BREAL – Big data REference Architecture and Layers

Towards a common Big Data architecture for the ESS

**Keywords:** Big Data, standards, enterprise architecture

# Introduction

The paper describes BREAL (**B**ig **D**ata **R**Eference **A**rchitecture and **L**ayers), a European reference architecture for Big Data, released as a result of the work carried out within the “Process and Architecture” Workpackage of the Eurostat-funded project ESSnet Big Data Pilots II [1].

BREAL serves the purpose of guiding Big Data investments by National Statistical Institutes (NSIs) and helping the development of standardized solutions and services to be shared within the ESS and beyond. In particular, intended users of BREAL are:

* NSIs that aim to introduce the use of Big Data in their production processes.
* Public and private organizations that would like to follow a defined and controlled way of producing Big Data-based statistics guided by the Official Statistics expertise.

From a practical point of view, BREAL can be used as follows:

* As a reference framework to be used at national and ESS-level by enterprise architects to align business and IT needs.
* As a language for IT/solution architects to describe information systems projects that make use of Big Data sources.
* As an instrument for NSIs’ top management to plan national investments related to Big Data projects, taking into account the economies of scale that are offered by European infrastructures and services for Big Data.

# Overview of BREAL

BREAL is an architectural framework including several architectural artefacts. The specific BREAL artefacts are represented in Figure 1, where they are placed on the specific architectural layer they belong to.

In the Business Layer, the BREAL artefacts are:

* List of principles, intended as guidelines and general rules that support (statistical) organizations implementing statistics based on Big Data sources.
* Business functions model, related to the development, production and deployment process necessary for a Big Data-based production line.
* Life Cycle model, sequencing BREAL business functions in a Big Data-based production line.
* Support functions model, including the functions that are thought to support the production process, like, e.g. human resource management.
* Stakeholder model, with the stakeholders playing relevant roles in a production system intended to manage Big Data sources.

The Application Layer consists of a:

* Generic Application Architecture, with a set of generic application services, proposed to show how the identified business functions can be implemented.

The Information Layer includes a:

* Generic Information Architecture, consisting of the three layers described by the ‘hourglass model’ proposed for Trusted Smart Statistics, namely raw data, convergence, and statistical layer.

In addition, BREAL includes an Operational Model, describing how data and services can be deployed in a Big Data solution. This can be considered as part of both the Application and Information layers.

The proposed Generic Application Architecture and Information Architecture are “generic” in the sense that they are not intended to be specific of a Big Data project or source. The use of the services and data that are specific to the Big Data projects of the Implementation Track of the ESSnet Big Data Pilots II are described in a set of solution architectures, namely (see Figure 1): (i) Solution architectures for *Online job vacancies* (WPB); (ii) Solution architectures for *Online based enterprise characteristics* (WPC); (iii) Solution architectures for *Smart energy* (WPD); (iv) Solution architecture for *Tracking ships* (WPE).



Figure 1: Overview of BREAL

# BREAL Artefacts at a Glance

## BREAL Life Cycle

Using existing reference architectures (in particular EARF [2], and GSBPM [3]), we designed the BREAL business functions model, including already defined business functions and several new business functions (see [4] for details). These business functions were used to build the BREAL Big data Life Cycle, shown in Figure 2. It encompasses three major areas:

* Development and Information Discovery – where the exploration of the Big Data source, its integration with other data and the discovery of information take place
* Production – actually creating statistical products through the use of Big Data sources
* Continuous Improvement – monitoring and assessing the Big Data source usage with a focus on the population coverage issues and the validity of the models used.

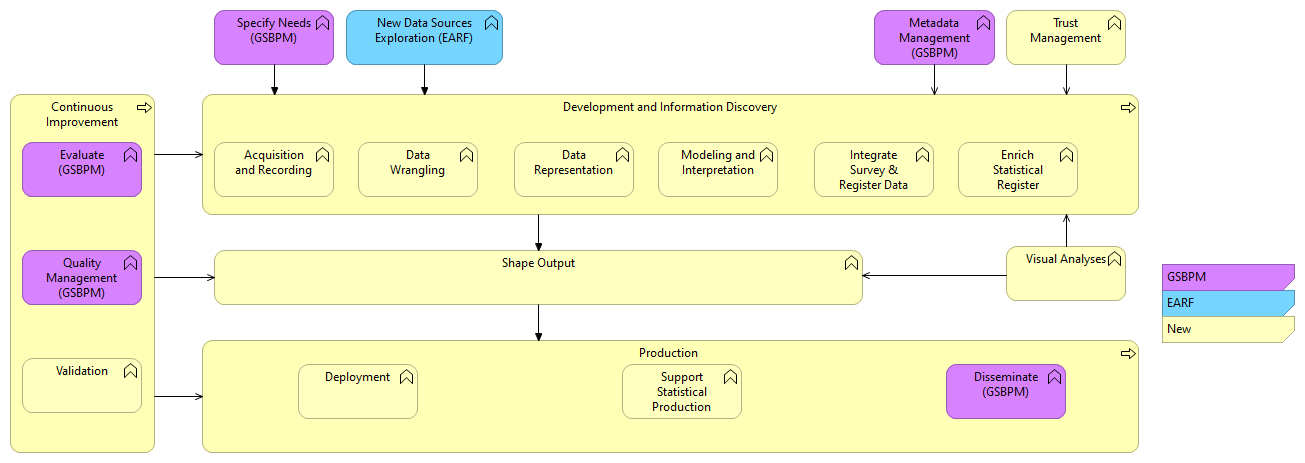


Figure 2: BREAL life Cycle

## Generic Information Architecture

The BREAL Generic Information Architecture for Big data (GIAB) consists of three layers, as described by the “hourglass model” proposed for Trusted Smart Statistics, namely raw data, convergence, and statistical layer:

* The *Raw Data Layer* includes data that are acquired and stored by the BREAL “Acquisition and Recording” business function. Let us remark that, at this stage, the GIAB just identifies the concepts, and no detail is provided on formats or other technical specifications that can be useful for raw data acquisition and storage.
* The *Convergence Layer* contains data represented as units of interest for the analyses. These data are produced as results of the BREAL life cycle functions of “Data Wrangling” and “Data Representation”.
* The *Statistical Lay*er includes those concepts that are the targets of the analysis. These data are produced (mainly) by “Modeling and Interpretation”, “Integrate Survey and Register Data”, “Enrich Statistical Registers” and “Shape Output”.

In addition to the data concepts, some metadata concepts are introduced for each of the three layers. In particular, one specific category of metadata has been selected as very specific of Big Data, namely Provenance metadata. These metadata have been specified for each of the three layers above mentioned. Moreover, in order to better highlight the integration of the GIAB with respect to existing standards for data modelling in Official Statistics, we included some key concepts from GSIM (Generic Statistical Information Model) [5]. The resulting composition of GIAB is shown in Figure 3. Each layer consists of: (i) specific BD data entities (blue color); (ii) GSIM entities (pink color) and (iii) specific provenance metadata entities (yellow color).

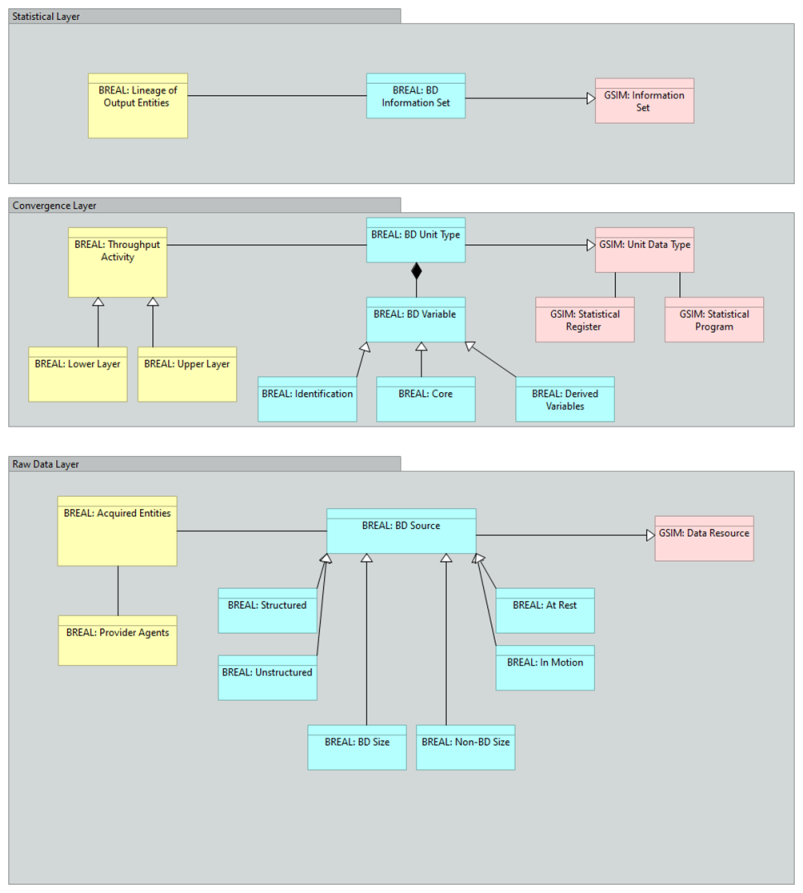


Figure 3: BREAL Information Architecture

# References

1. ESSnet Big Data Pilots II: https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/index.php/Main\_Page
2. Eurostat: ESS Enterprise Architecture Reference Framework: https://ec.europa.eu/eurostat/cros/content/ess-enterprise-architecture-reference-framework\_en
3. GSBPM vs 5.1: <https://statswiki.unece.org/display/GSBPM/GSBPM+v5.1>
4. Deliverable F1 “BREAL: Big Data REference Architecture and Layers Business Layer”: <https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/6/65/WPF_Deliverable_F1_BREAL_Big_Data_REference_Architecture_and_Layers_v.03012020.pdf>
5. GSIM vs 1.2: <https://statswiki.unece.org/display/gsim>