Modernstats Workshop 2019

CSDA project

Carlo Vaccari

The 2018 team

Team leads: Carlo Vaccari (PL) & Dick Woensdregt (Lead Architect)

Team members: representatives from from Canada, UK, Finland, Netherlands, Italy, Poland, Serbia, Mexico, Eurostat, Montenegro, Slovenia,

Face-to-face meetings: Belgrade (May) & Warsaw (Sept)

modernistats by HLG - MOS

UNECE

CSDA final deliverables

The CSDA project delivered:

- Reference Architecture (updated and expanded version)
- Set of guidelines, including a Maturity Model, for implementation of the Architecture
- Use-cases for testing the Architecture
- Leaflet for promotion

Reference Architecture: scope

- The focus of CSDA is on DATA (and MetaData), but not just any data. CSDA is restricted to *data that is valuable enough to be treated as an asset*

- The scope is the full statistical production process (end-to-end)

- It is not restricted to the physical boundaries of the statistical organization, but also includes any activities taking place outside the premises, but *under control of* the NSI, such as activities “in the cloud”.

Treating Data as an Asset: principles

CSDA is about data. Not just data, but primarily about valuable data, i.e. data that is worth **treating as an asset**. And as there is **no data without metadata**, CSDA considers both integrally as **Information**.

1. Information is **managed as an asset** throughout its lifecycle
2. Information is **accessible**
3. Data is **described** to enable reuse
4. Information is **captured** and recorded at the point of creation/receipt
5. Use an **authoritative source**
6. Use agreed **models and standards**
7. Information is **secured** appropriately

CSDA currently focusses on Capabilities

- In the context of modernization of statistical organizations, Capability is a rather new concept, mentioned in GAMSO, but not yet properly defined and used.
- CSDA proposes the definition and usage as described by TOGAF: as an instrument in strategic planning and systematic, iterative, renewal (modernization).

Capability: 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.

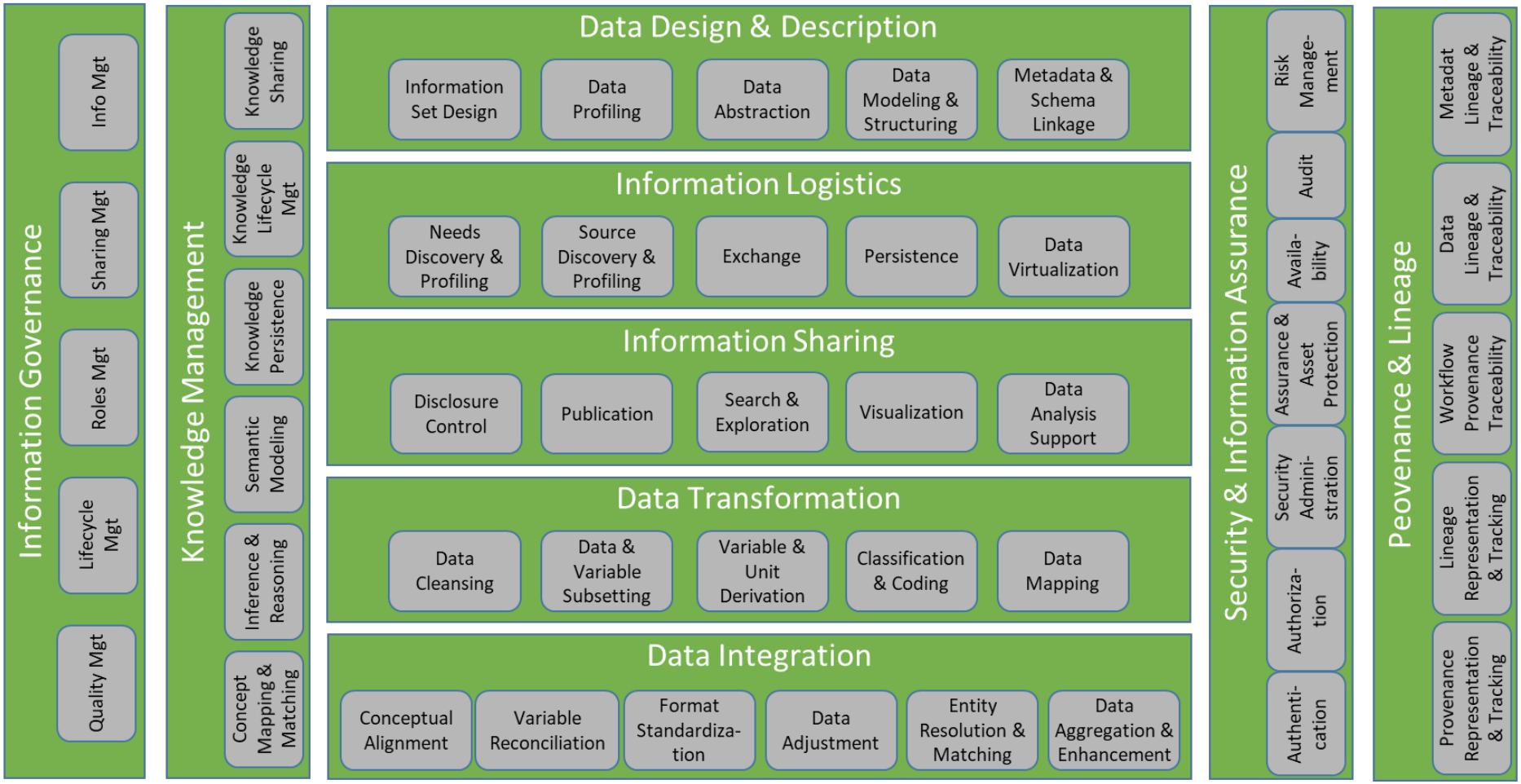
https://pubs.opengroup.org/architecture/togaf91-doc/arch/chap03.html

Capability Definition Principles

A set of principles, aiming to direct the way (new) capabilities are identified and defined:

- Capabilities are abstractions of the organization. They are the “what?” and “why?” not the “how?”, “who?”, or “where”
- Capabilities capture the business’ interests and will not be decomposed beyond the level at which they are useful
- Capabilities represent stable, self-contained business functions
- The set of capabilities should cover the space of interest, and no more
- Capabilities should be non-overlapping
- Two-levels Capabilities: top-level and lower-level
- For each top-level Capability: description and lower-level Cap

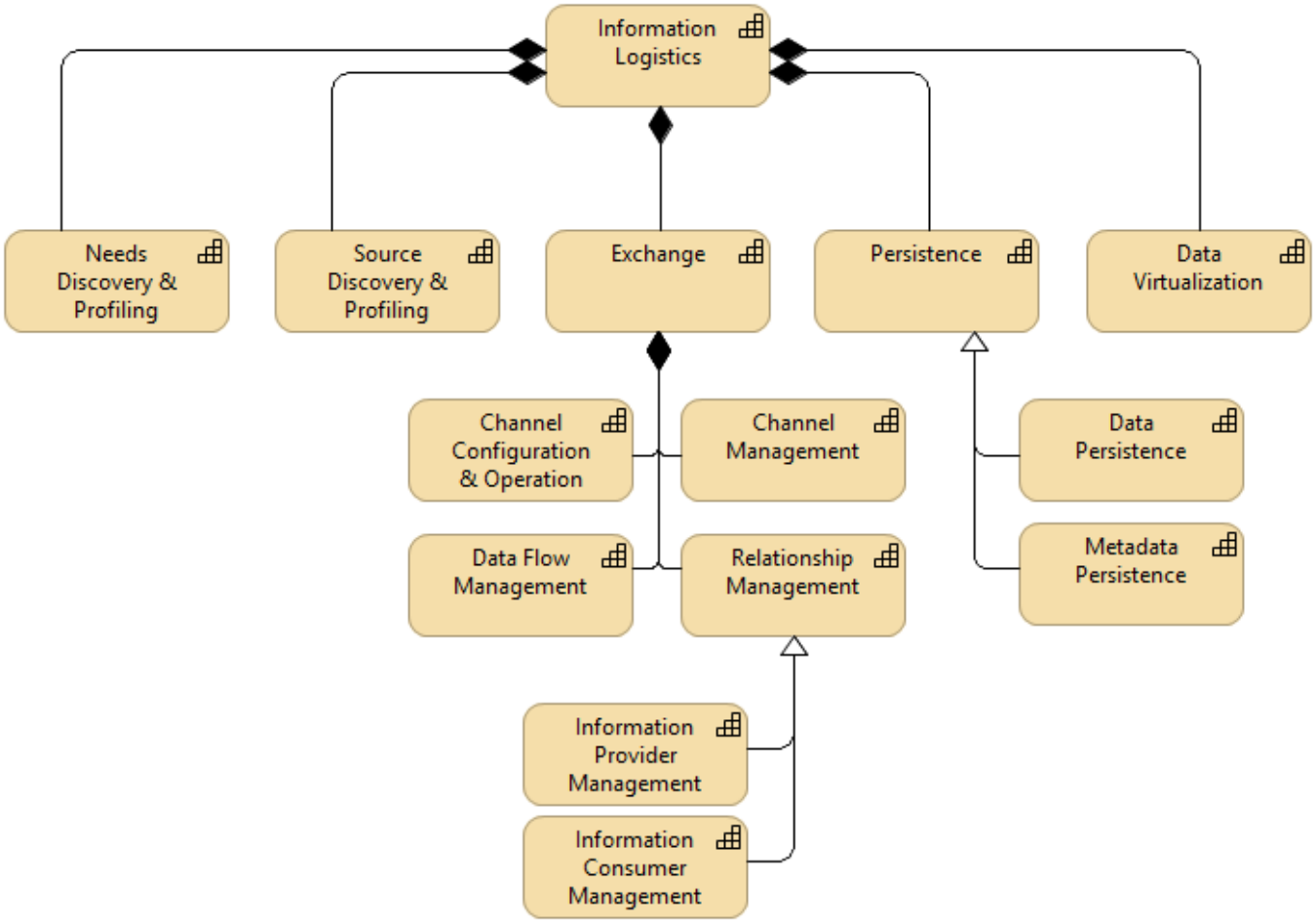
CSDA top-level Capabilities



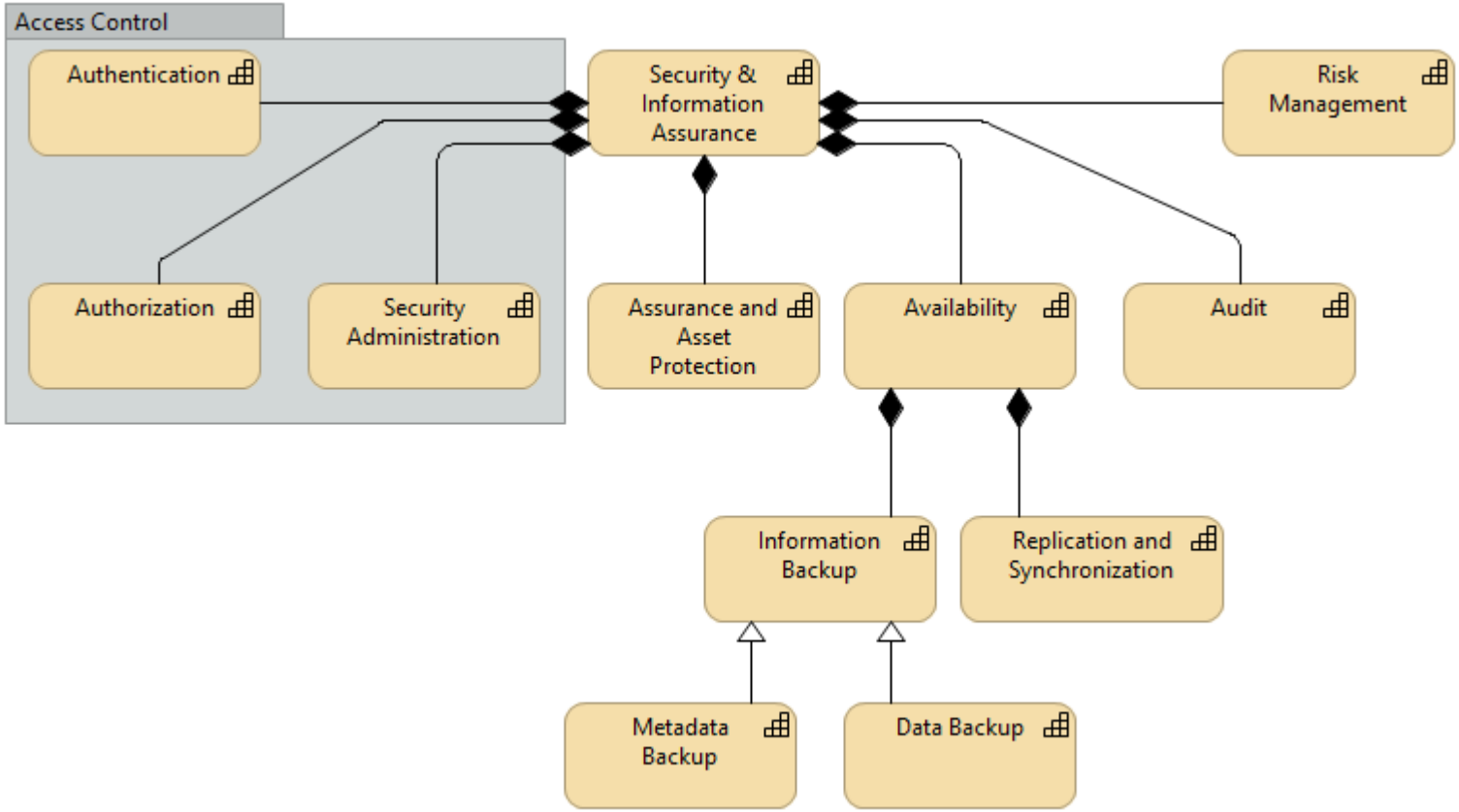
modernistats by HLG - MOS

UNECE

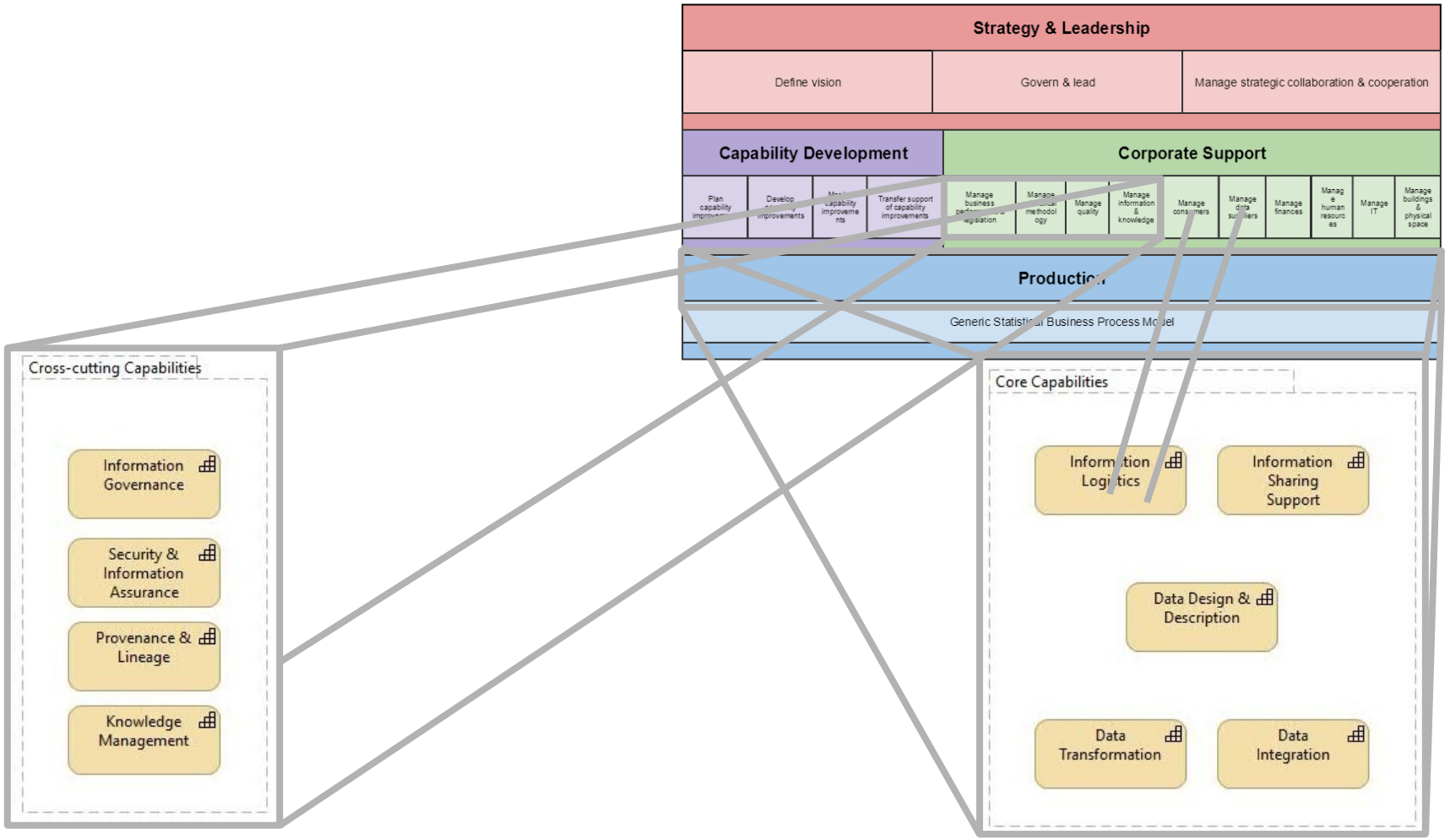
Example: Information Logistics



Example: Security and Info Assurance



CSDA & GAMSO

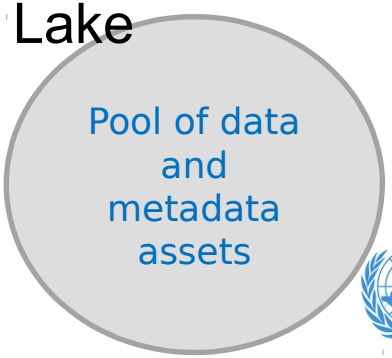


modernistats by HLG - MOS

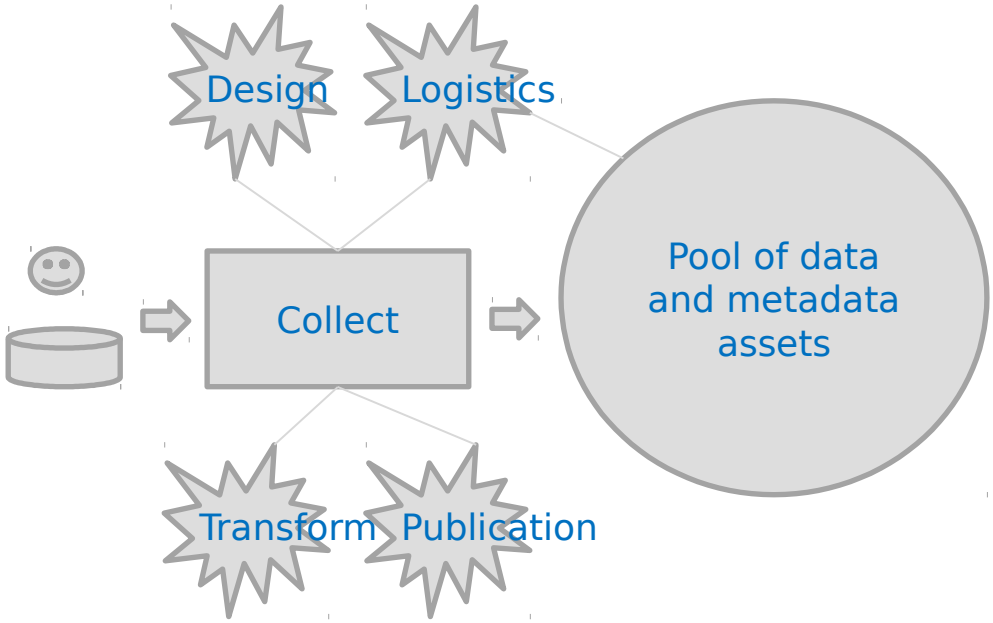
UNECE

Concept: the “pool”

- The “pool” is a (the?) collection of (all) data that is considered valuable enough to be treated as assets
- The “pool” is a collection of Data Sets, together with all available Metadata associated with those Data Sets
- The “pool” is a concept, not necessarily some form of physical storage!
- The “pool” may be segmented, e.g. separating different classes of data
- Data enters the “pool” through (input) Exchange channels (Data Logistics) and can be accessed only through suitable (output) Exchange channels
- Note: the “pool” contains ALL (statistical) data relevant to the organization
- Other terms for “pool”: Data Reservoir, Data Lake



CSDA and GSBPM: Data Collection

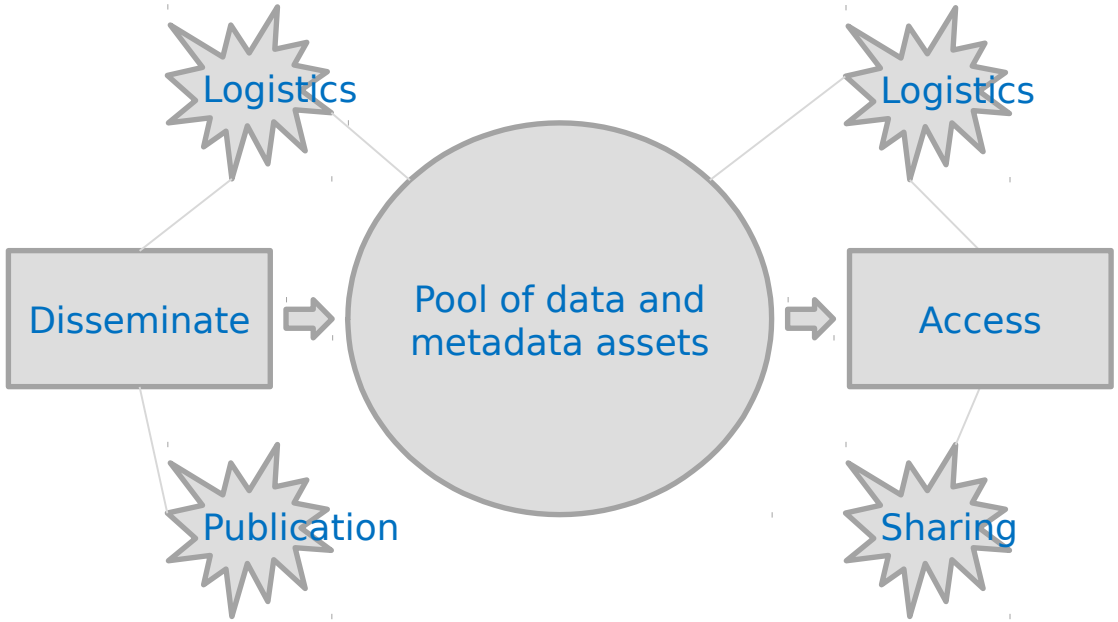


The “pool” only contains digital information, although it is conceptual and might therefore contain also “intangible” information such as data only existing in the minds of respondents.

Assuming the definitions from GSIM, we need a channel to collect data, such as a CAPI or CATI channel to collect such “intangible” data

In such cases, we will need internal persistence, in order to decouple the internal processing from the collection.

CSDA and GSBPM: Dissemination



The information to be published may come from the “pool” or from some internal process.

Publishing in the strictest sense only involves the Information Publication capability. In a broader sense, it may involve other capabilities such as Disclosure Control.

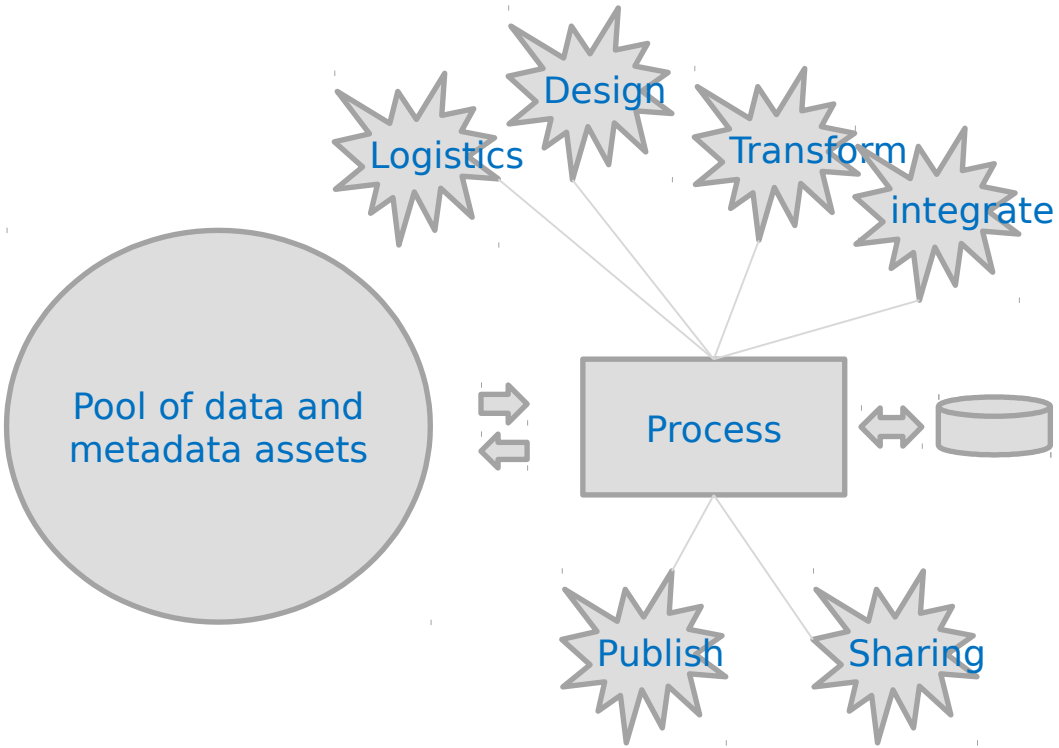
Information Publication includes: defining the composition of the Information Set, the channels available for access, the date & time of availability, the audience, etc.

CSDA and GSBPM: Processing

The process uses input data from the “pool”, and may produce data that is considered suitable to be released into the “pool”. This is a formal act of “publishing”, even if the data is NOT a statistical end-product.

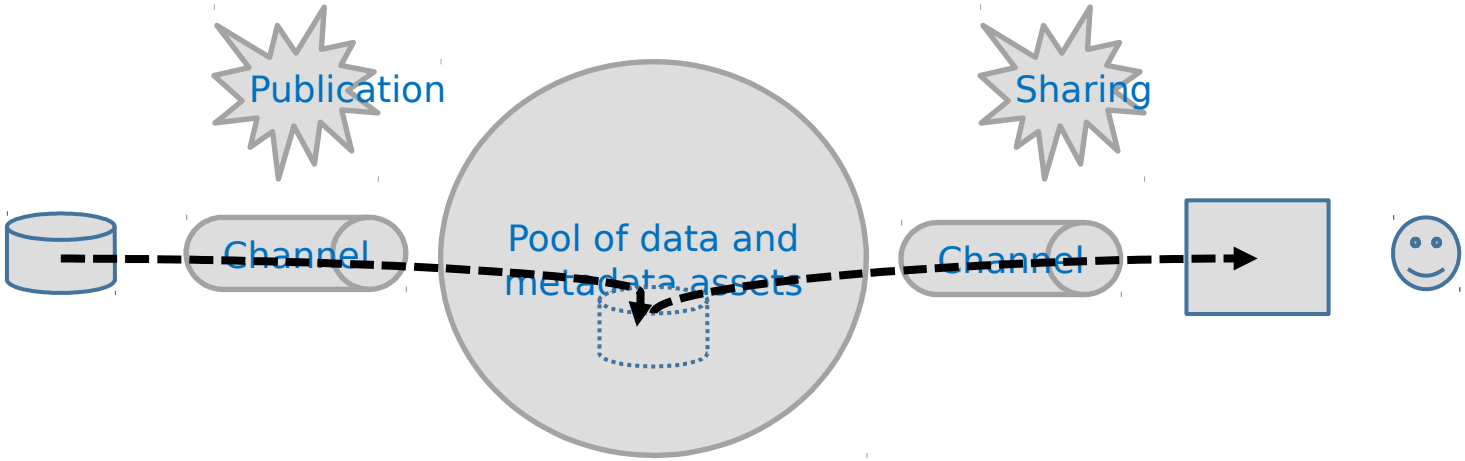
Accessing data from the “pool” involves both Sharing Support and the lower level capabilities from Info Logistics (channels).

The process may have internal persistence. Data stored there is NOT considered part of the “pool”.

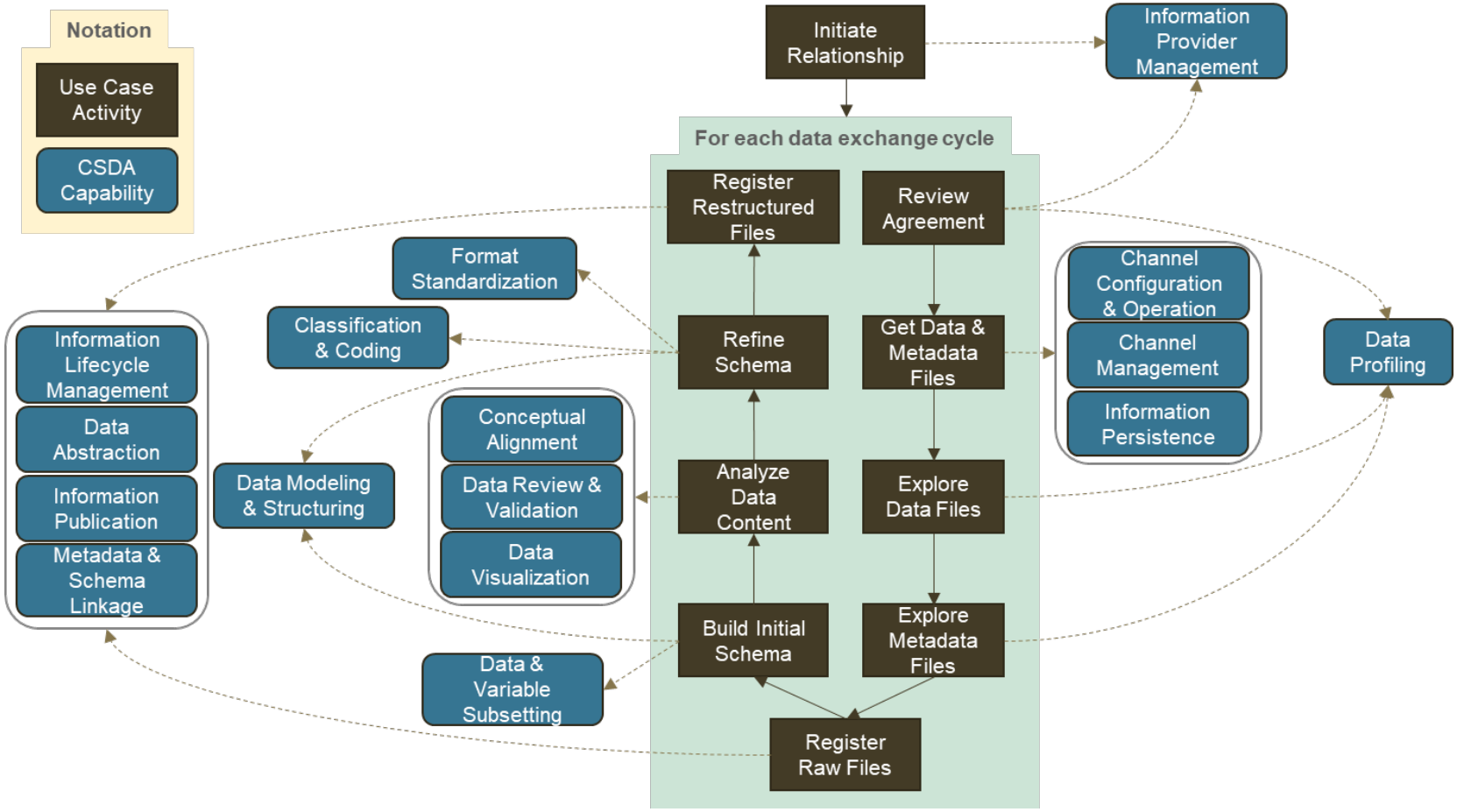


Collect vs Connect: paradigm shift

Collect vs Connect: capabilities



CUSIP use case – Statistics Canada



modernistats by HLG - MOS

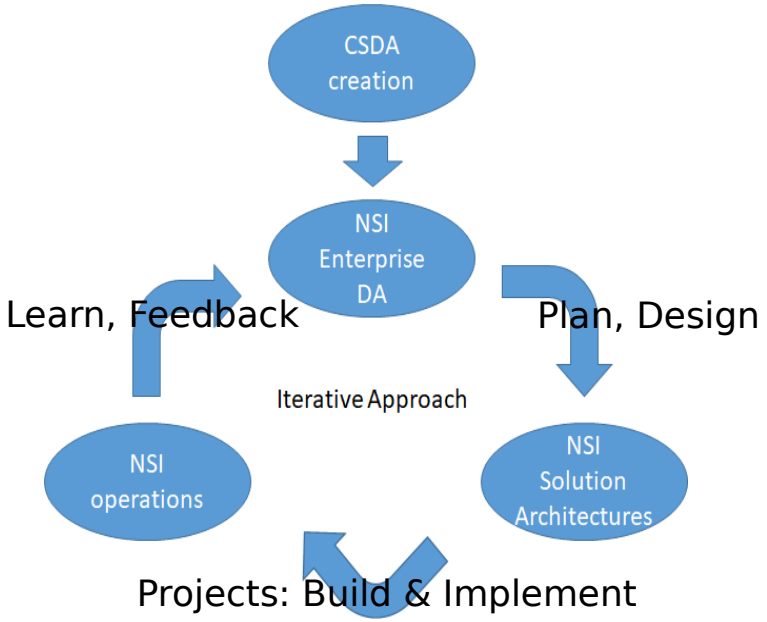
UNECE

CSDA guidelines: users and steps

Identify the target audience (those involved in the definition) and user groups (those that are informed) for Data Architecture

Identify steps for the introduction of a Data Architecture in NSI

Stress iterative approach: rather a model with feedbacks and loops, than sequential operations



CSDA guidelines: Maturity Model

Helping NSI to protect and exploit the value of data and metadata assets available along following dimensions:

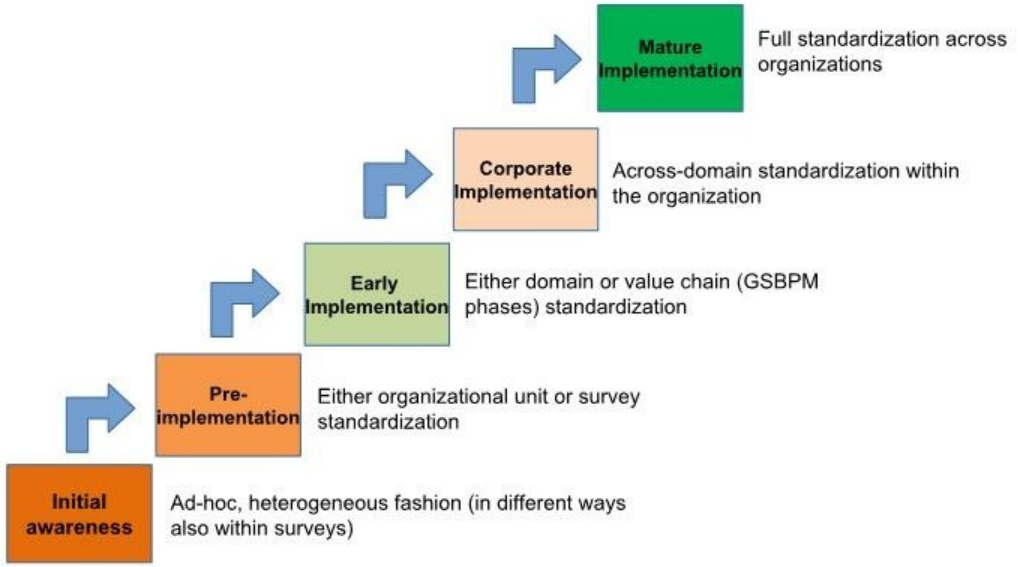
- Maintenance: the way assets are managed
- Protection: the level of protection against loss, disclosure, unavailability
- Sharing and re-use: quality of metadata, degree of re-use and promotion
- Growth: how to identify new needs, how to explore new data sources, ...
- Internal development: process of continuous improvement of the organization

CSDA guidelines: Maturity Model

Five levels (like CMMI and HLG MMM)

Five dimensions, the same as HLG-MMM:

- Business
- Methods
- Information
- Application
- Technology



modernistats by HLG - MOS

UNECE

CSDA guidelines: Maturity Model

Matrix 5x5 to:

- Assess current NSI situation
- Understand next steps to improve

Maturity levels can be applied also to specific Capabilities

Levels Dimensions	Initial implementation	Pre- implementation	Early implementation	Corporate implementation	Matu impleme
Business	Data are managed at survey level for specific goals. Lack of data standardization due to a stove-pipe data architecture, based on different tools and techniques.	Data are integrated within related areas or domains. Initial process standardization helps to reduce redundancy and to improve data governance.	The core business goals are achieved by a set of capabilities that allow to run standardized processes based on structured data objects. Reuse of shared solutions.	Adoption of a data-centric approach and Data architecture is one of the main components of the business strategy. Roles and responsibilities are clearly defined in the organisation.	Data capabilities i business goals, fo international stand used and shared b organizations and New challenges fr context, as new us don't cause drama strategic data mar
Methods	Standardized data input and output are limited to single methods, implemented for specific process steps.	Common methods are implemented at domain level and are supported by data and metadata integration and standardization.	A set of statistical services, based on standardized methods, is shared by different domains. Data processing and management is based on active metadata.	Methods, data and metadata are fully integrated in the core capabilities and managed and orchestrated by cross-cutting capabilities.	Standardized meth corporate level are used by several NS on the same topic countries are more and easier to comp
Information	Information is modelled for specific goals at survey level or for single process steps, resulting in redundant or duplicated data structures.	Information is managed and governed at domain level. Increasing standardization and integration of data and metadata.	Data and metadata integrated between domains reduce redundancy and increase information sharing, enriching statistical outputs.	Data-centric approach adopted at corporate level, according to international standards (GSBPM, GSIM, CSPA) improves data access and processing, enhancing the information	Data architecture a manage statistical providing solution external and intern Information is easi several NSOs, due of international sta

modernistats by HLG - MOS

UNECE

CSDA guidelines: principles

Assess: the current situation of your NSI

Choose priorities: for NSI, in term of Capabilities and/or domains

Highlight cross-domain analogies/differences

Improve: verify on Maturity levels which are steps needed to improve

Enhance standard compliance and re-usability

Verify prerequisites: for the desired level for each Capability/Domain

Everlasting self-assessment: capabilities need to be refreshed

CSDA guidelines: roadmap

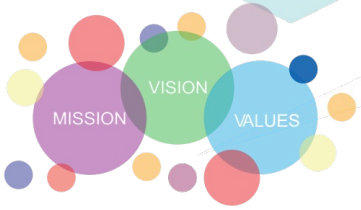
Once the current situation has been mapped, and the strategic objectives of the organization have been translated into a To-Be situation in terms of the CMMI reference framework, we suggest how to define a roadmap → how to get from As-Is to To-Be

To implement the roadmap also a couple of templates have been developed, similar to those used in HLG-MMM:

- Implementation Check-list
- Progress evaluation matrix

CSDA promotion leaflet

A simple tool for HLG to spread the meaning and the potential benefits of CSDA



Build your **C**reations

Exploit all your a**S**sets

Mo**D**ernize activities

Be data **A**rchitect

HLG MOS vision for CSDA

- This template statistical organizations can use to develop their own Enterprise Data Architectures. In turn, this will guide Solution Architects and Builders to develop systems that will support users to do their jobs and produce statistical products
- An important part of a suite of standards, developed and maintained by the international statistical community, led by HLG-MOS
- A unified statistical data architecture for all NSI
- An extended user chain to share knowledge with data providers

Modernising with the Common Statistical Data Architecture

As the complex world of data expands and demands on National Statistical Institutes grow, finding new and innovative ways to keep pace is challenging. To help navigate these challenges and take advantage of new opportunities, the High Level Group for Modernising Official Statistics has developed the **Common statistical data architecture - CSDA.**

What is CSDA?

The CSDA supports statistical organisations to design, integrate, produce and disseminate official statistics based on both traditional and new types of data sources.

CSDA is a tool that will help you:

- Recognize capabilities required to fully utilize your data assets
- Organise and structure processes and systems for efficient and effective management of data and metadata – from external data sources to internal storage, processing and dissemination.
- Make you independent from technology – processes and systems will better endure technological evolution
- Provide a clearer path for NSI growth
- Focus resources for building most important partnerships
- More easily manage new types of data sources such as Big Data
- Modernize your NSI to better react with your environment

CSDA is compatible with worldwide standards and with other HLG-MOS standards such as GSBPM.

How are these benefits achieved?

The scope of the CSDA includes all of the GSBPM phases – designing, building and use of processes and systems in statistical data collection, production, analysis and dissemination, based on external needs.

In order to help NSI's implement processes described in GSBPM, the CSDA defines capabilities that enable any NSI to undertake a specific data-related activity.

- A Capability is an ability an organization has or needs. It is an integration of methods, processes, standards and frameworks, IT systems and people skills
- The CSDA defines Core Capabilities and Cross-Cutting Capabilities as the basic elements to design and build solutions
- Lower level Capabilities include WHAT CSDA could do and HOW

UNECE

In summary ….

- Deliverables
  - Reference Architecture (in document, but also as an Archimate model)
  - Guidelines
  - Use cases
  - Leaflet
- Suggestions for future work
  - Integrate / further align with other HLG standards
  - Revise GSBPM (modernize Collect)
  - Add objects to GSIM to cover Knowledge
  - Develop the detailed architectures (according to TOGAF: Business, Info Systems) for implementing CSDA capabilities
  - Start applying CSDA in practice (ONS and StatCan already started …)
  - Use CSDA in future HLG projects / activities

    https://www.menti.com/ code 976880