ESSnet Big Data II

Grant Agreement Number: 847375-2018-NL-BIGDATA

https://ec.europa.eu/eurostat/cros/content/essnet-big-data-0_en

Work Package F

Process and architecture

Deliverable F2

BREAL: Big Data REference Architecture and Layers

Application Layer and Information Layer

Version 31-03-2021

Prepared by:

Frederik Bogdanovits, Tauno Tamm (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 3
1. BREAL Application Layer 4
1.1 Development, Production and Deployment Application Services 4
1.2 Support Application Services 12
2. BREAL Information Layer 22
3. BREAL Operational Model 25

Executive Summary

This document follows the document "BREAL: Big Data REference Architecture and Layers – Business Layer", where the Business Layer of BREAL was presented and focuses in particular on the Application Layer and the Information Layer of BREAL, i.e. the two architectural layers describing the "how" implementing the "what" defined by the Business Layer.

The **BREAL Application Layer** is presented in Section 1 and it consists of a set of generic application services, proposed with the purpose of showing how the identified business functions could be implemented. The proposed generic services are classified based on the specific business functions they implement, namely:

- Development, Production and Deployment Application Services as described in Section 1.1;
- Support Application Services as described in Section 1.2.

The **BREAL Information Laye**r is described in Section 2 and it consists of the three sublayers described by the "hourglass model" proposed for Trusted Smart Statistics, namely a raw data layer, a convergence layer, and a statistical layer. 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.

In addition to the Application Layer and the Information Layer, a further content of the document is a proposal for an **Operational Model** indicating deployment options for data, services and platforms. Specifically, the BREAL Operational Model presented in Section 3, includes three types of sharing, namely: sharing data, sharing application services and sharing infrastructure platforms.

The proposed Application Layers and Information Layers are generic in the sense that they are not targeted towards a specific big data project or big data 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 II are described in a set of **Solution Architectures** presented in the **WPB, WPC, WPD and WPE Appendixes**. In other words:

- The generic layers of the overall architecture have the purpose of identifying the services and the data that can be used in any big data project.
- The solution architectures take into account the specificity of each project, though maintaining the compliance to the generic level.

In this way, the use of BREAL Application and Information Layers can serve the following purposes:

- Ensuring the comparability and shareability of solutions through the generic layer, but also
- Providing enough details for reusing specific project solutions through the solution architectures.

1. BREAL Application Layer

The purpose of this section is to describe the Application Layer of BREAL, which consists of services implementing all the business functions defined (see Figure 1). Note that no special meaning should be attached to the order and relative size of the boxes representing the individual business functions, only the grouping (upper/lower) and color are meaningful. The business functions are grouped within 'Development, Production and Deployment' (section 1.1) and 'Support' (section 1.2).

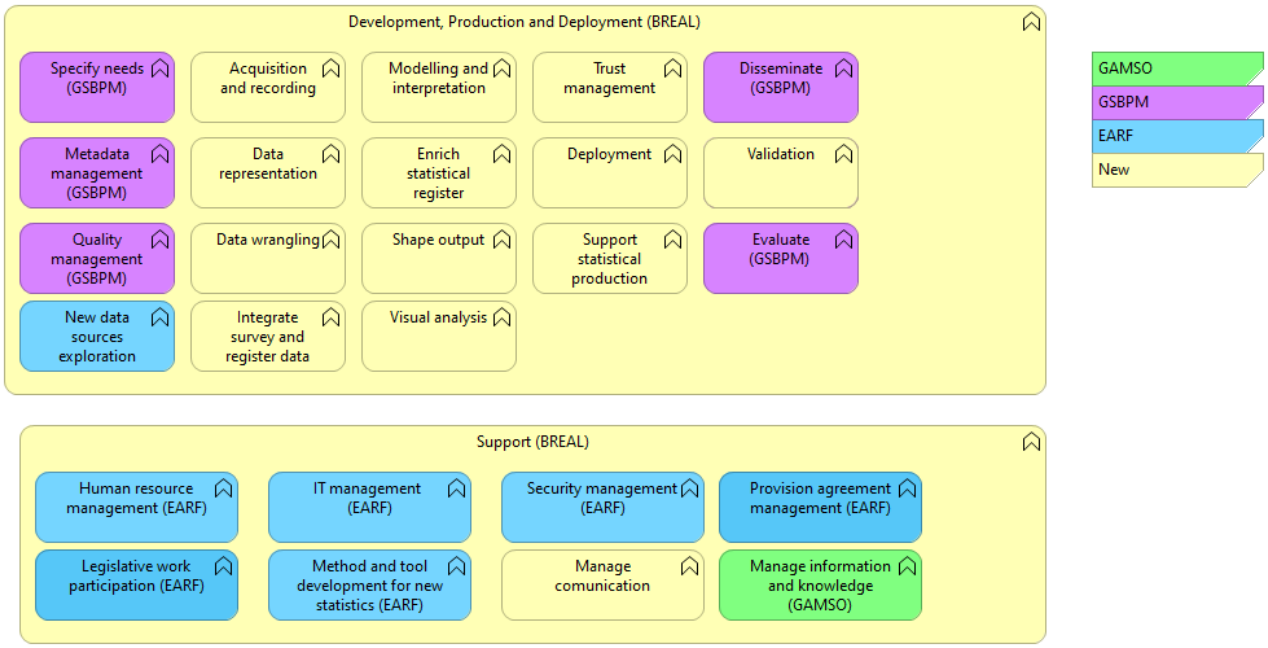

Figure 1: BREAL Business Functions

1.1 Development, Production and Deployment Application Services

The business functions of BREAL within the 'Development, Production and Deployment Application' as shown in Figure 1 can also be seen also according to their role in the big data life cycle as shown in Figure 2.

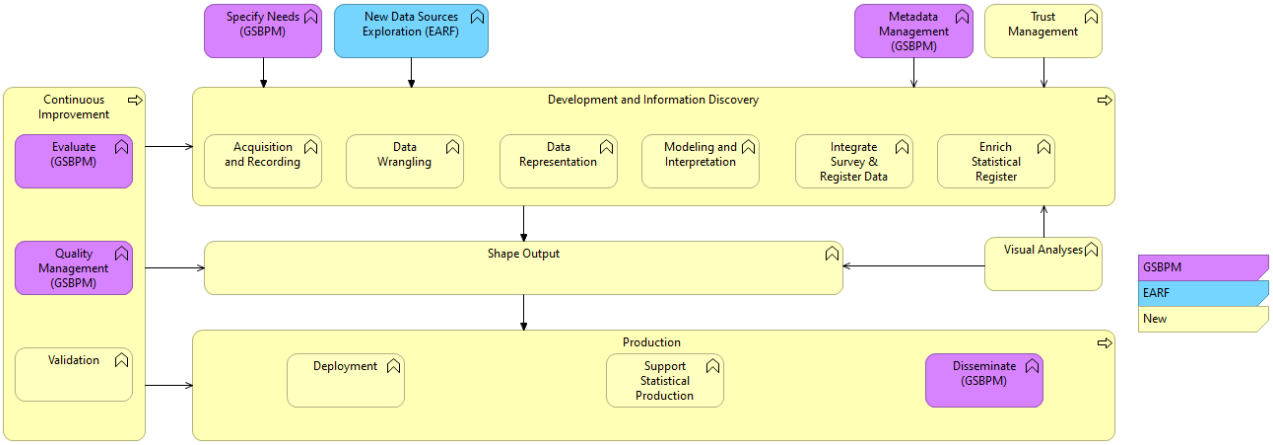

Figure 2: BREAL Business Functions in the Big Data Life Cycle

For some of the business functions, it is relevant to identify some Application Services, i.e. logical components belonging to a software application layer that implement the specific functions. For the remaining ones, instead, we choose not to provide any specification in terms of application services, mainly because of two reasons:

- Dependence on internal NSIs technical and infrastructural policies (e.g. BREAL Deployment Business Function)
- Dependence on internal NSIs production processes (e.g. BREAL Support Statistical Production).

The BREAL functions for which related application services have been specified are shown in Figure 3, with the business functions in yellow and the application services in blue. The application services have been formulated by researching the solutions for WPB through WPE. Even though we tried to provide a complete list, it might need be extended with future developments.

The detail of each application service is explained below.

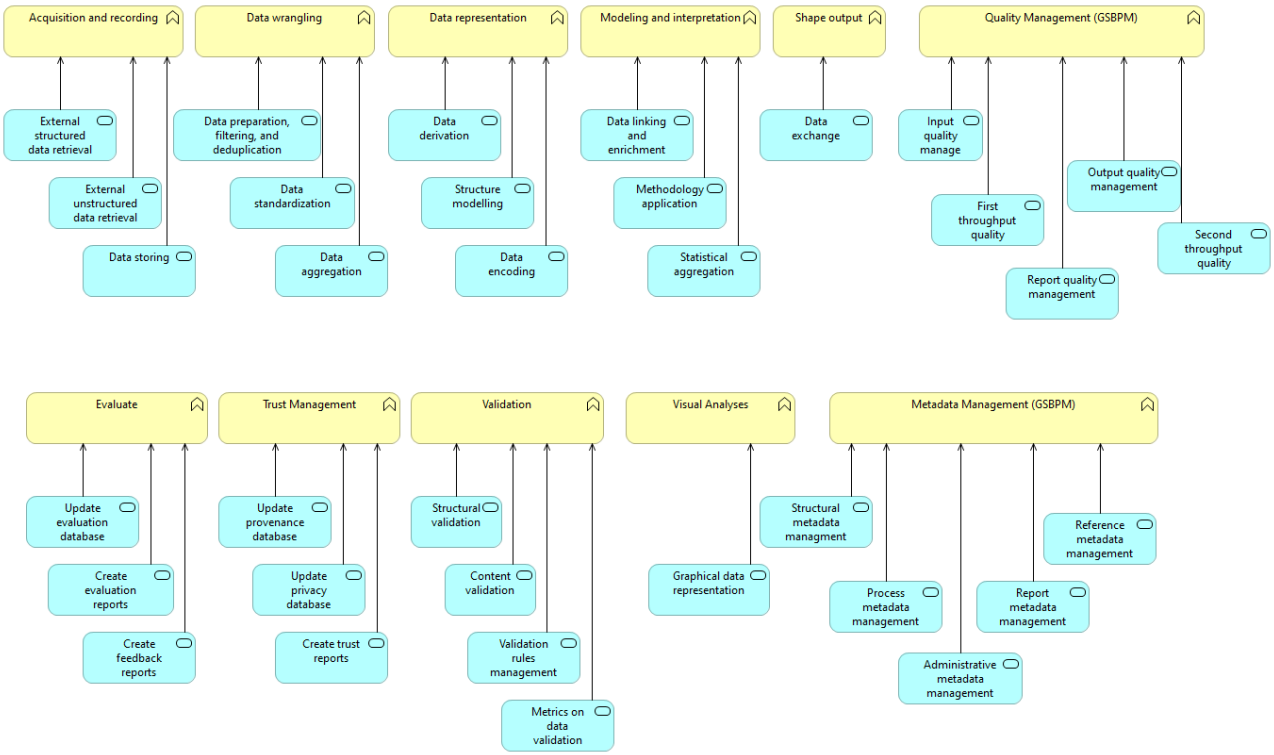

Figure 3: BREAL Application services for Development, Production and Deployment business function

Acquisition and Recording

This function controls two aspects necessary to create a basis of data to work on – data can be retrieved from a variety of data sources and there should be the ability to store them in any kind of data storage. The following services serve the implementation of this function.

External structured data retrieval: this service retrieves structured data from external sources. Structured data have a well-defined structure, usually in the form of tables. Sources for this service can be APIs, tables on websites or databases.

External unstructured data retrieval: this service retrieves unstructured data from external sources. Unstructured data have very little or no structure, for example text or images. Sources can be websites, satellite images, news articles, social media etc.

Data storing: the service provides storage options for data acquired via the retrieval services. Storage provided can be relational or NoSQL (document based, key-value storage, graph based etc.).

Data Wrangling

The data wrangling function manages the process of transforming and manipulating the raw data into a data format and volume that is easier to handle by the following services.

Data preparation, filtering, and deduplication: the service is responsible for bringing data in a machine-usable format, filtering relevant parameters out of unstructured data and removing duplicate data entries commonly occurring e.g. with web scraping.

Data standardization: data are transformed in order to make them comparable to data from other sources. The data are for example scaled to the same units; text is converted to the same encoding.

Data aggregation: data points are aggregated into larger intervals or groups in order to shrink the amount of data when the detailed resolution of the raw data is not necessary, and the available computational power is limited.

Data Representation

This function is responsible for adding context and structure to the processed raw data. This is necessary for function "Modelling and Interpretation" which can use this additional information for applying statistical methodology. The following services serve the implementation of this function.

Data derivation: the service is responsible for deriving additional variables from the data – e.g. deriving a birth date from an age variable.

Structure modelling: the service derives and adds structure to previously weak- or unstructured data by modelling categories or other structuring elements.

Data encoding: the service is responsible for adding meta-data like formats or classifications to data.

Modelling and Interpretation

This function applies statistical methods to the pre-structured data from the previous function in order to add meaning to these data. The following services serve the implementation of this function.

Data linking and enrichment: the data in the Big Data pipeline is linked and enriched with data from other sources (conventional statistics, other Big Data-derived statistics, etc.) in order to add more meaning to it.

Methodology application: domain-specific methodology (predominantly machine learning, but other methods are also in scope of this service) is applied to the data in order to derive information relevant for specific statistical products from it.

Statistical aggregation: results from the modelling and linking steps are aggregated in order to fit to the basis of other statistical outputs. This can for example be an aggregation on the timescale to make the data adhere to the same measurement intervals as other data sources.

Shape output

The shape output function is realized by one service. It takes the output of the preceding services, transforms it according to the necessary specifications and integrates it into statistical production.

Data exchange: The service is responsible for transforming and shaping the output of the Big Data pipeline in order to integrating the resulting information and models into statistical production.

Evaluate

This function is in charge of storing, updating and reporting information relevant for the evaluation of the process underling the execution of a Big Data pipeline. Some information is intended to improve subsequent executions by a feedback mechanism.

Update evaluation database: considering a specific execution of an application process and related services, this service collects information that centrally stores in a database for enabling specific evaluation functions.

Create evaluation report: this service accesses the evaluation database and creates specific reports that can include effectiveness evaluation (e.g. measurements of specific quality indicators), and efficiency evaluation (e.g. execution time for specific service instances).

Create feedback reports: the service is responsible for creating specific operational indications deriving from the analysis of evaluation reports. These can include change of parameters, revision of datasets, etc. The operational indications are meant to be taken into account in the subsequent execution(s) of the big data pipeline as a whole or of single services of the pipeline itself.

Trust management

This function is in charge of managing two important dimensions of trust, namely transparency and privacy. In combination with the Quality Management function, this function is responsible for tracking and reporting all the relevant trust information that are produced through the Big Data life cycle. The specific information related to the transparency are stored in a 'provenance database', while the 'privacy database' is in charge of collecting all the relevant information concerning privacy protection. The following services serve the implementation of this function.

Update provenance database: this service collects information that centrally stores in a database related to provenance information as produced during the execution of a Big Data pipeline. Provenance information can regard (see W3C Prov Model https://www.w3.org/TR/prov-dm/): entities (data) and how they are derived from each other, activities on entities and relationships among activities, agents to which activities are associated and entities are attributed.

Update privacy database: this service collects information that centrally stores in a database related to privacy measures that have been put in place during the execution of a Big Data pipeline. Possible privacy leaks are also stored within this database.

Create trust reports: this service accesses the provenance and privacy databases and creates specific reports that can have different uses like internal auditing, to trace how specific outputs have been actually produced and identify potential improvements, and external trust evaluation, to provide external users or data providers a transparent and auditable view of big data based production pipelines.

Validation

Data validation is a decisional procedure ending with an acceptance or refusal of data as acceptable. The decisional procedure is generally based on rules expressing the acceptable combinations of values. Rules are applied to data. If data satisfy the rules, which means that the combination expressed by the rules is not violated, data are considered valid for the final use they are intended to. As explained below, the Validation function can be implemented by the services Structural Validation, Content Validation, Rules Validation Management and calculating Metrics on the Validation Process.

Structural validation: for verifying the technical integrity of a data set, i.e., its consistency with the expected IT structural requirements that make possible its acquisition, recording and treatment by the existing services.

Content validation: checks the consistency of the data. This can be performed only within the data set, with other data sets within the same or in separate domains, within the same data source or between different data sources, and with data from other data providers.

Validation rules management: as statistical processes age and mature, the number of validation procedures and rules tend to grow organically, generating a need for maintenance. The validation rules management is of paramount importance to guarantee the performance of the validation procedures. They are responsible for managing the following aspects of the validation rules:

- Complexity, i.e. information needed to evaluate a rule, computational complexity and cohesion of a rule set.
- Feasibility, i.e. finding inconsistencies that compromise the evaluation of rules.
- Redundancy, i.e. implementing methods for redundant rules to be removed.
- Completeness, i.e. finding the balance between too few rules (resulting in non-valid data not being identified as erroneous) and too many rules (resulting in valid data being identified as erroneous).

Metrics on data validation: to analyse the performance of the validation process in respect to the data. These metrics are particularly useful for tuning the parameters of the data validation rules, providing quantitative information to assist the maintenance and monitoring, as well as measuring the quality, of a data validation procedure.

Visual analysis

The purpose of a visual analysis is to recognize and understand phenomena represented through graphical representations of data. It is also defined as the process for reaching a judgment about reliable or consistent effects by visually examining data. This analytical and graphical reasoning is supported by static or interactive visual techniques, which require interdisciplinary science integrating techniques from visualization and computer graphics, statistics and mathematics, knowledge representation, cognitive and perceptual sciences. Human factors in particular (e.g. perception, intuition, collaboration) play a key role both in the human-computer interaction as well as in the decision-making process. The visual analysis can also serve functions prior to reaching the final dissemination products:

- *Cleaning:* visual analysis can be a way to check plausibility of data and to detect possible errors. In many cases, available information from Big Data can be heterogeneous, conflicting or with many invalid or missing values. The use of data visualization may greatly help detecting abnormal data values through a visual "browsing" of data. The same kind of analysis can be produced to identify partially missing or duplicated values.
- *Exploratory data analysis:* visual analysis helps analyzing data to find implicit but potentially useful information. At the beginning, the analyst has no hypothesis about the data so the aim is to start an interactive and usually undirected search for structures, shapes and trends in data. Later, the same techniques can be used to test the likeliness of one or more hypotheses or to recognize the appearance of searched phenomena, quickly exploring multiple interacting data characteristics (variability, mean levels, trends, intercepts).
- *Confirmatory data analysis*: one or more hypotheses about the data are used as a starting point. The process can be described as a verification of these hypotheses. As a result, visualization can either confirm these hypotheses or reject them.

These cases are all considered as special cases of a general service, named Graphical data representation.

Graphical data representation: visual analysis helps in transferring to final users a graphical proof of a hypothesis or a graphical synthesis of a phenomenon. Filtering, aggregation, compression, principal component analysis or other data reduction techniques are needed to reduce the amount of data to be represented: the aim is efficient and effective communication of the results of an analysis.

Metadata Management

Creating, defining, and managing the metadata about a big data source will enable access and exploration of its content. To this end, several services must be in place to guarantee adherence to the metadata standards established:

Structural metadata management: describing how the components of an object, file, or element are organized and how the objects are identified. This service must articulate with the First Throughput Quality Management Service described in the Quality Management Services.

Process metadata management: describing how components are put together. Whenever possible the service must collect the active metadata during processing. Regardless of where the raw data is processed in house or off premises the process metadata must be kept and managed to guarantee not only the quality of the data but also the reproducibility of the methods used to treat it. This service may need to articulate both with First and Second Throughput Quality Management Services, described in this section.

Administrative metadata management: comprehends both the rights and preservation metadata management. The first part, who accesses which data where must be articulated with the Trust Management Service, described in this section. The second function of the service is to manage what information to preserve, when and how to save it, pull it and so on.

Reference metadata management: may expand or act upon the already existing Metadata Management system to relate the concepts and objects of the Big Data source with the statistical units, variables and classifications as well as the rules used for that purpose.

Report metadata management: creating specific reports for the first four Metadata Management Services described in this list.

Quality Management

Quality management is the act of overseeing all activities and tasks that must be accomplished to maintain a desired level of excellence. In statistical production, the quality management services target the well-established dimensions of quality: Relevance, Accuracy, Timeliness and Punctuality, Accessibility and Clarity, Comparability, and Coherence.

Quality management also faces special challenges when using a big data source. The most common are coverage, linkage (with statistical registers or other sources), and errors (measurement errors,

process errors, and model errors). These aspects are described according to the big data classes in the Quality Guidelines for the Acquisition and usage of Big Data[1].

Moreover, the usage of Big Data Sources for statistical production has new and in some cases different challenges to add to these, as it alters the processes for producing official statistics. Therefore, we must supplement with new quality management services as described below:

Input quality management: assessing and monitoring the acquisition and the recording of the data, that can be completely different from the most traditional path with survey or administrative data. It must include/use the Provenance information collected by the Trust Management Service.

First throughput quality management: assessing and monitoring the processing of potentially unstructured raw data into well-structured intermediate (statistical) data. It may have to validate data and procedures, prepared and executed off premises. It must use the information on entities, activities, relationships, and agents collected by the Trust Management Service.

Second throughput quality management: assessing and monitoring the upper layer in which the statistical data is used to produce statistical output.

Output quality management: assessing and monitoring the dissemination and evaluation.

Report quality management: creating specific reports for the first four services described in this list, where in particular the first two items will be different when using a big data sources for official statistical production. A recommended structure for such a report is the quality report template[2].

1.2 Support Application Services

As shown in the lower part of Figure 1, a set of support business functions have been identified as relevant to big data management. In Figure 4, the application services are represented as blue boxes for each of the business functions (yellow boxes). Relatively detailed descriptions for each service is provided in the remainder of this section.

---

1 https://ec.europa.eu/eurostat/cros/sites/crosportal/files/WP3_Deliverable_K3_Revised_Version_of_the_Quality_Guidelines_for_the_Acquisition_and_Usage_of_Big_Data_Final_version.pdf

2 https://ec.europa.eu/eurostat/cros/sites/crosportal/files/WPK_Deliverable_K6_Quality_report_template_2020_02_28.pdf

It is worth noting that the IT management business function is split into two sub-functions, namely 'Data Organization' and 'Infrastructure, Networking and Computing' with associated services.

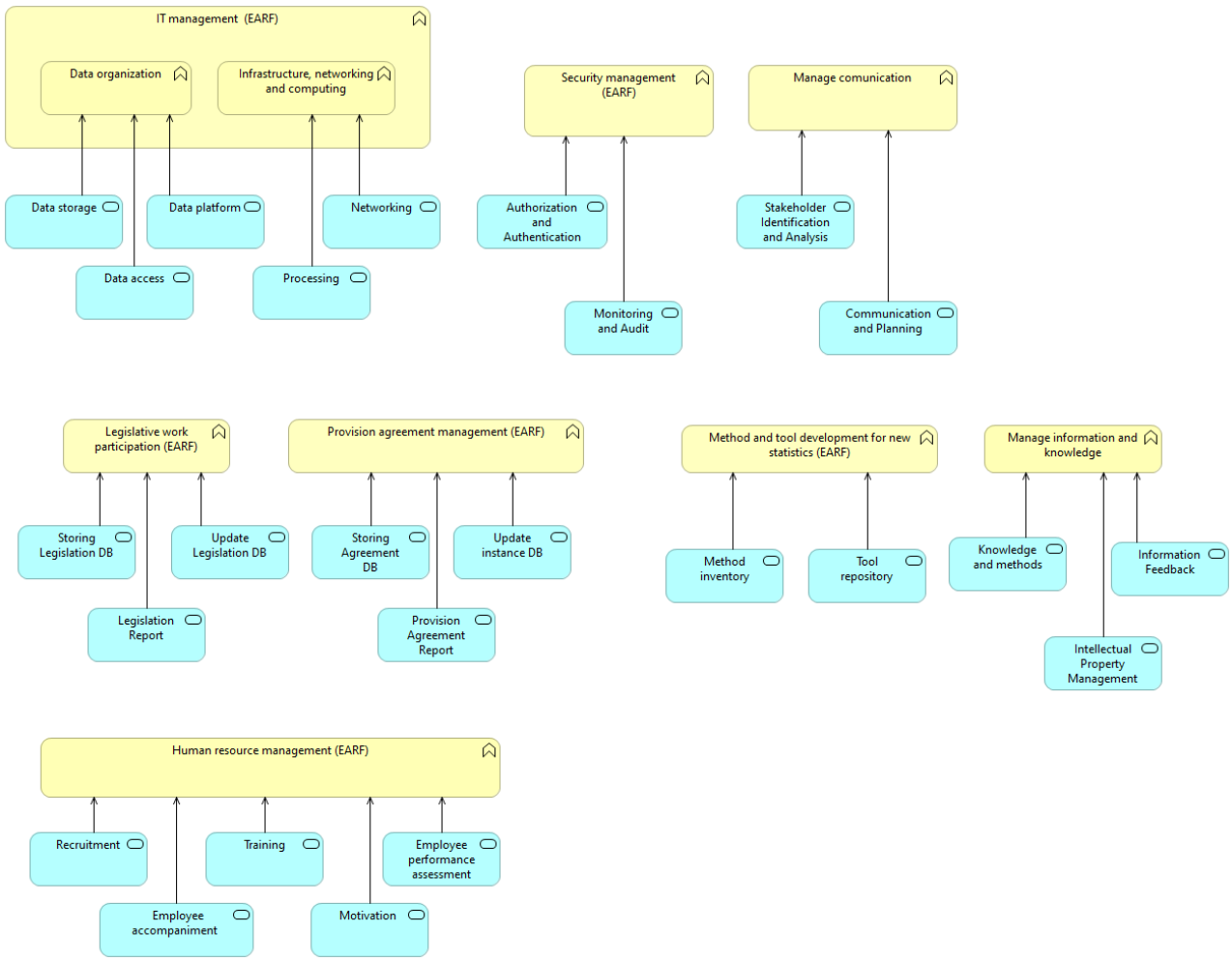

Figure 4: BREAL Application Services for Support Functions

Data organization

The implementation of a statistical process based on Big Data requires the mobilization of support services dedicated to the functions of storage and access to data, constituting altogether the data platform function. Data organization services are closely linked with computing and network services, both on technical and logical aspects.

Data storage: set of support services to manage the organization and distribution of data within the Big Data system. Since many Big Data systems are based on a close link between storage strategy and computing methods that will be used, these services are key elements to foster technical effectiveness of the statistical pipelines. As big data systems are usually distributed horizontally across multiple infrastructure resources, specific activities related to creating data elements can specify that data will be replicated across a number of nodes and will be eventually consistent when

accessed from any node in the cluster. Other activities should describe how data will be accessed and what type of indexing is required to support that access.

Two aspects of storage that directly influence their suitability for Big Data solutions are capacity and transfer bandwidth. Capacity refers to the ability to handle the data volume, while transfer bandwidth deals with the "flow" of information that can be accessed in a given time-lapse. Many Big Data implementations address these aspects by using distributed file systems within the platform framework.

Data access: they are focused on the communication/interaction with the statistical users and pipelines that require data access. Access service may be a generic service such as a web server or application server that is configured to handle specific requests from authenticated users and pipelines. In addition, the access activity confirms that descriptive and administrative metadata and metadata schemes are captured and maintained for access (see Metadata Application Service).

Data platform: they provide for the logical data organization and distribution combined with the associated application programming interfaces (APIs) or other access methods. The frameworks may also include data registry and metadata services along with semantic data descriptions such as formal ontologies or taxonomies. The logical data organization may range from simple delimited files to fully distributed relational data stores.

Infrastructure, networking and computing

Support services related to infrastructure are dedicated to manage the underlying computing and networking functions required to implement the overall system. Complementary to data organization services, they reflect the underlying operations performed on stored data within the system, which include transmission on the one hand, processing and computations on the other hand. Transmission services include descriptions of data transmission requirements that define the required throughput and latency. Computing services cover technical settings, which may deliver specific processing methods and having infrastructures adapted consequently.

Processing: Processing services are dedicated to set up relevant technical infrastructures and methods, given how data will be processed in Big Data statistical pipelines. This processing generally falls into a continuum, from long-running batch jobs to responsive processing, and supports interactive applications of continuous stream processing. The types of processing services described for a given architecture would be dependent on the characteristics (volume and velocity) of the data processed by the Big Data statistical pipelines and their requirements.

Networking: the connectivity of the architecture infrastructure is a crucial part of support services, as it affects the velocity characteristic and is a preliminary condition for deploying responsive processing for statistical pipelines. While some Big Data implementations may solely deal with data that is already stored in the data center and does not need to be accessed outside the local network,

others may need to deploy specific network capacities to be continuously connected to outside data providers (or may expose internal data to outside users)[3].

Networking services may also include various activities like dynamic supervision services and service priority handler, name resolution (e.g. domain name server), and encryption along with firewalls and other access control capabilities.

Security Management

Support services dedicated to Security and Information assurance implement the core activities supporting the overall security and privacy requirements outlined by the policies and processes of Big Data statistical pipelines. This includes authentication services and monitoring services.

Authorizations and authentication: these services must interface and interact with all other components within the Big Data system to support access control to the data and services of the system. This support includes authenticating the user or service attempting to access the system resource to validate their identity[4] .

Monitoring and audit: monitoring services are responsible for collecting, managing and consolidating information about events from across the system that reflect access to and changes to data and services across the system. The scope and nature of the events collected is based on the requirements specified by the security policies (see Trust Management). Typically, monitoring services will collect and store this data within a secure centralized repository within the system and manage the retention of this data based on the policies. The data maintained by these components can be leveraged during system operation to provide providence and pedigree for data to users or application components as well as for forensic analysis in the response to security or data breaches.

[3] On the external side, the availability of network connectivity and the latency of transmission protocol are two key characteristics to be handled by networking services in respect with the primary senders or receivers of big data processed by the statistical pipelines. On the internal side, volume and velocity characteristics of Big Data are also driving factors in the specification of a suitable internal network infrastructure as well. For example, if the implementation requires frequent transfers of large multi-gigabyte files between cluster nodes, then high speed and low latency links are required to maintain connectivity to all nodes in the network.

[4] For instance, authentication services may take the form of APIs to other services and components for collecting the identity information, and validating that information against a trusted store of identities. Frequently such authentication services will provide an identification token back to the invoking component that defines allowed access for the life of a session. This token can also be used to retrieve authorizations for the users/components detailing what data and service resources they may access.

Manage communication

Stakeholders of Big Data enabled statistical pipelines may cover various organizations impacting on the success of the process, including government authorities and regulators, Big Data providers, cloud and IT framework providers, academic community… and even the civil society which is at both ends of the spectrum (from individual data collection to final dissemination of results). Such a variety of stakeholders makes it necessary to take into account different interests, attitudes and priorities. Effective communication ensures that each kind of stakeholders receive information that is relevant to their needs and builds positive attitudes to the statistical project.

The stakeholders must be identified, actively managed, and communicated with to ensure their buy-in to the final product. Hence, communication services contain at least two components:

- Stakeholder Identification and Analysis
- Communication and Planning

Stakeholder identification and analysis: this service collects and stores centrally information about stakeholders, and must at least include the identification of each stakeholder, contact information, a categorization according to a project-adapted taxonomy of stakeholders, links with the projects and specific agreements that may have been set up between the NSI and each stakeholder.

Communication and planning: Communication with stakeholders builds dialogue. It may take various forms like newsletters, reports, conferences, or social media. By setting up these different forms of feedback, a better understanding of stakeholders' interests and attitudes is targeted so that one can fine tune communications. Using forums or other social media to communicate enables to respond to critical comments or correct any misunderstandings.

The communication management and planning service keeps track of how the stakeholder will be communicated with, the type, frequency, and medium. It records the content of the communication and what it intends to accomplish. It can be used to ensure that the Stakeholder Communications Plan is being followed (calendar and content).

Provision agreement management

The service to manage the provision agreement must store the stipulation in the agreements with information providers to provide data according to the requirements: timeliness, confidentiality, quality, transmission protocol, authorship and pre-processing characteristics. It must be closely related with the Trust and Quality Management services to insure that obligations are met. In this sense, it must also connect to the Validation Services.

The Provision Agreement Management Service should thus comprise the following components:

Storing agreement database: this service collects information centrally in a database related to all that is stipulated in the agreement. It must at least include dates where data must be provided, confidentiality and sensitiveness of the data that is provided, the transmission protocol through which data must be made available, structure agreed upon, and the pre-processing that may take place at the provider's premises.

Update instance database: this service collects the information of each instance/transmission pertaining to a specific provision agreement. Instances non-conformant with the base agreement provision must issue alerts. The sharing of confidential information requires that the flows of information are secured and access and usage is traceable.

Provision agreement report: this service accesses the provision agreement and instance databases and creates specific reports that can be destined to users/managers or machines/services:

- Trust Management monitoring appropriately the flow of information,
- Metadata and Quality management, to trace how specific outputs have been actually produced and identify potential improvements,
- Validation services information on the structure and content being provided
- Acquisition and recording services to support directly the statistical data production.

Legislative Work Participation

The Legislative Work Participation is understood as the ability to participate in and influence legislative work that forms the legislative basis of official statistical production. Big data sources exploration has the potential to change the data that is available to produce statistics. This change of paradigm may have a humongous impact on society. However, these big data sources pose new access questions. It is necessary to devise new protocols, agreements and forms of cooperation with the data producers/holders. Despite all the effort put on collaboration, in some cases, legislation supporting and enabling the official statistical production using Big Data sources may be required. Information is the key asset of official statistics and it cannot be compromised by external and internal stakeholders.

To foster legislative work participation supporting the Big Data sources usage the following services must be in place:

Storing legislation database: compiling the legislation, directives and their provisions related to the possibility of usage of relevant and reliable data for statistical production.

Updating legislation database: feeding database on new legislation and legal initiatives that may provide opportunities for other Big Data sources, not yet detected but emergent (for example from the digital transformation) and reporting it back to the New Data Sources Exploration services.

Legislation report: report coordinating the New Data Sources Exploration services with the existing legislation applicable to the new data sources that may be harnessed to produce meaningful statistics.

Note that an important special case of legislation reports is legislation gap reports. These are reports coordinating the Provision Agreement Management services with the lacking/possible legislation, especially where problems are consistently and continuously detected compromising the timeliness and quality of the official statistics production.

Method and tool development for new statistics

The method and tool development identifies possible activities, methods and tools that can accomplish the tasks necessary across statistical subject matter domains, statistical phases and process steps. When a need arises to face a new issue there are mainly three options to address the problem at hand. The first, 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 that already exists. These two options focus on reusing components with the goal of saving resources, like developers and development time. Only it the previous options are not viable, should acquiring an existing package or developing it be considered. Open source solutions should also be preferred, due the known advantages in terms of maintainability. To make this possible the Method and Tool development traditional IT support service must include a catalogue.

The development and management of such a portfolio of Building Blocks BB increases collaboration and reusability inside the NSI and/or across countries fostering the cooperation and creating synergies within the statistical community supporting the modernization and innovation.

Method and tool catalogue - contains references to various artefacts and services to support standardization and efficient sharing and reuse of process, information and services/applications. The catalogue must contain descriptions that are suitable to different to distinct types of users. The I3S[5] communication kit for services reuse identifies four main classes of users with different information requirements: subject domain expert, methodologist, IT and management. A tool may implement several methods and, reciprocally the same method can be used by several tools, leading us to divide such a catalogue in two services.

[5] https://ec.europa.eu/eurostat/cros/content/wp4-create-and-communicate-success-stories_en

Methods inventory: more research oriented, collecting evidence from used and proven methodologies to address specific questions. For Big Data such an inventory should be based on the Methodological Report for the usage of Big Data[6].

Tools repository: providing easy-to-deploy-and-test elements such as generic description and requirements, instructions for implementation, and testing data sets. Several such tools for big data have been made public available on platforms like GitHub that provide versioning. An example is the Starter Kit[7] to web scraping of enterprise characteristics (developed by WP C on ESSnet Big Data 2).

Human resource management

Human resources management services in an organization covers all the aspects related with its employees, like:

- Recruitment services
- Training services
- Assessment services
- Motivational services
- Compensation services
- Maintaining labour relations
- Employees healthy, welfare and safety services
- Compliance to labour laws services

Many of these are already software services but the usage of big data requires probably their extension. We will present the list, of software services that should be implemented whenever using big data, with a small description of each but before that we will introduce a 9th service that is of major importance and still does not exist in most cases.

At NSIs, one can traditionally find employees with very diverse profiles like methodologists, domain specialists, IT, communication, etc. Although it is already customary to have interdisciplinary teams to plan and address some of the projects, to handle a big data initiative it will be necessary that some of the people will fit not just one but several of these profiles simultaneously. This means, that it becomes necessary not only to master different methods and technologies but that these now come from such different fields as math, statistics, databases, programming, artificial intelligence,

6 https://ec.europa.eu/eurostat/cros/sites/crosportal/files/WPC_Deliverable_C7_Starter_kit_for_NSIs_V.2.pdf

7 https://github.com/EnterpriseCharacteristicsESSnetBigData/StarterKit/releases/tag/v1.1

data visualization, creative storytelling and communication. It is obviously impossible to find all of the mentioned skills in a single person. However, during their experience in an NSI every employee will be in contact with other borderline areas developing the sensibility for other fields. It is the manifestation of these capacities that need to be detected, fostered, and deepened. This can be achieved through the accompaniment of the employee and differs from the assessment services because it will keep track of interactions in a project independently of the project success. The need for this kind of service already exists but is more apparent because of the nature of big data.

Thus the first four subservices presented, that must be in place for a big data initiative, are extensions of the previous list but the last one, “Employee accompaniment service”, which is already described, emerges from the nature of big data itself:

- Recruitment services – Tracking enrolment and selection of new employees
- Training services – Collecting information on imparted training
- Assessment services – Appraising the performance of employees
- Motivational services – Gathering evidence for rewards and recognition, as well as strategizing the career planning

Employee accompaniment service– Tracking the capabilities displayed or developed by the employee in each project.

Information and Knowledge management

Information and knowledge management services cover the ownership or custody of records, documents, information and other intellectual assets held by the organisation and the governance of information to be capitalized in a process of continuous improvement and consolidation of knowledge. They include (i) organizing these pieces of knowledge, (ii) making them available so that they can be consulted, (iii) managing ownership of this information and knowledge, and (iv) maintaining the policies, guidelines and standards regarding information management and governance.

In the case of Big Data processes, knowledge and information management services may have specific characteristics. The first relates to the velocity of big data and the continuous entry of new data: this aspect indeed leads to build a knowledge and information management system, which is designed on the same velocity, therefore on a continuous pace. The second one is linked to feedback and permanent improvement. Because Big Data based statistical pipelines are dynamic, multidimensional, integrated and interactive process, knowledge and information management services need to include additional aspects within their scope, with a special attention to managing "feedback information". The knowledge management services are not only to learn about input and output of statistical pipelines but also to transform them and make them more and more accurate. In that view, these services are strongly linked with Quality management on the one hand (like throughput quality), and the Metadata management on the other hand (like process metadata). The third and final specificity is related to the somewhat "non-directed purpose" of a Big Data pool.

While a traditional survey is designed from the beginning for a well-defined statistical objective, Big Data deposits can serve several statistical processes, some of them being unsuspected at the start of statistical works. New use cases of a Big Data pool can therefore appear over time, and knowledge management must facilitate the detection and implementation of these new uses by keeping track of any useful information on the Big Data deposits as they are explored. Information and knowledge management can be broken down into different services, including:

Knowledge and methods: this involves making an inventory of the intangible assets produced by the organization as part of the implementation of its big data processes. These can be methodological documents on statistical techniques, algorithms, documentation on sources etc.[8] More generally, this sub-service can be in support of Metadata Management services.[9]

Information feedback: a key dimension of big data processes relates to the necessary retroactivity to be installed within the statistical pipeline, which can be designed in a learning way to gradually improve the algorithms applied to the inputs. Capitalizing on these feedback loops in order to be able to document them - and share them with other statistical pipelines - is therefore at the heart of knowledge management. More generally, this sub-service can support Quality Management Services but also Trust Management services.

Intellectual property management: the information and elements of knowledge mobilized or produced by Big Data statistical processes must be subject to qualification in terms of intellectual property and conditions of use. Regarding input data, such a topic is covered by the Provision Agreement in connection with the Data Providers; but other facets of intellectual property must be documented, like the use of algorithms developed by third parties, use of external application modules... Inside production can also be covered by a property management policy (status of the source code developed for the need of statistical pipelines, status of the settings resulting for training the models). Intellectual property management sub-service is designed to get record of knowledge legal status and make this information available for internal (or even external) users.

[8] This inventory differs from the methods and tools repository as it is more an index of available materials and the related usage, having a documentation purpose. The methods and tools repository is instead a catalogue with the purpose of making available assets to be used in production processes.

[9] Examples include Process Metadata Management sub-Service to link the steps of a process with the underlying knowledge elements, such as the fact that such and such a processing step is linked to such and such an algorithm, such methodological document, such source code, etc.

2. BREAL Information Layer

The Information Layer of BREAL consists of three sublayers[10], as described by the 'hourglass model' proposed for Trusted Smart Statistics, namely a raw data layer, a convergence layer, and a statistical layer.

- The *Raw Data Sublayer* includes data that are acquired and stored by the BREAL 'Acquisition and Recording' business function. At this stage, we just identify 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 Sublayer* 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 Sublayer* includes those concepts that are the targets of the analysis. These data are produced by 'Modelling and Interpretation', 'Integrate Survey and Register Data', 'Enrich Statistical Registers', and 'Shape Output'.

The intention of this model and the implied solution architecture is to hide the real-world complexity (the raw data sublayer) from the statistician and the statistical complexity (the statistical sublayer) from the data provider. In the ideal situation, the convergence layer contains the data that is actually shared between the parties.

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 mentioned above.

Moreover, in order to emphasize the integration of the Information Layer with respect to existing standards for data modelling in Official Statistics, we included some key concepts from GSIM (Generic Statistical Information Model).

The resulting composition of the Information Layer of BREAL is shown in Figure 5. Each sublayer consists of: (i) specific BD data entities (blue color); (ii) GSIM entities (pink color) and (iii) specific provenance metadata entities (yellow color).

[10] In the subsequent text we may refer to the sublayers as just layers unless there is a chance of confusion.

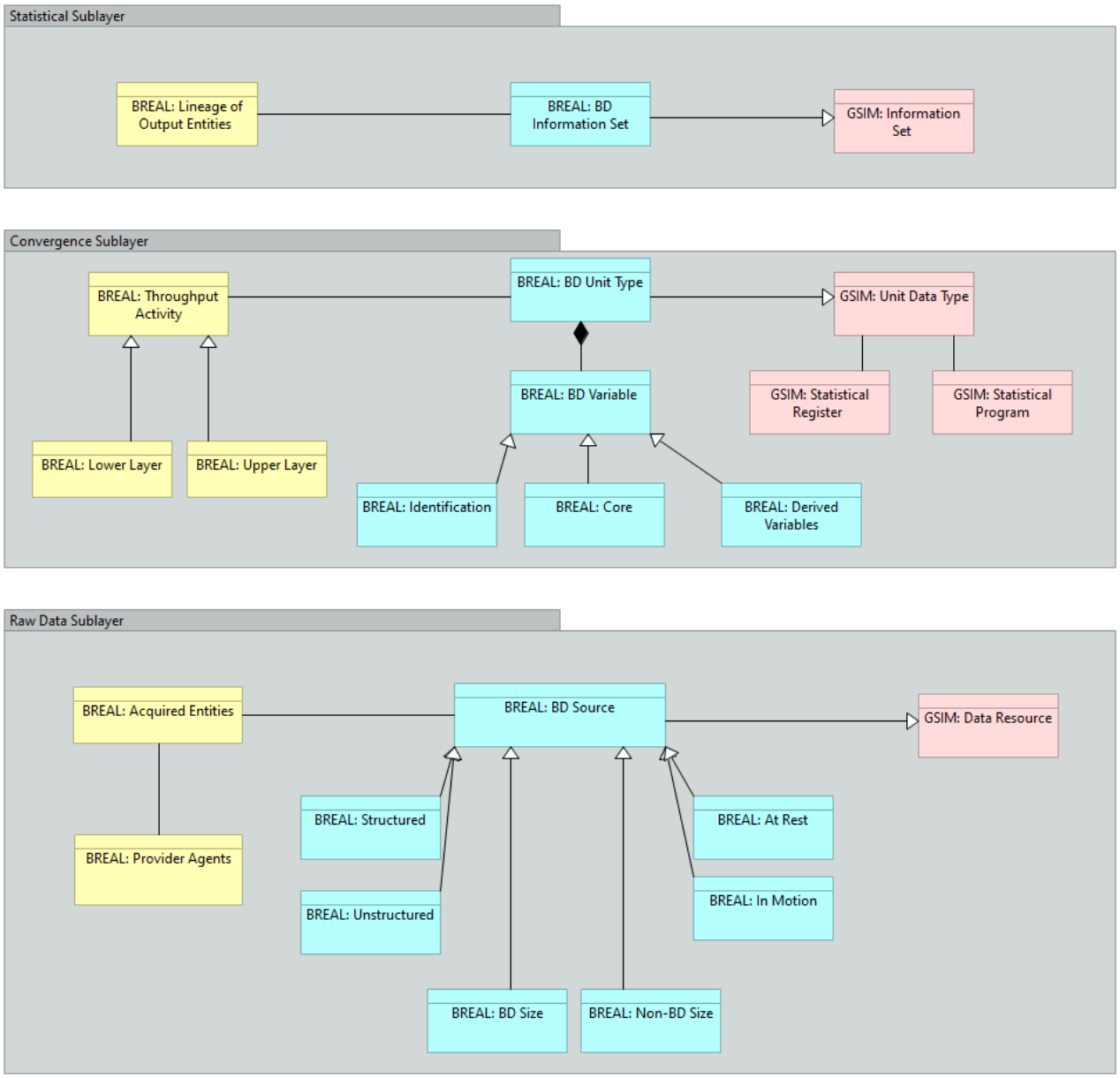

Figure 5: BREAL Information Architecture

The *Raw Data Sublayer* consists of:

- Specific BD data entities: a Big Data Source entity is introduced with three main characteristics that are considered as particularly relevant for characterizing the source itself, which indeed corresponds to volume, variety and velocity. Volume is taken into account by two entities, namely Big Data Size and Non-Big Data Size; variety, by the two entities Structured and Unstructured and velocity of data in acquisition phase, by the two entities In-Motion and At Rest.

- GSIM entities: GSIM Data Resource is introduced. A Big Data Source entity is a specialization of that.

- Specific provenance metadata entities: Two entities are introduced Acquired Entities, to take into account which are the specific data selected from the available source and Provider Agents, to consider who provides them.

The *Convergence Sublayer* consists of:

- Specific BD data entities: a Big Data Unit Type entity is introduced that is composed by BD Variable that can be of three types: Identification, Core and Derived.

- GSIM entities: GSIM Unit Data Type is introduced. Big Data Unit Type entity is a specialization of that. In addition, Statistical Register and Statistical Program (to be intended to take into account survey data are introduced as unit types to be potentially integrated with BD unit types.

- Specific provenance metadata entities: an entity Throughput Activity is introduced to take into account specific operations that are performed on raw data. This is specialized into two entities: Lower layer, which includes metadata possibly introduced by the "Data Wrangling" activities on data provided by the lower layer, and Upper layer, which includes metadata possibly introduced by the "Data Representation" and "Data Modelling and Interpretation" functions, for being used at the upper layer for data analyses.

The *Statistical Data Sublayer* consists of:

- Specific BD data entities: a Big Data Information Set entity is introduced that can be used either for internal or for external output.

- GSIM entities: GSIM Information Set is introduced. Big Data Information Set entity is a specialization of that.

- Specific provenance metadata entities: An entity Lineage of Output entities is introduced to take into account what is the overall process performed on raw data from the output perspective. This entity take into account the provenance metadata introduced at lower layers and is the result of an elaboration of such metadata for adding a contribution to make transparent the whole BD production process.

3. BREAL Operational Model

The purpose of an operational model is to give the possibility to describe solutions in terms of how they are deployed with respect some dimensions (platform, data, etc.) that have to be identified. In the following, a proposal is presented that relies on previous effort performed for the definition of the ESS Enterprise Architecture Reference Framework (EARF)[11].

Operational model

There are various possibilities for NSIs to combine their efforts, working together on a joint solution. The easiest solution is to share code, for example via GitHub. This is common practice within this ESSnet, and allows us to replicate application services. Another possibility is to share infrastructure, e.g. Eurostat's DataLab or the UN Global Platform. Another possibility is to share data sources. These three types of sharing are the basis of our operational mode:

- Sharing application services
- Sharing infrastructure platforms
- Sharing data sources

Sharing application services

The type of application service follows the definition of the ESS EARF. Here, the following approaches for consolidation are used:

- Autonomous service. The application services are designed and operated without coordination with other NSIs. This could be the case when the situation is country-specific.
- Interoperable service. In this case, the interface of the application services across the NSIs is aligned. Thus, the NSIs have similar application services, but the implementation may vary. Reason for this variety, might be for example that the data differs per NSI.
- Replicated services. The same application service is deployed (duplicated) at various NSIs.
- Shared service. There is a common, distributed application service, shared and accessible to all the NSIs.

[11] Eurostat: ESS Enterprise Architecture Reference Framework: https://ec.europa.eu/eurostat/cros/content/ess-enterprise-architecture-reference-framework_en

To indicate the type of application service in the various solution application architectures, the application services are color-coded as follows:

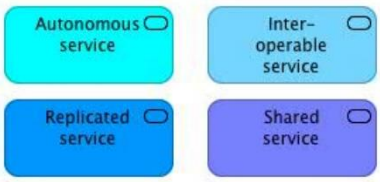

Figure 6: Application services

Sharing infrastructure platforms

The type of platform define where the platform is located and who supports that platform. There are three variants:

- Local platform (of the NSI). In this case, the Big Data platform is fully operated and/or controlled by the NSI.
- Local platform of the data provider. In this case, the data provider has its own Big Data platform, where statistical analysis of the NSIs can be operated. The data provider provides the data and the platform, and NSIs are involved in defining the methodology to be used, producing the statistical outcome.
- Shared platform. In this case a third party that operates and controls the Big Data platform. Examples are the UN Global Platform and Eurostat's EC Data platform. Both platforms can be used by NSIs and other agencies to derive statistics. Another example, is when a dedicated party within the country of the NSI, provides a platform that is shared between various (governmental) agencies within the country.

To indicate the type of platform in the various solution application architectures, the platforms are color-coded as follows:

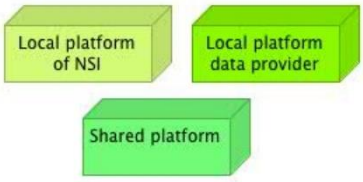

Figure 7: Platforms

Sharing data sources

The type of coverage of the data determines its usage and whether sharing adds value:

- Local data that can only be used by the NSI(s) of the country (or countries) to which the data applies. For example, data on smart meters provided by the energy companies, applies in most cases only for one country. Furthermore, privacy issues might prevent sharing this data, or processing this data outside the premises of the NSI.
- European data, which has Europe as coverage. An example is the European Maritime Safety Agency (EMSA), which has AIS data with European coverage. Another example is the job vacancy data gathered by Cedefop.
- Worldwide data with an implied global coverage. A good example is the AIS data from Orbcomm on the UN Global Platform. This data has global coverage.

To indicate the type of coverage of the data sources in the various solution application architectures, the coverage of the data are color-coded as follows:

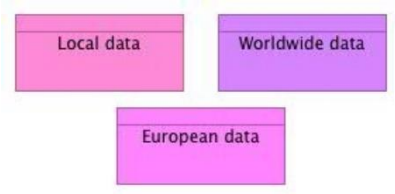

Figure 8: Type of coverage of the data sources

Combinations of sharing

Not all combinations of sharing provide added value. In the model below (Figure 9), we show the combinations that are valuable in our opinion. As a basis, we look at the type of platform:

- Local platforms of the NSI are running autonomous, interoperable and/or replicated application services. Of course, an NSI could provide a shared service to another NSI, but then the NSI would also become a shared platform, so from that perspective we do not consider that combination as logical.

  Local data can of course reside and be used on local platforms. Also, European and worldwide data can be used. However, these data sources are costly and it is better to share these sources on the platform of the data provides itself or on a shared platform. Therefore, these sources are depicted in grey.
- The main reason to execute your big data solution on the local platform of data providers is the access to their valuable data sources. The coverage of these data sources are in most cases, European or worldwide coverage, and the application services executed are mostly shared. Because it is an external party, it will be more difficult to share local data (due to privacy reasons) or to execute autonomous or interoperable application services (due to maintenance and trust reasons). Therefore, these are depicted in grey.
- A shared platform is the solution to share valuable resources. Again, the coverage of these data sources are in most cases, European or worldwide coverage, and the application services

executed are mostly shared. Local data can be used if supported by the platform provider and if allowed from privacy perspectives. Autonomous or interoperable application services can be used if supported by the platform provider.

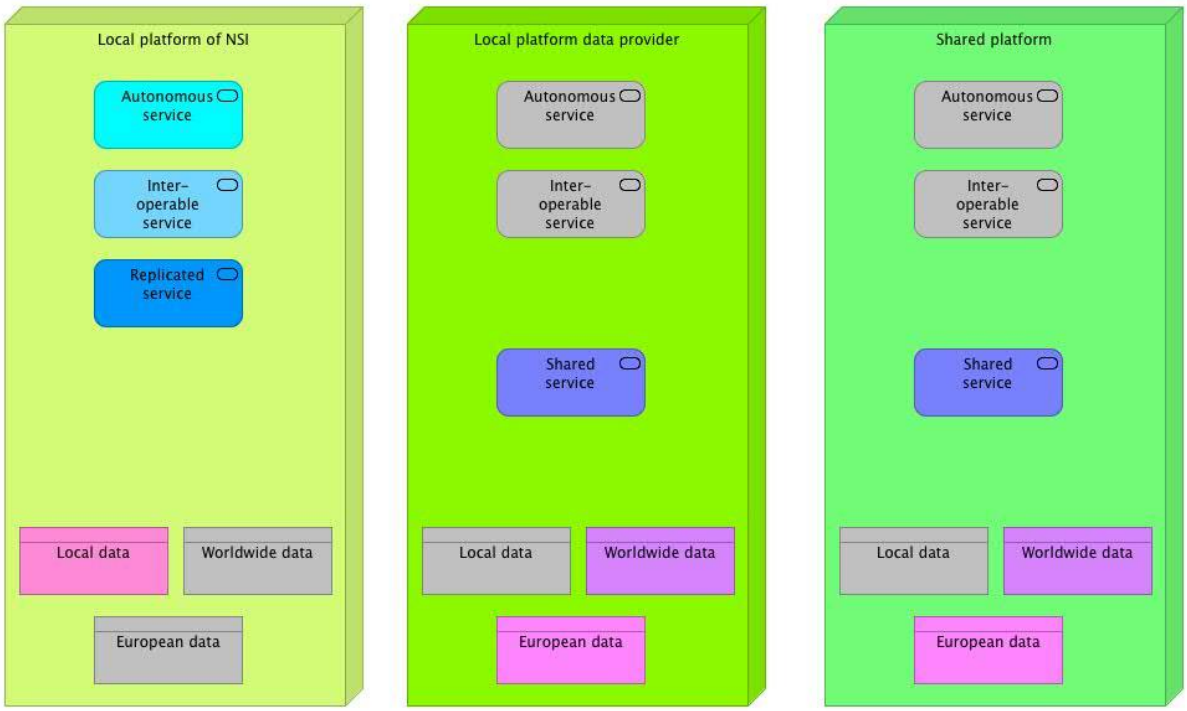

Figure 9: Combinations of sharing