Web Intelligence Network Conference
From Web to Data

4-5 February 2025

GDANSK - POLAND

A specialised architectural framework for web data: the BREAL extension and enhancement

Olav ten Bosch (CBS), Romain Lesur (INSEE), Sonia Quaresma (INE), Francesca Inglese, Annalisa Lucarelli, Renato Magistro, Giulio Massacci, **Giuseppina Ruocco (ISTAT)**

Web Intelligence Network

Funded by the European Union

Outline

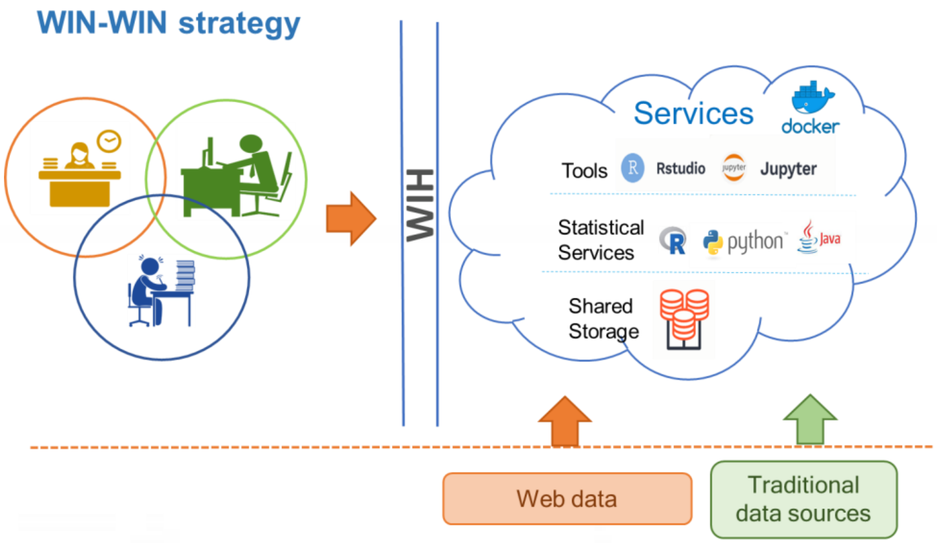
- Background overview
- Development of the WIH: Core concepts
- BREAL Business Functions
- BREAL Analysis
- BREAL and the project use cases
- BREAL and the OBEC/OJA Workflows
- BREAL Enhancement
- BREAL Extension
- Conclusions & lessons learnt

Funded by
the European Union

Background overview

Objectives of the ESSnet Web Intelligence Network (WIN) launched in 2021

- Establish the Web Intelligence Network (WIN) for integrating Web data sources in national statistical production systems
- Develop the Web Intelligence Hub (WIH), a common infrastructure, to share tools and technologies for collecting and processing web data services



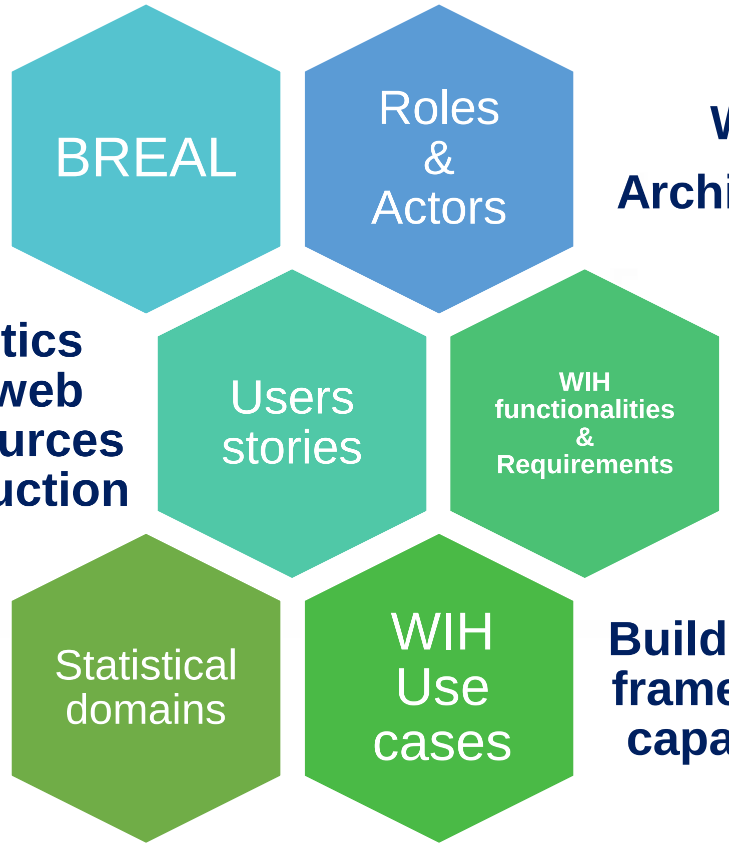
The WIH as:

- **Service provider** for Web data collection and processing
- Web data **repository**
- **Environment** for Web data processing

Web Intelligence Network

Funded by the European Union

Development of the WIH: Core concepts



BREAL

Roles & Actors

Users stories

WIH functionalities & Requirements

Statistical domains

WIH Use cases

WIH Architecture

Statistics from web data sources in production

Building TSS framework & capabilities

WIH implementation & official statistical standards

Business and functional requirements of the WIH

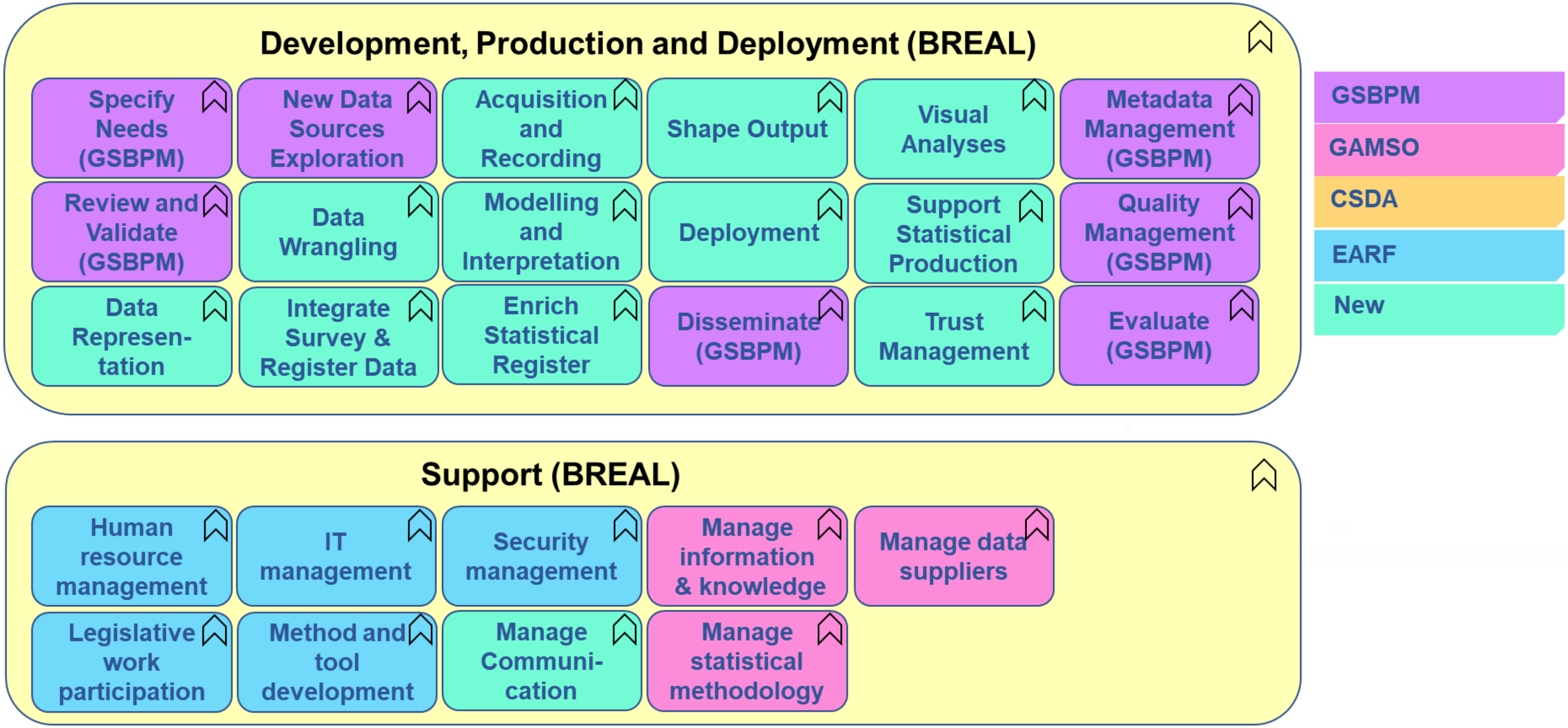
VS

BREAL framework

(Big Data REference Architecture and Layers)

Funded by the European Union

BREAL Business Functions



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 Represen-tation	Integrate Survey & Register Data	Enrich Statistical Register	Disseminate (GSBPM)	Trust Management	Evaluate (GSBPM)

Support (BREAL)

Human resource management	IT management	Security management	Manage information & knowledge	Manage data suppliers
Legislative work participation	Method and tool development	Manage Communi-cation	Manage statistical methodology	

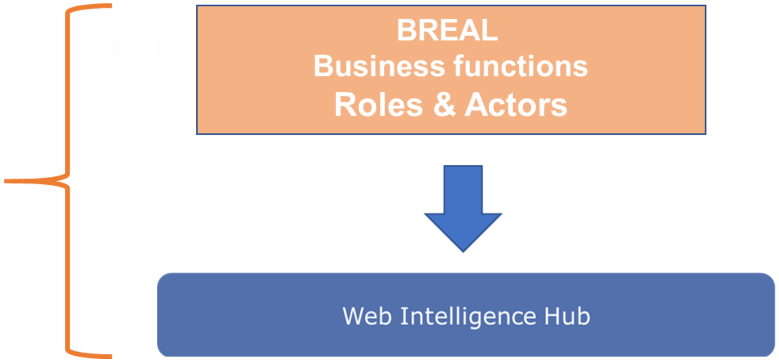
- GSBPM
- GAMSO
- CSDA
- EARF
- New

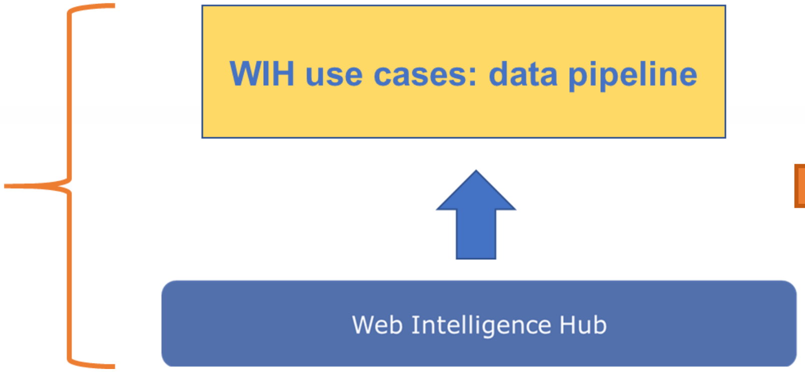
Source: *Scannapieco M., Bogdanovits F., Gallois F.; Fischer B. et al. (2019): BREAL. Big Data Reference Architecture and Layers. Version 2019-12-09. Edited by EUROSTAT*

BREAL Analysis

Top down
vs
Bottom-up approach

Analysis of the business layer (**WHAT**), to specify the **functional requirements** of the WIH services and **enhancing user experience**





Enhancement & Extension (E&E) of the BREAL framework

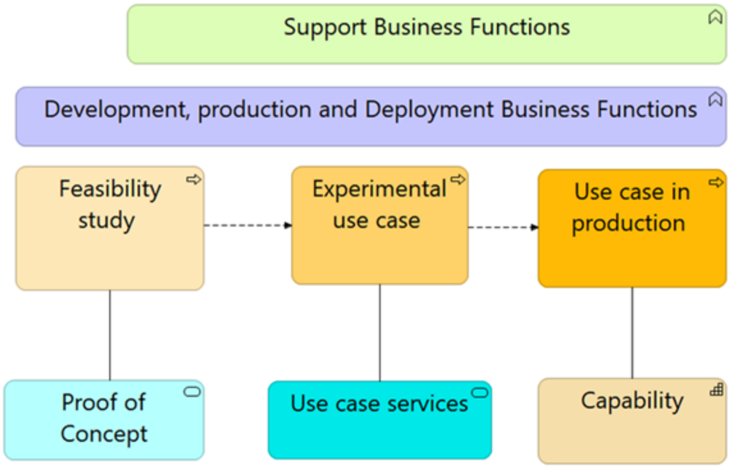
Web Intelligence Network

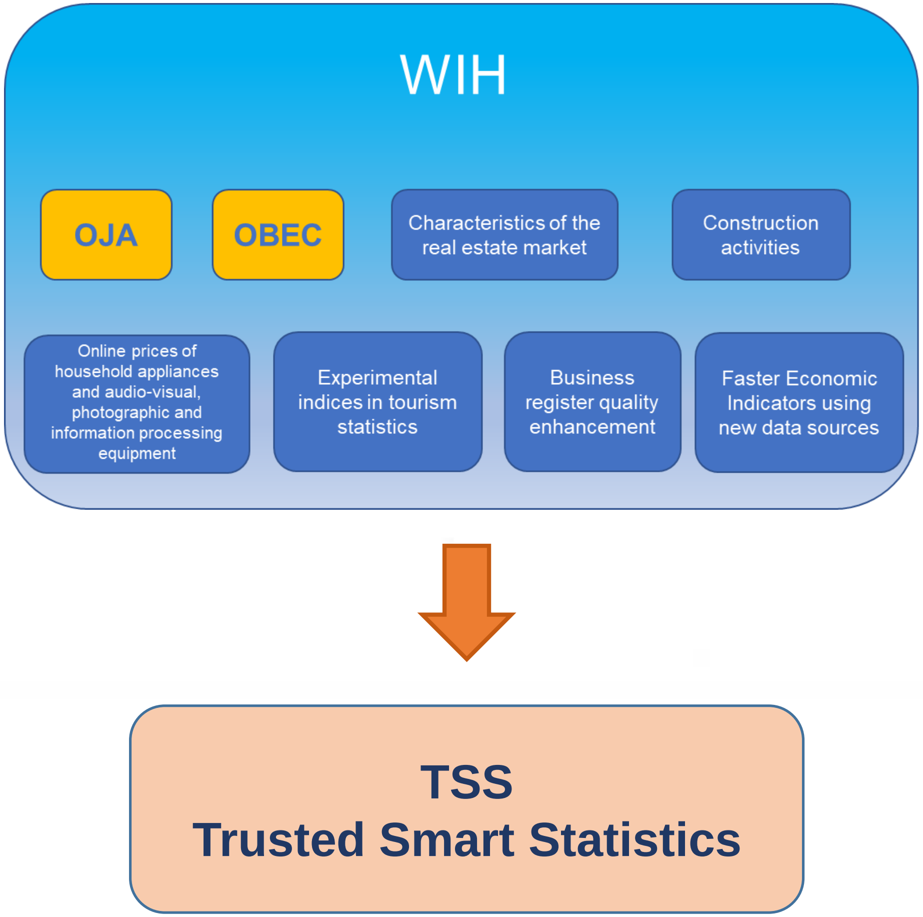
Funded by the European Union

BREAL and the project use cases

E&E of the BREAL framework according to Use case life cycle

Use case life cycle

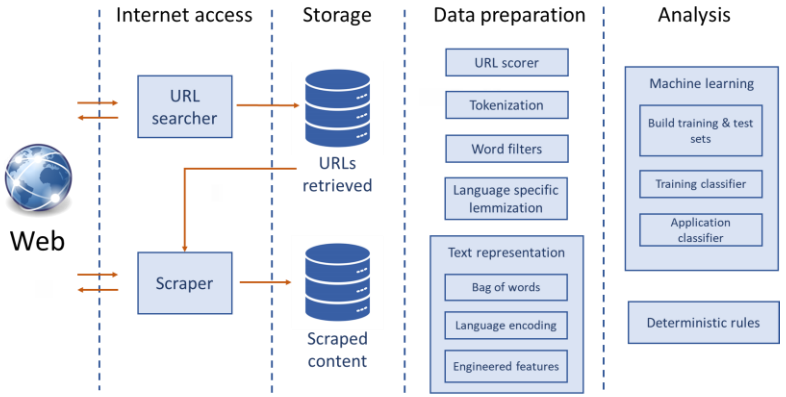


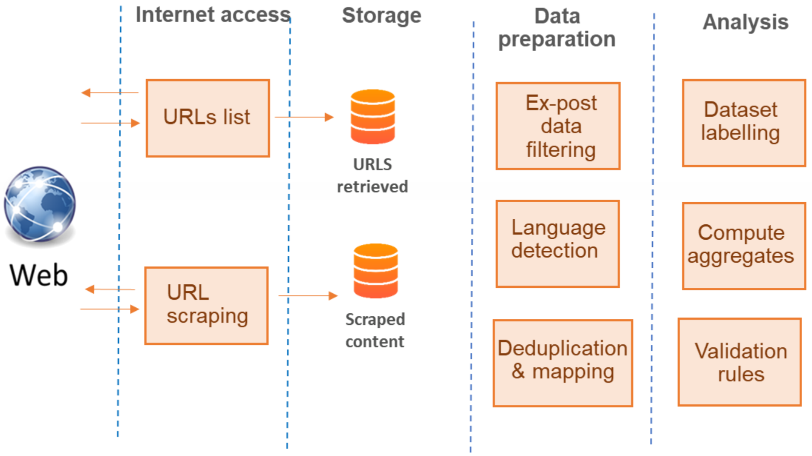


WIH
OJA
OBEC
Characteristics of the real estate market
Construction activities
Online prices of household appliances and audio-visual, photographic and information processing equipment
Experimental indices in tourism statistics
Business register quality enhancement
Faster Economic Indicators using new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

BREAL and the OBEC/OJA Workflows (1)

BREAL Business Functions (BBFs) and **Online Based Enterprise Characteristics** (**OBEC**) use case





BBFs and **Online Job Advertisements** (**OJA**) use case

Acquisition and Recording

Data representation

Data Wrangling

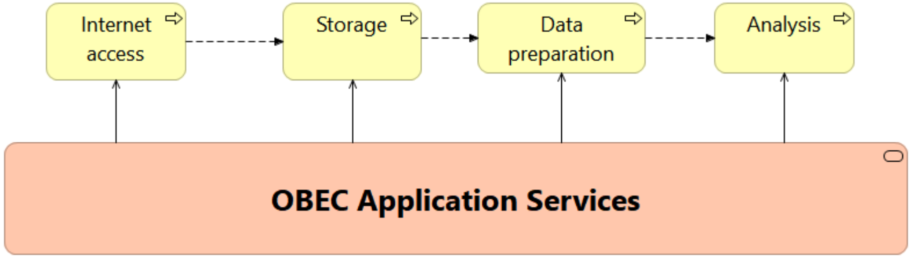
Modelling and Interpretation

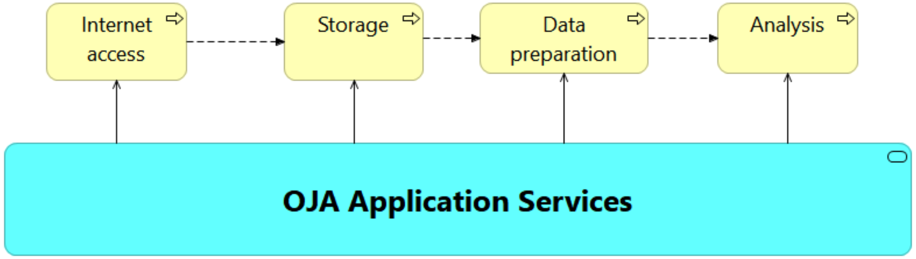
Shape output

Funded by the European Union

BREAL and the OBEC/OJA Workflows (2)

OBEC workflow





OJA workflow

Acquisition and Recording

Data representation

Data Wrangling

Modelling and Interpretation

Shape output

Standardisation of methods and tools

vs

Domain/Country specifics

Funded by the European Union

BREAL Enhancement

BREAL enhancement has resulted in the specialization of the main BBFs, included in the subset **“Development, Production and Deployment”**
The enhancement is not intended to replace the original description, but to enrich it

BREAL Business Functions	BREAL Original Description	BREAL enhancement for web data based on the project experience
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
Data Wrangling	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	The ability to transform web content into a target format and extract the relevant information from the website. This ability also involves: • 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

Funded by the European Union

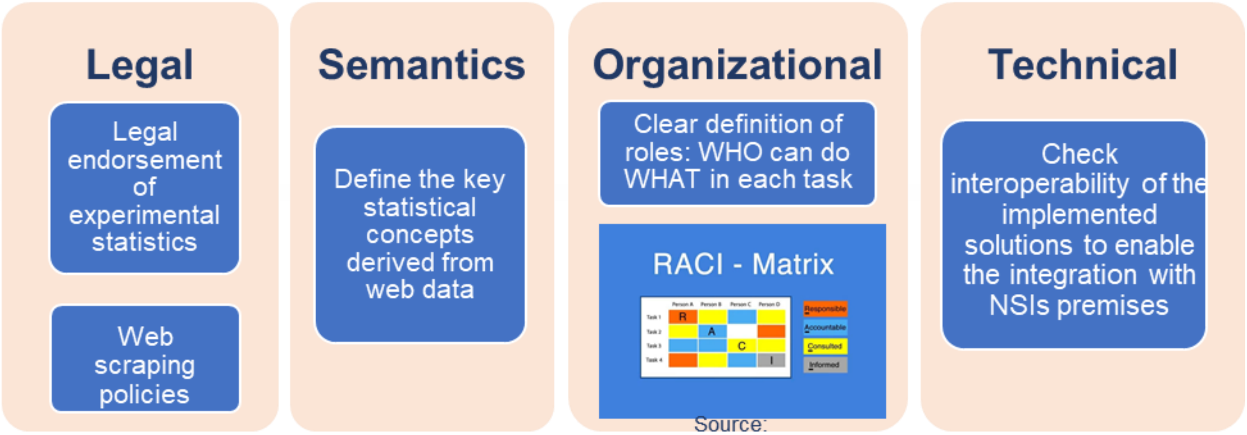
BREAL Extension

Based on the project experience…

Combining the **Bottom-up** and the **Top-down approaches**, addition of a new BBF to the BREAL in the ‘Support’ subset to:

- Promote process management and orchestration
- Deal with unexpected implementation issues
- Speed up and monitor the use case maturity, from the experimental to production phase

Strategy and Process management



Legal

Legal endorsement of experimental statistics

Web scraping policies

Semantics

Define the key statistical concepts derived from web data

Organizational

Clear definition of roles: WHO can do WHAT in each task

RACI - Matrix

Technical

Check interoperability of the implemented solutions to enable the integration with NSIs premises

Source: https://t2informatik.de/en/smartpedia/raci-matrix/

Funded by the European Union

Conclusions & lessons learnt

- There is no "one size fits all" production model. Each use case falls into **a specific production model** based on methods and tools that can be standardised
- **Increased sharing of tools and methods** between NSIs through the WIH
- **Real-world use cases can enrich official standards**, bridging the bottom-up and high-level approaches
- The specialization of the BREAL framework supports the **deployment of mature use cases in production**
- The E&E of the BREAL framework has underlined the **interconnection between methods, tools, data transformations and use case management**, and the need for a holistic approach to build common infrastructures for web data at EU level

Thank you for your attention!

Funded by the European Union