ESSnet Big Data II

Grant Agreement Number: 847375-2018-NL-BIGDATA

https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata

https://ec.europa.eu/eurostat/cros/content/essnetbigdata_en

Work Package F

Process and architecture

Deliverable F1

BREAL: Big Data REference Architecture and Layers

Business Layer

Version 2019-12-09

Prepared by:

Frederik Bogdanovits (Statistics Estonia, Estonia)
Arnaud Degorre, Frederic Gallois (INSEE, France)
Bernhard Fischer (DESTATIS, Germany)
Kostadin Georgiev (BNSI, Bulgaria)
Remco Paulussen (CBS, Netherlands)
Sónia Quaresma (INE, Portugal)
Monica Scannapieco, Donato Summa (ISTAT, Italy)
Peter Stoltze (Statistics Denmark, Denmark)

Work package leader:

Monica Scannapieco (ISTAT, Italy)
monica.scannapieco@istat.it
telephone : +39 06 46733319
mobile phone : +39 366 6322497

Outline

Executive Summary

This document describes BREAL (Big Data REference Architecture and Layers), a European reference architecture for Big Data. BREAL serves the purpose of guiding Big Data investments by 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, especially those that plan to use Web data and sensor data.
- In addition to NSIs, 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 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.
- As a 'reference framework' for enterprise architects to be used at national and ESS-level to align business and IT needs.
- As a 'language' for IT/solution architects to describe information systems projects that make use of Web data and sensor data.

This document describes the artifacts produced so far of the Business Layer of BREAL. These artifacts will be discussed and validated both within the ESSnet Big Data Pilots II and beyond the ESSnet, in all the venues that will be identified as relevant to this purpose.

The future work will entail the design of the Information Architecture (IA) and the Application Architecture (AA). For this purpose, implementation work packages will be grouped into two main clusters, namely:

- Web Intelligence, involving 'Online Job Vacancy' (WPB) and 'Enterprise Characteristics' (WPC).
- Sensor Data, involving 'Smart Energy' (WPD) 'Tracking Ships' (WPE).

Hence, by the end of 2020, BREAL will be fully defined and will include: (i) a general Business Layer, (ii) an Information Architecture Layer and an Application Architecture Layer for Web Intelligence and (iii) an Information Architecture Layer and an Application Architecture Layer for Sensor Data.

1 Introduction

Modern organizations have recognized the importance of having a defined and standardized architecture that is able to guide the implementation of the organization's vision and to harness external drivers and changes. National Statistical Institute (NSIs) have been investing in standardization projects for several years, which have produced several artifacts as constituting pieces of a 'reference architecture' for their specific business case. The ESS EARF (Enterprise Architecture Reference Framework)[1], developed under Eurostat's sponsorship, is an example of these kinds of artifacts. However, considering a reference architecture for Big Data in Official Statistics, there are few concrete results in terms of available standards, possibly due to the only quite recent attention that Big Data received by NSIs.

Work package F 'Process and Architecture' of the ESSnet Big Data II project has the purpose of creating a European reference architecture for Big Data, serving the purpose to (i) guide Big Data investments by NSIs and (ii) help the development of standardized solutions and services.

The outcome of such an effort is what we called **BREAL** (**B**ig Data **RE**ference **A**rchitecture and **L**ayers), which is a set of artifacts organized according to the different layers that typically compose enterprise architecture, namely:

- The Business Layer, dealing with 'what' NSIs do with respect to Big Data management. The artifacts of this layer are: (i) a set of principles, (ii) a set of *business functions*, (iii) a description of the Big Data based production process called *Big Data Life Cycle* and (iv) a set of Actors and Stakeholders.
- The Application Layer, dealing with 'how' NSIs could / should realize the business functions and the Big Data Life Cycle in terms of *application components* and *services*.
- The Information Layer, dealing with 'how' NSIs could / should realize the business functions and the Big Data Life Cycle in terms of *data models*.

As an example for an artifact of the Business Layer, let us consider the business function 'Acquisition and Recording' that is in charge of ingesting data from a Big Data source. An application component, belonging to the Application Layer, can be identified to realize such a business function, let's call it 'Scraping'. At the Information Layer, an 'Enterprise Web site' *information object* can be part of a data model and represents the input to the 'Scraping' application component.

The Technology Layer will not be part of BREAL. Providing technological solutions to be adopted for the implementation of BREAL-compliant projects is out of the scope of the WPF work.

BREAL has the additional objective of being used as a reference by work packages of the Implementation Track of the ESSnet Big Data Pilots II, namely: WPB – "Online Job Vacancy", WPC – "Enterprise Characteristics", WPD – "Smart Energy" and WPE – "Tracking Ships". In this respect, the

[1] https://ec.europa.eu/eurostat/cros/content/ess-enterprise-architecture-reference-framework_en

solution architectures that these work packages will put in place are supposed to be compliant to BREAL with respect to the three defined layers presented above.

The focus of this document is on the description of the Business Layer in terms of the defined artifacts, namely:

- Section 2 describes BREAL principles.
- Section 3 illustrates BREAL business functions and how they compose the Big Data Life Cycle.
- Section 4 is related to actors involved in the Big Data based production process represented by the Big Data Life Cycle.
- Section 5 will highlight ongoing and future developments on the Application and Information layers.

The next section is aimed to provide some background on existing standards and frameworks.

1.1 Existing architectures and frameworks used

There are already many architectures and frameworks[2] on producing official statistics and on Big Data. As BREAL promotes to share and reuse Big Data processes, methods and data, we of course also reused parts of these architectures ourselves. These architectures provided valuable inputs and ideas. From some of these architectures we reused principles, functions, processes, components and much more. We tried to include references everywhere.

The architectures and frameworks used as an input for producing the BREAL architecture are:

- Generic Statistical Business Process Model (GSBPM)
  The GSBPM describes and defines the set of business processes needed to produce official statistics. It provides a standard framework and harmonized terminology to help statistical organizations to modernize their statistical production processes, as well as to share methods and components. The GSBPM can also be used for integrating data and metadata standards, as a template for process documentation, for harmonizing statistical computing infrastructures, and to provide a framework for process quality assessment and improvement.
- Generic Activity Model for Statistical Organizations (GAMSO)
  The GAMSO describes and defines the activities that take place within a typical organization producing official statistics. It extends and complements the Generic Statistical Business Process Model (GSBPM) by adding additional activities needed to support statistical production. Like the GSBPM, the GAMSO aims to provide a common vocabulary and framework to support international collaboration activities, particularly in the field of modernization. While individual

[2] The difference between a reference architecture and a reference framework is that a reference architecture provides a common solution and vocabulary, while a reference framework is a more general concept that may include for instance principles and practices for creating specific solutions.

collaborations typically focus on modernizing a particular aspect of production (as described by the GSBPM), statistical production occurs within a broader context of corporate strategies, capabilities and support. The GAMSO helps to place collaboration in the wider context.

- Generic Statistical Information Model (GSIM)
  The GSIM is a reference framework for statistical information, designed to play an important part in modernizing and streamlining official statistics at both national and international levels. It enables generic descriptions of the definition, management and use of data and metadata throughout the statistical production process. It provides a set of standardized, consistently described information objects, which are the inputs and outputs in the design and production of statistics.

- Common Statistical Production Architecture (CSPA)
  CSPA is a reference architecture for the statistical industry. The scope of CSPA is statistical production across the processes defined by the GSBPM (it does not characterize a full enterprise architecture for a statistical organization).

- Common Statistical Data Architecture (CSDA)
  CSDA is a reference architecture supporting statistical organizations in the design, integration, production and dissemination of official statistics based on both traditional and new types of data sources, such as Big Data, Scanner data, Web Scraping, etc.

- ESS Enterprise Architecture Reference Framework (EARF)
  ESS EARF supports the implementation of the ESS Vision 2020 as the guiding frame for ESS development up to 2020. ESS EARF builds upon GSBPM, GAMSO, and CSPA.

- ESS Statistical Production Reference Architecture (SPRA)
  The SPRA outlines how the Collect, Process & Analyze, and Disseminate phases and subprocesses of the statistical production process should be supported by IT in the ESS. It does this by describing the Building Blocks and Services necessary to realize the architecture.

- Eurostat Big Data Task Force (BDTF)
  The BDTF of Eurostat is not an architecture, but a task force developing a strategy for the integration of Big Data in official statistics at EU level. They produced some interesting presentations about Big Data in the context of statistical processes and products. See for example, the presentation 'Towards a Reference Architecture for Trusted Smart Statistics'. We reused some of these ideas, including some principles.

- NIST Big Data Reference Architecture (NBDRA)
  The NIST Big Data Reference Architecture (NBDRA) is an open reference architecture for Big Data. It consists of a conceptual model discussing the roles and fabrics of the NBDRA, presenting an activities view to describe the activities performed by the roles and presenting a functional component view containing the classes of functional components that carry out the activities.

- Cross-industry standard process for data mining (CRISP-DM)

CRISP-DM is an open standard process framework for designing, creating, building, testing, and deploying data mining solutions, including machine learning. The process consists of six main phases: Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation, and Deployment. Note that the original standard is no longer maintained, but that the ideas are still valuable.

2 BREAL Business Layer: Description of the Principles

The BREAL business principles are guidelines and general rules that support (statistical) organizations implementing statistics based on Big Data sources. The selection or definition of the right principles is crucial in order to make a sound decision. During the implementation of a Big Data project, the guidelines are checked from time to time and it is determined whether the principles are correctly applied. Note that it is allowed to deviate from these principles if there are good reasons to do so. This is known as *comply or explain*.

Using existing reference architectures (both statistics-specific and non-statistical), the BREAL business principles are derived. Already defined business principles are re-used and several new ones are introduced. Below, we present and explain these business principles, which we have organized in a number of categories with most principles belonging to the 'Data capturing' and 'Data processing' categories. In case the business function was re-used, the original (shortened) description is used and, if necessary, the specific use for Big Data is included.

Data capturing

Principle	Use an authoritative source (CSDA principle, paragraph 5)
Statement	Within a business process, there should be an authoritative[3] source from which information should be sourced and updated. Where practical, existing information should be reused instead of recreated or duplicated.
Rationale	Maintaining fewer sources of information is more cost effective. Having one source of information supports discovery, reuse and a 'single version of truth'.

[3] According to the Online Dictionary for Library and Information Science by Joan M. Reitz, the definition of authoritative is: "A source that is official. Also, a work known to be reliable because its authenticity or integrity is widely recognized by experts in the field."

Principle	Information is captured and recorded at the point of creation/receipt (CSDA principle, paragraph 4)
Statement	Information should be captured and recorded at the earliest point in the business process to ensure it can be used by subsequent processes. Subsequent changes to information should be documented at the time of action.
Rationale	Information is captured and recorded at the time of creation/action so it is not lost. The amount of information reuse is maximised by capturing it as early as possible.

Principle	Data is only "used by" the Statistical Office (BDTF principle 2)
Statement	It is not always possible (nor desired) to receive all data and process it onsite. With new technologies like Secure Multi-Party Computation[4], it still possible to share and process secure data and retrieve the results without violating security rules and regulations like the General Data Protection Regulation (GDPR).
Rationale	As stated by BDTF, the goal is the output, not the input!

Principle	Standardise and harmonise data as quickly as possible
Statement	Standardise and harmonise the input data as early as possible (for example, the reference frame of maritime ships) to exclude non-used and non-interesting data.
Rationale	Standardising and harmonising a Big Data set makes it easier to handle and process the amount of data and remove errors. Be careful with filtering though. When using filtering it is important to understand the impact, like losing important details.

Principle	Capture metadata at source (Common Metadata Framework principle, Relationship to Statistical Processes, iv.)
Statement	Capture metadata at their source, preferably automatically as a by-product of other processes.
Rationale	Metadata is captured and recorded at the time of creation/action so it is not lost. Metadata is required in order to understand the data itself.

Data processing

Principle	Processing method (algorithm) transparent to all involved parties (BDTF principle 1)
Statement	Processing method (algorithm and process) is transparent to all involved parties • co-designed or at least agreed-upon (consensus-based design)

[4] According to Wikipedia: "A subfield of cryptography with the goal of creating methods for parties to jointly compute a function over their inputs while keeping those inputs private."

	• certified for privacy/ethical compliance, for example GDPR[5] certification
Rationale	Transparency results in better methods, because more people with different kinds of backgrounds and knowledge can verify that the method is correct. Transparency also results in trust, for example trust that the data is used correctly, and thus stronger relationships.

Principle	Push computation out (BDTF)
Statement	Let the data source provider perform as many processing steps as possible - if possible even up to providing intermediate products. Even though the BDTF explains the necessity to push the computation out, due to the complexity and even impossibility of handling so much data onsite, they did not define it as a principle. Because of its importance, we decided to do so.
Rationale	The size of the data and complexity of the processing can be too much to perform this at the statistical office only.

Principle	Consider all capability elements (CSPA principle, paragraph 69)
Statement	Consider all capability elements (e.g. methods, standards, processes, skills, and IT) to ensure the end result is well integrated, measurable, and operationally effective.
Rationale	All these elements are important and required in order to take the new statistical products into production.

Relationship management

Principle	Engage and partner with the input parties (BDTF principle 3)
Statement	Reward partners by giving back computational output – close the loop!
Rationale	To make it more desirable to share data and work together

Quality

Principle	Quality control is built in (EARF principle)
Statement	Information systems and statistical services generate the metadata required to track, monitor and continuously improve the quality of statistical outputs, statistical processes and the characteristics of the institutional environment.
Rationale	Support realizing the ambition of Vision 2020 stating that “We manifest ourselves as the statistical conscience, which guides society through the information overload".

[5] GPDR is the European Union’s (EU) General Data Protection Regulation

Security

Principle	Security is built in (EARF principle)
Statement	Information assets and systems are guaranteed to be available, cannot be compromised and their access is controlled implementing the following dimensions of security: Availability, Integrity, Non-Repudiation and Confidentiality.
Rationale	Implementation of security is imposed in different legal provisions requesting different levels of confidentiality. • Information is the key asset of official statistics. Information cannot be compromised by external and internal stakeholders • We should maintain the trust of our information providers. Ensuring security of data along their entire life cycle is the corner stone for building trust among ESS members • The sharing of confidential information requires that the information flow is secured and access and usage is traceable • External risk (attacks) are increasing

Reuse (and service-based approach)

Principle	Reuse before adapt before buy before build (EARF principle)
Statement	The first, preferred option to cover an identified need should be to reuse an existing generic component (methods, definition, package/module/component/service …). If such functionalities are not readily available, the second option is to adapt a solution which already exists ideally in an OSS framework. If not appropriate, the third option is to buy an existing package (Common-off-the-shelf, abbreviated as COTS). Only when no such packages are available, the functionalities should be built.
Rationale	• Reusing components minimizes development time and cost • Existing solutions which have already been tested in production are likely to be more robust and deliver to quality requirements

Principle	Reuse of data
Statement	In order to reduce the administrative burden and to prevent double work, all data already available should be reused, if possible.
Rationale	Reduce the administrative burden and prevent double work

3 BREAL Business Layer: Business Functions and Big Data Life Cycle

According to Archimate[6], a business function is a collection of business behaviours based on a chosen set of criteria (typically these are required business resources and/or competencies), closely aligned to an organization, but not necessarily explicitly governed by the organization. Just like a business process, a business function also describes internal behaviour performed by a business role. However, while a business process groups behaviour based on a sequence or flow of activities that is needed to realize a product or service, a business function typically groups behaviour based on required business resources, skills, competencies, knowledge, etc.

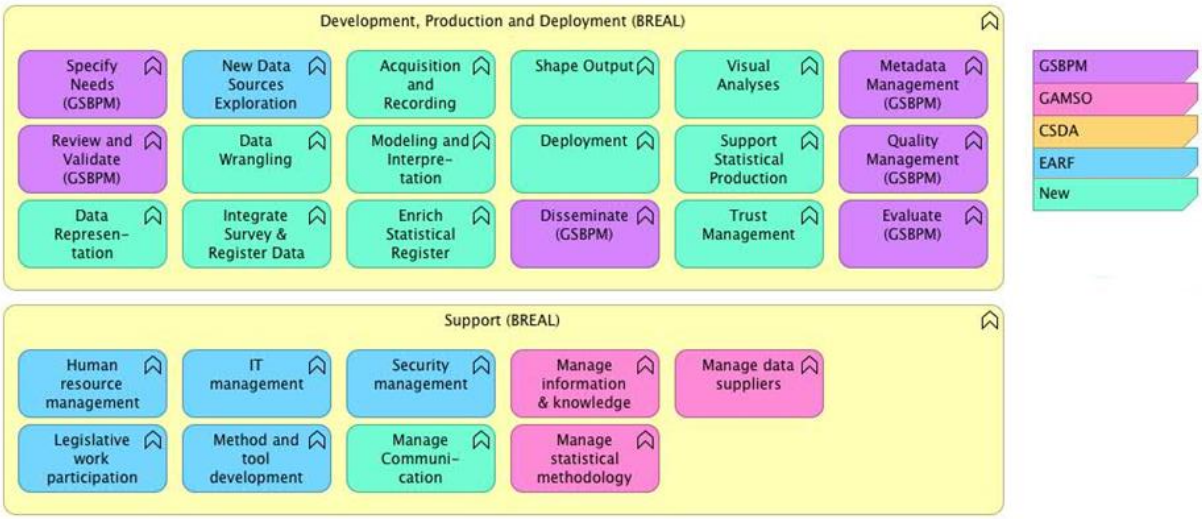

Figure 1: BREAL Business Functions

[6] ArchiMate is an open and independent enterprise architecture modeling language cf. Wikipedia.

Using existing reference architectures (both statistics-specific and non-statistical), the BREAL business architecture layer is derived. Already defined business functions are re-used in this layer and several new business functions are introduced. In Figure 1, we present all required business functions split by the two main areas: 'Development, Production and Deployment' and 'Support', respectively. The colours refer to the reference architecture where the business functions were originally defined. The business functions in light-green did not yet exist and are introduced as part of the BREAL architecture.

3.1 Business Functions: Development, Production and Deployment

In the table below, we describe the business functions that are part of development, production and deployment in an alphabetical order. In case a business function was re-used, the original (shortened) description is used and the specific use for Big Data is included when necessary.

Business Function	**Description**
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.
Data Representation	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.
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
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.
Disseminate (GSBPM)	For the original description, see GSBPM, paragraph 95 through 96: *This phase manages the release of the statistical products to users. It includes all activities associated with assembling and releasing a range of static and dynamic products via a range of channels. These activities support users to access and use the products released by the statistical organisation.* **Big Data specific** Original description applies also to statistical products, which are created using Big Data sources. The business function "Visualisation" is about creating the statistical products (GSBPM 7.2), while the business function

	"Disseminate" is more about providing the statistical products to customers.
Enrich Statistical Register	The ability to enrich the statistical register(s) with the information retrieved from the Big Data source.
Evaluate (GSBPM)	For the original description, see GSBPM, paragraph 103 through 105: *Evaluating the success of a specific instance of the statistical business process, drawing on a range of quantitative and qualitative inputs, and identifying and prioritising potential improvements.* **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.
Integrate Survey & Register Data	The ability to reuse and integrate other data like survey & register data in order to enrich the results derived so far.
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.* **Big Data specific** Original description applies as such.
Modelling and Interpretation	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).
New data sources exploration (EARF)	For the original description, see EARF, page 18: *The ability to explore the potential value of new data sources for improving existing statistics or innovating to obtain new statistics.* **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".
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. **Big Data specific** Original description applies as such.
Review and Validate (GSBPM)	For the original description, see GSBPM, paragraph 78: *This sub-process examines data to try to identify potential problems, errors and discrepancies such as outliers, item non-response and miscoding. It can also be referred to as input data validation. It may be run iteratively, validating data against predefined edit rules, usually in a set order.* **Big Data specific** Original description applies as such.
Shape Output	The ability to format and present data which is not foreseen in advance but instead emerges from the data patterns or is suggested during the data exploration.
Specify Needs (GSBPM)	For the original description, see GSBPM, paragraph 34 through 35: *This phase is triggered when a need for new statistics is identified or feedback about current statistics initiates a review. It includes all activities associated with engaging stakeholders to identify their detailed statistical needs (current or future), proposing high level solution options and preparing a business case to meet these needs.* **Big Data specific** 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.
Support Statistical Production	The ability to support the statistical production system(s) already in place.
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 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.
Visual Analysis	The ability to format and present data in such a way as to optimally analyse the results.

3.2 Business Functions: Support

In the table below, we describe the business functions that are part of support in an alphabetical order. In case the business function was re-used, the original (shortened) description is used and the specific use for Big Data is included, if needed.

Business Function	Description
Human resource management (EARF)	For the original description, see EARF, paragraph 2.3, page 16: *The ability to maintain the necessary human resources and optimize the value of human resources through hiring and development activities.* **Big Data specific** Original description applies as such. Big Data projects do require personnel with a specific skillset to manage and analyse obtained data effectively. Some of technical and business skills needed might include: data visualization and analytical skills, data structure and algorithms, hands-on implementation, theoretical knowledge.
IT management (EARF)	For the original description, see EARF, paragraph 2.3, page 16 *The ability to manage tasks and decisions surrounding the ingestion, annotation, cataloguing, storage, retrieval and distribution of IT assets.* **Big Data specific** Original description applies as such. However, specific processes and tools are required to manage Big Data. For example, machine learning, NoSQL databases, cluster computing, cloud-based computing and shared storage in cloud-based environments.
Legislative work participation (EARF)	For the original description, see EARF, paragraph 2.3, page 18 *The ability to participate in and influence legislative work that forms the legislative basis of official statistical production in a way that will support decision makers and is regarded as useful and important.* **Big Data specific** Original description applies as such. Compared to standard data projects, risk of privacy violations might be greater whilst carrying out certain Big Data projects (for example processing exact mobile positioning data in real time).
Manage Communication	The ability to communicate and involve all stakeholders. Note that the communication and involvement of data suppliers is already covered by the business function 'Manage Data Suppliers'.
Manage data suppliers (GAMSO)	For the original description, see GAMSO, page 12 *These activities cover the relationships with data suppliers, which could include public sector and/or private entities that supply data for*

	statistical activities. This includes cross-process burden management, as well as topics such as profiling and management of contact information. **Big Data specific** Original description applies as such.
Manage information & knowledge (GAMSO)	For the original description, see GAMSO, page 12 *These activities include the ownership or custody of records, documents, information and other intellectual assets held by the organisation and the governance of information collection, arrangement, storage, maintenance, retrieval, dissemination, archiving and destruction. They also include maintaining the policies, guidelines and standards regarding information management and governance.* **Big Data specific** Original description applies as such.
Manage statistical methodology (GAMSO)	For the original description, see GAMSO, page 11 *These activities manage the statistical methodology used to design and carry out the statistical production process. These include initiating and ensuring that standard statistical methods and practices for the processes and sub-processes are identified, put in place in the organisation, and reviewed, to continuously improve efficiency of the production process.* **Big Data specific** Original description applies as such.
Method and tool development for new statistics (EARF)	For the original description, see EARF, paragraph 2.3, page 18 *The ability to effectively develop methods and tools to support the exploration and innovation of new statistical products.* **Big Data specific** Original description applies as such. In case of Big Data, methods and tools are different from what is used for typical official statistics production (specific skills and a different approach are required). For dealing with Big Data we need to enlarge the scope of 'standard statistical methods'.
Security Management (EARF)	For the original description, see EARF, paragraph 2.3, page 16 *The ability to ensure the confidentiality, integrity and availability of the ESS information, data and IT services through administering and deploying adequate security measures, managing risks and assuring and controlling how security policies have been implemented.* **Big Data specific** Original description applies as such.

3.3 Big Data Life Cycle

In the previous section, all business functions were shown and described in a detailed manner. The goal of the current section is to show how they are used in the Big Data Life Cycle. Note however, that business process groups behave different from business functions and potentially there can be many-to-many relationships between business processes and business functions.

The Big Data Life Cycle for official statistics production encompasses three major business process 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.

Note that the business functions making up 'Support' in the lower panel of Figure 1 (and further described in Section 3.2) are *not* included in the Big Data Life Cycle.

The main drivers to the first phase within the Big Data Life Cycle are two business functions which group different activities requiring Big Data examination skills, knowledge and resources: **Specifying Needs** (GSBPM) and **New Data Sources Exploration** (EARF). The development and information discovery is supported and served by the business functions **Metadata** (GSBPM) and **Trust Management**. This latest function is crucial to the Big Data lifecycle as the overall consistency of measurements is central to the adoption of Big Data in official statistical production and although the accuracy and consistency of a Big Data source may be high at any given moment its reliability over time may prove to be problematic.

Within the development and information discovery area, several business functions take place. Those overlap sometimes with business processes, but the focus remains on the functions due to the absence of sequence or flow of activities implied by the business processes. The functions comprised in development and information discovery are not mandatory and are not meant to be understood in a linear fashion. Which functions are necessary depends on the type of Big Data source being addressed and the agreement with the specific data provider. Most of the business functions emerge from Big Data usage:

- **Acquisition and Recording** – Ensuring access to the data (may not be necessary)
- **Data Wrangling** – Couples three business processes: extraction, cleaning and annotation
- **Data Representation** – Necessary when data is unstructured or semi-structured
- **Modelling and Interpretation** – Using algorithms and models specific to Big Data
- **Integrate Survey & Register Data** – Providing support to statistical operations as surveys
- **Enrich Statistical Register** – Improving, extending complementing and/or enhancing existing Registers

Using Big Data sources poses the issue that the right statistical model results out of the data discovery step and is not known beforehand. For example a supervised machine learning process may be adopted, and two sets of data for learning and testing will be established. However once we have trained the model this step may not need to be executed at least for a while until the model is tested again and/or readjusted. This means part of the development cycle will possibly not be replicated in the deployment or production cycle.

The **Review and Validate** (GSBPM) is of course indispensable during the whole development and information discovery and it is required at several points during this sub-cycle, for example before and after data wrangling. Although there are validation capabilities and processes already in place at the statistical offices they may have to be expanded / adapted to tackle Big Data sources.

During the development and information discovery sub-cycle **Visual Analysis** will support the user. This may occur as soon as data is acquired and recorded, after data wrangling or even after modelling and interpretation - whenever it can provide insights on the data and/or facilitate the comprehension of the data and its patterns. **Visual Analysis** provides functionality to the development and information discovery as do **Metadata** and **Trust Management**.

The results of the development and information discovery business process group are selected and formatted through the **Shape Output** business function. Shaping outputs differs from Design Outputs (GSBPM phase 2.1) and Prepare Draft Outputs (GSBPM phase 6.1) which address the detailed design of statistical outputs, products and services. The outputs shaped by this new business function will be used in the production business process group and may not be statistical outputs on its own.

The business process **Production** includes three major business functions:

- **Deployment** – Deploy new processes that make use of the Big Data source. This business process is challenging as it may require new infrastructure to cope with large volumes of data in production.
- **Support Statistical Production** – Enrich or replace the currently established statistical outputs
- **Disseminate** – Deliver the statistical products and services to customers

The process and the (quality of the) results are constantly analyzed, evaluated and improved through the **Evaluate** (GSBPM) and **Quality Management** (GSBPM) functions. The continuous improvement business process directly impacts:

- the development and information discovery – for example if the models used - i.e. machine learning - need to be updated because the previous model is no longer valid

- the shaping of outputs – as the statistical products that may be created by the exploration of a Big Data source are difficult to know a priori and will emerge from the patterns in the data (the models used may also change over time)
- the production – as the possibility to enrich other data sets may change over time

In order to correctly interpret the Big Data Life Cycle model as visualized in Figure 2, please note the following:

- All business functions part of the same business function (see for example **Development and Information Discovery**) are depicted in the normal causal sequence. However, from any point in the sequence, the process can loop back and start over. To make the diagram more readable all trigger relationships have been left out and only the business functions as such are presented.
- Lines with solid arrows are trigger relationships. According to Archimate, the triggering relationship describes a temporal or causal relationship between elements.
- Lines with open arrow are serving relationships. According to Archimate, the serving relationship models that an element provides its functionality to another element.

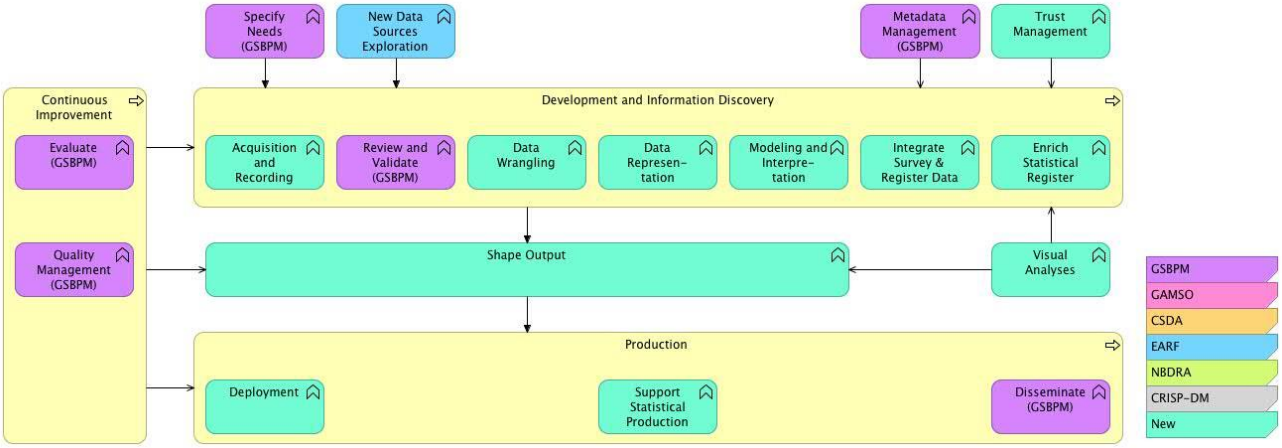

Figure 2: BREAL Big Data Life Cycle

4. BREAL Business Layer: Actors and relationships

4.1 Actors

One particularity about Big Data based projects is that resources are usually shared between several clients, like for cloud-based data systems. Contrary to traditional data systems which tend to be hosted, developed, and deployed for only one organization, Big Data systems are likely to be distributed among various organizations and used for different topics. Hence, actors in Big Data systems can come from multiple organizations, even though they are in charge of specific processes.

In Figure 3 a number of actors are visualized in a seemingly complex and interwoven fashion. In this section we describe the roles of the different actors and how they interrelate. To aid the readability of the text, the roles are in italics, e.g. *Data Scientist*. Also, the many roles are described in subsections in line with the figure, namely:

- IT & Statistical Pipeline Actors
- Capacity Providers
- Global roles (*Statistical Institutions* and *System Orchestrator*)
- Audit, Control, and Compliancy Actors
- Further Actors Linked to (but not involved in) the Process

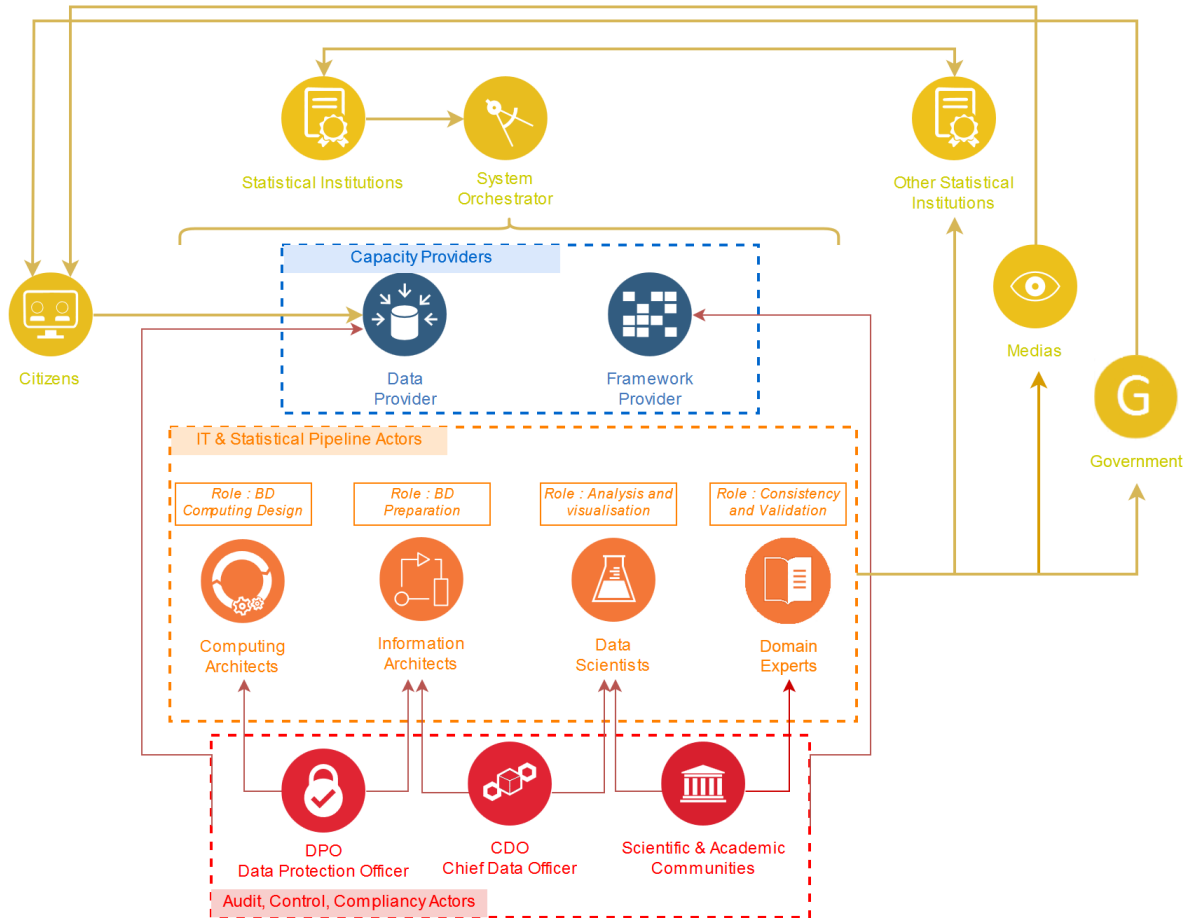

Figure 3: BREAL actors

IT & Statistical Pipeline Actors

According to the BREAL Big Data Life Cycle (see Figure 2), the first kind of actors is linked to the development and information discovery.

The very first role in that view is dedicated to build and deliver computing capabilities that are necessary to capture and use raw Big Data in a statistical process. This role is labelled as 'Big Data Computing Design', with actors being *Computing Architects***.** The capabilities linked to this kind of actors deal with cloud computing, storage management and, more generally, IT architecture. *Computing Architects* interact mainly in the 'Acquisition and Recording' step, but should be considered also with a transversal role throughout the whole Big Data Life Cycle, enabling the technical prerequisites for both 'Discover' and 'Production' major phases.

Computing Architects are first in charge to define detailed architectural requirements for the data system, including the following: Data process requirements; Software and hardware requirements; Logical data modeling and partitioning; Data import and export requirements; Scaling requirements.

Their activities cover both infrastructure frameworks and data platform frameworks, which include networking (data transfer from one resource to another), computing (physical resources, operating system, virtual implementation, logical distribution), storage (resources which provide persistence of the data). In comparison with traditional data systems used in statistical operations, Big Data processes require special storage capacities, both on physical storage (e.g. distributed and non-distributed file systems and object stores) and logical storage (e.g. distributed file systems, NoSQLtables, indexed document store, graphs, triple stores) .

Computing Architects skills are related to operating systems, servers, cloud platforms and technologies, orchestration (e.g. Docker and VMware), but also programming languages (e.g. Python and Java).

Once general computing framework has been set up and delivered, a specific role is dedicated to 'Big Data Preparation'**,** with actors being *Information Architects*. These actors are mainly dedicated to steps 'Data Wrangling' and 'Data Representation'.

In comparison with traditional data collection, Big Data processes require specific steps covering both data capture and data preparation because data stored in its raw form typically comprise a very large volume. It also implies data extraction tasks, like data summarization, helping having an understanding of the distribution of the entire dataset and preparing data visualization. List of main tasks include (but is not restricted to):

- Data validation (e.g. format checks) and data cleaning (e.g. eliminating bad records, deduplication)
- Data conversion (e.g. standardization, reformatting, and encapsulating)
- Data aggregation and summarization

Information Architects are also in charge of modelling data: They organize entities/objects into models with combination of attributes to represent information. The data requirements are recorded as a conceptual data model which is used to discuss initial requirements with the business stakeholders. This model is then translated into a logical data model, which documents how Big Data can be accessed in a way that fulfill the conceptual requirements. Note that the effective access to data is handled jointly with *Computing Architects*, because of IT topics necessary to guarantee access, performance and storage needs. Data modeling defines not just data elements, but also their structures and the relationships between them, and is one major sub-task within the data wrangling subphase of the Big Data Life Cycle.

Finally, *Information Architects* handle metadata and various categories of contextual information, including the origins and history of the data, the processing times etc. In addition, data can be described semantically to better understand what the value represents and to make the data machine-operable.

The skills of *Information Architects* are related to technical tools such as database management, programming languages (e.g. Python and Java), ETL (Extract Transform and Load), but also a good knowledge of Data Modeling and Metadata standards (like DDI and ontologies).

Once the data is available and properly managed, another kind of actor can use it to identify its potential use for statistics and build new statistical pipelines: This is the 'Analysis and Visualization' function, which is managed by *Data Scientists*. The latter can achieve the 'Modeling and Interpretation' step and the 'Visual Analyses' step, and may also work together with *Information Architects* in the 'Data Wrangling' and 'Data Representation' steps.

The *Data Scientist* role usually refers to an end-to-end data life cycle. Here we focus on statistical tasks, assuming that e.g. computing capabilities and data management are supported by dedicated actors. (Obviously, organizations may mix these functions into cross-functional jobs, but it serves a purpose to look at these tasks as handled by different roles). Hence, *Data Scientists* should focus on the discovery of the statistical potential of Big Data (i.e. run rapid hypothesis-test cycles) for finding value, doing this by combining two approaches: Analytics and visualization.

Data Scientists have to implement analytical methods using parallel data distribution across a cluster of independent nodes. In relation with the storage paradigm built up by the *Computing Architects* and the modeling of data done by the *Information Architects*, the *Data Scientists* deploy analytics techniques adapted for different types of data access. They manage the following subtasks:

- Human-in-the-loop analytics life cycle (e.g. discovery, hypothesis, hypothesis testing)
- Statistical method to the implemented (e.g. machine learning, deep learning, natural language processing, image processing and neural networks)
- Development of algorithms and their optimization

All of the above subtasks require strong statistical skills as the choice of methodology and methods is key to the results obtained and to the estimates produced.

The analytical steps described are often heavily dependent upon visualization from which the *Data Scientist* forms a hypothesis. Visualization may serve dual purposes: On the one hand, it helps in understanding large volumes of repurposed data and check quickly in which way further exploration should be done ; on the other hand, it helps in creating a simplified representation of the results, suitable for assisting a decision or communicating the knowledge gained (through simple visuals or infographics). Thus, *Data Scientists* also manage the following subtasks:

- Exploratory data visualization for data understanding (e.g. browsing, outlier detection, boundary conditions)
- Explicatory visualization for analytical results (e.g. confirmation, near real-time presentation of analytics, interpreting analytic results)
- Explanatory visualization to 'tell the story' (e.g. business intelligence)

Data Scientist skills are related to both statistical and computing domains. They should master programming and statistical languages (e.g. Python and R) but also web languages used in datavisualization (e.g. javascript) and may also manage core programming languages (e.g. C++ and Java). They have a strong background in statistics (descriptive and inferential statistics, probability theory) and can develop tailored algorithms (e.g. machine learning algorithms).

The functions described here as separate actors can also be linked for common tasks. For instance, the steps dedicated to integrating surveys and register data to the Big Data process, and enriching statistical register with Big Data outputs, are likely to be conducted by all three types of actors mentioned so far:

- *Computing Architects* for the technical aspect (linking capacities and data flows management between Big Data and registries)
- *Data Architects* for the logical aspects (data modeling and pairing between different schemas)
- *Data Scientist* for the statistical aspects (data editing, classification matching)

As for any statistical process, Big Data processes need a high level of expertise in the topic being covered, e.g. economics, society, demographics or health. *Domain Experts* are required so as to design relevant indicators and models that should be produced or applied to the data analytic pipeline. The function is dedicated to design valid statistical models with regard to the topic being studied and the various data being analyzed (Big Data, registry, surveys). The role is defined as safeguarding 'Consistency and Validation' based on domain knowledge. Their deep backgrounds and rigorous training in the disciplines and domain areas linked with the topic covered are essential to guarantee the quality of the overall process. *Domain Experts* must also develop an advanced knowledge of both traditional and new data on the topic (surveys, registers, and of course Big Data), and distil the nuances of the data or the assumptions of the domain they are working in, so as to improve global quality of the deliverables. Driving a Big Data pipeline is an interactive work with many exchanges between *Data Scientists* and *Domain Experts*, in order to validate and make sense of the results. *Domain Experts* should particularly be involved in

validation steps, i.e. validate whether the data at hand can be transformed into something that will actually support the desired analysis. This should cover both looking on the representativeness (is my target population correctly covered by the Big Data pipeline?) and the concepts (can I use the same concepts as the one used in surveys / registers for the phenomena I want to catch through the Big Data pipeline?).

Capacity Providers

The four types of actors described so far (*Computing Architects*, *Information Architects*, *Data Scientists*, and *Domain Experts*) are at the very heart of the IT and Statistical stack within the BD process (see Figure 3). On the top of these actors, note that two specific entities have to be described, which can be internal or external to the organization orchestrating the system: *Data Provider* and *Framework Providers*. They play the general function of 'Capacity Providers'. In a large majority of situations, these functions are played by external actors, which deliver these services to numerous clients. However, in some statistical processes data providers can be internalized. This can be the case for instance with public data hubs, like the Cedefop online job vacancies data collection center, who do web scraping from various places and deliver it to national statistical institutes. To some extent, the Big Data platforms built by international organizations (United Nations, European Commission) for the use of national actors can also be seen as an internalized function of a *Framework Provider*.

The *Data Provider* captures primary data from its own sources or others (e.g. web browsers, mobile devices, sensors). It takes charge of data persistence, often accessed through mechanisms such as web services. In the NIST Big Data Interoperability framework, this kind of data persistence is named 'DaaS' (Data as a Service). Finally, the *Data Provider* defines policy for others' use of the accessed data, as well as what data will be made available. They are likely to be enterprises, for instance web companies, network operators, or commodity agencies with smart captors, but it may also be a public agency.

The *Framework Provider* delivers resources or services to be used by the various actors in the creation of the specific application. The *Framework Provider* manages three activities: Infrastructure frameworks, data platform frameworks, and processing frameworks – each activity being in relation with process actors previously described. For instance, for processing, the *Framework Provider* create services related to the requirements of the *Data Scientist* (e.g. linear algebra, Map/Reduce, or Bulk Synchronous Parallel).

Global Roles

For integrating both internal and external functions that have been described so far, a global function – named *System Orchestrator* – is dedicated to provide the overarching requirements that the system must fulfill, including policy, governance, architecture, resources, as well as monitoring or auditing activities to ensure that the system complies with those requirements. This function is fulfilled with the help of a high-level body within the NSI (e.g. scientific councellors or an advisory

board). Acting as the business owner of the system, the *System Orchestrator* oversees the business context within which the system operates, including specifying the business goals, the *Data Provider* contracts, leading negotiation with *Framework Provider*, and designing the hiring plan for human resources (internal needs for *Computing Architects*, *Information Architects*, and *Data Scientists*).

The *System Orchestrator* is acting on behalf of the *Statistical Institutions,* from which they inherit an operational delegation for designing and managing the brand new Big Data process. *Statistical Institutions* are in charge, at the very beginning of the story, i.e. the 'Specify needs' business function. They are also responsible for the availability of general resources, and also for connecting the Big Data process to generic functions like 'Support Statistical Production' and 'Dissemination' which are not really different whether these statistics rely on Big Data or not.

Audit, Control, and Compliancy Actors

Three transversal functions are also to be specified: Overall Data Management, Security and privacy, Scientific and data relevance.

Using Big Data changes the approach. Instead of looking for the data to achieve a survey or a statistical operation which are precisely defined before, the data is now collected in order to aim a goal, which is less precise than before. Thus collected data need to be consistent with each other. These changes implies the reinforcement of the *Chief Data Officer* (abbreviated CDO). He makes easier the way to access the data, and also he spots useful data. Especially in the 'New Data Sources Exploration' business function the *Chief Data Officer* is the key to a data centric process. In charge of the overall data management for the NSI, the *Chief Data Officer* is expected to be an internal actor.

In the future, the 'New Data Sources Exploration' function should be more guided by user needs. As there will be more new data sources, the identification and prioritization of user needs will be the key to focus on exploring the sources with the biggest potential. The *Chief Data Officer*, supported by actors in charge of external relations, will monitor the new needs in terms of statistical information. This monitoring could be based both on traditional channels dedicated to user needs (e.g. customer satisfaction research, online surveys, forums) but also on an examination of emerging topics that may be brought by new data sources. This is the big topic of Big Data: It is now possible to address former requests just as much as it creates new requests. Future expectations of civil society are thus partly shaped by the availability of new sources. In relation with the *Chief Data Officer*, listening to potential needs may mobilize *Domain Experts* and *Data Scientists*, who can appreciate together how these materials address social and economic matters, and new topics that can be covered (see Figure 4).

The *Data Protection Officer* (abbreviated DPO) responds to the need of building a set of safeguarding measures for all types of Big Data platforms. Those safeguarding measures include security and privacy controls to protect the critical and essential operations and assets of organizations and the personal privacy of individuals. The ultimate objective is to check the information system meets the

security and privacy requirements that have been defined by laws, as well as complementary measured decided by the *System Orchestrator*.

The continual improvement of the statistical process cannot be a reality without some exchange outside of the statistical community. This is why S*cientific & Academic Communities* act in the process to improve it and play a function of scientific and data relevance. The mutual exchange of knowledge and expertise is a good way to progress, both in methodological topics (algorithms) but also about data consistency (both input and output data from the Big Data process) and its ability to reflect the economic and social phenomena that the statistician seeks to study.

Further Actors Linked to (but not involved in) the Process

After explaining the actors directly involved in the process, we now finally look at different kinds of actors that are present in Figure 3 but not directly involved in the process.

First to be mentioned are *Other Statistical Institutions*. Obviously these may be ONAs which is short for Other National Authority, where *other* refers to the fact that they produce statistics but is not an NSI. Other examples of *Other Statistical Institutions* are the ESS or central banks. In a statistical process using Big Data, *Other Statistical Institutions* are naturally linked to the NSI works. First of all because the statistical institutions (will) share reference frameworks, processes, models or tools. So this kind of actors are interested in sharing experience first, and in comparing methods and tools to improve upon them.

In addition *Other Statistical Institutions* deal with using Big Data in statistical processes as they may not be able to process the data themselves, but are mainly interested in the results of analysis. In this case, we can mention as an example the use of health data: A health research center cannot use a big amount of data and is looking for a capability which may be available by a statistical office for example.

The next example refers to another way of being linked to the process for *Other Statistical Institutions*, namely as 'Sponsors'. In this case, *Other Statistical Institutions* or (for example) research centers, academies or companies support a team working on a specific subject, especially at the beginning of the process when the idea is becoming reality. Such sponsors can act by giving financial support, human resources or computing capabilities. Once a potential new statistical process using Big Data is identified, the sponsors may provide support (in the form of IT, data, experts) to making a proof of concept.

Another highly relevant actor is the *Government*, which may be perceived as either the national government or some supranational entity like the European Union. The *Government* is linked to statistical processes using Big Data for at least two reasons. The first comes from the legal part of the institution. One of the goals of the laws is to protect the citizens as an individual, especially regarding the privacy policy and ethics. The laws, issued by the *Government* and the deputies in order to protect the individuals, give some rules and constraints when using Big Data, even in statistical processes.

The second link to the process for the *Governments* is a result of the expression of new needs. As statistics are a tool in order to help public policies, they have to suit with the needs. In this way the *Government* may need the same statistics but with a higher degree of actuality. This change in needs sometimes means the the NSI has to change the methodology or the precision, but may also imply to think about a whole new statistical process using Big Data. The other change coming from the *Government* deals with the human (and financial) resources given to the NSI. When a survey cannot be realized because of a lack of human resources, one solution consists in using Big Data.

The *Medias* are also an actor related to statistical processes using Big Data. *Medias* link statistical offices and *Citizens* by publishing the results of statistical processes. By publishing such results, *Medias* may also give explanations about the methods or the sources that were used during the process. However, *Medias* can also act with new needs, in the same way as the sponsors. In order to realize the publications, *Medias* use a lot of data sources. Working with these sources may give some ideas for new processes using Big Data especially in order to check the information.

Last but not least, the *Citizens* are related to the statistical processes using Big Data in different ways. Using their mobile phone or their connected watch, paying with their credit card are some examples of the implication of the citizens in processes using Big Data. Without these actions, there would not be any Big Data. Using connected objects and participating to the process in a passive way is one thing. Another one consists in sharing data. A part of the *Citizens* also contribute in an direct manner by sharing data they are collecting as a hobby, e.g. genealogical research or weather studies. These data can contribute to statistical processes as an input or even as citizen science.

4.2 Actors and Big Data Life Cycle

The iterative dimension in processes based on Big Data, i.e. alternating exploration phases and production phases, also leads to numerous interactions between actors. In this section, we describe these interactions by gradually extending from two basic business functions (‘Specify Needs’ and ‘New Data Sources Exploration’) to the full Big Data Life Cycle.

The development of a complete Big Data process arises from the encounter between the expression of a need (e.g. to compile statistics on a new economic or social phenomenon) and the identification of a possible solution among Big Data Sources to meet this need (see Figure 4). The specification of needs emanates from the *Statistical Institution*, which itself builds upon expectations from *Citizens*, *Medias*, and *Government* (though not necessarily in that order). The needs are discussed in broad within the statistical community, but the need specification phase is formalized by *Statistical Institutions* on the basis of stakeholder expectations. In parallel, the first explorative analysis on which sources may address these emerging needs is as the heart of the work done among NSI by *Domain Experts* (review of literature, analysis of current survey and administrative data) and by *Data Scientists* (new data sources exploration), in relation with *Scientific & Academic Communities*.

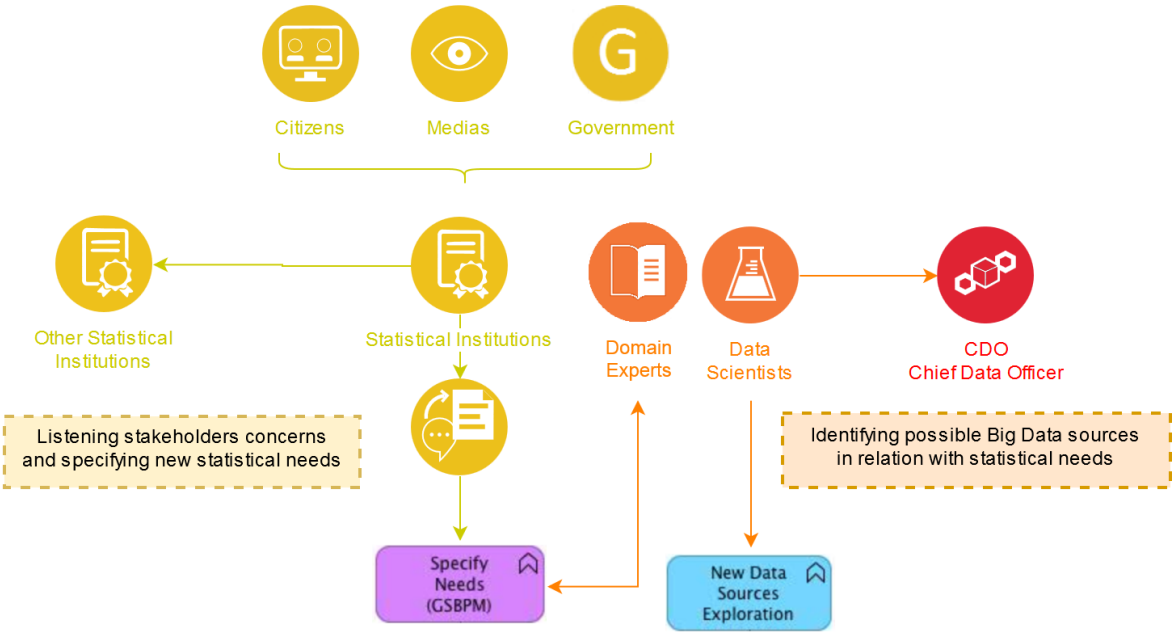

Figure 4: Specify needs and data exploration steps & Actors

Once a Big Data source has been identified as possibly providing a partial solution to the new statistical needs, a preliminary work to define the conditions of access to the data is initiated by the *Chief Data Officer* and the *Data Protection Officer*. This happens in a relation with the *Data Provider* (which may be private companies), so as to define the contractual and legal framework for experimental access to data. At this stage, data access is limited to exploration and prototyping purposes (see Figure 5). The technical aspects of data access are addressed by *Computing Architects* together with the *Framework Provider* (e.g. setting up a testing platform with the appropriate database systems and corresponding software).

Once a data source is available for development and information discovery, *Information Architects* are in charge of preliminary data cleaning. They can also initiate a first step of data representation, even if the discovery phase will lead to regular reworking of the data contours retained in a production process. This representation will lead to a conceptual data model that later will be used to discuss with *Data Providers* the final needs for accessing data sources.

Basing their work on data representation done by *Information Architects*, *Data Scientists* deploy analytics techniques adapted for different types of data access. They set up analytical treatments based on working hypotheses and shape the treatments using methods adapted to the nature of the data, but also to the type of model tested. One particularity of the Big Data Life Cycle may rely on the frequency of data, i.e. when the Big Data source is a continuous flow of data rather than a static database, this call for specific statistical treatments. This is why *Data Scientists* may work iteratively with *Computing Architects*, tuning the statistical pipeline that is being tested according to, firstly, the volume and frequency of data, secondly, the kind of method that is applied to data.

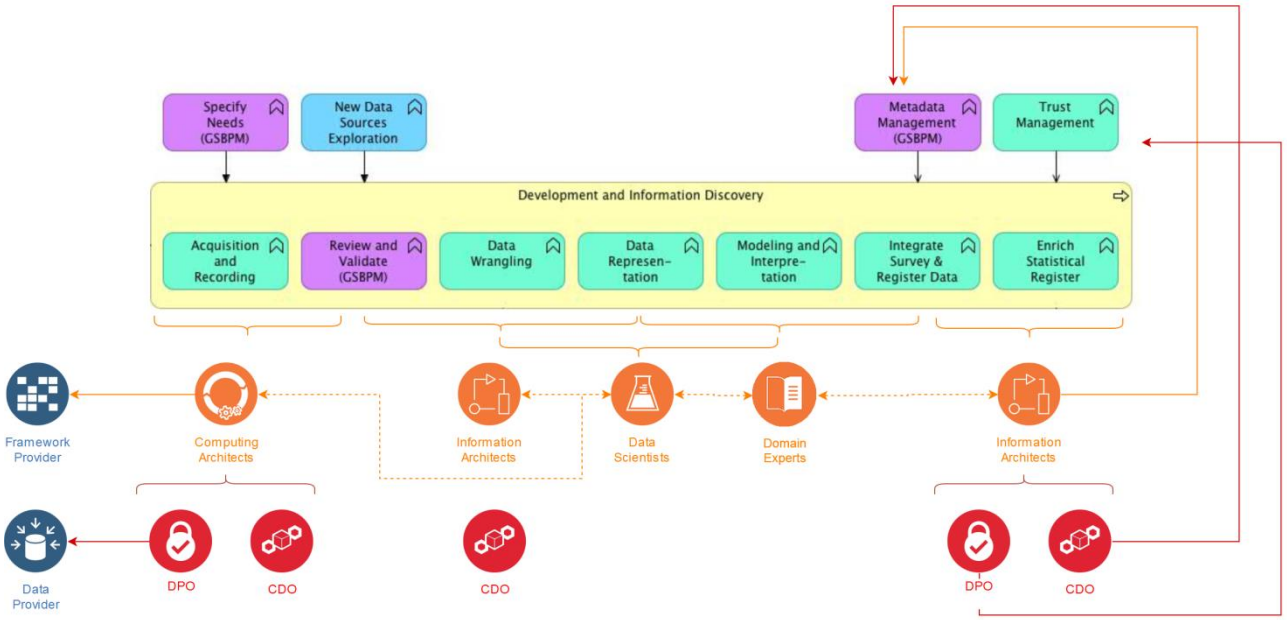

Figure 5: Development and Information Discovery steps & Actors

Data Scientists also work in concert with *Domain Experts*, who design statistical models inspired by economic and social theories that have been developed on the topic, and help to appreciate the likelihood of results obtained by data scientists.

Whenever Big Data sources can be merged and enriched with other statistical sources (like survey or register data), the examination of the legal conditions of matching (ultimately compliance with GDPR) is examined in relation with the *Data Protection Officer*. Source integration procedures are also discussed by the *Information Architects* with the *Chief Data Officer* to ensure optimized management of the pairings (avoiding data duplication, looking at data integration within data warehouse or data lake, filling metadata information).

Modelling and interpretation done jointly by *Data Scientists* and *Domain Experts* give rise to visualizations and first results which, besides their contribution to appreciate the richness of the data and to test theoretical hypotheses on the phenomena data are covering, help defining the final outputs of the statistical process (see Figure 6). In the case of Big Data processing, final outputs may be more than just summary indicators, and include infographics and interactive data visualization modules for end-users. This includes especially explicatory visualization to 'tell the story'. This may imply that *Domain Experts* check that explications are congruent with main theories on this topic. Shaping outputs may again include common work with *Computing Architect*, so as to build not only processing pipelines but also delivering pipelines from data sources to final outputs may will be compliant with technical constraints, hence a specific work on the analytical framework (configuration of the cloud platform). Particular attention must be paid to the respect of the confidentiality in the data visualization that will be delivered to the final user, in particular if interactive data visualization includes functions of data exploration.

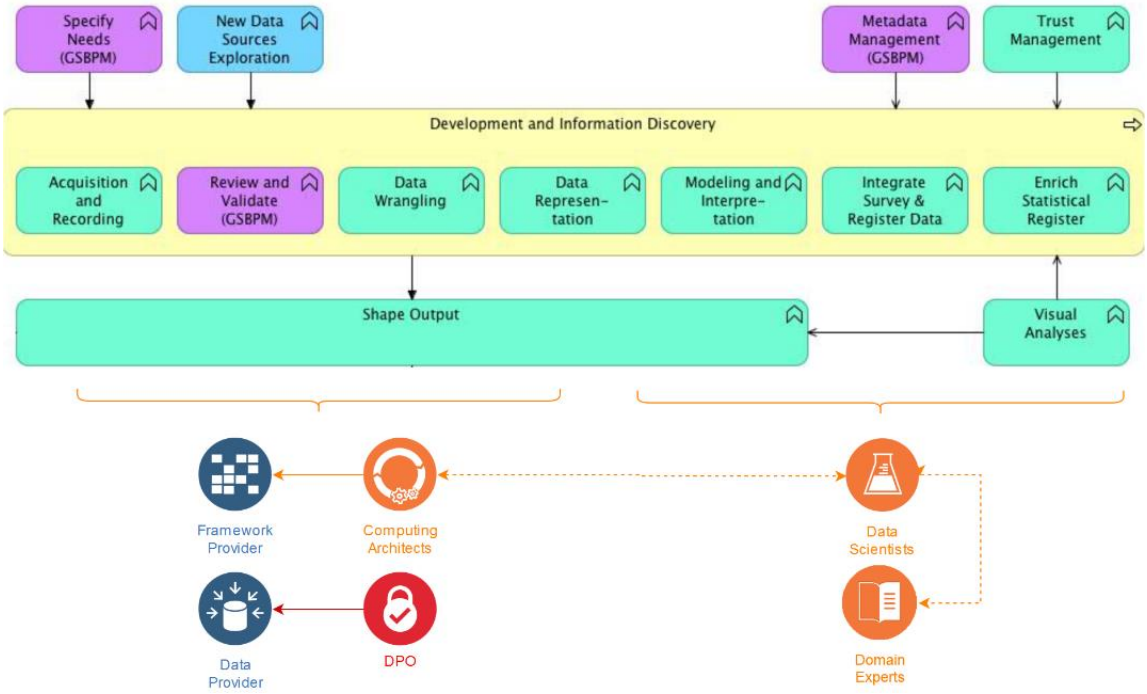

Figure 6: Shape Output Step & Actors

The tasks described at this stage are subject to an improvement loop, which may lead to modified data deliveries, up to the final (and stable) definition of the statistical process and deliverables. The evaluation of the process and results can be conducted within the statistical community (e.g. with the other NSIs of the ESS) but also in relation with the academic communities, in relation with Domain Experts (see Figure 7).

The quality supervision associates *Information Architects* and the *CDO* to ensure an optimal integration of the new process into the entire management and data processing system of the Institute. The final contracts of the data delivery is negotiated by the NSI with the data supplier, while the DPO guarantee the compliance with the legal framework. Similarly, the final production pipeline is defined in conjunction with the Framework Provider. The dissemination stage makes it possible to make the results available to the target audiences (media, authorities, general public).

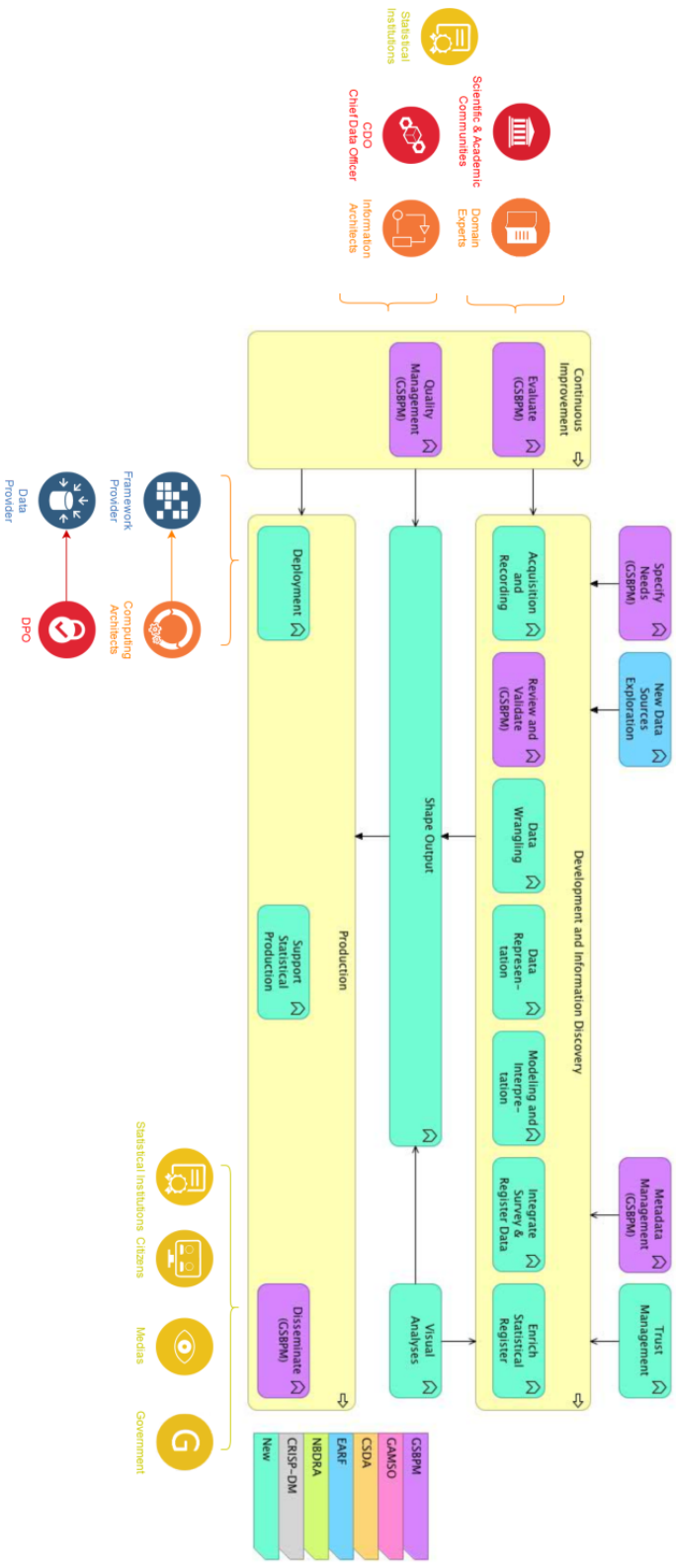

Figure 7 : Continuous Improvement Loop & Actors

4.3 Business Principles and Big Data Life Cycle

The Business Principles introduced in section 2 and the Big Data Life Cycle introduced in Section 3.3 combines nicely in the sense, that specific business principles are typically applicable at certain points within the Big Data Life Cycle. Thus, in Figure 8 we have attempted to overlay the life cycle model with the business principles. Not surprisingly, most business principles apply to the main process in the life cycle model (i.e. 'Development and Information Discovery). As a result, we recommend that at each evaluation (see the 'Evaluate' business function) of this phase, the business principles are checked and validated.

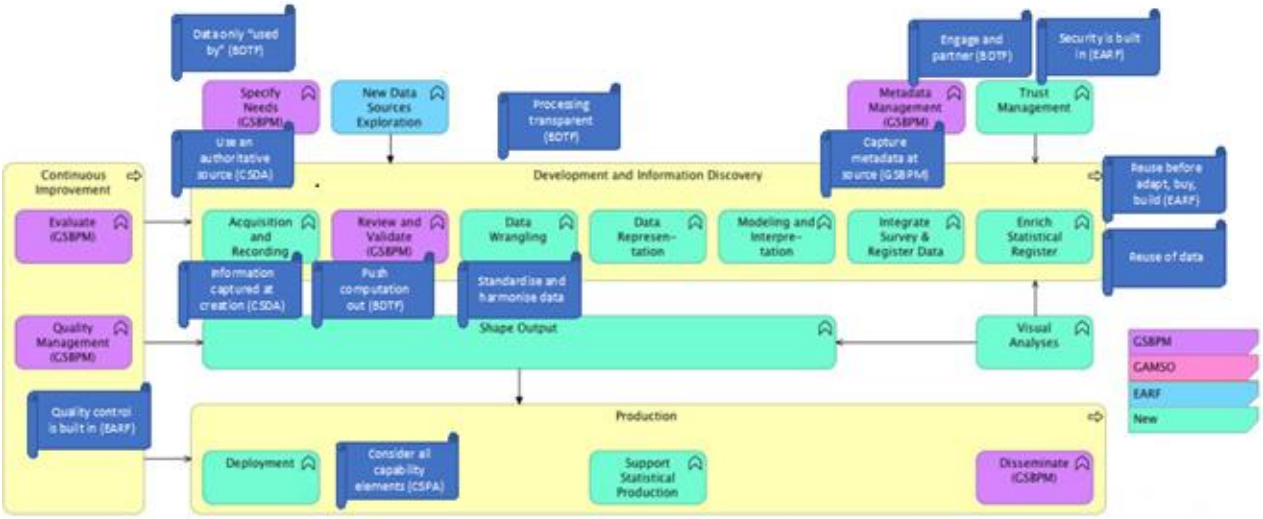

Figure 8: BREAL Mapping of Business Principles on the Big Data Life Cycle

5. Conclusions and Future work

This document describes the artifacts produced so far of the Business Layer of BREAL. The artifacts will be discussed and validated within the ESSnet Big Data II pilots (i.e. work packages B, C, D, and E) and at other venues identified as relevant to this purpose.

The next steps are related to the design of the Information Architecture (IA) and the Application Architecture (AA). To this scope, implementation work packages can be grouped into two main clusters, namely:

- Web Intelligence: Online Job Vacancy (WP B) and Enterprise Characteristics (WP C)
- Sensor Data: Smart Energy (WP D) and Tracking Ships (WP E)

As an example of how the Information Architecture related to the Web Intelligence cluster could look like at a very high level let us consider Figure 9 . Three main layers are shown, inspired by the

'hourglass model' proposed for Trusted Smart Statistics[7], namely a raw data layer, a convergence layer, and a statistical layer.

- The Raw Data Layer includes data that are scraped from the web as they are, e.g. enterprise website and job portals. Please consider that at this stage the IA just identifies the concepts, and no detail is provided on formats or other technicalities.
- The Convergence Layer contains data represented as units of interest for the analysis, hence enterprises with identification and core attributes and OJV[8], with identification and core attributes as well.
- The Statistical Layer includes those concepts that are the targets of the analysis, i.e. Ecommerce or Skills.

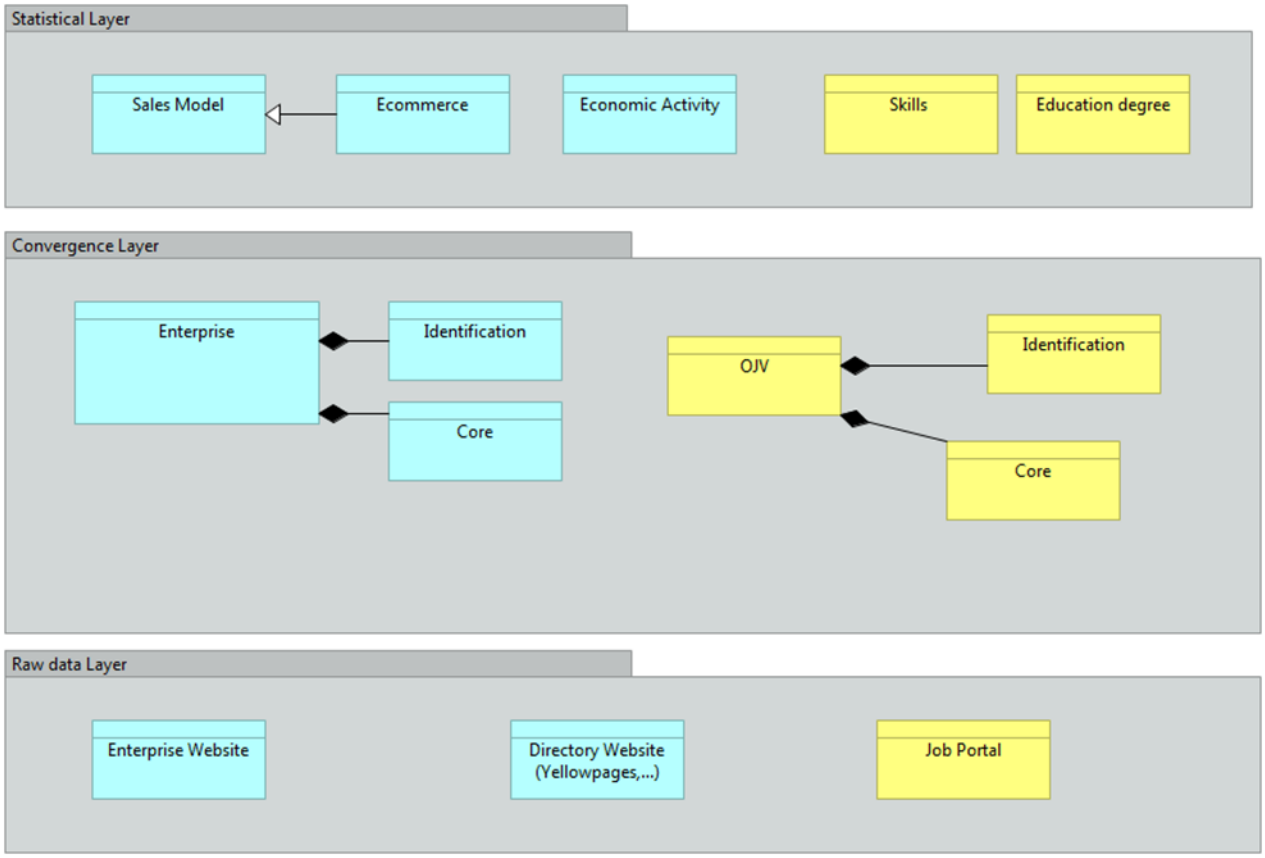

Figure 9: Draft Information Architecture (IA) for projects within the Web Intelligence cluster

An example of what the AA could look like for the Web Intelligence cluster is shown in Figure 10. In particular, the application components and services implementing the business function Acquisition

[7] https://zenodo.org/record/3066061#.XS3bnPIzaUl

[8] Online Job Vacancy portal.

and Recording are shown. Two services are exposed by the Web Scraping application component, namely ad-hoc scraping and generic scraping. In addition, two services are identified for the application component API Querying, namely Website retrieving and API Access.

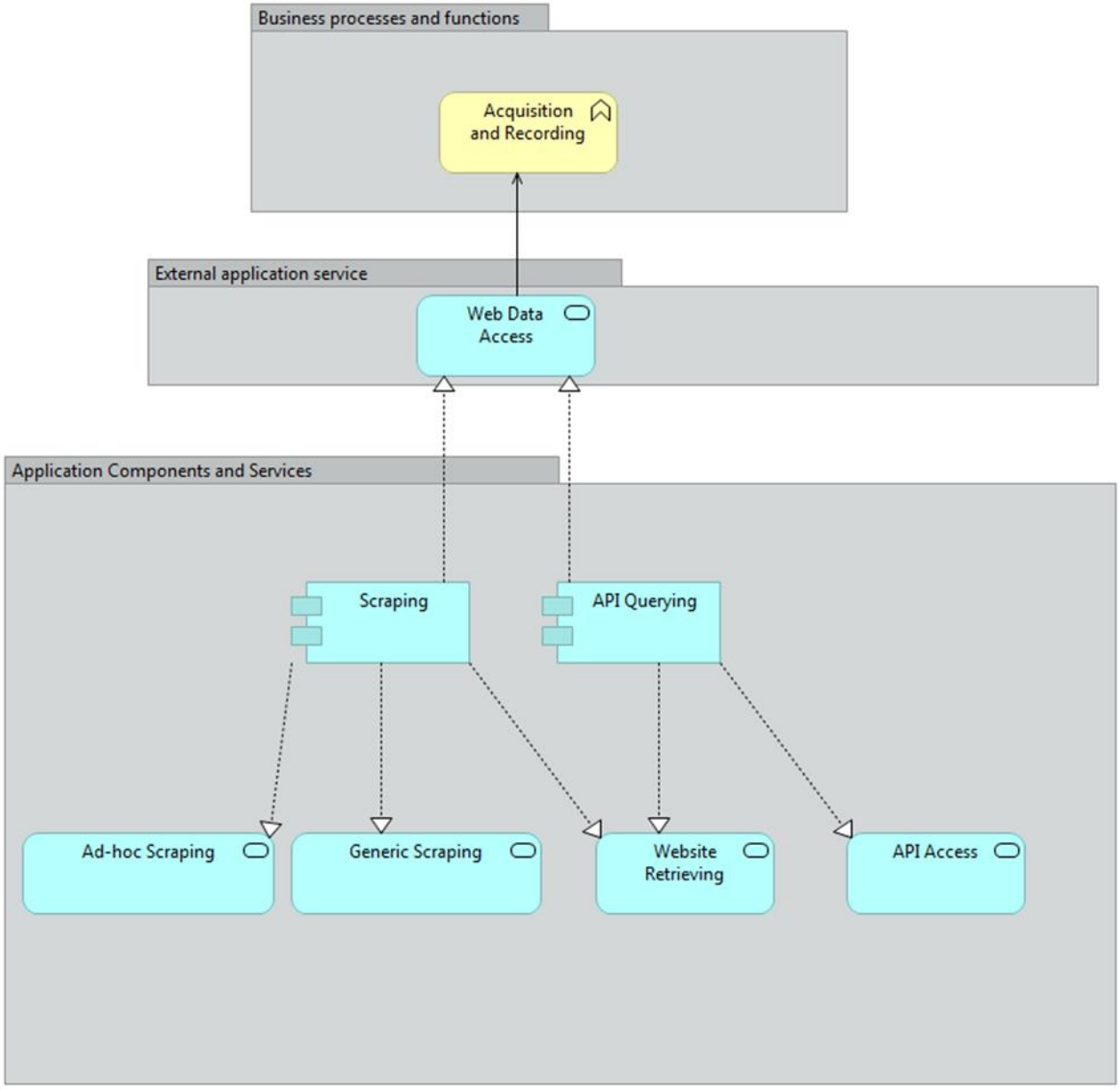

Figure 10: Draft Application Architecture (AA) for projects within the Web Intelligence cluster.

Concerning future work, our plans are, by the end of 2020, to fully define BREAL.

In particular, by that time BREAL will include: (i) a general Business Layer, (ii) an Information Architecture Layer and an Application Architecture Layer for Web Intelligence and (iii) an Information Architecture Layer and an Application Architecture Layer for Sensor Data.

References

BDTF EUROSTAT, Big Data Task Force, Towards a Reference Architecture for Trusted Smart Statistics, Fabio Ricciato

CMF Common Metadata Framework

CRISP-DM Cross-industry standard process for data mining

CSDA Common Statistical Data Architecture, version 1.0

CSPA Common Statistical Production Architecture, version 1.5, December 2015

EARF ESS EA Reference Framework, version 1.0, August 2015

GAMSO Generic Activity Model for Statistical Organisations, version 1.2, January 2019

GSBPM Generic Statistical Business Process Model, GSBPM, version 5.1, January 2019

GSIM Generic Statistical Information Model, version 1.2, 2019

NBDRA NIST Big Data Interoperability Framework: Volume 6, Reference Architecture, version 2, June 2018

SPRA ESS Statistical Production Reference Architecture, version 0.4, August 2015