Common Statistical Data Architecture (CSDA)

(Version 1.0)

Table of Contents

List of abbreviations

Term	Meaning
CSDA	Common Statistical Data Architecture
CSPA	Common Statistical Production Architecture
DCAT	Data Catalog Vocabulary
DDI	Data Documentation Initiative
EIRA	European Interoperability Reference Architecture
FAIR	Findable, Accessible, Interoperable and Reusable
GSBPM	Generic Statistical Business Process Model
GSIM	Generic Statistical Information Model
HLG-MOS	High Level Group for the Modernisation of Statistical Production and Services
OWL	Ontology Web Language
PROV	A W3C activity. The goal of PROV is to enable the wide publication and interchange of provenance on the Web and other information systems
PROV-O	The PROV Ontology
SDMX	Statistical Data and Metadata eXchange
SKOS	Simple Knowledge Organization System (W3C Semantic Web)
TOGAF	The Open Group Architectural Framework

Preface

This document is the result of the 2017 Data Architecture project, the first project to specifically address the data aspects of statistical production. It is to be expected that other, more mature versions of CSDA will be released in future.

This 2017 version includes an explanation of the purpose and use of the Reference Data Architecture, but focusses on the functionality that statistical organisations will need for the design, integration, production and dissemination of official statistics based on both traditional and new types of data sources.

The main content of the CSDA in this 2017 version therefore is the description of the Capabilities and Building Blocks. Capabilities are abilities, typically expressed in general and high-level terms that an organisation needs or possesses. Capabilities typically require a combination of organisation, people, processes, and technology. Building Blocks represent (potentially re-usable) components of business, IT, or architectural capability that can be combined with other Building Blocks to deliver architectures and solutions.

In this 2017 version, a first attempt is made to describe the characteristics and structure of the actual data and metadata. This is an area that is particularly difficult to tackle - as most statistical organisations still struggle to find the right approach. Part of this section is the semantic layer, where ontologies (semantic models) have their place. Future work will be needed to further develop this (important) component of CSDA.

The 2017 project also defined a number of Use Cases for checking and testing the CSDA. These Use Cases (and the outcome of their execution) are described in a separate document.

Future work will also include Guidance on the use and implementation of the CSDA.

I. Introduction

1. Statistical organisations have to deal with many different external data sources. From (traditionally) primary data collection, via secondary data collection, to (more recently) Big Data. Each of these data sources has its own set of characteristics in terms of relationship, technical details and semantic content. At the same time the demand is changing, where besides creating output as "end products", statistical organisations create output together with other institutes.

2. Statistical organisations need to find, acquire and integrate data from both traditional and new types of data sources in an ever increasing pace and under ever stricter budget constraints, while taking care of security and data ownership. They would all benefit from having a reference architecture and guidance for the modernisation of their processes and systems.

3. A Data Architecture is defined as:

- **"**A **data architecture** is *[an architecture that is]* composed of models, policies, rules or standards that govern which data is collected, and how it is stored, arranged, integrated, and put to use in data systems and in organizations." (Wikipedia[1])

- "A description of the structure and interaction of the enterprise's major types and sources of data, logical data assets, physical data assets, and data management resources." (TOGAF 9, Part I[2])

4. Although CSDA is (loosely) based on TOGAF, it should be stressed that "data" to statistical organisations means something different from what is understood by most industries. "Data", to statistical organisations, is the raw material, the parts and components and the finished products, rather than the information needed to support and execute the organisation's primary processes (although, also in statistical organisations, there is data that plays that role, of course). Although the definition still applies, "data architecture" as meant in this document also has a (slightly) different scope.

II. Purpose

5. A reference architecture is described as **"**in the field of software architecture or enterprise architecture [a reference architecture] provides a template solution for an architecture for a particular domain. It also provides a common vocabulary with which to discuss implementations, often with the aim to stress commonality." (Wikipedia[3])

6. In the context of this document, the domains are the (data aspects of) individual statistical organisations around the world.

7. The purpose and use of the Common Statistical Data Architecture as a reference architecture is to act as a template for statistical organisations in the development of their

[1] https://en.wikipedia.org/wiki/Data_architecture
[2] http://pubs.opengroup.org/architecture/togaf9-doc/arch/chap03.html
[3] https://en.wikipedia.org/wiki/Reference_architecture

own Enterprise Data Architectures. In turn, this will guide Solution Architects and Builders in the development of systems that will support users in doing their jobs (that is, the production of statistical products).

8. The CSDA supports statistical organisations in the design, integration, production and dissemination of official statistics based on both traditional and new types of data sources.

9. The CSDA shows the organisations how to organise and structure their processes and systems for efficient and effective management of data and metadata, from the external sources through the internal storage and processing up to the dissemination of the statistical end-products. In particular, in order to help organisations modernise themselves, it shows how to deal with the newer types of data sources such as Big Data, Scanner data, Web Scraping, etc.

10. The CSDA must be seen in the context of a whole suite of standards, developed and maintained by the international statistical community, led by HLG-MOS. Among these are GSBPM, GSIM, and CSPA. Where applicable, the CSDA also links to other international standards such as TOGAF, DDI, SDMX, etc.

11. Another useful reference is the European Interoperability Reference Architecture (EIRA[4]). EIRA focuses on building blocks and distinguishes architecture and solution building blocks. In EIRA terms, an architecture building block "represents a (potentially re-usable) component of legal, organisational, semantic or technical capability that can be combined with other architecture building blocks".

III. Scope

12. The scope (or focus) of the CSDA includes all of the GSBPM phases, specifically the designing, building and use of processes and systems in statistical data collection, production, analysis and dissemination, based on external needs, in statistical organisations. It addresses Data Analysis related to (new) sources of data. There is no restriction as to the types of data. Besides the operational phases, the CSDA addresses cross cutting issues: Data Governance, Traceability, Quality and Security.

13. The CSDA is restricted to conceptual and logical descriptions of components. Where there is reference to technical implementation, this is done is a conceptual or logical way (light touch).

14. From the above, it follows that supporting non-statistical, business processes such as HR and Finance, are out of scope for the CSDA.

IV. Benefits

15. The benefits of having a CSDA are:

- Independence from technology: Statistical organisation processes and systems will, eventually, become more robust to technological evolution

[4] https://joinup.ec.europa.eu/solution/european-interoperability-reference-architecture-eira

- Sustainability: A reference architecture that is shared by the worldwide statistical community is necessary as a common vocabulary for exchanging ideas and collaboration on development and maintenance of new solutions, processes and systems.
- Maintainability: Maintenance of statistical organisation architectures and solutions is facilitated by the availability of a reference architecture that is shared by a larger community.
- Cost saving to global optimisation strategies/solutions: By referencing a shared framework, statistical organisations can better collaborate in the development, maintenance and use of common solutions.

V. Use of the CSDA

16. Different users at different steps in the process can use the CSDA. We see four different, but related, steps (see Figure 1):

1. **CSDA**: a document created and maintained at the statistical community level;
2. **Statistical Organisation Data Architecture**: a document created and maintained at the statistical organisation enterprise level;
3. **Solution (Data) Architecture**: a document created and maintained at a lower level inside a statistical organisation such as a statistical business division.
4. **Solution**: a process or system used by a user (statistician) to do his daily job: producing statistical products.

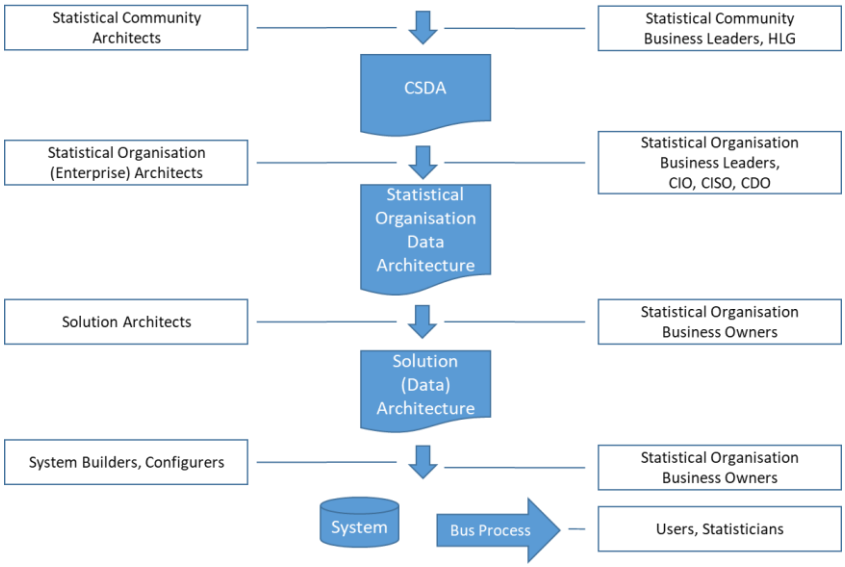

Figure 1. Use of CSDA

17. In each step, there are two broad groups of people involved. On the one hand, there is the business management, as **stakeholders** that will use (data) architecture as a means to formulate, communicate and enforce their strategies and policies. On the other hand, there are the **experts**, such as (business and IT) architects, methodologists, etc. that are the actual users and creators of (data) architectures and solutions.

18. Business Leaders may use the data architectures in two ways, and as a result, they appear on two levels in Figure 1. The data architecture is the vehicle for business leaders to

formulate and communicate their policies and strategies. As such, they are the owners of the document. Business leaders may (should) also use it as the book of law to enforce their policies. As such, they act one level below the previous role, because they now must ensure that the next level conforms to those policies as laid down in the data architecture.

CSDA step

19. In this step, the UNECE Data Architect group defines the guidelines for implementing a statistical organisation Data Architecture. The CSDA includes a complete overview of Capabilities (based upon GSBPM), and a growing overview of Conceptual Building Blocks.

20. Roles involved in this step:

- Stakeholders: Senior level managers, such as (deputy) Chief Statisticians, CIO's, CISO's, etc., acting as representatives of statistical organisation, collaborating on globally, representing the statistical community as a whole;
- Experts: Enterprise architects, methodologists, senior advisors, etc., collaborating in (project) groups.

Statistical Organisation Data Architecture step

21. In this step, the statistical organisation Data Architects will use the CSDA as a basis for the statistical organisation Data Architecture. Not all parts of the CSDA need to be described in a statistical organisation data architecture (although this is strongly recommended). In this step, the Logical Building Blocks (derived from the Conceptual Building Blocks) will be further defined. New projects could initiate the definition of new Logical Building Blocks or reuse of existing Building Blocks based on the Business Requirements. The CSDA should provide the guidelines and knowledge that help in creating new Logical Building Blocks.

22. Roles involved in this step:

- Stakeholders: Senior level managers, such as (deputy) Chief Statisticians, CIO's, CISO's, and business division managers that lead and sponsor the development of their organisations, including the development and maintenance of processes and systems. These managers need to ensure that the architecture reflects and supports the strategic goals and policies for the domain.
- Experts/Users: statistical organisation Enterprise architects, methodologists, senior advisors, developing (data) architectures on behalf of the senior leaders leading the business domains to which those architectures apply.

Solution (Data) Architecture step.

23. In the Solution Architecture step the statistical organisation's Data Architecture will be used to implement new services/components based on the described Logical Building Blocks. If Capabilities/Building Blocks are not yet developed in the statistical organisation's Data Architecture, the CSDA could be used to start developing/implementing the missing Building Blocks. In addition, existing CSPA services and/or Logical Building Blocks of other statistical organisations can be considered here for inclusion in the Solution.

24. Roles involved in this step:

- Stakeholders: Business leaders for domains needing (new) solutions such as (business) processes and systems.
- Experts/Users: Solution architects, methodologists, etc. involved in the design, development and maintenance of processes and systems.

25. In general, the target audience (those involved in the definition) and user groups (those that are informed) for Data Architecture, consist of the following types of people:

- Audience:
  - Managers
  - Business people (subject matters, process owners)
  - Design authority, high-level committee on investments
  - Enterprise Architects
  - Solutions/IT Architects (logical layer, mapping with applications)
- Users:
  - External data providers for Data Collection
  - Audit agencies
  - External solution providers

26. The following use cases show how the CSDA can be used by the various roles:

- As a manager, I want to understand how a Data Architecture can deliver on strategic outcomes (like data integration, new data sources etc.) so that there is a common high level approach
- As a manager or expert, I want to be able to share solutions between different agencies
- As a business leader or expert, I want to (be facilitated to) influence managers to get Data Architecture and Enterprise Architecture adopted as a discipline
- As a business person, I want to be guided by a common framework so I can design my solutions without reinventing any wheels, keeping the cost down and lead-times minimal
- As the Design Authority, I want to review and endorse so that a Data Architecture is delivered that is aligned to objectives and policies but that can also be translated into something that can be implemented.

VI. Key principles

27. Data in this context is the raw material, semi-finished and finished product of the statistical organisation, rather than (as well as) the information, the organisation needs to manage its processes. Metadata is information about Data; Information is the general term meaning both Data and Metadata.

28. Metadata is data that describes data. Meta is a prefix that in information technology means “an underlying definition or description”. Metadata summarises basic information about data, which can make finding and working with particular instances of data easier.

29. These principles are compatible with FAIR data principles[5]: Findable, Accessible, Interoperable, and Reusable.

[5] Ref: http://www.fairdata.org.uk/

Table 1. Key Principles

Principle	Statement	Rationale	Implications
1. Information is managed as an asset throughout its lifecycle	• Information includes both the data and the metadata describing that data; • Information includes all objects that describe the context, content, controls and structure of data and metadata; • Information is an organisational asset that all employees have a responsibility to manage; • Information must be actively managed throughout its lifecycle from creation to disposal; • The ownership, status, quality and security classification of information should be known at all times.	• The statistical organisation has a responsibility to manage the data and metadata it acquires in accordance with relevant legalisation; • Managing the information is necessary to guarantee constant quality of statistical products; • Information needs to be managed to ensure its context and integrity is maintained over time; • As information is increasingly shared across business processes it is important to understand the dependencies of its use.	• The statistical organisation will take an enterprise approach to managing information as an asset; • Organisational policies and guidelines will be put in place to ensure data will be managed in accordance with this principle; • All data assets will have an owner responsible for their management; • Staff will be trained to understand the value of data and their individual responsibilities; • Data quality and sensitivity will be documented where required for business processes; • Data will be protected against loss; • Data and metadata must not be kept longer than necessary in order to protect privacy; it should be deleted at the end of its lifecycle.
2. Information is accessible	• Information is discoverable and usable; • Information is available to all unless there is good reason for withholding it; • Data and metadata is accessible to humans as well as machines.	• Ready access to information leads to informed decision-making and enables timely response to information needs; • Users (internal and external) can easily find information when they need it, saving time and avoiding repetition.	• The organisation will foster a culture of information sharing; • Information will be open by default; • The way information is discovered and displayed will be designed with users in mind; • Systems will be designed to ensure that the minimum amount contextual information required to understand information is captured; • Staff will create and store information in approved repositories;
3. Data is described to enable reuse	• Data must have sufficient metadata so it can be understood outside its original context;	• Data can be easily understood and used with confidence without requiring further information;	• Staff will document data with reuse in mind; • Staff will consider reuse when designing systems for capturing information.

	• Connections between data objects must be documented; • Restrictions to data usage must be documented.	• Data and its related metadata can be easily reused by other business processes reducing the need to transform or recreate information; • The dependencies and relationships between data objects can be easily known.	
4. Information is captured and recorded at the point of creation/receipt	• Information should be captured and recorded at the earliest point in the business process to ensure it can be used by subsequent processes; • Subsequent changes to information should be documented at the time of action.	• Information is captured and recorded at the time of creation/action so it is not lost; • The amount of information reuse is maximised by capturing it as early as possible.	• Systems will be designed to automatically capture information resulting from business processes; • Staff will need to prioritise and be given time to capture information when it is fresh in their minds.
5. Use an authoritative source	• Within a business process, there should be an authoritative source from which information should be sourced and updated; • Where practical, existing information should be reused instead of recreated or duplicated.	• Maintaining fewer sources of information is more cost effective; • Having one source of information supports discovery, reuse and a 'single version of truth'.	• There will be authoritative repositories for different types of information; • Information needs will be satisfied using existing sources where possible.
6. Use agreed models and standards	• Key information should be described using common, business-oriented, models and standards, agreed by the organisation.	• Having agreed models and standards will enable greater information sharing and reuse across the business process; • Having agreed models and standards will enable staff to communicate using a common language.	• There will be responsibility assigned for creating and maintaining agreed models and standards; • Staff will be made aware of what the approved models and standards are and how to use them; • Agreed models and standards will enable external collaboration but also be fit for business purposes; • Agreed models and standards will form the basis of system and process design, deviations from the standards and models will be by agreed exception only.

VII. Capabilities and building blocks

30. CSDA is (loosely) based on TOGAF. According to TOGAF 9 definitions[6]:

- A capability is "an ability that an organization, person, or system possesses. Capabilities are typically expressed in general and high-level terms and typically require a combination of organization, people, processes, and technology to achieve. For example, marketing, customer contact, or outbound telemarketing".
- A building block "represents a (potentially re-usable) component of business, IT, or architectural capability that can be combined with other building blocks to deliver architectures and solutions".

31. This aligns with the definition provided by CSPA: "Capabilities provide the statistical organisation with the ability to undertake a specific activity. A capability is only achieved through the integration of all relevant capability elements (e.g. methods, processes, standards and frameworks, IT systems and people skills)".

32. In CSDA, capabilities are expressed at the conceptual level and building blocks at the conceptual and logical levels. Capabilities and building blocks are restricted to activities regarding the management and use of statistical data and metadata.

33. In general terms, building blocks realise capabilities and are the basic elements to design and build solutions.

34. In order to realise (implement) a certain Capability, a statistical organisation may use “reusable components” or Building Blocks. In this document, Building Blocks are defined and described on a conceptual (the “what”) and on a logical (the “how”) level.

35. On the highest level, the CSDA distinguishes 8 Capabilities, divided into 2 groups: Core Capabilities and Cross-Cutting Capabilities. Each of these 8 capabilities is described in the following sections, where also the Conceptual and Logical Building Blocks for each Capability are listed.

36. The core capabilities are:

- Data Ingestion (section 1A)
- Data Transformation (section 1B)
- Data Integration (section 1C)
- Provisioning (section 1D)

37. The cross-cutting capabilities are:

- Metadata Management (section 1E)
- Data Governance (section 1F)

[6] http://pubs.opengroup.org/architecture/togaf9-doc/arch/chap03.html

- Provenance & Lineage (section 1G)
- Security & Information Assurance (section 1H)

38. Figure 2 shows an overview of the Capabilities and Building Blocks

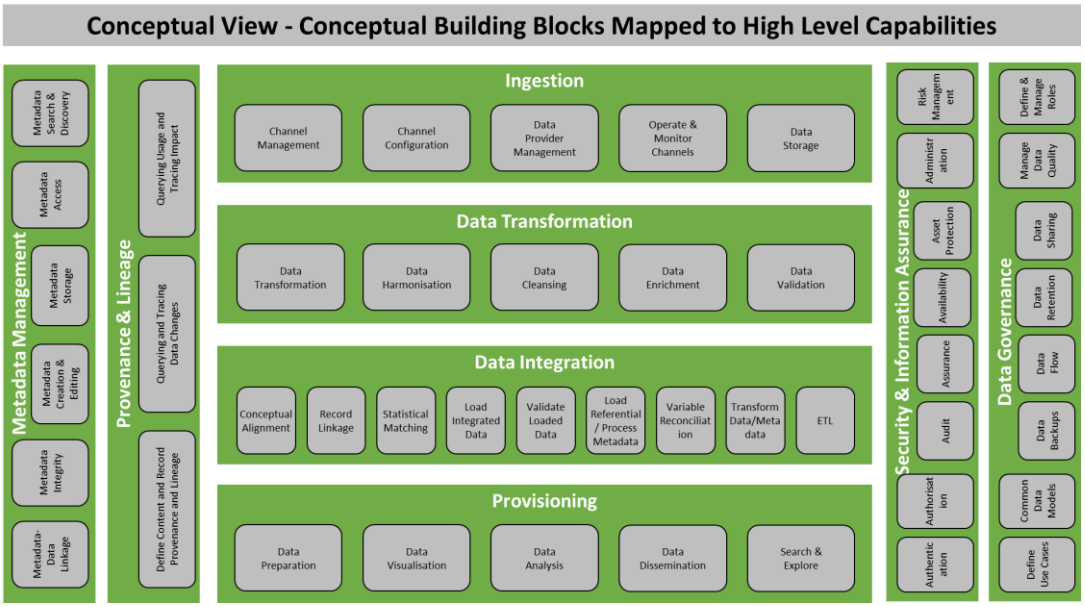

Figure 2: Overview of Capabilities and (conceptual) Building Blocks

39. The remainder of this section describes each of the capabilities in turn, including the conceptual and logical building blocks that it may use.

A. Data Ingestion

Summary

40. The ability to receive data from external sources or providers. This includes:

- Accessing data sources through a variety of channels (for example API, web questionnaires, administrative archives, streaming data, etc.)
- Loading data into a storage facility (for example relational database, NoSQL, big data storage, etc.)
- Utilising the metadata management capability to capture the relevant information about the ingested data
- Managing the relationship with data providers (for example respondent management and Service Level Agreements for administrative data sources)

41. Note that sample design and drawing a sample are out of scope for this capability.

Description

42. Data ingestion is the first step in the typical statistical production process, and an essential one: no statistical information can be built without data. We call "data ingestion" the capability that consists in establishing and securing data provision agreements with relevant providers, creating reliable data provision channels with existing sources, acquiring the data under specified formats and loading it in a controlled data storage environment, and creating (or

using existing) metadata to describe the data and acquisition process.

43. 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 and corresponding collection instruments (typically survey questionnaires) have to be designed and created. In other cases, existing data can cover the needs and the problem is to secure access to it.

44. In the case of data coming from external sources, we have to assess the reliability of the source and its stability in terms of time and format. For data coming from private companies, additional actions needs 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.

45. 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. Some modification treatments can be applied to facilitate data integration, such as filtering, transcoding, normalisation, translation, codification, etc.

46. The Data Ingestion capability is very 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.

Conceptual Building Blocks

47. The conceptual building blocks for the Data Ingestion capability are:

- Channel Management
- Channel Configuration
- Data Provider Management (e.g. Respondent, Admin register owner)
- Operate & Monitor Channels
- Data Storage

Logical Building Blocks

48. The logical building blocks for the Data Ingestion capability are:

- Channel Management
  - Create, maintain and withdrawing (data collection) channels
  - Generate metadata describing new channel
- Channel Configuration
  - Configure channel for specific data streams (for instance surveys)
  - Generate metadata to describe dataset types (data streams)
- Data Provider Management

    - Securing data (establishing contracts, legislatively securing data provision, etc.)
    - Case Management: ensuring data provider agreements are being executed, in a timely fashion, for each scheduled delivery.
- Operate & Monitor Channels (a channel can be an interviewer; in that case, it consists on assessment, training, assignment, etc...)
    - Operate & Monitor Channel
    - Managing capacity
    - Generate basic operational metadata/paradata about the individual datasets
- Data Storage
    - Load, search, access
    - Provision space for data
    - Manage authorisation

B. Data Integration

Summary

49. The ability to connect/integrate different data sets in order to create a coherent set of information.

Description

50. Data integration is a key capability of the target architecture supporting our ability to fulfil information needs from different and existing sources.

51. 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 (Mongo DB); 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

Conceptual Building blocks

52. The conceptual building blocks for the Data Integration capability are:

- Conceptual alignment
- Record linkage
- Statistical matching (deterministic and probabilistic)
- Load integrated data
- Validate/cleansing of loaded data
- Load referential and process metadata
- Variable reconciliation
- Transform data/metadata between different sources for integration purposes

- ETL (Extraction/Transformation/Loading) functionality designing data integration/fusion jobs

Logical Building blocks

53. The logical building blocks for the Data Integration capability are:

- Conceptual alignment
  - Access Unitary Metadata System to retrieve structural metadata
  - Concept and semantic mapping
- Record linkage
  - Intra-source deduplication
  - Apply decision model
  - Human review possible matches
- Load integrated data
  - Apply the transformations to the integrated
- Variable reconciliation
  - Variables profiling
  - Editing integrated data
  - Apply a reconciliation method
- Validate loaded data
  - Apply a validation method
- Load Referential/process metadata
  - Update Unitary metadata system

C. Data Transformation

Summary

54. The ability to transform raw data stored within the organisation to make them correct and usable for statistical production.

Description

55. Data Transformation is the ability to transform data (already digested and stored within the organisation) in a format that is (re-)usable for the provisioning 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 appropriate application. This includes value conversions or translation functions, as well as normalising numeric values to conform to minimum and maximum values.

56. 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.

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

58. 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.

59. 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.

60. 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.

61. 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.

62. The required capability includes:

- 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

<u>Conceptual Building Blocks</u>

63. The conceptual building blocks for the Data Transformation capability are:

- Data Transformation

- Data Harmonisation
- Data Cleansing
- Data Enrichment
- Data Validation

Logical building blocks

64. The logical building blocks for the Data Transformation capability are:

- Data Transformation
  - Formatting data following standards and/or external archives and/or metadata
  - Editing data that shows errors like: missing values, duplicates, wrong codes
  - Imputation of missing values
  - Cleansing the data from wrong values
  - Filtering of rows
  - Selection of columns
- Data Harmonisation
  - Coding and Decoding of datasets to make them comparable and linkable
  - Weighing, Fitting, Seasonal Correction
- Data Cleansing
  - Data Editing: Correcting data with help of validated data sets
  - Duplicate elimination: Duplicate detection requires an algorithm for determining whether data contains duplicate representations of the same entity. Usually, data is sorted by a key that would bring duplicate entries closer together for faster identification
  - Parsing: for the detection of syntax errors. A parser decides whether a string of data is acceptable within the allowed data specification. This is similar to the way a parser works with grammars and languages.
  - Statistical methods: used to handle missing values that can be replaced by one or more plausible values, which are usually obtained by extensive data augmentation algorithms.
- Data Enrichment
  - Data Enhancing: adding new data to existing datasets by joining data with other datasets.
  - Data Aggregation: moving from elementary (micro) data to aggregated (macro) data
  - Derivation: deriving new variables from existing ones
- Data Validation
  - Validity checking: The degree to which the measures conform to defined business rules or constraints (see also Validity (statistics))
  - Accuracy checking: The degree of conformity of a measure to a standard or a true value
  - Checking on Completeness: The degree to which all required measures are known
  - Consistency checking: The degree to which a set of measures are equivalent in across systems
  - Uniformity checking: The degree to which a set data measures are specified using the same units of measure in all systems
  - Check data on following security or statistical significance reasons

D. **Provisioning**

Summary

65. The ability to make data and metadata available to authorised internal and external users

and processes. The required capability is:

- Metadata-driven access to data sets, for example APIs or through a data catalogue (provided via the metadata management capability)
- Providing direct access to data through data analysis tools or APIs or query languages
- Access and disseminate data using open standards

Description

Metadata-driven access to data sets

66. 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

67. 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

68. 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

69. 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.

Conceptual Building Blocks

70. The conceptual building blocks for the Provisioning capability are:

- **Data Preparation** is the act of preparing (or pre-processing) "raw" data or disparate data sources into refined information assets that can be used effectively for various business purposes such as analysis. Data Preparation is a necessary, but often tedious, activity that is a critical first step in data analytic projects. It is in fact similar to data transformation but plays at the (end) user side.
- **Data Visualization** is viewed by many disciplines as a modern equivalent of visual communication. It involves the creation and study of the visual representation of data, meaning "information that has been abstracted in some schematic form, including attributes

or variables for the units of information". A primary goal of data visualisation is to communicate information clearly and efficiently via statistical graphics, plots and information graphics.

- **Data Analysis** is a process of inspecting and modelling data with the goal of discovering useful information, suggesting conclusions, and supporting decision-making. Data analysis has multiple facets and approaches, encompassing diverse techniques
- **Data dissemination** is the distribution or transmitting of statistical, or other, data to end users. The exchange of data should preferable be done using open standards. SDMX (stands for 'Statistical Data and Metadata Exchange') and Data Documentation Initiative (DDI) are standards that should be considered in this context.
- **Search & Explore** is the ability to search and explore data sets based on their metadata. For this, the access to the data is not strictly necessary, as long as all metadata is accessible. In addition to finding the most relevant data sets, the ranking of these data sets is also essential so that the user can easily distinguish the relevant (important) data sets from the less relevant data sets. For exploring the data, a quality indicator is also invaluable.

Logical Building Blocks

71. The logical building blocks for the Provisioning capability are:

- Data Preparation
  - Cleansing
  - Imputation
  - Joining
  - Conversion
  - Reformatting
  - Aggregation
  - Sorting
- Data Visualisation
  - Tables
  - Different charts
  - Info graphics
  - Dashboarding
  - Scorecarding
- Data Analysis
  - Regression
  - Correlation
  - Trends analyses
  - Deep learning
  - Machine learning
- Data dissemination
  - Overview of datafiles, catalogues, dataflows,
  - API's, data streams, web services, open data, etc.
- Search & Explore
  - Exploration of datasets
  - Search capability on metadata

E. Metadata Management

Summary

72. The ability to record, maintain, validate and query at the semantic level all metadata relevant to the statistical organisation, and connect them with data sets.

Description

73. Metadata has a dual aspect in the framework of a data architecture. On one hand, metadata is data, and as such all the capabilities and building blocks 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 the organisation's most precious asset.

74. Metadata lives at the semantic level, and thus specific features apply to its management. Particularly relevant aspects for metadata are:

- semantic consistency
- conformance to standards
- actionability

75. Semantic consistency means, for example, 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.

76. 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.

77. Actionability is an important characteristic of metadata. It means that metadata is not only documentation, but that it is used actively in the statistical production process. For example:

- Descriptive metadata used for generating parts of the statistical process or given tools like data collection instruments.
- Discoverability metadata used to localise and access data sets, as mentioned in 5.4 – Provisioning

78. Actionability usually implies that metadata is stored in specific machine-actionable formats, which requires particular expertise.

79. 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.

Conceptual Building Blocks

80. The conceptual building blocks for the Metadata Management capability are:

- Data-metadata linkage
- Metadata integrity

- Metadata creation and editing
- Metadata storage
- Metadata access
- Metadata discovery and search

Logical building blocks

81. The logical building blocks for the Metadata Management capability are:

- Data-metadata linkage
  - Create links between metadata and data
  - Maintain those links
  - Navigate (in both directions) on those links
- Metadata integrity
  - Enforce integrity of data as defined in the metadata.
- Metadata creation and editing
  - Create and edit metadata in conformity with the organisation wide metadata model
- Metadata storage
  - Persist metadata
  - Safeguard against loss
  - Restore if necessary
- Metadata access
  - Access and retrieve metadata from the repository
- Metadata discovery and search
  - List all the metadata in the repository
  - Search in different ways some metadata in the repository

F. Data Governance

Summary

82. 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.

- Allocation of ownership and stewardship roles around the data
- Data sensitivity classification
- Security and access policies management
- Data Quality management
- Retention management
- Usage and performance monitoring
- Business Continuity / Disaster Recovery

Description

83. 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).

84. Security is a capability that is closely linked to data governance; this is described in more detail in another capability (see 5.8 Security and Information Security). However, data governance has a responsibility to manage aspects of the policies relating to security and access.

85. 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.

86. 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.

87. 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.

88. 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.

Conceptual Building Blocks

89. The conceptual building blocks for the Data Governance capability are:

- Define use cases for different category of users
- Common data models
- Data Backup
- Data flows
- Defining and managing roles for data
- Managing data quality
- Data retention
- Data sharing

Logical Building Blocks

90. The logical building blocks for the Data Governance capability are:

- Define use cases for different category of users:

    - Grant access,
    - Define views
    - Profiling users
- Common data models:
    - Data restructuring
    - Data versioning
    - Data reorganisation
- Data Backup:
    - Saving data
    - Restoring data
    - Defining backup strategies
- Data flows:
    - Defining
    - Managing
    - Monitoring
- Defining and managing roles for data:
    - Owner
    - Steward
    - Custodian
- Managing data quality:
    - Defining indicators for data quality
    - Checking and monitoring indicators
- Data retention:
    - Define policy for historical data management,
    - Historical data maintenance
- Data sharing:
    - Provide knowledge about data availability to improve reuse (metadata?)

G. Provenance and Lineage

Summary

91. The ability to manage and obtain provenance and lineage of data.

Description

92. 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.

93. 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.

94. 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.

95. World Wide Web Consortium (W3C)[7] provides an ontology to express provenance and lineage data.

Conceptual Building Blocks

96. The conceptual building blocks for the Provenance and Lineage capability are:

- Defining the content and recording of data provenance and lineage e.g. provider, conditions, licences, transformation, use and reuse
- Querying and tracing all the changes made to the data: user, date/time, before/after images,
- Querying usage and impact: trace the impact of a change to a datasets on other datasets and products

Logical Building Blocks

97. The logical building blocks for the Provenance and Lineage capability are:

- Defining provenance/linage model
- Tracking automatically provenance/lineage information
- Accessing and querying the provenance and the lineage of a data item
- Defining and obtaining provenance and lineage information for data obtained by combining data sources different data models (e.g. corporate data and semi-structured web data)
- Audit Trail

H. Security & Information Assurance

Summary

98. 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.

[7] https://www.w3.org/TR/prov-o/

Description

99. 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.

100. 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.

101. 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).

102. 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.

103. 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

104. 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.

105. 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.

106. 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.

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

108. 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.

109. 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.

110. 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.

Conceptual Building Blocks

111. The conceptual building blocks for the Security and Information Assurance capability are:

- **Authentication**: The substantiation of the identity of a person or entity related to the enterprise or system in some way.
- **Authorisation**: The definition and enforcement of permitted capabilities for a person or entity whose identity has been established.
- **Audit**: The ability to provide forensic data attesting that the systems have been used in accordance with stated security policies.
- **Assurance**: The ability to test and prove that the data architecture has the security attributes required to uphold the stated security policies.
- **Availability**: The ability to keep data despite abnormal or malicious events.
- **Asset Protection**: The protection of information assets from loss or unintended disclosure or use.
- **Administration**: The ability to add and change security policies, add or change how policies are implemented in the enterprise, and add or change the persons or entities related to the data access.
- **Confidentiality:** The ability to apply a set of rules limiting access to certain types of information

Logical Building Blocks

112. The logical building blocks for the Security and Information Assurance capability are:

- Authentication
  - Maintain Authentication
- Authorisation:
  - Maintain security classification of data objects
  - Maintain access of roles on security classification
- Audit:
  - Create Logging

    - Maintain Security Policies
- Assurance:
    - Monitor security attributes required to uphold the stated security policies.
- Availability:
    - Monitor data availability during different events (can be abnormal / malicious)
- Asset Protection:
    - Monitor assets on unintended disclosure
    - Monitor assets usage
- Administration:
    - Maintain security policies (add or change how policies are implemented in the enterprise,)
    - Maintain user list
    - Maintain Role membership of users
    - Ensure separation of duties
- Risk Management:
    - Maintain risk catalog
    - Train users on organisation's attitude and tolerance for risk.
    - Identify Cyber threat
    - Provide expert IA and Security advice

VIII. Semantics

113. Statistical organisations wanting to share data internally or between themselves, need to share and agree on the meaning (semantics) of that data. This includes concepts and definitions, most importantly units and populations, variables and codes.

114. An important future expansion of CSDA, therefore, will be a semantic framework that allows a common way of modelling and sharing the semantics of data and metadata. Such a framework could be used at statistical organisation level as well as community-wise, and for example provide the ability to:

- share data and (semantic) models within and between organisations
- create coherent views of data and metadata assets
- build metadata-driven systems

115. In regard to CSDA, some specific use cases are:

- data integration, including conceptual and instance-level alignments
- conceptual data access, which is the ability to access (and query) data at the semantic level independently of implementation details
- quality checking

116. The features expected from this semantic modelling framework are:

- in line with GSIM;
- a well-recognised standard, facilitating the reuse of semantics already defined by other authorities;

- machine-actionable, at least for the metadata representation;
- supporting multi-linguistics, for human consumption;
- comprehensive enough to cover the whole domain of interest;
- covering microdata, macrodata and metadata (including structural, quality & provenance);
- prescriptive enough to enforce share and reuse by default at a global community level.

117. In reference to these requirements and use cases, and awaiting the development of the framework itself, CSDA strongly recommends the adoption of the Ontology Web Language (OWL2[8], a W3C recommendation) and associated vocabularies. The re-use of existing OWL vocabularies such as SKOS[9], PROV-O[10], Annotation vocabulary[11], DCAT[12] and its applications profiles is also strongly recommended.

118. In addition, a role of "keeper of global definitions", or ontology custodian, should be defined in the CSDA management framework, at the highest level of the statistical community.

[8] https://www.w3.org/TR/owl2-overview/
[9] https://www.w3.org/TR/skos-primer/
[10] PROV-O = the PROV Ontology, ref https://www.w3.org/TR/prov-primer/
[11] https://www.w3.org/TR/annotation-vocab/
[12] https://www.w3.org/TR/vocab-dcat/