sites
- Acquisition of information, mainly textual, from the selected URLs
- Pre-processing the acquired data, to select the useful information for statistical purposes and preparing it for further processing, after de-duplication and data structuring.

Figure 3 – *The WIH use cases.*


WIH
OJA
OBEC
Characteristics of the real estate market
Construction activities
Online prices of household appliances and audio-visual, photographic and information processing equipment
Experimental indices in tourism statistics
Business register quality enhancement
Faster Economic Indicators using new data sources

During the project, the architectural task specialized BREAL's business layer, building on experience with more mature use cases, mainly OBEC and OJA, to ease their transition into production environments.

The objective was to standardize the main process steps and develop shareable, reusable tools, with a focus on methodological aspects and quality assessment.

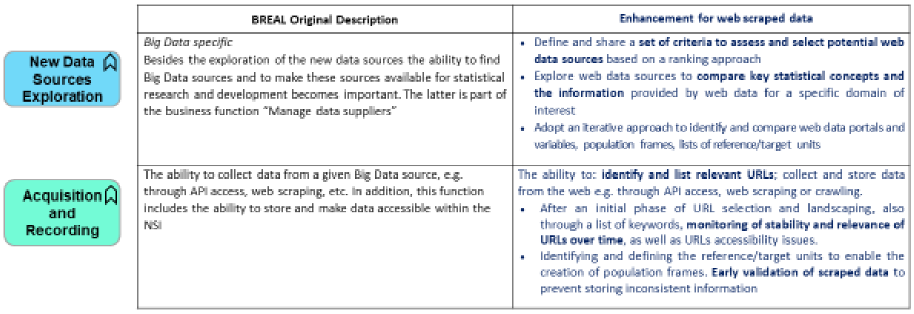
BREAL's specialization mainly concerned the subset 'Development, Production and Deployment'. It is important to note that the approach adopted was bottom-up, based on experience in the field rather than theoretical standardization. This approach not only enabled existing functions to be refined, but also allowed the BREAL model to be extended in a coherent and practical manner while maintaining alignment with other existing standards. In summary, the specialization is an evolution of official standards driven by experience, not a replacement. As an example, Figure 4 shows the results obtained for the main exploration and acquisition phases of new data sources.

More in detail:

- New Data Sources Exploration: This function involves identifying and evaluating new web data sources that could be useful for statistical purposes. A set of criteria has to be defined to evaluate and rank potential sources according to their relevance of content. Web data sources are explored to evaluate the key statistical concepts that can be derived from them and the information they provide on a specific domain of interest.

- Acquisition and Recording: This function concerns the automatic collection of data (via scraping or API) and the consequent transformation into a usable format. Adaptive solutions are required because web data sources are liable to change their structure. In addition, it is particularly important to monitor the URL stability and relevance over time, as well as accessibility issues. It is also necessary to pre-process the scraped data to avoid duplicate information and select the information that is useful for statistical purposes.

Figure 4 – *Examples of enhanced Business Functions from the ‘Development, Production and Deployment’ subset.*



	BREAL Original Description	Enhancement for web scraped data
New Data Sources Exploration	*Big Data specific* 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"	• 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	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	The ability to: **identify and list relevant URLs**; collect and store data from the web e.g. through API access, web scraping or crawling. • 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

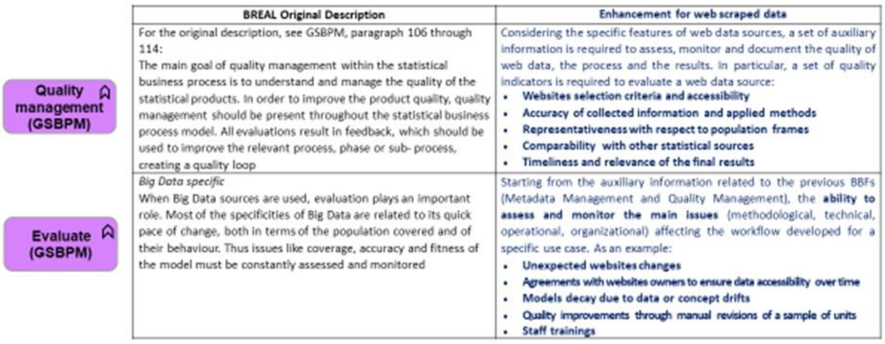
Source: For the descriptions - G. Ruocco et al. (2025): Deliverable D4.7 BREAL - Big Data REference Architecture and Layers for web scraped data. Final version 2025-03-31. Edited by EUROSTAT

As an example, the specialisation of overarching Business Functions relating to the evaluation phase and quality management has emphasised the importance of using additional data sources as benchmarks for evaluating, monitoring and documenting the quality of web data. This includes a set of quality indicators relating to:

- The selection criteria and the degree of accessibility of websites
- The methods applied in the different phases of deriving statistical information, and their representativeness with respect to reference populations in the different statistical domains. The aim is to ensure compliance with minimum standards of accuracy, consistency, completeness and timeliness.

In order to evaluate the implemented steps, it is important to be able to assess and monitor the main issues (methodological, technical, operational and organizational), for example unforeseen changes to websites or the need to enter into special agreements with website owners to guarantee long-term data accessibility. The evaluation phase also identifies possible areas for improvement and contributes to progressive standardization, implementing the iterative vision of the process, in full consistency with the GSBPM principles, while adapting to the flexibility required by web data. The figure below shows the original BREAL description and the enhancement version of these Business Functions for web data.

Figure 5 – *Examples of enhanced Business Functions from the 'Support' subset.*



	BREAL Original Description	Enhancement for web scraped data
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: • **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)	*Big Data specific* 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: • **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**

Source: For the descriptions - G. Ruocco et al. (2025): Deliverable D4.7 BREAL - Big Data REference Architecture and Layers for web scraped data. Final version 2025-03-31. Edited by EUROSTAT

In addition to methodological and architectural guidelines, the WIN project has produced a series of empirical results that demonstrate the validity and feasibility of the proposed approach in supporting web data acquisition and management processes in real-world contexts.

Applying the standard has helped to highlight which stages of the process need to be monitored over time to improve input data quality. For instance, in the OJA use case, the assessment of the web data sources (in the landscaping phase of New Data Sources Exploration) resulted in the definition of indicators to evaluate and monitor the relevance and stability of sources over time[2]. Overall, this evidence shows that the project has provided not only a conceptual model, but also an operational platform capable of producing concrete results and driving innovation in statistical processes across various domains.

The main lessons learnt from the initial applications are as follows:

- As previously mentioned, the processes that perform specific Business Functions are influenced by the particular characteristics of the national contexts in which they are implemented. This is particularly evident in ethical and legal aspects, as well as those relating to environmental and infrastructure

[2] These results are documented in Deliverable 4.8, *Quality Assessment for the Statistical Use of Web-Scraped Data*, which provides further details on performance metrics, critical issues encountered, and solutions adopted.

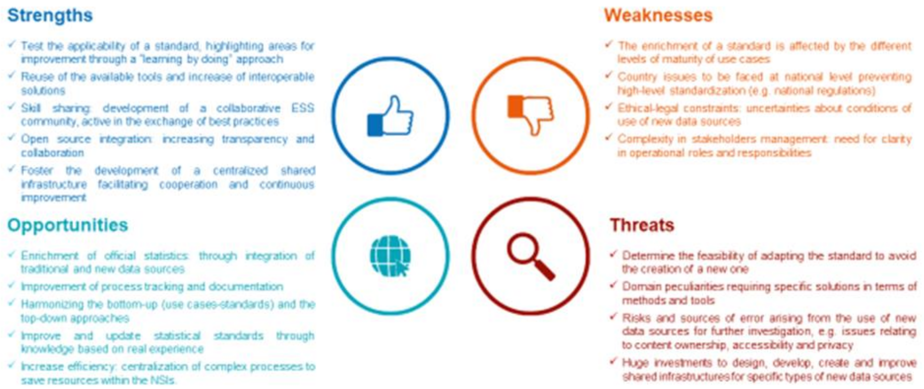
management. This limits the applicability of a framework and highlights areas for improvement

- Sharing experiences, methods, and tools creates synergies that benefit the entire statistical community. Therefore, it is crucial to monitor various use cases over time and support the development of shared infrastructures to facilitate the integration of new data sources into official statistics
- From an architectural perspective, combining a top-down approach — from Business Functions to the implementation of processes to achieve them — with a bottom-up approach — from implemented use cases to Business Functions — fosters process standardization based on the type of source and data. This promotes the sharing of tools and methods between NSIs, starting from a common framework.

5. Specialization of official standards based on use cases experience: a SWOT analysis

In order to assess the pros and cons of specializing official statistical standards and frameworks based on use cases experience, the SWOT analysis reported in Figure 6 provides an overview of the benefits and challenges.

Figure 6 – *Specializing official standards through use cases experience: SWOT Analysis*.



Strengths

- ✓ Test the applicability of a standard, highlighting areas for improvement through a 'learning by doing' approach
- ✓ Reuse of the available tools and increase of interoperable solutions
- ✓ Skill sharing: development of a collaborative ESS community, active in the exchange of best practices
- ✓ Open source integration: increasing transparency and collaboration
- ✓ Foster the development of a centralized shared infrastructure facilitating cooperation and continuous improvement

Weaknesses

- ✓ The enrichment of a standard is affected by the different levels of maturity of use cases
- ✓ Country issues to be faced at national level preventing high-level standardization (e.g. national regulations)
- ✓ Ethical-legal constraints: uncertainties about conditions of use of new data sources
- ✓ Complexity in stakeholders management: need for clarity in operational roles and responsibilities

Opportunities

- ✓ Enrichment of official statistics: through integration of traditional and new data sources
- ✓ Improvement of process tracking and documentation
- ✓ Harmonizing the bottom-up (use cases-standards) and the top-down approaches
- ✓ Improve and update statistical standards through knowledge based on real experience
- ✓ Increase efficiency: centralization of complex processes to save resources within the NSIs

Threats

- ✓ Determine the feasibility of adapting the standard to avoid the creation of a new one
- ✓ Domain peculiarities requiring specific solutions in terms of methods and tools
- ✓ Risks and sources of error arising from the use of new data sources for further investigation, e.g. issues relating to content ownership, accessibility and privacy
- ✓ Huge investments to design, develop, create and improve shared infrastructures for specific types of new data sources

Specifically, the main Strengths (S), Weaknesses (W), Opportunities (O) and Threats (T):

S

- Testing the applicability of a standard and highlighting areas for improvement through a 'learning by doing' approach
- Reusing available tools and increasing interoperable solutions

- Competence sharing: developing a collaborative ESS community that is active in exchanging best practices
- Open-source integration: improving transparency and collaboration
- Fostering the development of a shared, centralized infrastructure to facilitate cooperation and continuous improvement.

W

- The maturity level of use cases influences the enrichment of a standard
- Problems that need to be addressed at a national level, e.g. national regulations, may prevent a high degree of standardization
- Ethical and legal constraints: uncertainties about the conditions of use of new data sources
- Complexity in stakeholders' management, necessitating clarity in operational roles and responsibilities.

O

- Enriching official statistics by integrating traditional and innovative data sources
- Improvement of traceability and documentation of processes
- Harmonization of bottom-up (use cases) and top-down (standards) approaches
- Improvement and updating statistical standards based on experience
- Increase of efficiency through the centralization of complex processes to save resources within NSIs.

T

- Determining the feasibility of adapting the standard to avoid creating a new one
- Peculiarities of the domain that require specific solutions in terms of methods and tools
- Risks and sources of errors that arise from using new data sources for further investigation (e.g. issues related to content ownership, accessibility and privacy)
- Large investments that are required to design, develop, create and improve shared infrastructures for specific types of new data sources.

6. Conclusions

The WIN project demonstrated the benefits of applying and enhancing the BREAL architectural framework, in order to foster the implementation of solutions in the production environments. This approach could be extended to other types of new data sources with the aim of identifying common aspects that could be standardized. This strategy is useful and feasible across different statistical domains, facilitating the implementation of shared infrastructures at a European level to integrate new data sources into the statistical process and enrich official statistics. Adopting this approach increases process standardization and the development of shareable tools that can be reused by other statistical organizations, and/or adapted to other use cases and/or statistical domains.

In addition to the operational benefits demonstrated in the use cases, the WIH is not only a common infrastructure, but also a strategic element if considered within the data ecosystem strategy conceived for the process of modernising and innovating the ESS.

Firstly, the WIH is fully in line with European strategic guidelines, pursuing the objectives of efficiency, cooperation and reduction of production costs. The centralisation of complex functions – such as the acquisition and pre-processing of web data – allows NSIs to free up internal resources, focusing on analyses and adaptations specific to national priorities.

Secondly, the WIH provides resources for addressing emerging legal and ethical frameworks.

In a context characterised by complex regulations (e.g., the GDPR and the recent EU AI Act on artificial intelligence), the hub provides a space in which common guidelines can be developed and shared for ethical scraping, data anonymization and transparent governance. In this sense, the WIH reduces regulatory uncertainty and strengthens user and citizen trust in official statistics.

Another aspect to consider is the potential for future developments. Although the WIH is currently designed for web data, its operating model can be extended to other innovative sources (e.g., data from sensors, satellite images) as well as new-generation administrative sources. Thus, the WIH can evolve from a thematic platform into a reference infrastructure for integrating heterogeneous sources, thereby ensuring methodological continuity and consistency with official statistical standards.

Finally, in terms of the relationship between national priorities and European cooperation, the WIH acts as an enabling tool. On the one hand, it enables NSIs to respond quickly to specific national requirements (e.g. labour market indicators), and on the other, it fosters the exchange of best practices and the reuse of tools, thereby promoting innovation at the ESS level.

In summary, the WIH is not only the outcome of a project, but also an innovation platform for official statistics. It strengthens the ability of the European system to respond to the challenges posed by new data sources by integrating methodological rigour, institutional cooperation, ethical responsibility, and regulatory compliance.

References

AA.VV. 2025. Chapter 5 National data ecosystems and governance. In *The Handbook on Management and Organization of National Statistical Systems.*

Edited by UN, pp. 97-111. https://projects.officialstatistics.org/hb-mgnt-org-nss/handbook/intro.html .

ASCHERI A., MUSEUX J. M., WIRTHMANN A., GIANNAKOURIS K., KARLBERG M., BALDACCI E.2022 Innovation in the European Statistical System: Recent achievements and challenges ahead, *Statistical Journal of the IAOS*, Vol. 38, No. 3, pp. 805-813. https://doi.org/10.3233/SJI-220053.

AUNO V. *et al*. 2024. *Deliverable 4.8: Quality Assessment for the Statistical Use of Web Scraped Data. Final version, 2024-11-20*. Edited by Eurostat.

DAAS P. J. H. 2015. Big data as a source for official statistics, *Journal of Official Statistics*, Vol. 31, No.1, pp. 249–262. https://doi.org/10.1515/jos-2015-0016.

Generic Statistical Business Process Model – GSBPM. https://statswiki.unece.org/display/GSBPM

Generic Statistical Information Model – GSIM. https://statswiki.unece.org/display/gsim

Generic Activity Model for Statistical Organizations – GAMSO. https://statswiki.unece.org/spaces/GAMSO/pages/105580149/Generic+Activity+Model+for+Statistical+Organizations

KOWARIK A. *et al.* 2021. *Deliverable 4.1: Minimal guidelines and recommendations for implementation. Version 2021-09-23.* Edited by Eurostat.

RUOCCO G. *et al.* 2025*. Deliverable D4.7 BREAL - Big Data REference Architecture and Layers for web scraped data. Final version, 2025-03-31.* Edited by Eurostat.

SCANNAPIECO M. *et al*. 2021. *(Deliverable F2) BREAL. Big Data Reference Architecture and Layers. Application layer and Information layer. Version 2021-03-31.* Edited by Eurostat.

SCANNAPIECO M. *et al*. 2019. *(Deliverable F1) BREAL. Big Data REference Architecture and Layers. Business Layer. Version 2019-12-09.* Edited by Eurostat.

SIX, M. *et al*. 2025. *Deliverable 4.6: WP4 Methodology report on using web scraped data. Final version, 2025-03-25.* Edited by Eurostat.

SIX M. *et al*. 2025. *Deliverable 4.5: Quality Guidelines for acquiring and using web scraped data. Revised final version, 2025-02-20.* Edited by Eurostat

STRUIJS P, BRAAKSMA B., DAAS P. J.. 2014. Official statistics and Big Data, *Big Data & Society*, Vol. 1, No. 1 https://doi.org/10.1177/2053951714538417.

Giuseppina RUOCCO, Istat, giruocco@istat.it
Renato MAGISTRO, Istat, renato.magistro@istat.it
Giulio MASSACCI, Istat, giulio.massacci@istat.it