

**26<sup>th</sup> Meeting of the Wiesbaden Group on Business Registers  
- Neuchâtel, 24 – 27 September 2018**

Irene Saleminck  
Statistics Netherlands

Integrated Statistical Register Systems

**How the Data Lake approach can strengthen the SBR role and vice versa**

### **Introduction**

Statistics Netherlands' mission is to publish reliable and coherent information that adapts to the need of society. The demand for information on the one hand asks for fast and flexible information with high quality. On the other hand there rises a social necessity to join one's own data with data of SN. To meet both requirements SN wants to be able to access, share, combine and re-use data, without endangering privacy sensitive data. In order to meet this challenge SN is developing a "Data Lake" solution. Chapter 1 deals with the Data strategy of SN and the rationale behind the Data Lake.

The rationale behind the Data Lake concept is to "connect" statistical processes to the Data Lake in order to access the potential of reusable data. This implies that the Data Lake should contain all usable data, at least from a consumers point of view. In real the data are not stored physically in one place (e.g. a database). The SN Data Lake is a so called logical data warehouse. It's design and architecture will be described in chapter 2.

The concept of sharing data applies for both internal (statistical) data, on premises, as well as data stored externally at the source owners' location. The Data Lake contains functionalities to enable faster and easier data handling, like for example searching, finding and understanding data, as well as for deriving, combining and transformation of data. Also logging, monitoring, authentication and security functionalities are foreseen. In order to create the look and feel of "data at your fingertips" SN is applying data virtualization, metadata-modelling and semantic technologies. How the Data Lake works, it's building blocks and the pillars underneath it is explained in chapter 3.

A special place in the Data Lake is kept for the "source layer", the holy grail where virtually all data sources will be "accessible". Next to all kinds of statistical- and register data also the Statistical Business Register can be approached by the Data Lake as a data source. The Data Lake approach facilitates to access the SBR-data as well as to join and combine various (register) sources with it, in a virtual manner, thus without copying, moving, duplicating or transforming data physically. This concept was tested in various Proof of Concepts of which the implementation of the use case "Family Businesses" was the pièce the resistance thus far. The results are described in chapter 4.

*Keywords: Data Lake, meta data, data virtualization, Statistical Business Register, Family Businesses*

## 1. Statistics Netherlands Data strategy and the Data Lake solution

Goal of the Statistics Netherlands (SN) Data strategy is to realize the ambitions formulated in SN Vision on data. In the 21<sup>st</sup> century we experience the emergence of new (large) data sources and a rapid development of new methodologies, tools and technologies in order to use existing and new kinds of information. The result of which is, almost invincible, that these new and existing data sources are spread out over various systems within and between different organizations and countries. At the same time policy makers, researchers, entrepreneurs and the general public have a need for high quality information concerning a large diversity of often unforeseen subjects. Where it's crucial for the information to be; transparent, factual, up-to-date, accurate and detailed, to be used in evidence based decision making and as a basis for research.

Statistics Netherlands role to provide society with statistical information has been fulfilled by producing a constant information packet (supply oriented) and a service for custom fit solutions (demand-driven). In order to facilitate SN to produce this statistical information and to decrease administrative burden SN has access, by law, to governmental data sources. All data collected by other governmental bodies can be re-used for statistical purposes. However, in these new times in this data driven society SN has, next to its role as statistics producer, more and more also the function of a data hub. It is SN ambition to strengthen this role and to make the increasing offer of new and existing data sources accessible for use i.e.; statistics data hold by SN, governmental data and privately held data. SN has a vast amount of data, a lot of knowledge about data and wants to share that with others thereby respecting the FAIR data principles; Findable, Accessible, Interoperable, Reusable. By sharing data and knowledge SN wants to further strengthen its social relevance. SN defined 3 aspects of data needs to keep and/or improve and 3 new aspects to be explored and expanded. See figure 1.










 <b>Production</b>	Regular operational process that produces the statistical product packet of SN. Processes are designed (once) in advance and are executed on a regular basis. Typical are the processing steps with excellent methodological applications.	 <b>Information dialog</b> 	The ID is actual custom fit statistical information that is processed fully automated. In an automated dialog the information need of the customer is analyzed and answered real time based on present content.
 <b>Research</b>	Purpose of research is to explore and try new possibilities that subsequently can be used in statistics production or custom fit products. It concerns often new (big) data sources, improved editing strategies or improvement of the output.	 <b>Datacenters</b> 	One or more (external) parties want to combine their own data with data of SN for research purposes. For these parties a (shielded and secured) environment is set up.
 <b>Custom fit</b>	Starting point and basis is the customer that formulates information needs. Via an interactive dialog the user need is formulated. The (open) data is compiled based on the existing information at SN.	 <b>Shared Services</b> 	Minimal two external parties want to combine their data amongst each other and ask SN for help. No SN data is used, SN has the role of consultant and/or Trusted Third Party to promote information services.

Fig.1 Existing and new aspects of data needs

The ambitions emerging from the Vision on data have been formulated as follows;

- Stimulate #MoreDataCoupling
- Become a #DatacentricOrganization
- Support #FAIRprinciples
- Develop users #FromConsumerToProducer
- Stimulate #SupportToThirdParties
- Give priority to #MetaDataFirst
- Start an #Informationdialog

In this paper the emphasis will be on #MoreDataCoupling and #DatacentricOrganization.

Concerning external (micro) data research three “Architectural patterns” of bringing data together can be distinguished;

1. External data is brought to the SN environment and all data is processed there. Usually data source has no steady state and/or history → on request of data owner;
2. SN microdata is brought to an “external” environment to be processed i.e. a High Performance Computing (HPC) facility administered and managed by SN → cooperation with universities and in order to connect to existing large (international) networks between R&D institutes;
3. Data is left in its own environment i.e. the environment of the data owner and in order to process the data Data-virtualization techniques are applied (see chapter 3) → when multiple data sources are in place and/or data source are too big to copy (HPC may be needed in addition), when data lineage (more control) for data owner is mandatory and data redundancy is crucial to prevent.

The first architectural pattern is common practice for many years at SN. However as mentioned before, times are changing and introducing new patterns is a necessity in order to secure “access” to and usage of data sources. This has to do with various developments. Organizations are sometimes not yet prepared to send data, which has been reinforced by the newly introduced General Data Protection Regulation (GDPR). Also, data sources become too big and too complex to be copied and an HPC capability is needed as well as the possibility for extreme scalability. When we look for example at Internet-of-Things (IoT) applications so much data is produced that processing needs to be done at the end (Edge computing). This asks for specific solutions in the environment where the data is processed (abstraction layer, see chapter 3). This also applies for real-time streaming data. The GDPR on the other hand put renewed emphasis on data usage and proportionality i.e. to use only the minimum set of data needed for a specific goal.

The Data Lake solution that is being developed at SN fulfills the requirements for the environment needed when data is left at its own environment, i.e. architectural pattern 3. The qualification “own” can be read both as the environment of a data owner outside SN or the SN environment where data is kept internally but at different network locations and/or in different technical formats. The application and benefits are similar. In this paper the main focus will be on the internal SN environment and data sources.

## 2. The SN Data Lake design and architecture

SN performs direct data collection, uses extensively secondary data sources including administrative data (governmental registers) as well as emerging big data (for example road sensors and web-scraped data) for regular statistical production and ad-hoc analysis. This results in wide variety of input data sources and formats. Datasets needed for statistical production and analyses are, historically, often stored and managed in some central data stores similar to Data Warehouses/Data Marts as well as in decentral data stores and data files often “hidden” within survey processing platforms and systems. As a result many statistical datasets are described with their local data taxonomy and metadata vocabulary and are stored in various formats using different storage techniques, e.g. file base storage using CSV or fixed width, relational database tables etc. In order to easily access and (re-)use the datasets, SN is developing a Data Lake<sup>1</sup> solution. The Data Lake main goal is to empower users (statisticians, researchers, data scientists) by allowing them to find the datasets that they need for analysis, to use these datasets and to perform operations on them.

From the perspective of the end user the Data Lake should:

- Enable more phenomenon based output (a phenomenon is a striking event that you want to explain with data, usually comprising multiple datasets)
- Enable more current and coherent statistics
- Stimulate the reuse of data
- Accelerate the statistical processes
- Stimulate the access to a large number of existing and new data sources
- Provide faster response and output to requests from external clients
- Accelerate the design process around collecting and storing data

The main idea behind the Data Lake is to connect all statistical processes to the Data Lake in order to make the full potential of (re)usable data available and accessible. This means that the Data Lake should comprise all (re) usable data. This does however not mean that all data are stored physically at the same location, for example in one data base. However, from the user point of view it has the look and feel it does (see Fig. 2). In fact, data streaming out the Data Lake are not physically in the Data Lake. The Data Lake organises these data streams of ingoing and outgoing data where the ingoing data always has its origin in a local data source. Such an organised stream is called a “view”.

The Data Lake facilitates the design of such a view, authorisation of its use, the composition from local data sources and access to the statistical production process of the user in order to use the result of the view. On top of that the Data Lake governs the designed views, facilitates

---

<sup>1</sup> Definition of ‘Data Lake’ in the context of SN is somehow different and much broader than the usual definition of ‘Data Lake’ and corresponds more closely to the term ‘Logical Data Warehouse’, introduced in 2011 by Mark Beyer and further defined by Gartner.

finding and understanding these views. The Data Lake fulfils a bridging function between supply and demand. See figure 2 below.

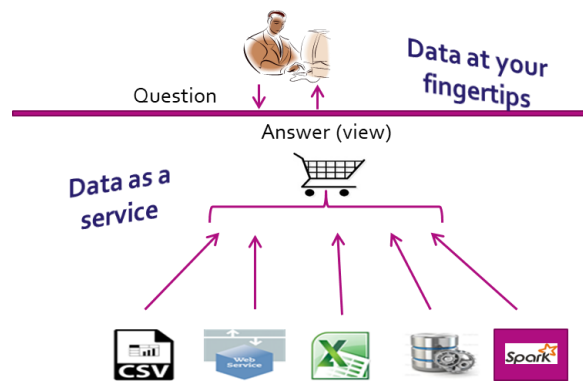


Fig. 2 Data Lake as a “broker” between supply and demand

In order to achieve the before mentioned goals and the broker function a list of capabilities (see appendix 1) has been defined that can make possible that data, stored in a distributed way, can be transformed and integrated in a logical layer, presented as building blocks (views as virtual datasets) and provisioned to users and applications via a standard interface. The capabilities were modelled into the SN Data architecture.

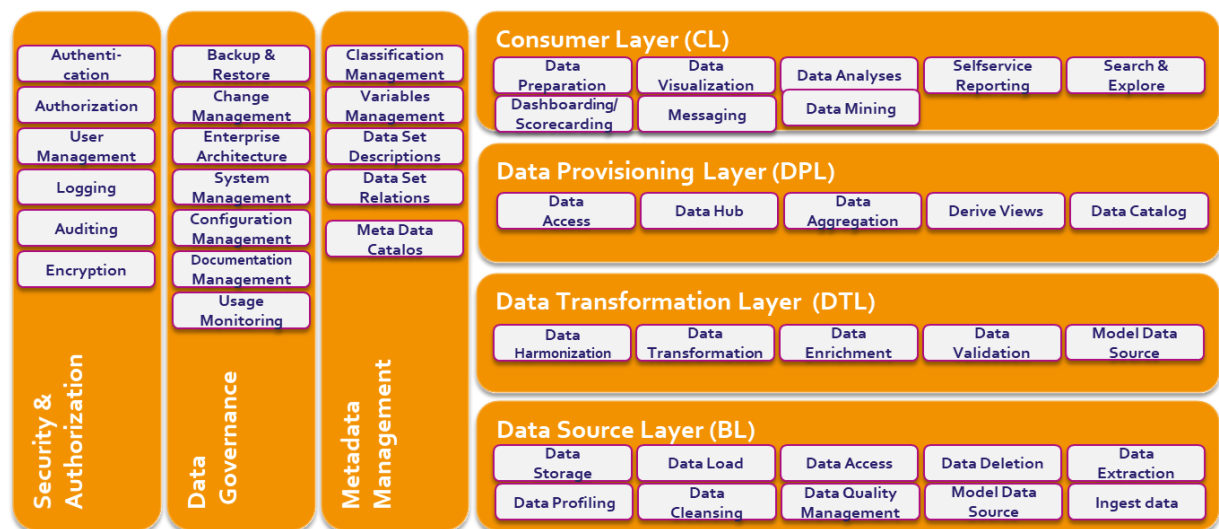


Fig. 3 Architectural layers containing the capability groups<sup>2</sup>

Capabilities required to achieve the goals described in the vision are described in more detail in appendix 1. The capabilities identified during the project have been placed into a layered scheme that defines the new Data architecture and consists of the following logical layers:

1. **Data Source Layer (DSL)** provides data storage for the “Data Lake”. This can be implemented using the different types of databases management systems for example SQL, noSQL, Hadoop but also various file formats or access to web services.

<sup>2</sup> SN recently adopted the UNECE Common Statistical Data Architecture (CSDA) capability model. See appendix 2. The CBS Architectural layers model will subsequently be updated in time.

2. **Transformation Layer** enables data transformations for example data format transformation capability and the capability to create new data sources from already attached data sources. Each data set in the Data Source Layer will in the proposed data virtualization solution (chapter 3) have a representative in the Transformation Layer (aka base data sets). There is no physical data to be found in this layer, only the definition of virtual data sets with associated mappings / transformations.
3. **Provisioning Layer** makes logical data structure available to users and systems using open standard interfaces and protocols. In comparison with the data sets in the Transformation layer, redundancy in the provided data is allowed (certain objects will be included in multiple data sets – one can find data in de-normalized form). The data sets are designed on the bases of a phenomenon (or topic). Data can be provided as a data set or as a web service (e.g. as a oData4 dataset), which makes it possible to include the Data Lake in a Service Oriented Architecture (SOA).
4. **Consumer Layer** consists of tools and systems that are used for processing or analysis of the data. They communicate with the provisioning layer to get access to the chosen logical datasets. Data Preparation, Data Analytics, Data Visualization are main sub-domains of this layer but we can also include here interfaces for automated system-to-system data access and presentation of data in web interfaces.

These layers are accompanied by three cross-cutting layers:

1. Security and Authorization,
2. Corporate Metadata Management and
3. Data Governance.

Cross-cutting layers are key part of the Data Lake concept and provide for the majority of new capabilities needed in this Data Lake approach.

In developing the Data Lake, User stories and analysis of capabilities were crucial for the implementation of this new architectural vision and delivery of goals. Priority was given therefore to capabilities that make it possible that data, stored in a distributed way, can be transformed and integrated in a logical layer, presented as building blocks (virtual datasets) and provisioned to users and applications via standard interface. Proofs of Concept (PoC's) run for Data Lake program therefore focused on the multi-layered data architecture shown in Fig. 3, data virtualisation technology and a semantic metadata model.

### **3. Building blocks of the Data Lake**

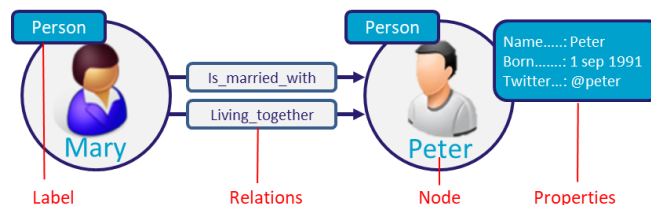
The layered architecture has been initially implemented in a project that has delivered a database designed to provide Statistical Business Register data to interactive users and systems (a Unit base derived from the SBR, result presented previously at Wiesbaden group meeting in Tokyo 2016). This project confirmed the feasibility of this architecture and additional Proof of Concept projects tested key technologies that need to be combined to prove that the architecture can be efficiently implemented at a large scale: Metadata Management and Data Virtualisation.

In the first (minimal) version of the Data Lake the project realizes the broker function by the implementation of a Data Virtualisation (COTS) platform (Fig. 4). This platform makes local sources like the Statistical Business Register, a Data warehouse containing statistical output data and several registers available, via views, to the users of the Data Lake. With these views data sets of local sources can be disclosed and subsequently combined, enriched (aggregated, filtered) to new datasets. Also users will be provided with basic functions to collect these views with for example Alteryx, R, Python, SPSS and to use them in their business processes or analysis purposes. Essential security measures are tuned in accordance with current security policy.

Data virtualisation; The big advantage of this technology is that there is no need to copy data into a new structure and therefore no effort is needed to load and unload data during maintenance. Changes in the data model can be applied quickly with a minimal impact and with a very short time-to-market. The owner of the data will stay fully in control improving the willingness of data providers to share their data. Furthermore the virtual datasets can be fully customized to the customer needs (no limitations concerning storage requirements and actuality). As the virtual data sets only consists of meta data, versioning of the data sets is also very easy. During a transition period multiple versions of the same data set can exist, providing more time for the receiving (legacy) systems to change their model accordingly. By positioning the Data Lake as a single data platform, future capabilities can focus on this platform. It is not needed to implement these new capabilities on all existing (legacy) data providers.

Next to the data virtualisation platform, in this first version of the Data Lake a metadata model and a metadata handling system is introduced (cross cutting layer in Fig. 4). With this model and system the Data Lake facilitates users to search for available views in the Data Lake based on a meta data driven approach and to understand which views are needed to fulfill their data need. The data of the view found can be retrieved via the Data Virtualization platform using the metadata id.

The purpose of development of the metadata model is to help users to find the right data, understand its semantic characteristics and use this information in data integration, data analysis and other statistical activities. The model is based on a graph representation of characteristics that describe statistical dataset as well as relationships between the datasets.



Work on the metadata model has proven that it is possible to design such graph for any statistical dataset, map it to the semantic network and subsequently create a map of semantically linked data<sup>3</sup>. A PoC also proved that this map can be stored in a graph database and

<sup>3</sup> Description of the semantic model exceeds the purpose of this paper but it is described in Datameer: Uitwerkingen user stories en model metadata by Tjalling Gelsema, Robbert Renssen, Ilona Armengol Thijs, Ger Sloodbeek en Peter Zandbergen (2016).

further work is underway to understand how metadata can be captured (automated) and used in conjunction with the data virtualisation.

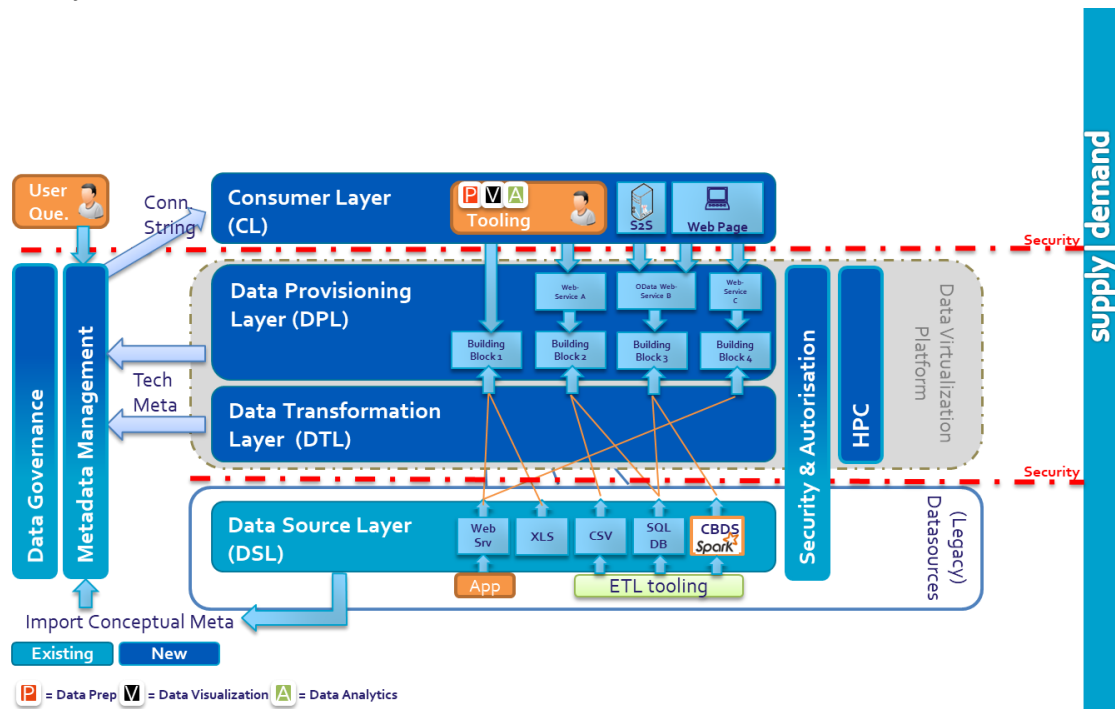


Fig. 4 The Data Lake “Data as a Service” Architecture

#### 4. Proof of Concept; family businesses and the Data Lake

SN has been researching the possibilities for detecting Family Businesses (FB) by using registers aiming for the situation that it will not be necessary to conduct surveys for detecting family businesses. For this research the GEEF-definition for family businesses & has been used.

A firm, of any size, is a family business, if:

- The majority of decision-making rights are in the possession of the natural person(s) who established the firm, or in the possession of the natural person(s) who has/have acquired the share capital of the firm, or in the possession of their spouses, parents, child or children's direct heirs.
- The majority of decision-making rights are indirect or direct.
- At least one representative of the family or kin is formally involved in the governance of the firm.
- Listed companies meet the definition of family enterprise if the person who established or acquired the firm (share capital) or their families or descendants possess 25% of the decision making rights mandated by their share capital.

This definition mentions that a business can be a family business regardless of size. However we did not want to include every Freelancer and Self Employed Person, since they often run businesses that are not transferable. Their business is totally dependent on the knowledge and skills of self-employed person and is very likely to end when the self-employed person stops working. Therefore we have tried to make a distinction between Self Employed Persons and Freelancers on one hand and sole proprietorship on the other hand.



It's been shown to be possible to automatically detect about 260.000 Family Businesses with size class SBS > 20 from the Statistical Business Register in combination with several other registers. Subsequently a PoC has been executed using the Data Lake technology. The sources used were;

- The Unitbase (3N-normalized database derived from the SBR)
- The Unit Environment (Denormalized version of Unitbase where for instance Enterprise groups together with their attributes and their legal units with their attributes are presented in 1 table)
- Fiscal Wage Data (FDR)
- Trade Register of Dutch Chamber of Commerce
- Management of Relations of Tax authorities
- Corporate Tax Data containing information about major shareholders

Unfortunately the parent-child register, that contains almost every parent-child relation since 1945, and the engagement hub (marriages and registered partners) could not be used "as such" in the PoC with the Data Lake technology. The reason for this is that these registers were only available in an encrypted format due to the way of implementing privacy rules with these registers. Therefore additional actions needed to be executed that are not yet supported by the minimal versions of the Data Lake.

In the PoC de following family businesses where detected;

- One man businesses size class 21 or >
- Partnerships with the majority of the partners to be family in which family-relations are based on family names of partners in the partnership and family names of spouses or registered partners as known at the chamber of commerce
- Family businesses of legal persons with the limitation of
  - Enterprise Groups (EG's) with only 1 single shareholder
  - EG's with more legal persons with a single shareholder where the majority is family based on family name
  - EG's where 1 person is employed with income coed = 17 (i.e. officials that have great influence on their own termination of employment and therefore are not insured for employees insurance)
  - EG's where 1 person has the majority of shares where this person is also the official
  - EG's with as a head a centralized accountancy office / trust company with 1 official or more official that are family of each other

The result of the actions described above where 16 views which in turn where based on another 20 views. The total duration to develop all the views was approximately 80 hours. The 16 views mentioned above consist of views (9) that contain the family businesses above, but also of the views (7) with the data that needed to be encrypted (in this PoC an action to be executed still outside the Data Lake environment) in order to detect family relations based on the parent-child relations and the engagement hub.

The basic idea to detect a FB in the SBR by linking additional administrative data to the existing administrative and statistical units stored in the SBR was proven. In this way a 'satellite' could be created to characterize a FB.

Obviously for this satellite a lot of sources need to be analyzed and coupled, however after establishing the initial views, the 9 views comprising the methodology behind FB were executed in a production setting in less than 40 minutes. The 7 views containing the data that had to be encrypted first, executed in between 40 and 180 minutes depending on the Data Virtualisation-tool used. In order to achieve this performance some of the 20 intermediate views had to be materialized. Also worth the investment is that new sources can be analyzed additionally and added (for example after de-anonymization) without having to do the work on the algorithm again, a new view could be added replacing another. Although in this PoC we merely reproduced a previously executed experiment on creating a sub population on Family Businesses it became clear that an improvement in duration of data handling, quality of process and product development and flexibility can be reached by introducing data virtualisation technologies and the Data Lake approach. In turn, the Statistical Business Register in combination with additional administrative sources showed to be a valuable contributor to the source layer of the Data Lake approach.

The research on detecting Family Businesses in the SBR fits into a broader field of research on 'profiling' enterprises thereby differentiating businesses based on certain characteristics. Also policy makers show interest in different typologies for Small and Medium Enterprises. Besides Family Businesses and the Self Employed there is interest for Hidden Champions, Almost Failed Firms, Ambitious Entrepreneurs, (Un)-Consciously Constraint Entrepreneurs and Corporate Social Responsibility. These "sub-populations" can only be derived by combining the SBR with a multitude of various data sources (registers, administrative data, internet data etc.) At the Register Department of SN methodologies are developed to use several (new) sources like internet data and logged user data about the use of our national NACE-application to distinguish for example enterprises involved in internet economy.

Also the international dimension in characterizing enterprises (also small and medium) is unabated important. As stated by Timothy Sturgeon<sup>4</sup>:

*"Clearly, the assumptions behind current data regimes have changed and statistical systems are struggling to catch up. While it will be exceedingly difficult to fill data gaps without new data, and progress that relies only on existing data resources will always be limited, the most efficient approach will be to develop systematic links between key existing data, supplemented with a few additional variables, with data on enterprise characteristics drawn from administrative sources, all tied together by enterprise identifiers that make ownership clear, even when it extends across borders."*

The future lies in extensively coupling of a multitude on data sources. The Data Lake approach is envisioned to be the enabling technology to facilitate this work. Chapter 5 summarizes the benefits that were identified by introducing the next generation data management architecture.

---

<sup>4</sup> Global Value Chains and Economic Globalization- Towards a new measurement framework, 2013, Timothy J. Sturgeon, Industrial Performance Center MIT

## 5. Benefits and challenges of the Data Lake approach

Work to date identified the following benefits;

- Increased accessibility to available datasets physically stored in different locations and multiple formats without the need for costly data migration and standardisation required to centralise data in single physical data store;
- Logical separation between data sources layer and data consumer layer allows flexibility for example adding (registering) new data sources while protecting analytical and processing systems from changes that happens in the data sources layer;
- Decoupling of different layers provides mature versioning capability for every data set, meaning a data set can have multiple versions enabling a controlled change management process without time pressure on data set related applications (a timeframe can be provided for the transition from version x to version x+1);
- Data transformation layer can be used to define transformations and other functionality and make it available across all users/systems therefore reducing development time and fostering reuse;
- Different tools and applications in Consumer layer have access to predefined (virtual) datasets, transformations and other common functions (data versioning, data quality, security confidentiality) without the need to copy datasets and replicate these functions in each individual tool or application;
- Faster phenomenon-based analysis and reporting based on semantic relationships that help find and understand relevant data faster and reduce time to market;
- Improved query performance (especially for file based data sources) because of query optimisation and cache-ing methods of data virtualisation technology in transformation & provisioning layer;
- Possibility to add new types of data sources (unstructured, high volume etc.) that require different storage, have various levels of data quality etc.

While research and implementation of PoC's already confirmed the feasibility and benefits of the Data Lake approach also areas that will require additional work are identified.

- Data Governance: new architectural approach and underpinning technologies and models alone aren't enough: with the possibility to provision virtual datasets in a simple and expedient way we have to make sure that we apply appropriate governance mechanisms.
- Security: a lot of current security and confidentiality controls are based on physical datasets and their location. With the new architecture there is possibility to apply security at two levels: data sources layer and provisioning layer. In combination with technological possibilities like audit trails and more centralised monitoring we have potential to increase security however we need to carefully investigate best approach and change policies accordingly.
- Confidentiality control: similarly, it will be possible to design virtual aggregated datasets directly based on microdata however the use of this approach in some areas for example external researchers would require implementation of confidentiality on the fly to replace manual control of outputs.
- Metadata enrichment and linking to the new semantic metadata model: this is perhaps the most difficult area. Capture and management of Metadata has always been time consuming

and not a very popular task (particularly as benefits of metadata are often accrued by data users and not by data owners / producers who have to do most of the work). Technology offers some new possibilities to automate extraction of technical metadata from data sources however we still don't have all answers how to efficiently and effectively define and maintain semantical and other added-value metadata.

The author wishes to thank; Ger Slootbeek, Paul Grooten, Robert Rensen, Matjaz Jug, Rico Konen, Leon Custers, Marien Vrolijk and Tjalling Gelsema for their contributions that formed the basis of this paper.

## Appendix 1: CBS capabilities model

### Capabilities:

- are the building blocks of the business;
- represent stable business functions;
- should be self-contained;
- are abstractions of the organisation;
- capture the business' interests and will not be decomposed beyond the level at which they are useful.
- may be defined in terms of (be decomposed into) more detailed views (lower level capabilities).

### Description of Capabilities belonging to Fig. 3

- **Assure Metadata Quality:** each data source must be described with high-quality metadata so users can have all information they need for their use of datasets. For example, statement about the quality and reliability of the data.
- **Secure Data Source:** this enables authorization of users to get access to the data they need.
- **Protect Data Source:** protect confidentiality of data, such as ability to make certain data unrecognizable (for example Data masking).
- **Usage Monitor:** this provides monitoring and measures about the use of the data (how often are data sources used, who is accessing them, etc.)
- **Manage content types:** this concerns the management of names and definitions of object types, such as person, ride, transfer, etc.
- **Manage classifications:** this concerns the management of classifications and associated code lists.
- **Manage variables:** this concerns management of names and definitions of variables. These variables are always related to an object type with the relevant unit of measure (number, weight, currency, value, etc.), or link to classification.
- **Manage data source descriptions:** this concerns management of the (logical) structure of all data sets. This structure describes which object types and variables are in a relevant data set.
- **Manage data source relations:** datasets can have common object types or variables. This capability enables the management of such relationships.
- **Data Sources Register:** catalogue with powerful search functionality that enables users to quickly find the desired variables.
- **Relate data sources:** many data requests cannot be delivered by a single source (dataset), but rather require integration of multiple data sources. This coupling must be facilitated in such a way, that it is clear to the user which fields are eligible for this purpose and the conditions which apply to it.
- **Derive Views:** enables extraction of new information (often not as a "copy" of a resource, but a selection and / or aggregation or clustering). Merging data provides an added value.
- **Standardize:** it is important to make arrangements in regards to definition of metadata, quality of data sources, allowable source data formats, etc.
- **Reformat Data Source:** makes it possible to map any kind of data format for the storage of the data source to the standard format.

## Appendix 2: The UNECE Common Statistical Data Architecture capability model (CDSA v 1.1)

The CBS capabilities model is based on the UNECE CSDA capability model and will in time be replaced by it.

On the highest level, CSDA distinguishes 11 Capabilities, divided into 2 groups: Core Capabilities and Cross-Cutting Capabilities. Core capabilities are those capabilities that are used in day-to-day (statistical) operations. They are self-contained but rely on support from the cross-cutting capabilities. Cross-cutting capabilities are used to set, maintain and enforce general policies that apply across all capabilities. They also provide supporting services (e.g. managing metadata) for Core Capabilities, and provide additional functionality for stakeholders (e.g. data governance).

The CSDA **core capabilities** are:

- Data Ingestion
- Data Description & Organization
- Data Transformation
- Data Integration
- Data Sharing

The CSDA **cross-cutting capabilities** are:

- Data Management
- Data Governance
- Security & Information Assurance
- Metadata Management
- Provenance & Lineage
- Knowledge Management

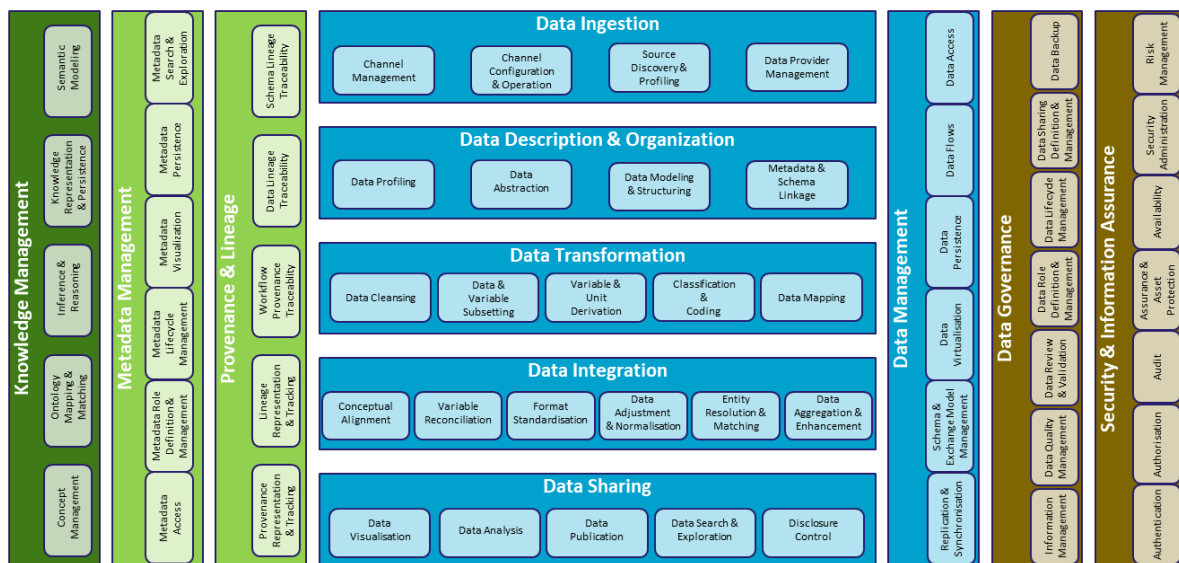


Fig. 5. shows an overview of the CSDA Capabilities, with their second level capabilities. In the following sections, each of the 11 high level Capabilities is described in more detail.

**Data Ingestion:** The ability to receive data from internal and external sources or providers.

- Finding and characterizing relevant data sources for statistical production purposes
- Accessing data sources through a variety of channels (for example API, web questionnaires, administrative archives, statistical registers, streaming data, etc.)
- Managing the relationship with data providers (for example respondent management and Service Level Agreements for administrative data sources)
- Managing the relationship with data consumers by defining service level agreements for downstream usage.
- Making data available by maintaining channel operations for downstream consumption
- Note that sample design and drawing a sample are out of scope for this capability.

This capability consists of (i) establishing and securing data provision agreements with relevant providers, internal or external, (ii) creating reliable data provision channels with existing sources, (iii) acquiring the data as-is or under specified formats, (iv) describing the data and acquisition process and (v) making data available for downstream processing within the constraints of service level agreements.

Data ingestion is the first step in the typical statistical production process, and an essential one: no statistical information can be built without data.

Deciding what sources of data are interesting for creating or enriching statistics is a preliminary step. It builds on a good evaluation of users' needs and implies a good knowledge of the data and information landscape. In some cases, new data needs to be acquired externally and corresponding collection instruments (typically survey questionnaires) have to be designed and created. In other cases, existing external and internal data can cover the needs and the problem is to secure access to it.

In the case of data coming from external sources, the reliability of the source and its stability in terms of time and format should be assessed. For data coming from private companies, additional actions need to be performed regarding the assessment of the reliability of the provider and its willingness to provide the data. In some cases, it can be necessary to act on the legislative level to secure data provision, which requires specific competences. In any case, clear contracts need to be set up and managed in time with the providers.

Once access to data is ensured, it is necessary to design and implement the operational solution that will bring the data to a storage or processing facility controlled by the statistical organisation. This includes technical (network, file transfer, data capture, web scraping, etc.) as well as organisational (monitoring, verification, etc.) aspects.

Note that some modification treatments can be applied later on to facilitate data integration, such as filtering, transcoding, normalisation, translation, codification, etc. Those modifications are not part of this capability, they belong to Data Transformation and Data Integration.

Data can be made available in different ways, e.g. by moving it into a persistent environment (for example relational database, NoSQL, data lakes, big data storage, etc.), or by setting up a data pipeline for stream processing or configuring other types of data flow execution jobs via the Data Management capability.

The Data Ingestion capability is tightly connected to the Metadata Management capability. First, it is important to capture metadata about the structure, quality, provenance of the data, and about the ingestion process itself. Second, metadata can be used in an active way in the data ingestion process, for example for the automatic generation of collection instruments or of parts of the ingestion process.

**Data Description & Organization:** The ability to describe, prescribe, and assess data and the structures that organize them.

- Enabling consumers of information assets to understand their contents regardless of their technical implementation.
- Maintaining mappings between data assets and relevant metadata structures to support other capabilities in the architecture, e.g. data search, exploration and analysis.
- Assessing quality of data assets in terms of their data and associated metadata.
- Providing data models at multiple levels of abstraction using standards, when possible, to support both metadata-driven processes and communication.
- Providing mechanisms to validate instances against schemas and other prescriptive metadata.

This is a capability that structures, organizes and abstract information assets so that they can be made available, and findable, to other capabilities in the architecture.

It includes capturing data requirements, or reverse-engineering information assets, and formally representing them in a precise form in data models at various levels of abstraction. Data models are not only diagrams but also any associated documentation that helps to understand information contents and drive data profiling activities.

Models at higher levels of abstraction, e.g. conceptual, are usually in business language to enable consumers to understand the information contents regardless of their implementation. They describe business objects and their relationships. An example of a data model at the conceptual level is the Generic Statistical Information Model (GSIM). Models at lower levels of abstraction, e.g. physical, tend to be implementation-dependent and machine-actionable, e.g. database schemas. They describe how data points are organized into a variety of structures, e.g. records, tables, trees, graphs, etc. Examples of such models are the Data Documentation Initiative (DDI) and the Statistical Data and Metadata Exchange (SDMX) (see Data Management).

This capability also maintains the mappings between information assets and associated structures, i.e. metadata and schemas, including data models. The structures associated with the information assets could be prescriptive, normative or descriptive. Prescriptive metadata establishes what "must be", i.e. there are no exceptions and all data instances must comply.

Examples of those are integrity constraints and data types as defined in database schemas, where a data instance cannot even be loaded unless it is validated against the schema.

Normative metadata is a weaker form of prescriptive metadata in which there is a notion of "should be" instead with several levels of compliance. Examples of those are business rules establishing the metadata objects required to ensure the best possible data quality to data consumers, but where having incomplete metadata is still acceptable. Finally, descriptive metadata provides information about "what is", i.e. it describes a single data instance without any general rule. Examples of those are the record layouts describing CSV files.

**Data Transformation:** The ability to transform data stored within the organisation to make them suitable for specific purposes downstream. [FR: Data transformation is not uniform for all statistical production, the same data can be transformed in multiple ways at different steps for multiple purposes and uses]

- Clean the data to preserve internal coherence of data: correcting input errors, checking duplicates, imputing missing data, verify and correct data formats, checking inconsistent data and unaccepted values
- Check the coherence of data with data definitions coming from metadata system used in the organisation
- Check the harmonisation with classifications coming from national and international standards, correcting and imputing the right values
- Support the Process phase of the GSBPM by allowing for general data processing: creation of new derived variables, use of standard codes
- Define and create aggregations to be used for dissemination system
- Code data coming from textual or non-structured sources, looking up data from classifications and codelists
- Match data from different sources, standardising codes (ref Data Integration)
- Reduce the amount of data, filtering rows and selecting columns
- Alter the data following security or statistical significance reasons
- Ensure consistency synchronising data between different repositories

Data Transformation is the ability to transform data (already ingested and stored within the organisation ) in a format that is (re-)usable for the Sharing capability. During the transformation process, the data can be cleansed, reformatted, harmonised, enriched, and validated. Data transformation allows the mapping of the data from its given format into the format expected by the consuming application . This includes value conversions or translation functions, as well as normalising numeric values to conform to minimum and maximum values.

Data cleansing or data cleaning is the process of detecting and correcting (or removing) corrupt or inaccurate records from a data set and refers to identifying incomplete, incorrect, inaccurate or irrelevant parts of the data and then replacing, modifying, or deleting the dirty or coarse data. Data cleansing may be performed interactively with data preparation (data wrangling) tools, or as batch processing through scripting. The actual process of data cleansing may involve removing typographical errors or validating and correcting values against a known list of entities. Data Cleansing will not change the data model.

Reformatting is often needed to convert data to the same (standard) type. This often happens when talking about dates or time zones.

Harmonisation of the data is the process of minimising redundant or conflicting dimensions that may have evolved independently. Goal is to find common dimensions, reduce complexity and help to unify definitions. For example, harmonisation of short codes (st, rd, etc.) to actual words (street, road, etc.). Standardisation of data is a means of changing a reference data set to a new standard, ex, use of standard codes.

After cleansing and reformatting, a data set should be consistent with other similar data sets in the system. The inconsistencies detected or removed may have been originally caused by user entry errors, by corruption in transmission or storage, or by different data dictionary definitions of similar entities in different stores. Data cleansing differs from data validation in that validation almost invariably means data is rejected from the system at entry and is performed at the time of entry, rather than on batches of data.

Data Validation may be strict (such as rejecting any address that does not have a valid postal code) or fuzzy (such as correcting records that partially match existing, known records). Validation of data quality (no impact on data model and data). The result will mainly produce quality indicators. Data Validation will not change the data model. Some data cleansing solutions will clean data by cross checking with a validated data set.

A common data cleansing practice is data enrichment (aka enhancement), where data is made more complete by adding related information. For example, appending addresses with any phone numbers related to that address. Data Enrichment can change the data model.

**Data Integration:** The ability to connect/integrate different data sets in order to create a coherent set of information.

Data integration is a key capability of the target architecture supporting the statistical organisation's ability to fulfil information needs from different and existing sources.

It is supported by:

- Metadata-driven (schema-driven) data discovery within sources
- Data mash up and blending of heterogeneous sources (dataset, relational data bases, data warehouses, Big Data, Linked Open Data) using different techniques
- Transformation/normalisation of data available in different format : e.g. a unstructured format (NoSQL Data Base ); structured (Relational Data Base)
- Access and connection to sources APIs independently from their location (local/remote/cloud environments)
- Agile acquisition/processing and delivery data workflows with automation / batch features
- Agile data modelling and structuring allowing users to specify data types and relationships
- Generation of semantic models and ontologies

**Data Sharing:** The ability to make data and metadata available to authorised internal and external users and processes.

**Metadata-driven access to data sets**

The available data needs to be findable by using a data catalogue that provides an identifier that gives access to the referenced data set. The data catalogue capability is covered by the metadata management capability. The information provided by the Provision capability should be enough to access the data sets via tools or machine-to-machine interfaces. All data sets should be findable based on their metadata. The overview of all found relevant data sets should be ranked to distinguish the most relevant from the less relevant data sets.

**Providing direct access to data through data analysis tools or APIs or query languages**

This capability makes it possible to access data through analysis tools (e.g. statistical analysis, data preparation, visualisation, dashboarding, business intelligence) using standard protocols (e.g. ODBC, JDBC, web services (SOAP / Restful)) or through APIs or using standard query languages like SPARQL or GraphQL. This also includes a description of the used/needed authentication and authorisation to access the data sets and the selection and filtering options to retrieve the data.

**Providing access to data using open standards**

Standards for accessing data e.g. open standards. All data should be (if applicable) accessible for external parties using open data standards. Public data should be freely available to everyone to use and republish as they wish, without restrictions from copyright, patents or other mechanisms of control. For data exchange, the OData protocol is recommended. In computing, Open Data Protocol (OData) is an open protocol that allows the creation and consumption of queryable and interoperable RESTful APIs in a simple and standard way.

**Enforcing security**

It must be possible to enforce security on all data sets; the building blocks to make this possible are described in 5.8 - Security and Information Assurance (Current work version). Reduce quantity of data following security or statistical significance reasons.

**Data Management:** The ability to manage physical data organizations to support other capabilities reliably and efficiently.

- Providing data migration and replication
- Consolidating data physically and virtually
- Enabling data sharing between applications and consumers
- Facilitating interoperability via data interfaces using common exchange models
- Making data available to applications and consumers in the format and time frame needed

This capability deals with all aspects of data management at the physical level. In a way, Data Management complements Data Description and Organization: whereas the latter deals with logical and conceptual representations, the former does it with physical representations used by applications.

Data Management supports a variety of types of persistent stores, including the following: hierarchical, e.g. XML, folders; multidimensional, e.g. data cubes, star schemas; relational, e.g. SQL databases; and non-relational, e.g. column stores, spatial DBs, document stores, graph DBs, triplestores (RDF), etc. It also supports two main classes of processing styles: ACID and BASE. ACID is a set of properties that characterize transactions in traditional, relational DBMSs: Atomicity, Consistency, Isolation and Durability. In contrast, BASE is a type of processing that relaxes some or all four ACID properties and has become the processing style in newer data environments. Accessing the data from persistent repositories can be done with a variety of standard protocols and languages, including SQL, ODBC, JDBC, ADO.NET, C, C++, REST, XML, XPath and Java.

Schemas and data exchange models are fundamental for describing, and prescribing, how the data is physically organized. Data exchange models enable the organization to share and exchange enterprise information that is consistent, accurate and accessible. These models are not intended to replace the disparate set of heterogeneous schemas used to persist data across the organization. Both data consumers and producers can continue to use their own schemas for data at rest within their own environments and just translate them to data exchange models only when data needs to be shared between applications. These exchange models can be easily used by Data Services integrated with an Enterprise Service Bus or by Data-as-a-Service (DaaS) solutions. Such models can be based on the Data Documentation Initiative (DDI), the Statistical Data and Metadata Exchange (SDMX) model or other standards. Replicating and synchronizing data instances is fundamental in high-availability environments. It means maintaining copies of data for disaster recovery and to guarantee performance during peak usage. Different replication patterns can be used depending on the non-functional requirements stated in the service level agreements.

Data pipelines consists of a coherent and integrated mechanisms to manage data flows and data in motion in general, including event management, messaging, connections to persistent stores, workflow management and serialization frameworks for data exchange.

**Data Governance:** The ability to manage the life cycle of data through the implementation of policies, processes and rules in accordance with the organisation's strategic objectives.

One of the main aims of data governance is to ensure that the data has consistency and that it is trustworthy. It is important that the allocation of the ownership and stewardship need to be maintained through implementing a data governance framework, this involves defining the owners or custodians of the data assets in the enterprise. This role is called data stewardship. The data sensitivity classification is key step to building a secure organisation. Classifying the data is the process of categorising data assets based on nominal values according to its sensitivity (e.g. impact of applicable laws and regulations).

Security is a capability that is closely linked to data governance; this is described in more detail in another capability (see Security and Information Security). However, data governance has a responsibility to manage aspects of the policies relating to security and access.

The quality of statistical data is also of paramount importance, as poor quality data will undoubtedly inflict reputational damage upon the organisation. Therefore, it is imperative that quality management processes and controls are applied at all appropriate stages of processing. Data Quality management should ensure that the appropriate quality framework is used to guide these controls. As well as the possibility of reputational damage, poor data quality management could also affect the reliability of business analytics and business intelligence reporting.

Data retention management defines the policies of persistent data and records management for meeting legal and business data archival requirements. These policies should outline the criteria for archiving data (which would probably be quite numerous, considering the numbers and types of datasets managed by statistical organisations), and the processes for managing historical data.



The usage and performance monitoring need to include data logging and visualisation tools that can monitor and analyse network performance, usage patterns. The data should always be hosted in secure environment relevant to the correct security qualification of that unique data set, these facilities can be on-premise, hybrid or in the Cloud, all with an approved back up and business continuity policy.

The effort and resource put into business continuity and disaster recovery will normally be appropriate to the risk presented by the loss of the data in question. For example, if there is loss of data for a particularly important financial indicator which prevents publication at the appropriate time, the impact on not only institutional reputation, but also on the international financial markets themselves could be extreme – therefore, in this situation, business continuity is extremely important.

**Security & Information Assurance:** Security and Information Assurance is the ability to grant security and continuity to the information system, and will provide the following controls:

- Granting access to authenticated and authorised users and successfully deny access to all others
- Applying security to data in transit and at rest, to an appropriate level in line with the relevant official security classifications and Privacy Impact Assessments (if applicable)
- Ensuring the preservation of the integrity and availability of data
- Ensuring the business continuity of the system, putting in place the capability to overcome temporary problems and ensuring the availability of alternative sites in the event of a disaster
- Detecting hardware and software errors and bring the system back to a consistent state
- Managing security rules, also in connection with external systems providing data (either administrative sources or Big Data)
- Monitoring user actions to identify security breaches
- Providing intrusion detection and intrusion prevention to the hosted infrastructure
- Protecting user privacy
- The use of data encryption techniques where applicable.

The provision of Information Assurance and Security in an ever changing statistical data world has to be fluid. This is due to the changing IT landscape with an ever increasing drive to Big Data.

The fundamental ethos for Security and Information Assurance to protect the confidentiality, integrity and availability of data remains unchanged, regardless of the sources of the data. It is important that the security of the statistical organisation engenders trust from the stakeholders, whether it be data suppliers (whose interest would be maintaining security of data which is probably confidential), or data consumers (who would be interested in the integrity and quality of the data).

With increased access to sources of Big Data, and forging partnerships with other public and private organisations, security is essential. Working with big data is becoming ever more important to national and international statistical systems for fulfilling their mission in society.

In order to advance the potential of official statistics, statistical organisations will need to collaborate rather than compete with the private sector. At the same time, they must remain impartial and independent, and invest in communicating the wealth of available digital data to the benefit of stakeholders. We must consider the (now wider) range of data sources, which will include: Traditional (paper based) surveys, On line surveys (in house hosted in cloud), On line surveys direct to businesses and individuals, On line surveys hosted and run by 3rd Parties, Data purchased from commercial organisations, Web-scraped data, or other internet-based data sources, Shared Government data.

Each of these will have their own inherent security risks associated with them, and each must have the appropriate security controls applied to them. The use of a series of data zones with various levels of security controls can help to cater for the variety of requirements and needs of the different datasets.

A major objective of IA and security is to facilitate access to Big Data sources as input into official statistics production. As these sources have their own potential security risks associated with them (e.g. unknown provenance, unknown virus status etc.), particular care needs to be taken to ensure the appropriate level of security controls are applied.

Where data is being shared with other organisations, there will be a need to provide assurance that the statistical agency will protect shared data to an acceptable level. This assurance can be facilitated by forming partnerships with the other organisation(s), whether they are public or private sector organisations, and setting up some form of service level agreements where the security controls to be applied to the datasets in question can be agreed.

Other data security risks can be realised when data from different sources is matched and linked, especially when applied to person information.

Additionally, data should undergo disclosure checking where there is a risk of revealing information about an individual or organisation, especially where, for example, it could lead to detriment to the individual, or commercial damage to a business. This is particularly important for data that is being prepared for publication or dissemination.

There will be a need for data to undergo stringent checks when it is being brought into an organisation, regardless of its source and method of ingestion (e.g. streaming, batch, etc.). Multi-AV scanning should be adopted to reduce the risk of infection by viruses.

It is important that security and information assurance needs to be considered in the context of the data stored and used by the statistical organisation all through the statistical process.

**Metadata Management:** The ability to record, maintain, validate and query all metadata relevant to the statistical organisation.

Metadata has a dual aspect in the framework of a data architecture. On one hand, metadata is data, and as such all the capabilities defined in CSDA apply to metadata. On the other hand, metadata has a special status since it conveys all the context needed to understand the data: without metadata, data is useless. Therefore, metadata represents a particularly valuable type of data, and is actually one of the organisation's most precious asset.

Metadata describes technical and business processes, business rules and constraints, and data structures at both logical and physical levels. Metadata also describes the concepts represented in the data, such as business processes, provenance and lineage, statistical variables, unit types, classifications, etc. Metadata enables other capabilities in the architecture and is integral to the management of databases and other technical components, like collection instruments, coding tools etc. and to the localization and access of data sets.

Metadata management has three important aspects: semantic consistency, conformance to standards and actionability. Semantic consistency means that metadata is precisely defined according to a well-known modelling framework, and that coherent naming rules are established and applied throughout the organisation. This could require specific skills and organisational measures. Conformance to standards is a good way to achieve semantic consistency, since standards usually undergo collaborative production processes that confer them a high quality level. It is also essential to interoperability between organisations, notably at an international level. It may be useful for statistical organisations to participate in the governance of the standards that they use. Actionability means that metadata is not only documentation, it is also used to automate parts of the statistical production process. This usually implies that metadata is stored in specific machine-actionable formats, which requires particular expertise.

The metadata associated with data assets could be prescriptive, normative or descriptive. Prescriptive metadata establishes what "must be", i.e. there are no exceptions and all data instances must comply. Examples of those are integrity constraints and data types as defined in database schemas, where a data instance cannot even be loaded unless it is validated against the schema. Normative metadata is a weaker form of prescriptive metadata in which there is a notion of "should be" instead with several levels of compliance. Examples of those are business rules establishing the metadata objects required to ensure the best possible data quality to data consumers, but where having incomplete metadata is still

acceptable. Finally, descriptive metadata provides information about “what is”, i.e. it describes a single data instance without any general rule. Examples of those are the record layouts describing CSV files.

Specific attention must be given to metadata for data coming from external sources, like Big Data. In this case, metadata must be obtained in collaboration with data provider, trying to set up common standards and vocabularies.

**Provenance & Lineage:** The ability to manage and obtain provenance and lineage of data.

Official Statistics will increasingly use data from different sources (both corporate and external). In order to be able to assess the quality of the data product built on these data, information on data's origin is required. The provenance and lineage data can be information on processes, methods and data sources that led to product as well as timeliness of data and annotation from curation experts.

Provenance is information about the source of the data and lineage is information on the changes that have occurred to the data over its life-cycle. Together they both provide the complete traceability of where data has resided and what actions have been performed on the data over the course of its life.

This capability entails the recording, maintenance and tracking of the sources of data, and any changes to that data throughout its life-cycle, in particular it should include date/timestamps, and who/what carried out the changes.

The World Wide Web Consortium (W3C) provides an ontology to express provenance and lineage data.

**Knowledge Management:** The ability to manage intellectual capital (knowledge) in all its forms.

- Capturing and formalizing knowledge in semantic models and actionable formats, like RDF, SKOS/XKOS, OWL, etc.
- Maintaining multiple versions of semantic models and knowledge representations at different levels of abstraction, and the lineage between them.
- Maintaining mappings between different models to support translations between vocabularies.
- Maintaining supporting artifacts, like architecture documents, best practices, guidelines, etc.
- Supporting inference and reasoning to derive new knowledge from existing one.

This is a capability to create, organize, augment, and share intellectual capital (knowledge) relevant to an organization or domain. It includes the creation and management of an environment to turn information into actionable knowledge, maintained in a virtual repository, to benefit all aspects of the statistical production. Implementing this capability requires understanding the agency's information flows and implementing knowledge acquisition and representation practices to make key aspects of its knowledge base explicit in a usable form.

Knowledge is ubiquitous: it resides not only in documents and databases, but also in experts' minds and the agency's routines, processes and practices. That's why its capture and formalization is difficult but at the same time critical to the agency's success.

Knowledge management deals not only with business knowledge but also with “support” knowledge that helps the organization to function, e.g. architecture documents, methodological approaches, data quality guidelines, security policies, etc.

Models are abstract description that hide certain details and emphasize others. A semantic model is an interconnected network of concepts linked by semantic relationships. The RDF graph data model is a semantic model consisting of a collection of triples of the form subject, predicate, object. Each triple can be viewed as an assertion about a relationship (the predicate) that holds between the subject and the object. RDF was developed by the W3C for the Semantic Web and provides a mechanism to make knowledge actionable and to derive (or infer) new knowledge from explicitly represented knowledge.

RDFS is a simple ontology language, or vocabulary, built on top of the RDF data model. An RDFS ontology also consists of a collection of triples, but this time subject and object are RDFS resources. In other words, the RDFS ontology is a collection of assertions between resources. In addition, properties in RDF are grouped into a class, which means they can also be extended.

A multitude of ontologies, classifications, taxonomies and thesauri exist to organize knowledge in RDF. Many of the commonalities among them are captured by the Simple Knowledge Organization System (SKOS), which is extended by the Extended Knowledge Organization System (XKOS) to cover the statistical classifications domain.

OWL is a knowledge representation language for building ontologies that represent complex domain models. OWL is more expressive than RDFS and has many advantages, including a clear separation between classes and individuals, a classification of properties (object, data and annotation), richer built-in datatypes and a variety of axioms to express logical statements about class relationships, property constraints (domain, range), etc. The latest version of the language is OWL 2.