Common Statistical Data Architecture (CSDA)

(Version 2.0, 2018)

Table of Contents

List of abbreviations

Term	Meaning
CSDA	Common Statistical Data Architecture
CSPA	Common Statistical Production Architecture
DAMA	Data Management Association
DCAT	Data Catalog Vocabulary
DDI	Data Documentation Initiative
DMBOK	Data Management Body Of Knowledge
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

I. Preface

This document is the result of the 2018 Data Architecture project. The 2018 DA project was a continuation of the 2017 project, which was the first project to specifically address the data aspects of statistical production.

The main reason for putting the focus on Data is the fact that data is both the raw material, the components and the finished product of Statistical Organisations. Data therefore deserves to be treated as the asset is truly is. The Data Architecture defined by the 2018 project therefore is based on a set of Principles that stress the importance of treating data in appropriate ways.

The main deliverable of the 2018 project is a revised version of the **Reference Data Architecture** that focusses on the data-related functionality that statistical organisations will need for the design, integration, production and dissemination of official statistics allowing those organisations to use and exploit both traditional and new types of data sources. The main content of the 2018 version is the description of the Capabilities. Capabilities are abilities that an organisation needs or possesses, typically expressed in general and high-level terms. Capabilities typically require a combination of organisation, people, processes, and technology for proper implementation.

Another important deliverable of the 2018 project are the **Guidelines for the use and implementation** of the Reference Data Architecture. These Guidelines also include a Maturity Model that will allow statistical organisations to assess their current situation and to plan their migration to a more mature state.

And finally, the 2018 project defined a number of **Use Cases** for checking and testing the Reference Data Architecture.

II. 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. Let us start by defining data architecture:

- **"**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.

A. CSDA, a special kind of Data Architecture

5. CSDA is not a normal Data Architecture, at least not according to the definition of TOGAF. According to TOGAF, a data architecture is an integral part of the Information Systems architecture and "describes the structure of an organization's logical and physical

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

data assets and data management resources". CSDA is focused on capabilities related to data and metadata, which can be seen as "data management resources", rather than on the structure and organization of data assets. Capabilities are strategic elements and are the starting points for the incremental development of the business architecture, the information systems architecture (including the data architecture) and the technology architecture.

6. In fact, CSDA is a "data centric" view of an NSI's architecture, putting emphasis on the value of data and metadata, the need to treat data as an asset. Both the CSDA architecture and the companion Maturity Model have their focus on the way a statistical organization could/should treat their data and metadata.

7. Because CSDA is "just" a specific view of a general architecture, it has (or should have) all the components of a general architecture. According to TOGAF, this should include the strategic level (the capabilities and roadmap planning) as well as the business, information systems (including data) and technology architecture. The current version only defines the strategic elements, the capabilities. An effort has been made to show how (elements of) GSBPM can be used to create the business architecture, whereas CSPA (and specifically certain, still to be developed, CSPA services) should be the basis for the Information Systems architecture. Because of the currently still very divers situations of NSI's, it is difficult to say much relevant concerning the general technology architecture, but certain guidelines can be formulated, specifically with respect to security measures.

8. 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, GAMSO, GSIM, and CSPA. Where applicable, the CSDA also links to other international standards such as TOGAF, DDI, SDMX, DAMA DMBOK, etc.

9. Specifically, for this version of CSDA, reference is made to the following versions of these other standards:

- GAMSO version 1.1
- GSBPM version 5.0
- GSIM version 1.1
- CSPA version 1.5
- TOGAF version 9

10. Another useful reference is the European Interoperability Reference Architecture (EIRA[3]). 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".

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

III. Purpose

11. 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[4])

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

13. 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 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). More about this is described in the companion document "CSDA Guidelines".

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

15. 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.

[4] https://en.wikipedia.org/wiki/Reference_architecture

IV. Scope

16. 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 in statistical organisations. It addresses Data Analysis related to traditional and 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.

17. 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).

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

V. Benefits

19. The benefits of having a CSDA are:

- **Independence from technology**: By using a technology-agnostic reference architecture, statistical organisation processes and systems will, eventually, become more robust to technological evolution
- **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.

VI. Key principles

20. Data in this context is the raw material, semi-finished and finished product of the statistical organisation, rather than the information the organisation needs to manage its processes, as is the case in other industries.

21. Metadata is data that describes other data. ("Meta" is a prefix that in most information technology uses means “an underlying definition or description”.) Metadata is data with 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.

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

23. Information is the general term meaning both Data and Metadata.

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

[5] https://www.dtls.nl/fair-data/fair-principles-explained/

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 NSI 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 NSI 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.

Principle	Statement	Rationale	Implications
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; • Connections between data objects must be documented; • Restrictions to data usage must be documented.	• Data can be easily understood and used with confidence without requiring further information; • 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.	• Staff will document data with reuse in mind; • Staff will consider reuse when designing systems for capturing information.

Principle	Statement	Rationale	Implications
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.

Principle	Statement	Rationale	Implications
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.

Principle	Statement	Rationale	Implications
7. Information is secured appropriately	• Information should be classified according to its level of sensitivity, secured and managed accordingly.	• The NSI has a legal responsibility to adhere to regulations governing the safeguarding of national security and personal information; • Our reputation and ability to collect information relies on our ability to protect that information; • NSI's need to balance making information available with the risks to information providers and the organisation.	• The organisation will foster a culture of awareness and respect for sensitive information; • Staff will adhere to organisational policies governing the use of sensitive information; • Risks relating to sensitive information will be considered, monitored and managed; • Security will be built into the design of systems and processes; • Information will be classified appropriately and reviewed as required.

VII. Types of Data

25. Not all data an NSI has at its disposal or produces has the same importance and value and therefore not all data needs to be cared for and managed in the same way. In order to avoid unnecessary effort and cost, and also in order to be compliant with legislation such as privacy laws, it is advantageous to distinguish certain classes of data, where each class can be managed in its own way. The idea is that it might be helpful for NSI's to distinguish such different types and to develop and adopt different policies for the management of each type.

26. Most probably, policies regarding the management of the data could be quite different from one type to another. It is to be expected that the various types will have different

- security policies, such as access authorization, back-up, etc.;
- retention periods;
- metadata quality (completeness, correctness) requirements;
- etc.

27. As a first approach, the following types or classes are defined:

- **Explorative**: Data that is obtained from outside sources, is usually "sampled" and is used to assess the nature, structure and quality (usability) of that data source. After the exploration, this data in most cases loses its value.
- **Organizational**: The true (data) assets of the organization, that are to be treated as such and must be protected and shared where possible.
- **Temporary, local**: Data that is produced as an intermediate product in a statistical process and has no real value outside that process. This data usually loses its value after the process (cycle) is completed, but may have value for the next cycle as a reference. May be persisted within the process space.

28. An important sub-type of "Organizational" is the Master Data such as statistical registers, back-bones of populations, collections of statistical units. For instance: Company register, People Register, Buildings register. At least the stable "snapshots" that come out of the ongoing maintenance process and reflect the state of the population at (statistically) relevant moments in time. The unstable, continuously updated, collection from which the "snapshots" are taken, could be classified as "temporary, local".

29. "Organizational" data is the main input for statistical production processes. In addition, it will be needed in explorative research as well. The final output from statistical processes definitely will be "Organizational", intermediate results may be typed "Organizational" or "Temporary, local", depending on its value for the organization. The true purpose of exploration is insight or knowledge (of new data sources or about new use cases for existing data sources). Therefore it is expected that explorative research will not create data output, but rather

knowledge, for instance in the form of metadata for future “Organizational” data. Statistical products (the output of the NSI) is “Organizational” only.

30. On first sight, “Explorative” and “Temporary” may seem very similar, but “Explorative” always comes from outside sources, whereas "Temporary" is derived from "Organizational".

31. Note that the classification proposed here is certainly not a recommendation to physically separate the data of the various types. In fact, the need to be able, where relevant, to integrate all kinds and types of data, will make it counterproductive or even futile to try to keep the various types of data separated. Organizations will need to develop policies and means for managing the policies defined for the various types, no matter where that data is located and how many copies of a particular dataset are “floating around”.

32. In order to define the classification more precisely, the following questions (not necessarily disjoint) are still to be answered:

- When does data (a dataset) become a data asset?
- What criteria must be fulfilled for data to be(come) a data asset?
- Which type(s) of data are worth to be(come) data assets?

33. Now even if one succeeds in properly defining each of the classes of data, and then defining the set of policies to be applied for each of those classes, it is not so simple to actually manage the proper application of such policies. The main reason for this is that data literally is everywhere. Specifically digital data is copied easily, and in the course of daily operations, information of different classes is used together in the same processes, making it very hard to ensure the proper application of the policies defined for each class.

VIII. Exchange Channels

34. As explained in the previous chapter, managing the proper application of policies to different classes of data (or information in general) is rather complex. Statistical organizations have therefore tried to simplify things. One of the measures organizations traditionally have taken is to turn themselves into "strongholds", with "walls" and "mounts" around them. Inside, the data was considered to be safe, and usually, all data was treated in more or less the same way, i.e. no differentiation of policies for different classes. And then they built the equivalents of gates and drawbridges to allow data to flow into and out of their stronghold in manners that could be controlled.

35. Nowadays, we find that these traditional measures no longer are adequate[6]. More and more, we need closer interaction with external parties, both providers and consumers. And so, we need to allow these external parties inside our strongholds and we also want to participate (and use our data) outside, in the cloud or even on systems controlled by our allies. But still, we need to be able to control the flow of information into and out of areas under the control of separate organizational units. So there still is a need for "toll gates". Exchange Channels provide such mechanisms.

36. GSIM defines an Exchange Channel as “An abstract object that describes the means to receive (data collection) or send (dissemination) information”. GSIM is referring to the information object describing the “real thing”, i.e. the actual Exchange Channel being “the means to receive or send information”. CSDA is concerned with the latter. Contrary to GSIM, CSDA does not restrict Exchange Channels to Data Collection and Dissemination (in GSBPM terms), as will be explained in the following.

37. As explained before, the original purpose of channels is to control the flow of information in or out of an area of control, i.e. an area controlled by some power (e.g. an organizational unit). Channels are owned and provided by the one responsible for the controlled area. But they are used by those that want (or need) to provide or access information. They are the points of exchange between different areas of responsibility. And thus they are the natural places for controlling the flow. More and more, however, these touchpoints also become the focus of user friendliness, and therefore channels must provide alternatives to consumers and providers. Alternatives that must fit the needs and concerns of those users. As we will see further on, channels have several responsibilities or tasks related to this control of flow of information.

---

[6] Since 2003, the Jericho Forum (merged with The Open Group Security Forum in 2014) is working on this topic[6], called “de-perimeterisation”.

38. There are two kinds of Exchange Channels: those that disclose information sources and those that allow consumers to access the information assets (treasures) of the statistical organisation. For ease of reference, we'll call the first kind "input" and the second kind "output". By making information out of sources accessible through "input" channels, that information is thereby considered valuable enough to be treated as an asset.

39. It should be stressed here that information sources and information consumers not necessarily are meant to be "external to the organisation". They are "external to the sphere of control" exercised by the statistical organisation (or one of its agents[7]) on all its information assets. Information is placed under such control by opening up an input channel to that information. Information leaves such control by allowing it to "exit" through one of the output channels. The collection of information under such control is also referred to as "the pool". It is important to understand that "the pool" is a concept, denoting the collection of information, not necessarily a storage pool. And most certainly not a centralised storage. As will be explained in a later chapter, the information considered "in the pool" may never actually be physically stored there, but only pass through from some input channel directly to some output channel. Still, this information is considered "under control".

40. The next figure schematically shows these concepts:

Figure 1: Pool & Channels

41. Equally important is to understand that even the channels themselves, and (part or whole) of "the pool" may be located outside of the premises of the statistical organisation. That is, they may be located "in the cloud", or even on the premises of some other (private or public) organisation, as long as the statistical organisation has sufficient control to ensure that its interests in terms of protection of the information assets is safeguarded.

[7] For the definition of "agent", ref GSIM

42. Exchange channels are also the organisation's interfaces to the outside world. And because of that, they need to be flexible enough to follow the developments in that outside world. New technologies that open up new opportunities, new demands from providers or consumers, new types of information that require new ways of connecting, these are all examples of developments that channels must be able to adapt to. To put it in other words: exchange channels must be replaceable, in a "plug and play" manner. And they must provide a variety of different interfaces to the outside world in order to satisfy the variety of different requirements. On the inside, however, channels should present a standardized interface. How else can they be "plug and play", if every new channel requires changes to the internal systems?

43. Turning to the tasks of an exchange channel, it is obvious that its main task is to transport ("channel") information. And, as we have already seen, it must be able to present some interface on the outside, and (most likely another) one on the inside. Part of an interface is a protocol, so a channel is also a protocol converter. In addition, a channel must be able to bridge the distance (both in space and in time) between external source or consumer and the internal side. Furthermore, a channel must be able to reliably establish the identity (authenticate) and the authorization of the external party (Access Control). And lastly, the channel may need to do translation, in some cases even natural language translation, but more likely transformation of format. The channel will need some internal process, to control its behaviour. In most cases, in the traditional data collection, this is done through training of interviewers and by providing questionnaire instruments. Both are examples of "configuring" the channel for specific surveys. The next figure shows the necessary capabilities of a generic channel.

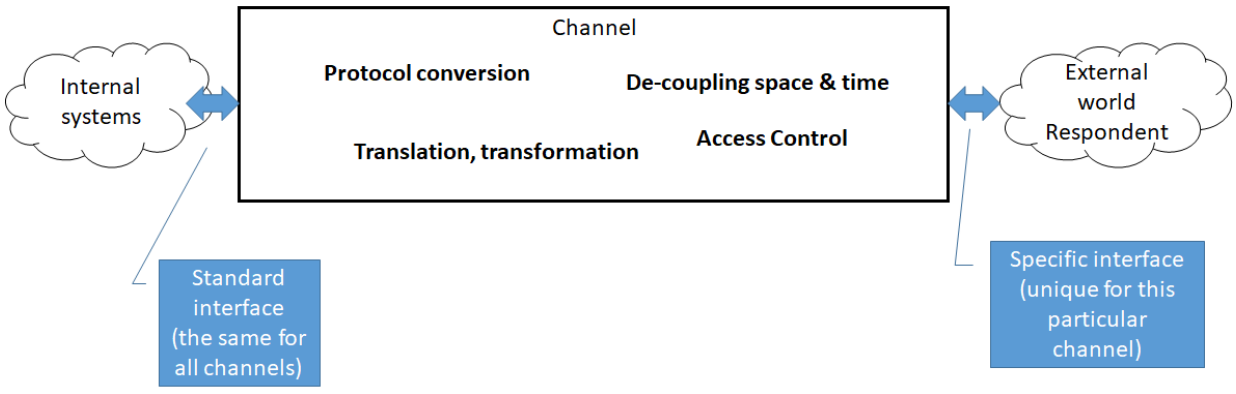

Figure 2: Generic Exchange Channel

44. More details about the internal workings of Exchange Channels can be found in the Annex: Examples of Exchange Channels.

IX. Modes of Operation and Data Management Styles

A. Information Logistics Modes of Operation

45. Information Logistics has two basic modes of operation. Traditionally, NSOs have gathered all the data they needed for statistical production, either capturing it via questionnaires and other collection instruments or acquiring it from other organizations (admin data) which is then physically stored on the NSO's premises. In other words, data used to be moved to where the processing capabilities were. More recently, with the advent of Big Data, cloud platforms and IoT, NSOs have started to consider more distributed data and processing approaches in which data stays in place (where it was produced, or in the owner's environment) and the processing is moved to where the data is. The reason for this change is not only technical, e.g. minimize data movement, but also contractual, political and regulatory, e.g. privacy concerns, legal implications, national and transnational regulations, etc. Gartner named these two different modes of operation: the centralized data and processing mode is called "**collect**", while the decentralized data and processing mode is called "**connect**". [8]

46. CSDA capabilities are flexible enough to support both information logistics modes of operation. The main capabilities that need to be aware of where the data resides and whether it's a connect or collect scenario are Publication within Information Sharing and Exchange and Persistence within Information Logistics. In particular, Channel Management will have to create channels in the mode specified by the SLA's defined in Relationship Management, which will then be configured and operated by Channel Configuration & Operation.

47. In terms of implementation, the simplification of data logistics in a connect mode scenario will likely produce an increase in complexity of data processing, since data processing will have to be shipped to where the data resides and its results integrated back for downstream consumption. The trade-offs of each approach and what degree of decentralization is required need to be evaluated on a use case basis. For example, Data Transformation and Data Integration might be optimally implemented in a centralized way to serve traditional analytics and statistical production based on data collected by surveys via questionnaires (collect mode). In other situations, these capabilities can be implemented in a decentralized manner (connect mode) when

[8] Modern Data Management Requires a Balance Between Collecting Data and Connecting to Data; https://www.gartner.com/doc/3818366/modern-data-management-requires-balance

the volume, or the number of sources (e.g. IoT), of admin data could create information logistics issues the NSO may want to avoid.

48. Decentralization also affects metadata management. Metadata describes where the data resides and how the channels operate, how the data relates to the rest of the ecosystem, who accesses it and how business processes use it. In addition, the implementation of Metadata & Schema Linkage becomes more complex when data and its metadata live in entirely different environments.

B. Data Management Styles for Modern Data Integration Challenges

49. Modern business requirements demand a more proactive and coordinated data integration and data provisioning strategy founded on a portfolio-based data integration approach encompassing:

- **Bulk/batch:** This incorporates a single or multipass/step processing that includes the entire contents of a data file, after an initial input or read of the file has been completed from a given source or multiple sources. All processes take place on multiple records within the data integration application, before the records are released to any other data-consuming application.
- **Message-oriented data movement:** This utilizes a single record in an encapsulated object that may or may not include internally defined structure (XML), externally defined structures (electronic data interchange), a single record or other source that delivers data for action to the data integration process.
- **Data replication:** This involves simple copying of data from one location to another, always in a physical repository. Replication can be a basis for all other types of data integration, but specifically does not change the form, structure or content of the data it moves.
    - **Change data capture (CDC)** is a form of data replication that delivers the capability to identify, capture and extract modified data from the data source and apply this changed data throughout the enterprise in near real time. CDC minimizes the resources required for ETL processes because it deals only with data changes.
- **Data synchronization:** This can utilize any other form of data integration, but focuses on establishing and maintaining consistency between two separate and independently managed create, read, update, delete (CRUD) instances of a shared, logically consistent data model for an operational data consistency use case. Data synchronization also maintains and resolves instances of data collision, as it can establish embedded decision rules for resolving collisions.
- **Data virtualization:** This involves the use of logical views of data, which may or may not be cached in various forms within the data integration application server or the

systems/memory managed by that server. Data virtualization may or may not include redefinition of the sourced data.

- **Streaming/event data delivery:** This involves datasets that have a consistent content and structure over long periods of time and large numbers of records, and that effectively report status changes for the connected device or application, or that continuously update records with new values. Streaming/event data delivery includes the ability to incorporate event models, inferred row-to-row integrity, and variations of either of those models or the inferred integrity with alternative outcomes that may or may not be aggregated and/or parsed into separate event streams from the same continuous stream. The logic for this approach is embedded in the data stream processing code.

X. Capabilities

50. CSDA is (loosely) based on TOGAF. According to TOGAF 9 definitions[9], 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".

51. 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)".

52. In CSDA, capabilities are expressed at the conceptual level, which means they are defined in terms of "what" and "why", rather than "how", "who" and "where". In scope, the CSDA Capabilities are restricted to activities regarding the management and use of statistical data and metadata.

53. In the following sections, the set of capabilities selected for CSDA is presented.

A. Capability Definition Principles

54. The principles below are designed to direct the effort of defining a proper set of capabilities, specifically for CSDA, but potentially also in other areas. Their purpose is to help in answering the question "**How does one properly define a capability?**". These principles are an extension of the description of what capabilities are in other standards such as TOGAF. This set of principles is based on earlier work by Stats New Zealand.

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

Table 2: Capability Definition Principles

Statement	***Description***	***Rationale***	***Consequences***
Capabilities are abstractions of the organization. They are the “what?” and “why?” not the “how?”, “who?”, or “where”	Capabilities are abstractions of "what" an organization does. Capabilities are conceptual. They are completely separated from "how" the organization chooses to implement them.	Business capabilities are about defining what needs to be delivered and why it needs to be delivered but does not define the how it will be delivered.	A clear understanding is required of what the abilities are, as opposed to how they are implemented or executed. The “How” is postponed to later stages in the development process.
Capabilities capture the business’ interests and will not be decomposed beyond the level at which they are useful	Capabilities are expected to be defined to a level at which it is useful to the stakeholders, i.e. strategic level management. Top level capabilities may be decomposed into more detailed lower level capabilities, but decomposition will stop at the level where further decomposition brings no clear advantage.	An appropriate level of detail is required for a good (enough) understanding. Further detail does not help at the strategic level and should be postponed until real development (i.e. design & build). At that stage, other architectural constructs are needed.	The set of capabilities will be restricted to a reasonable number of capabilities. Capability definitions and descriptions need to be concise, i.e. brief but comprehensive.

Statement	*Description*	*Rationale*	*Consequences*
Capabilities represent stable, self-contained business functions	Both the function and the capability must be enduring and stable unlike the processes, strategy & initiatives which can change frequently.	Capabilities should be self-contained pieces of a business that potentially could be sourced from another business unit, another agency, or an external provider. If it cannot be logically separated and shifted it might be too detailed.	Any capability should be essential and inherent to the business, now and in the foreseeable future
The set of capabilities should (completely) cover the space of interest	The set of capabilities should be exhaustive and should include everything the organization needs to be able to do to be successful.	Leaving gaps means running the risk of oversight, i.e. missing potentially important facets in strategic decision making. Including capabilities outside the scope directs focus away from the relevant area.	A clear scope of the "space of interest" is required.
Capabilities should be non-overlapping	No two capabilities should cover the same abilities.	Overlapping implies redundancy and potential conflicts in implementation, which also hampers strategic decision-making.	Higher level capabilities cannot share lower level capabilities.

B. Overview

55. The set of capabilities selected for CSDA represents those capabilities that an organization would need to be able to fully live up to the Key Principles presented in chapter VI. These are the capabilities that collectively enable the organization to properly treat its data as an asset.

56. A subset of these capabilities is used to formulate and implement the policies that the organization chooses for its internal operations. This subset is called "cross-cutting capabilities", because the policies direct and manage how all (information related) work is done. The Cross-

cutting capabilities represent abilities that GAMSO calls “Corporate Support Activities”, and GSBPM calls “Overarching processes”.

57. The remaining capabilities are collectively called “Core capabilities” because these are the capabilities that the organization needs to execute its core business, the production of statistics. These capabilities therefore align with phases of GSBPM.

58. Each capability is further decomposed into lower level capabilities.

59. There are 5 (top level) Core Capabilities:

- Data Design & Description
- Information Logistics
- Information Sharing
- Data Transformation
- Data integration

60. Data Design & Description, that is, the creation of metadata, both prescribing and describing, is separated out as a capability in its own right, because it is clear that data cannot be used and shared unless it is properly described.

61. Information Logistics and Sharing always work together. Information Logistics deals with all the nitty-gritty of storing and moving data around. Information Sharing is about the more business oriented activities of sharing data. This has two sides to it: the “giving” and the “taking”, i.e. making data and metadata available versus the searching, finding and retrieving.

62. Data Transformation and Data Integration is what makes up the essence of statistical production: the refinement, linking and aggregation of data into statistical output products.

63. There are 4 (top level) Cross-cutting capabilities:

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

64. Information Governance comprises all the abilities that have to do with defining and enforcing policies around proper governance of data and metadata.

65. Security and Information Assurance deals with defining and enforcing policies that ensure appropriate protection of data and metadata, against various threats and risks.

66. Provenance and Lineage is the ability to maintain and provide information about the origin and derivation of data and metadata.

67. Knowledge Management, in this version, is positioned as a separate (cross-cutting) capability, because it is a rather new domain in the world of official statistics that is still somewhat vague. But the expectation is that this will become better understood and integrated with the rest in future.

68. The next figure shows the 9 top level capabilities, each with its second level components. All of these capabilities are further described in the following sections.

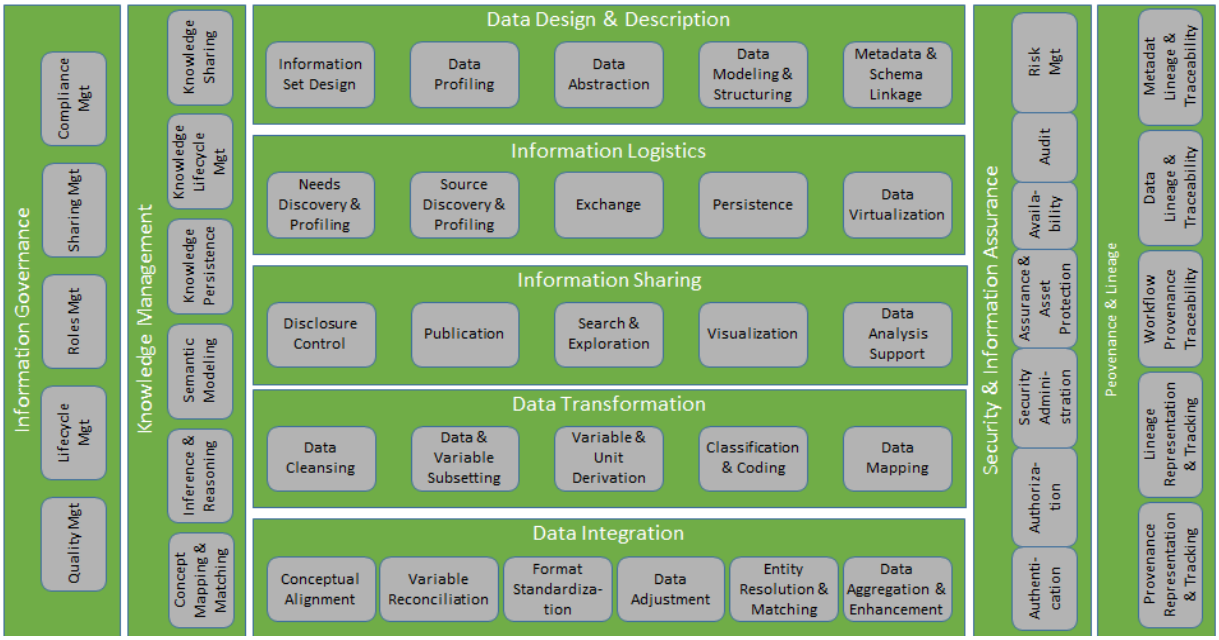

Figure 3: Capability Overview

C. Data Design & Description

Summary

69. The ability to design and describe, prescribe, and assess data and the structures that organize them.

70. This includes:

- 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.

Lower Level Capabilities

71. The Lower Level Capabilities for the Data Description & Organization capability are:

- **Information Set Design**
  Create conceptual, logical and physical information set designs, i.e. designs of organized collections of statistical content.
  Specialisations are:
  - **Dataset Design**
  - **Metadata Set Design**
- **Data Profiling**
  Assess and summarize datasets to determine data quality levels, accuracy and descriptive metadata completeness.
- **Data Abstraction**
  Describe data and its structures in such a way that information consumers can understand the data contents without any detailed knowledge of the underlying physical and technical implementation.
- **Data Modeling and Structuring**
  Represent datasets in terms of information objects, relationships and constraints to formally describe information contents and their organizations.
- **Metadata and Schema Linkage**
  Maintain links between datasets, schemas, and other types of metadata to describe (and prescribe) information contents and support metadata-driven data processing.

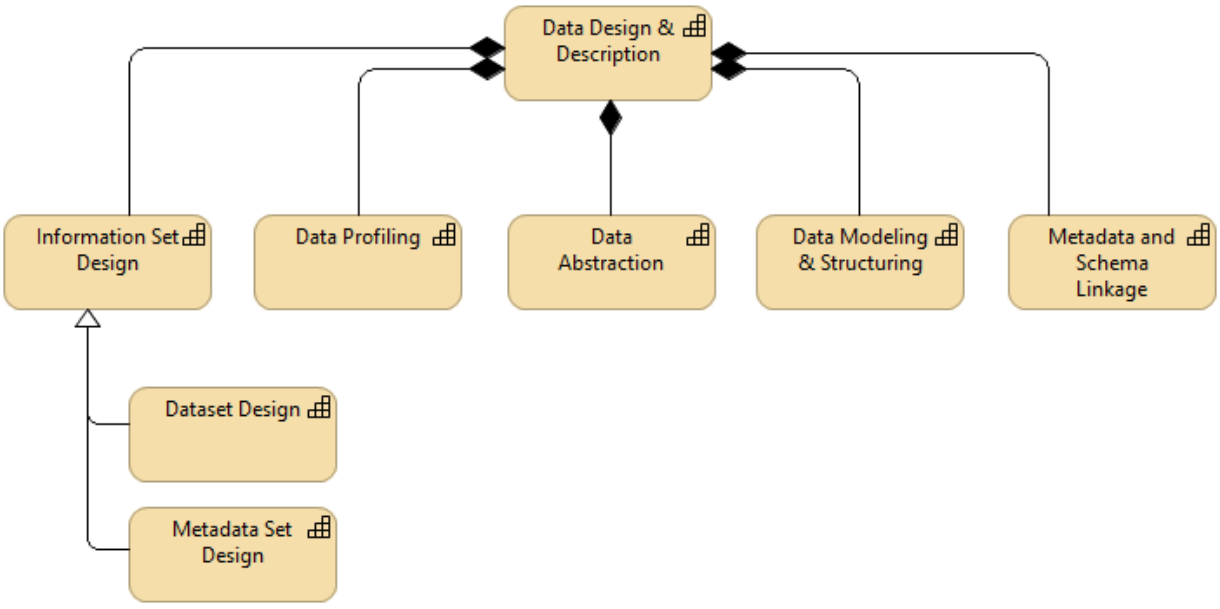

Figure 4: Data Design & Description

Description

72. Data Design and Description is a capability that structures, organizes, abstracts and describes information (specifically: data) assets so that they can be made available, and findable, to other capabilities in the architecture.

73. It includes both the purposeful design of datasets, based on known information requirements, and the reverse-engineering of information assets. In both cases it entails formally representing data 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.

74. 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), which in fact is a meta-model. 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).

75. 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* or *descriptive*. *Prescriptive* (or *normative*) metadata establishes what “must be” via rules that need to be applied to achieve full compliance. In some cases the rules are fully enforced, e.g. integrity constraints (rules) as defined in database schemas, where a data instance cannot even be loaded unless it satisfies all constraints in the schema. In other cases there is a relaxation of those rules, and the "must be" becomes a "should be" with multiple 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. In contrast, *descriptive* (or *informative*) 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.

76. Metadata 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. This also includes building ontologies and dictionaries and semantic analysis of data to ensure semantic consistency.

- **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 (i.e. a passive role), it is also used to drive the manual and automated parts of the statistical production process (the active role). This active role can also be seen in driving (enforcing) conformance to standards and semantic consistency. For metadata being able to take this active role, it usually means that the metadata is stored in specific machine-actionable formats, which requires particular expertise.

D. Information Logistics

Summary

77. The ability to make and keep information available to consumption points and manage information supply chains between providers and consumers.

78. This includes:

- Finding and characterizing relevant information sources for statistical production purposes.
- Connecting to information sources through a variety of channels (for example API, web questionnaires, administrative archives, statistical registers, streaming data, etc.).
- Managing the relationship with information providers (for example respondent management and service level agreements for administrative data sources).
- Ensuring information provisioning by maintaining channel operations.
- Enabling information exchange between applications and consumers.
- Manage information at-rest and in-motion, both physically and virtually.

Lower Level Capabilities

79. The lower level capabilities for the Information Logistics capability are:

- **Needs Discovery & Profiling**
  Identify, assess and summarize information needs at different granularities and within a variety of scopes, both internally and externally.
- **Source Discovery & Profiling**
  Find existing sources (enterprise data hubs, admin data registers, operational data stores, social networks, etc.) that may hold information of interest and characterize the data and

metadata they contain. Sources may be found outside, but also inside the own organization, such as other statistical processes, internal (backbone) registers, etc.

- **Exchange**
  Exchange statistical information between parties, both external and internal to the own organization.
  - **Channel Management**
    Create, maintain and withdraw information (exchange) channels, including their descriptive and/or actionable metadata.
  - **Channel Configuration & Operation**
    Configure information (exchange) channels for specific types of (data and metadata) traffic and maintain channel operation in accordance with service level agreements and other types of contracts.
  - **Relationship Management**
    Manage relationships with information exchange partners, both internal and external, by establishing and monitoring service level agreements and other types of contracts.
    Specialisations are:
    - **Information Provider Management**
    - **Information Consumer Management**
  - **Data Flow Management**
    Manage data in-motion as streams and setup and maintain data pipelines for real-time usage.
- **Persistence**
  Keep information available in the selected physical or virtual location (i.e. at-rest) over time.
  Specialisations are:
  - **Data Persistence**
  - **Metadata Persistence**
- **Data Virtualization**
  Dynamically create and maintain virtual (i.e. not physically stored) representations of data to simplify delivery, access and persistence and minimize data movement.

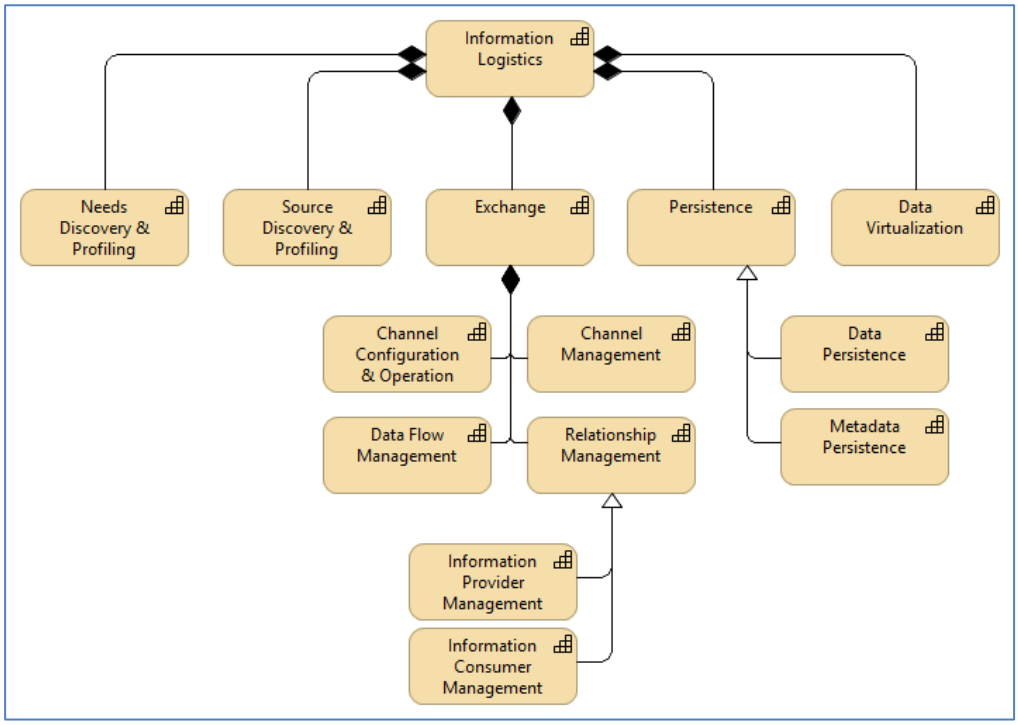

Figure 5: Information Logistics

Description

80. Information Logistics deals with the nitty-gritty of accessing, storing and moving data and metadata. This ability works closely together with Information Sharing.

81. Managing information storage (persistence) and exchange is the core of Information Logistics. This includes establishing and securing information provision agreements with relevant providers and consumers, both internal and external to the organization. It also includes ensuring the channels can operate with all necessary data pipelines and technologies.

82. Data pipelines consists of 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 pipelines connected to the channels might modify the received data contents and/or their structures to facilitate downstream consumption in general, and data integration in particular.

83. Schemas and information exchange models are fundamental for describing, and prescribing, how the data in-motion and at-rest is organized. Information exchange models enable the sharing and exchange of enterprise information that is consistent, accurate and accessible. These models are not intended to replace the disparate set of heterogeneous schemas used to persist information across the organization -- both data consumers and producers can continue to use their own schemas for data at-rest within their own environments.

Additional Notes

84. 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, key-value stores, spatial DBs, document stores, graph DBs, triple stores (RDF), etc. It also supports two main classes of processing styles[10]: 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, Open Data Protocol (OData), ADO.NET, C, C++, REST, SOAP, XML, XPath and Java.

85. 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 and other standards.

86. 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, e.g. to assess 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 all cases, clear contracts need to be set up and managed in time with the providers, which will include various degrees of harmonization and/or structure agreements that can range from loose to strict.

87. 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.), setting up a data pipeline for processing, or configuring other types of data flow execution jobs.

E. Information Sharing

Summary

88. The ability to make (statistical) information (i.e. data and metadata) available to authorized internal and external users and processes. This includes:

[10] Ref DAMA DMBOK

- Making data available via Catalogues and other API-based data publication mechanisms to allow consumers to find and browse data based on multiple criteria and methods
- Enabling the production of new insights via a combination of visual representations, traditional statistical modelling, machine learning and other advanced analytics methods
- Publishing data, with its pertinent metadata, via a range of internal and external channels
- Anonymizing data to ensure appropriate levels of confidentiality

Lower Level Capabilities

89. Specialisations of Information Sharing are:

- **Data Sharing**
- **Metadata Sharing**

90. The Lower Level Capabilities for the Data Sharing capability (inherited by both Data Sharing and Metadata Sharing) are:

- **Disclosure Control**
  Obfuscate, anonymize and/or redact information deemed sensitive by security and privacy policies by applying data suppression, perturbation, summarization or other techniques to ensure the appropriate level of confidentiality while preserving the usefulness of the data outputs to the greatest extent possible.
- **Publication**
  Make information sets available, either immediately or at a specific future date and time, to authorized consumers, both internal and external, via a range of (output) Exchange Channels.
- **Search and Exploration**
  Provide support for authorized internal and external users and processes in finding, ranking and browsing information sets based on their affinity to a set of target concepts or keywords.
- **Visualization**
  Create and maintain visual representations of information in order to communicate its content clearly and efficiently.
  Representations include table or cube structures, but also diagrams, maps, graphs, etc., including the use of colours.
- **Data Analysis Support**
  Enable authorized users to perform Data Analysis, i.e. inspecting data and producing statistical models with the purpose of finding new insights and validating assumptions about that data.

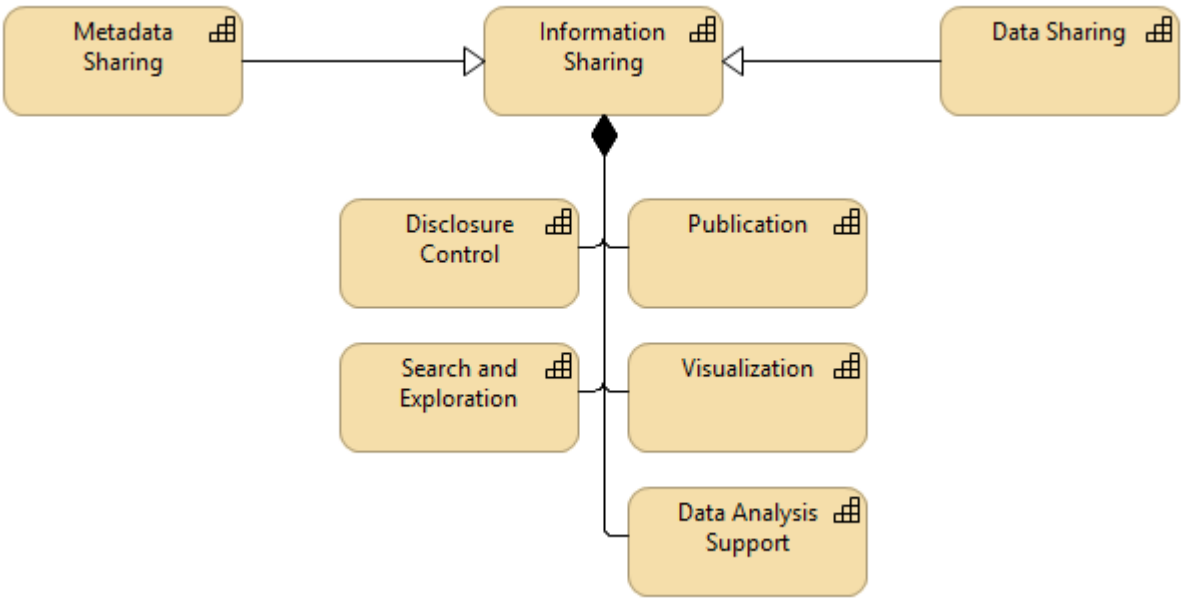

Figure 6: Information Sharing

Description

91. This capability is probably the most complex of all CSDA capabilities. This is because of the duality of functionality in this capability. Information Sharing has two sides to it: the "giving" and the "taking".

92. "Giving" deals with making data and metadata available to a selected audience, by opening up selected information sources through an appropriate set of exchange channels. Most of the "giving" takes place in Publication, but in some cases Disclosure Control is involved in order to preserve the privacy of persons or companies in the population. But much of the work involved here is actually around setting up and configuring the connections through Information Logistics. This involves making decisions about the mode of operation (CONNECT vs COLLECT, as explained in chapter IX), and (in case of COLLECT) setting up the necessary persistence mechanisms and related policies, taking into account the type and anticipated usage of the information as explained in chapter VII. And lastly, "giving" also involves deciding about, setting up and configuring the necessary mechanisms (exchange channels) for allowing information consumers to find and access ("take").

93. The "taking" has to do with the finding and accessing of information, previously made available through "giving". The ones actually doing the "taking" are the consumers (human beings, processes, systems) of data and metadata, internally (as part of same other capability) or externally. And in fact, the lower level capabilities for the "taking" included in Information Sharing are actually abilities *to support* such activities of information consumers. "Taking" again involves exchange channels, set up previously in "giving".

94. Data and metadata need to be exposed to both internal and external consumers (processes, people, and applications) via publication (exchange) channels. Information published this way becomes available to authorized user for downstream consumption.

95. Published information needs to be findable and discoverable, and this capability provides the abilities to browse and search. The ranking of search results is essential so that the user can easily distinguish the most relevant information sets from the less relevant ones. Once the desired information sets are identified, the information provided in the search result should be enough to access the actual data and/or metadata.

96. Visualizing data and metadata in different ways and at different granularities is also critical. 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 (and metadata), meaning "information that has been abstracted in some schematic form, including attributes or variables for the units of information"[11]. A primary goal of visualisation is to communicate information clearly and efficiently via statistical graphics, plots and information graphics.

97. This capability also provides the ability to support the analysis of data. This includes inspecting the data and applying statistical models with the goal of discovering useful information, suggesting conclusions, and supporting decision-making. Data analysis has multiple facets and approaches, encompassing diverse techniques, including statistical analysis, dashboarding, business intelligence, machine learning, and query languages.

98. When data assets are shared with consumers that don't have full access to the underlying data points, some form of disclosure control is required. A wide variety of methods can be applied as part of this capability, including suppression, perturbation, summarization or other techniques to ensure the appropriate level of confidentiality and anonymity.

Additional Notes

99. Search and exploration is metadata-driven – access to the actual data is not required. Usually, a data catalogue capability provides the ability to browse and search. This catalogue may include a description of the used/needed authentication and authorisation to access the data assets and the selection and filtering options to retrieve the data. For exploring the data, a quality indicator is also invaluable.

[11] Ref Wikipedia: https://en.wikipedia.org/wiki/Data_visualization

100. Publication channels often rely on virtual or persistent data management solutions, e.g. data hubs, for consolidating and registering data to be published.

101. Once a data asset of interested is found in the catalogue, the actual data can be accessed using standard protocols, e.g. ODBC, JDBC, SOAP/Restful services, through APIs, or using standard query languages like SPARQL or GraphQL. For data exchange, a range of protocols can be supported, e.g. Open Data Protocol (OData), an open protocol that allows the creation and consumption of queryable and interoperable RESTful APIs in a simple and standard way. Standard exchange models could also be supported, including the Data Documentation Initiative (DDI) and the Statistical Data and Metadata Exchange (SDMX).

102. 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.

F. Data Transformation

Summary

103. The ability to transform data to make them suitable for specific purposes downstream.

104. This includes:

- Checking data coherence against data descriptions coming from metadata system used in the organisation and flagging issues.
- Cleaning data by correcting input errors, deleting duplicates, imputing missing data, correcting data formats, and applying other data cleansing techniques.
- Harmonising data values using classifications coming from national and international standards.
- Coding data coming from textual or non-structured sources by using coded, enumerated value domains, e.g. classifications and code lists.
- Reducing the amount of data by filtering rows and selecting relevant columns for specific uses.
- Altering the data following security or statistical significance reasons.

Lower Level Capabilities

105. The Lower Level Capabilities for the Data Transformation capability are:

- **Data Cleansing**
  Detect and correct data errors and inconsistencies in structure, representation and content, remove duplicates (after performing entity resolution), and fill in missing values with estimates.

- **Data and Variable Subsetting**
  Select from a dataset those records and variables that are relevant for specific purposes downstream, eliminating those that are not.
- **Variable and Unit Derivation**
  Create new statistical units and variables derived from those existing in a dataset.
- **Classification and Coding**
  Assign codes to textual descriptions according to pre-determined classification schemes.
- **Data Mapping**
  Establish correspondences between two data models for the purpose of delivery, access or persistence.

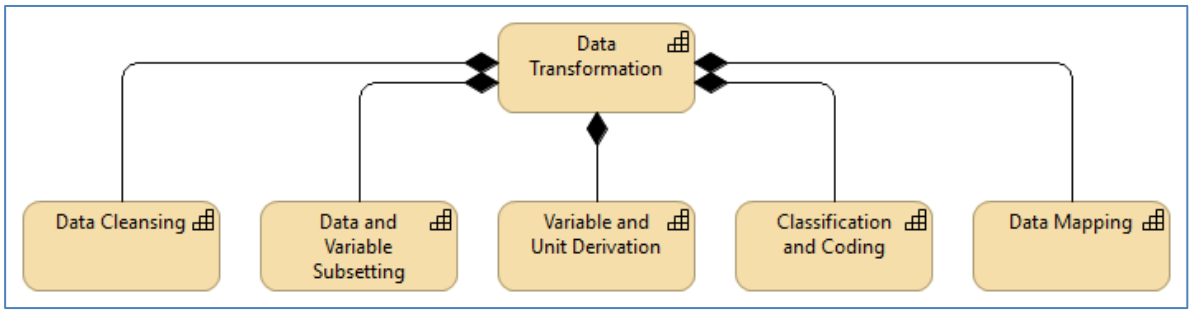

Figure 7: Data Transformation

Description

106. Data Transformation is the ability to transform data 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.

107. 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.

108. Data cleansing, or data cleaning, is the process of detecting and correcting (or removing) corrupt or inaccurate records or values from a data set. It 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. The inconsistencies detected or removed may have been originally caused by user entry errors, by corruption in transmission or storage, or by different definitions and semantics used in different stores/sources. 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.

109. Reformatting is often needed to convert data to the same (standard) type. This often happens when talking about dates, time zones and other common data types.

110. After cleansing and reformatting, a data set should be consistent with other similar data sets in the system.

111. Data harmonisation is the process of minimising redundant or conflicting information across dimensions that may have evolved independently. The goal here 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, e.g., use of standard codes.

G. Data Integration

Summary

112. The ability to combine, link, relate and/or align different data sets in order to create an integrated information set.

113. This includes:

- Harmonizing variables and concepts across data sets and repositories.
- Standardizing data types across multiple platforms, technologies and solutions, including unstructured/semi-structured (NoSQL Databases) and structured (Relational Databases).
- Identifying multiple representations of the same entity across data sets and repositories.
- Aggregating, blending and creating mashups of heterogeneous sources (data sets, databases, data warehouses, Big Data, Linked Open Data) using a variety of techniques.

Lower Level Capabilities

114. The Lower Level Capabilities for the Data Integration capability are:

- **Conceptual alignment**
  Map (or harmonize) concepts between multiple conceptual frameworks, taxonomies, ontologies and data/metadata exchange standards to support semantic integration.
- **Variable Reconciliation**
  Map (or harmonize) variables across datasets to improve comparability and linkage.
- **Format Standardization**
  Modify physical representations and data types to conform to standards and to best support readily shareable information.

- **Data Adjustment and Normalization**
  Perform seasonal and other statistical adjustments to data values to improve accuracy and align data values to a notionally common scale or a probability distribution to make datasets comparable.
- **Entity Resolution and Matching**
  Identify different instances or representations of the same entity (real-world object) across different datasets by inferring relationships and remediating conflicts to support data integration, record linkage and entity deduplication.
- **Data Aggregation and Enhancement**
  Clean, improve and extend information contents by joining records from multiple datasets and/or heterogeneous sources, with potentially different representations, to produce a more comprehensive view of an entity, concept or other subject of interest.

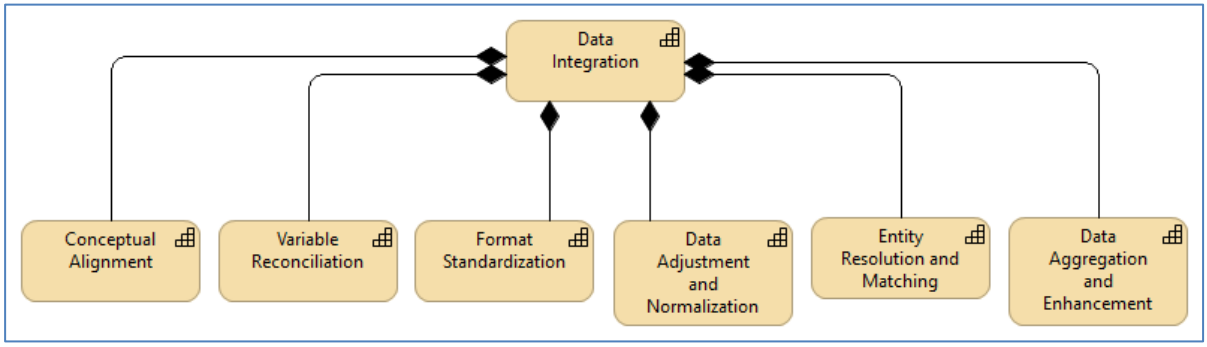

Figure 8: Data Integration

Description

115. Data integration is a key capability of the target architecture supporting the statistical organisation's ability to fulfil information needs from multiple data sets. According to GAMSO, data integration is "the process of combining data from two or more sources."

116. At many points in statistical production data needs to be linked, integrated and consolidated from multiple sources. This requires harmonization at multiple levels, e.g. concepts, variables and formats, together with agile data modelling and structuring allowing users to specify data types and relationships as needed, especially in schema on-read scenarios.

117. The results of the integration could become new data sources that need to be persisted and managed independently in statistical registers, master data solutions, and other types of data hubs, e.g. data registries, data warehouses, etc. They could also include the generation of semantic models and ontologies.

118. A data integration technique commonly used in the statistical domain is record linkage, i.e. finding records that refers to the same entity across multiple data sources. Record linkage

benefits from entity resolution and matching, since it requires to infer relationships and remediate data conflicts. It produces as a result and enhanced, more comprehensible view of a set of entities of interest.

H. Information Governance

Summary

119. The ability to manage information (assets) through the implementation of policies, processes and rules in accordance with the organization's strategic objectives.

120. This includes:

- Defining roles and responsibilities.
- Defining and assessing data quality indicators.
- Defining and enforcing rules for disclosure control.
- Setting and enforcing policies related to data and metadata lifecycle management, including schemas and data exchange models.
- Ensuring the organization can meet data-related regulatory compliance requirements.

Lower Level Capabilities

121. Specialisations are:

- **Data Governance**
- **Metadata Governance**

122. The lower level capabilities for the Information Governance capability (inherited by both Data Governance and Metadata Governance) are:

- **Roles Management**
  Define information-related roles (e.g. steward, custodian, etc.) and maintain their allocations to people and organizational structures over time.
- **Quality Management**
  Assess information sets against relevant standards (format, semantics and level of quality), and define and monitor quality indicators. This includes:
  - **Review and Validation**
    Examine information in order to identify potential problems, errors and discrepancies, in terms of validity (against business rules), accuracy (conformity to a standard or true value), completeness, and consistency (with other referential info) to ensure the appropriate level of quality.
- **Sharing Management**
  Define and enforce rules for sharing information that is deemed sensitive by

implementing security and privacy policies and manage how obfuscation, summarization, anonymization and other disclosure control techniques are applied.

- **Lifecycle Management**
  Manage the release and configuration of information changes, from creation/acquisition to persistence, enrichment, usage, sharing, archival and destruction. This includes the ability to define and implement retention rules for controlling the persistence (i.e. at-rest) of information. It also includes:
  - **Schema Management**
    Maintain persistent schemas (e.g. relational, XML, etc.) for information at-rest.
  - **Exchange Model Management**
    Maintain physical exchange models for information flows (in-motion), i.e. common vocabularies for interoperability,

- **Compliance Management**
  Manage (measure, enforce) the degree of compliance with governance policies regarding standardization, security, privacy and lifecycle management.

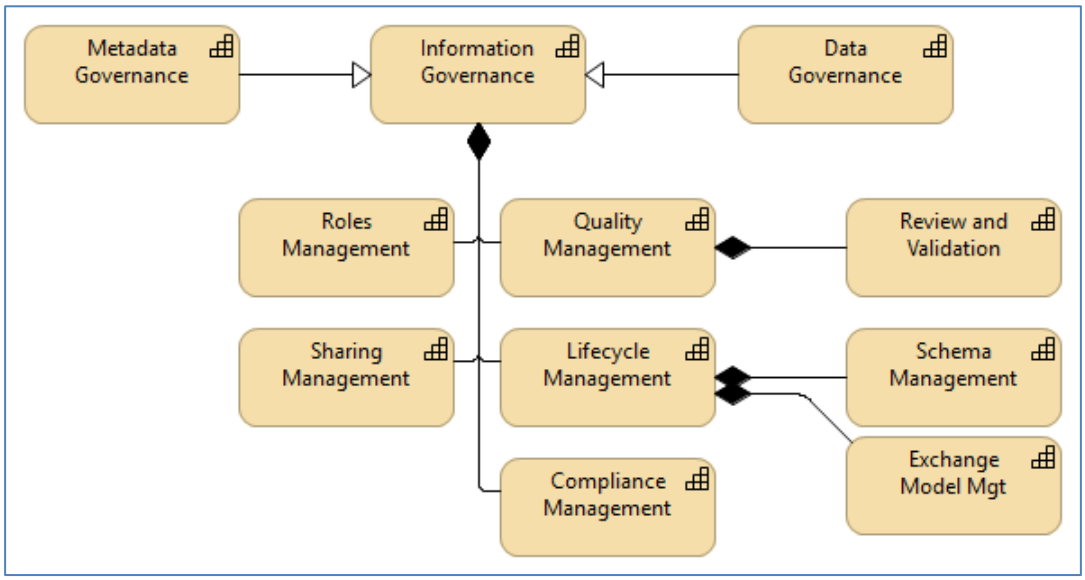

Figure 9: Information Governance

Description

123. GAMSO defines "Governance" as "Establishment of policies, mechanisms, and continuous monitoring required for proper implementation by an organization."

124. Gartner defines information governance as the specification of decision rights and an accountability framework to ensure appropriate behaviour in the valuation, creation, storage, use, archiving and deletion of information. It includes the processes, roles and policies, standards and

metrics that ensure the effective and efficient use of information in enabling an organization to achieve its goals.

125. Information Governance is the ability to define, implement, monitor and enforce the execution of policies concerning all information the statistical organisation somehow has a responsibility for (e.g. as user, creator or provider). Within this capability, several layers can be distinguished. The top layer is where policies are defined. As explained before, it is recommended to do this by "type of data", i.e. sets of policies that apply to specific classes of data. The next layer is the implementation of these policies. This specifically includes being involved in the design and implementation of new or changed capabilities (GAMSO: Capability Development), as policies often have to be ingrained into the fabric of the organisation. Yet another layer is where the monitoring of compliance occurs, i.e. the actual behaviour of the organisation (in this context, specifically in the CSDA Core Capabilities) is measured against the norms set out in the policies. And lastly, there is the layer of correction, where deviations of the norms are rectified, and measures are taken to prevent re-occurrence. In fact, these layers correspond to the 4 steps in the Deming circle, PDCA (Plan, Do, Check, Act).

126. As explained before, the specific problem in this day and age is the fact that information is extremely dispersed, easily copied and moved through digital networks. We all are probably aware of the struggle creators of intellectual property such as music, songs, etc. have to control the (paid) use of their work. The task of a Statistical Organisation in managing and governing the information acquired or created for the execution of its mission is truly a daunting challenge.

127. In addition, the sheer volume of information that is available and the increasing appetite of NSI's for tapping into the potential value of that enormous volume, leads to new types of problems as well.

I. Security & Information Assurance

Summary

128. The ability to maintain security and continuity of (the availability of) all data and metadata under control of the Organization.

129. This includes:

- 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.

Lower Level Capabilities

130. The Lower Level Capabilities for the Security and Information Assurance capability are:

- **Authentication**
  Define and implement procedures and mechanisms for secure identification of users.
- **Authorization**
  Define and implement procedures and mechanisms for access control to (statistical) information.
- **Security Administration**
  Add and change security policies and their implementations, manage security groups and access control to information assets.
- **Assurance and Asset Protection**
  Monitor security attributes to uphold the stated security policy, data usage, loss and unintended disclosures.
- **Availability**
  Maintain availability of data and metadata according to existing service level agreements and other contracts and policies despite abnormal or malicious events. This includes:
    - **Information Backup** (of which **Data Backup** and **Metadata Backup** are specializations)
      Define information backup strategies, policies, procedures and responsibilities based on information requirements.
    - **Replication and Synchronization**
      Maintain multiple copies of information contents to satisfy non-functional requirements while keeping them consistent by physically moving and reconciling them across disparate systems and models.
- **Audit**
  Provide forensic information attesting that datasets have been created, read, updated or deleted in accordance to stated security policies.

- **Risk Management**
  Identify, evaluate, and prioritize risks, including cyber threats, provide mechanisms to monitor and minimize them, and train users on risk-related topics.

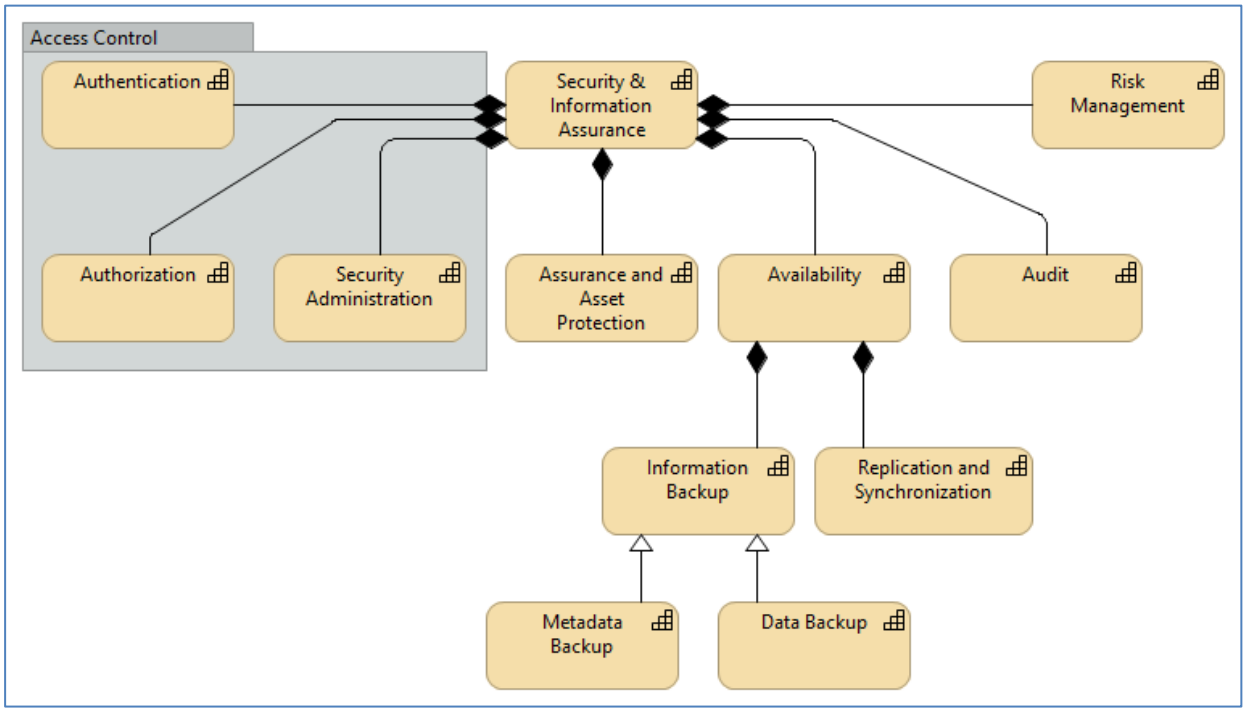

Figure 10: Security & Information Assurance

Description

131. 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.

132. 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.

133. 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 continued availability, integrity and quality of the data).

134. With increased access to an ever larger variety of data sources, 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.

135. 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

136. 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.

137. A major objective of Information Assurance 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.

138. 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.

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

140. 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.

141. 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.

142. 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, whether this process is executed on-premises or somewhere outside, in the cloud or even on systems owned and operated by other parties.

143. Because of the inter-operability aspects of such scenarios, It is strongly recommended that NSI's, in the implementation of their IA&S policies and mechanisms, apply the many international standards available. One in particular, in the area of access control, is the application of the PEP/PDP principle, i.e. the separation of Policy Enforcement from Policy Decision making, as defined by the architecture underlying the XACML standard.

J. Provenance & Lineage

Summary

144. The ability to capture and maintain provenance and lineage of data and metadata.

Lower Level Capabilities

145. The Lower Level Capabilities for the Provenance and Lineage capability are:

- **Provenance Representation and Tracking**
  Keep track of the origins of information entities and the activities and agents that interact with them over time.
- **Lineage Representation and Tracking**
  Determine which CRUD (Create, Read, Update and Delete) operations are deemed relevant and trace how they affect the information contents and their causal relationships over time.
- **Workflow Provenance Traceability**
  Query and explore how information entities of interest flow thru activities and agents over time.
- **Data Lineage and Traceability**
  Query and explore data contents of interest that are causally related, trace their origins and how they are used and/or changed over time.
- **Metadata Lineage and Traceability**
  Query and explore schemas (and other metadata) that are causally related, trace their origins and how they are used and/or changed over time.

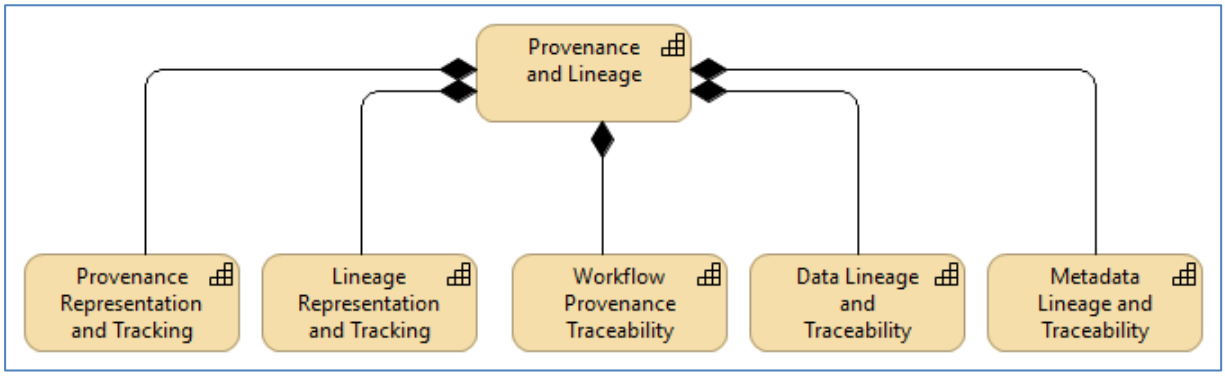

Figure 11: Provenance & Lineage

Description

146. 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.

147. 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.

148. 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.

149. The World Wide Web Consortium (W3C)[12] provides an ontology to express provenance and lineage data.

K. Knowledge Management

Summary

150. The ability to manage intellectual capital (knowledge) in all its forms.

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

151. This includes:

- 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 artefacts, like architecture documents, best practices, guidelines, etc.
- Supporting inference and reasoning to derive new knowledge from existing one.

Lower Level Capabilities

152. The Lower Level Capabilities for the Knowledge Management capability are:

- **Knowledge Lifecycle Management**
  Manage the release and configuration of semantic models and knowledge representations changes.
- **Concept Mapping and Matching**
  Find and maintain correspondences between concepts in different semantic models and knowledge representations for the purpose of delivery, access or persistence.
- **Inference and Reasoning**
  Derive new knowledge (implied facts) from explicitly represented, actionable knowledge (asserted facts).
- **Semantic Modelling**
  Capture and organize knowledge (concepts, terms, definitions) in domain-specific ontologies, classifications, taxonomies and thesauri to be used for communication purposes and/or to create actionable knowledge representations.
- **Knowledge Representations and Persistence**
  Formalize knowledge into actionable representations using standards, e.g. RDF, OWL, SKOS/XKOS, PROV-O, and keep that knowledge available over time.
- **Knowledge Sharing**
  Make knowledge available to authorized internal and external users and processes.
- **Knowledge Set Design**
  Create conceptual, logical and physical knowledge set designs, i.e. organized collections of statistical knowledge content.

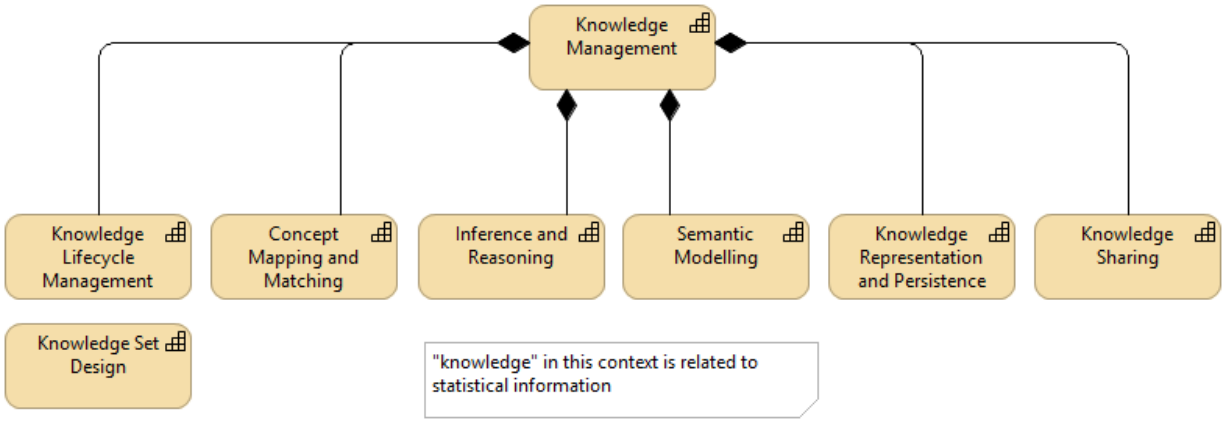

Figure 12: Knowledge Management

Description

153. 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 physical or 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.

154. (Davenport, 1994): *"Knowledge Management is the process of capturing, distributing, and effectively using knowledge."*

155. 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.

156. 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.

157. 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.

158. 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.

159. 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.

160. 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.

161. In the context of statistical production, knowledge associated to data is organized into *Knowledge Sets*, an extension of GSIM Information Sets that consists of datasets, their referential metadata, other types of statistical content and *actionable semantic descriptions*, e.g. linked open data (LOD) and semantic web knowledge content.

Additional information

162. Knowledge is a fairly new topic in this context, and not yet well understood. For that reason it is kept separate in this version of CSDA. In future versions, knowledge may become more integrated, specifically with Metadata.

XI. Relation to other standards

163. In this chapter, the position of CSDA with respect to some other standards is explained. In particular, the relationship with TOGAF, GAMSO, GSBPM and GSIM is presented.

A. TOGAF

164. CSDA accepts and embraces the way The Open Group's TOGAF standard defines and positions capabilities. Capabilities are the content of an organization's strategic planning effort. Capabilities are the "what", TOGAF provides the "how" of capability design and development. TOGAF also describes the use of architecture (business, information systems, data and technology) in that development.

B. GAMSO and GSBPM

165. Probably the most important relationship between CSDA and GAMSO is in Capability Development. The initial implementation and further (iterative) development of CSDA capabilities falls into the GAMSO activity area Capability Development. The use of TOGAF, and more specifically the application of the (iterative) ADM development method is strongly recommended in this context.

166. The relationship between CSDA capabilities and GAMSO is shown in Figure 13: CSPA vs GAMSO. As explained before, the CSDA core capabilities are the capabilities that directly enable statistical production, that is, the storing, exchanging, processing of data and metadata including the description (prescriptive or descriptive) of data. All this overlaps with the Production activity area of GAMSO, and therefore with GSBPM. The CSDA cross-cutting capabilities are a further elaboration of (data related aspects of) some of the GAMSO Corporate Support activities, notably Manage Information & Knowledge, Manage Quality, Manage Business Performance & Legislation and Manage Statistical Methodology.

167. CSDA considers the two GAMSO Corporate Support activities Manage Consumers and Manage Data Suppliers as more operational level activities. For that reason, CSDA places these two in the Core capabilities (as lower level capabilities of Information Logistics).

168. Figure 13 shows these relationships graphically.

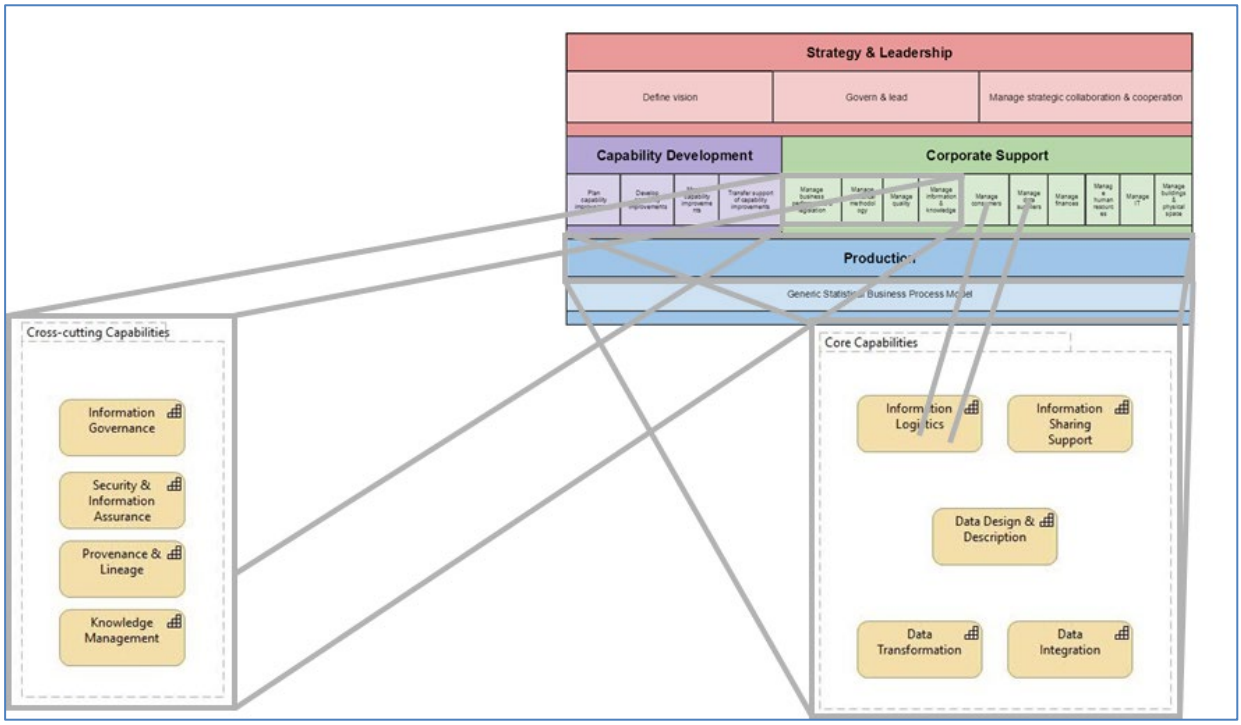

Figure 13: CSPA vs GAMSO

169. CSDA (core) capabilities are used throughout the statistical process. For that reason, there is no one-to-one relationship between CSDA capabilities and GSBPM process phases. Most CSDA capabilities are needed in many GSBPM (sub) phases, and any GSBPM (sub) phase needs more than one CSDA capability. Some further elaboration of this is presented in the examples in chapter XII.

C. GSIM

170. There is a strong relationship between CSDA and GSIM. As explained before in this document, CSDA is about exchanging information and therefore models concepts that are also modelled in GSIM, albeit from different angles. CSDA is looking to model the behaviour (of the real-world object), whereas GSIM models the same as information objects, in order to be able to capture and exchange information **about** those objects. Nonetheless, also GSIM has a notion of the (expected) behaviour of such objects. This is most notable in active objects such as "Exchange Channel" (and its specializations) and "Process", but also "Information Provider", "Information Consumer". And there are quite a few other GSIM objects that are mentioned in CSDA.

171. Hopefully, CSDA will eventually provide sufficient information for NSI's to be able to design, build and implement capabilities. GSIM then will provide attributes to describe the (relevant) characteristics of such capabilities. Where the implementation still contains elements

that can be configured, GSIM will provide the attributes to specify the parameters that define such configurations.

XII. Examples

172. In this chapter, we present a couple of examples of how the CSDA capabilities can be used to design or analyse solutions such as business processes. The first three examples show which CSDA capabilities most likely play a role in the GSBPM phases Collect, Process and Disseminate. The two other examples show which CSDA capabilities are involved in two of the CSDA use cases that are described elsewhere.

A. Data Collection

173. Let's take a look at the GSBPM phase "Data Collection". This takes care of bringing data in from the external world. In CSDA's philosophy, that data then enters "the pool", thus, depending on authorization, of course, becoming available to users both internal and external to the organization. In line with GSIM, all of the work done in this phase takes place in the Exchange Channels, thus in Information Logistics and Information Sharing (Publication). The data related activities of the GSBPM phases Design and Build are collapsed into "Design & Description". Publication is the capability that takes care of designing and setting up the collection process, as well as preparing "the pool" and any output channels for future access to the information collected. Preparing "the pool" may or may not include setting op persistence mechanisms, but in any case it includes setting up the mechanisms to protect the data in accordance with the policies laid down by the Cross-Cutting capabilities (Governance, Security).

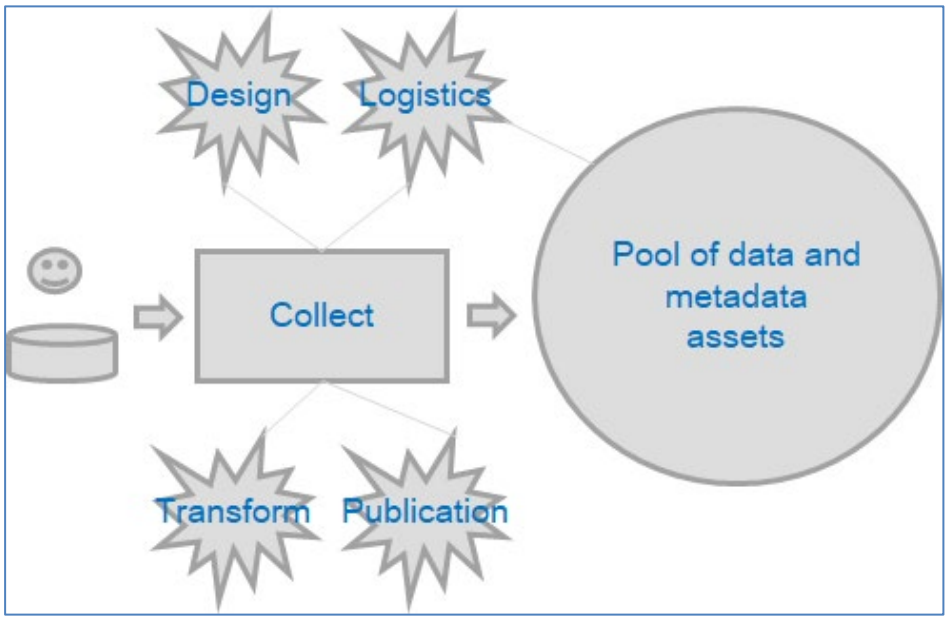

Figure 14: Example GSBPM Collect

174. Although we may want to collect information from sources that contain information in all kinds of forms, the pool of data and metadata contains only digital information. We often even

want to collect "intangible" information, such as the facts, ideas and opinions in the heads of people. It is the task of the Exchange Channels (as explained in chapter VIII) to collect and digitize such "intangible" information.

175. Collecting information of such nature, internal persistence will be required in order to decouple the internal processing from the collection.

176. This way, non-digital sources can be treated the same as digital ones. All sources are connected to the “pool” through channels responsible for digitizing any non-digital data.

B. Processing

177. The actual data processing of GSBPM’s “Process” phase takes data from the Pool (through Information Sharing and Information Logistics), and the output is placed back into the pool, again through Sharing and Logistics. Both the final result and any Intermediate results are designed (using Design & Description) and the actual processing happens in Transformation and/or Integration. Any data that is “transient” and not considered of enduring value, is not shared and kept locally in the process.

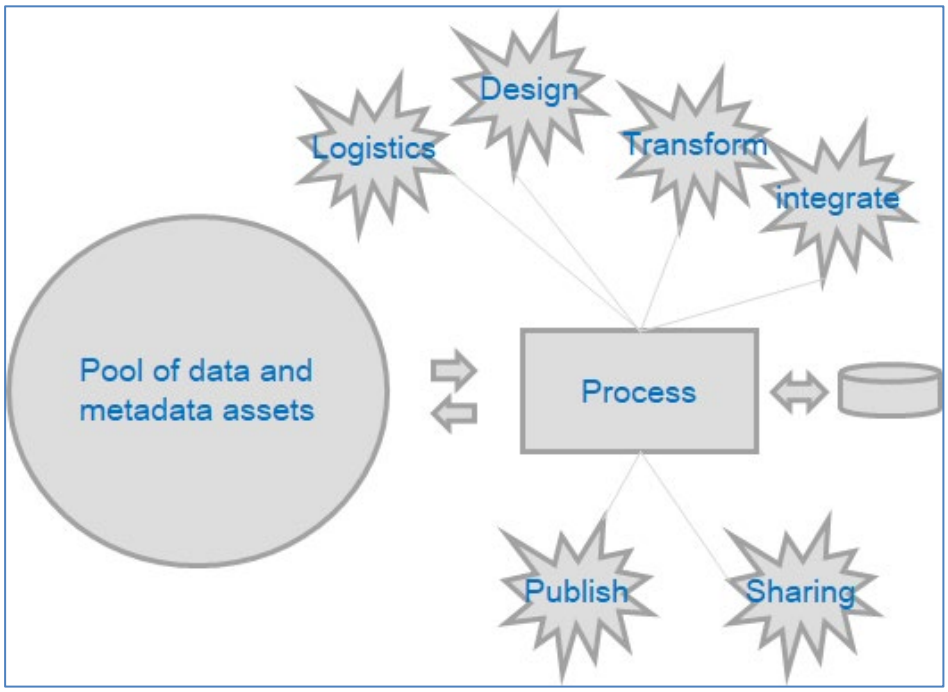

Figure 15: Example GSBPM Process

178. 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.

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

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

C. Dissemination

181. GSBPM’s Dissemination is the opposite of Data Collection: the final product is made available to “the world”, “the general public” or any subset thereof, again by publishing to “the pool”. External consumers use Sharing and Logistics to search, find and extract the data.

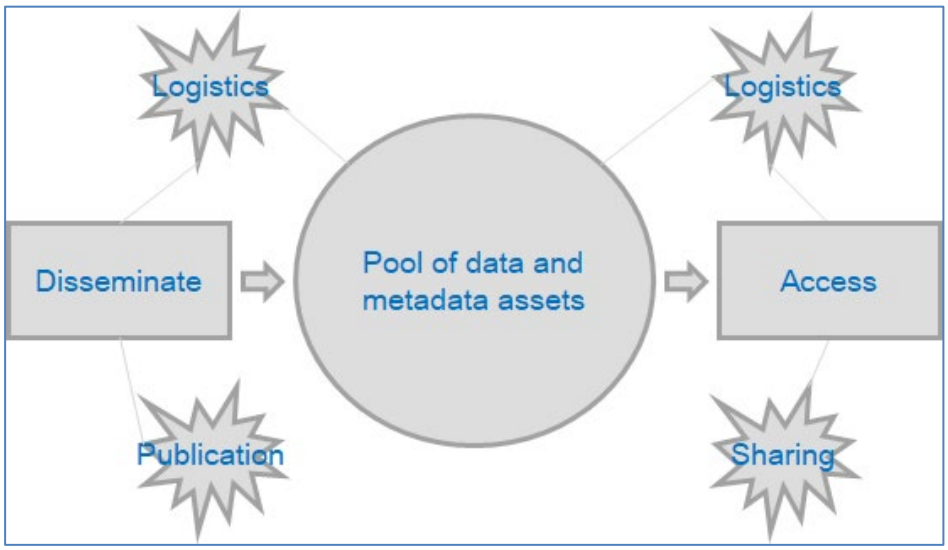

Figure 16: Example GSBPM Disseminate

182. Publishing a statistical end-product is (conceptually) the same as publishing any other Information Set. The information to be published may come from the “pool” or from some internal process.

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

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

185. This, in a nutshell, is how the Core Capabilities can be used in designing the GSBPM (sub-)processes. What we did not show is the role of the Cross-cutting capabilities. Their role is to define and enforce the policies that govern the way that data is handled, protected, assured, etc. It is the responsibility of the Core Capabilities to act in accordance with those policies.

D. Use case: Data Collection (StatCan)

186. The process depicted in the figure below (the dark boxes) is the data collection and initial treatment of a complex set of datasets, which provides 9-character CUSIP numbers, standardized descriptions and additional data attributes for over 6 million corporate, municipal and government securities offered in the United States and Canada. Although the top level structure of this data is rather simple, the details are very complex and are changing over time. That's why, in this process, a lot of data modelling takes place. Data is published in "almost" raw format as well as in a "more refined" form.

187. As you can see from the mapping to the CSDA capabilities (the lighter boxes), there's most often an n-m relationship, that is, a process step implements multiple capabilities and the same capability is used in multiple steps.

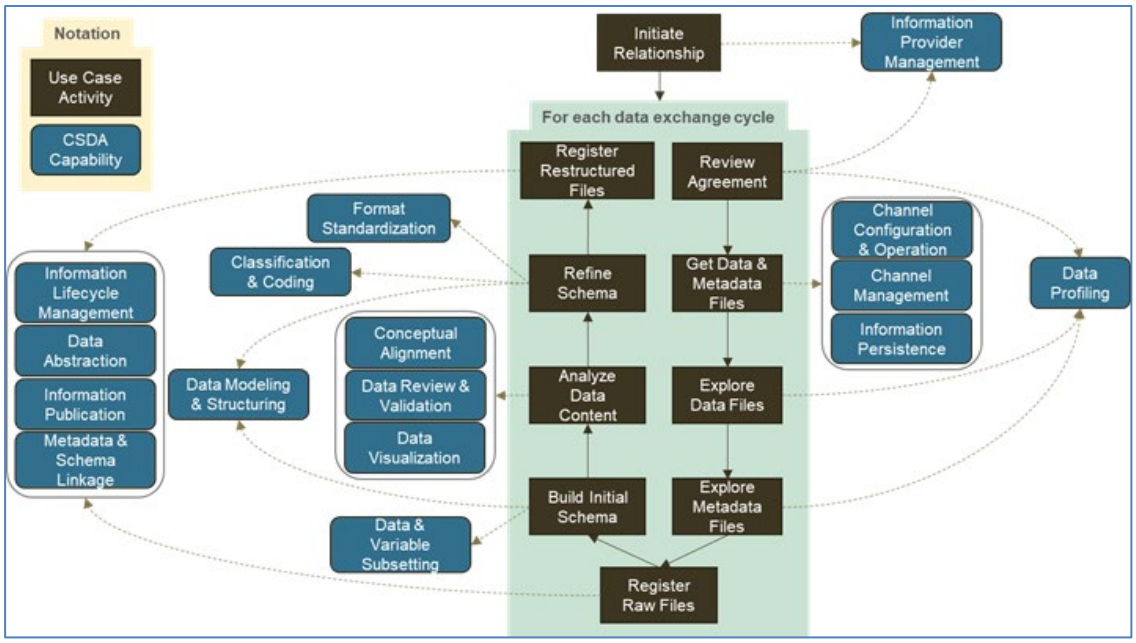

Figure 17: Example StatCan UseCase

E. Use case: Privacy Preserved Data Sharing (Stats Netherlands)

188. The second use case is from Statistics Netherlands and is called "privacy preserved" sharing of data. Data from two sources (one of which is CBS) are brought together and integrated without disclosing the identity of the persons described by the data to the other party. A lot of complex encryption, digital signing and secure communications takes place, but conceptually, we can map the CSDA capabilities involved.

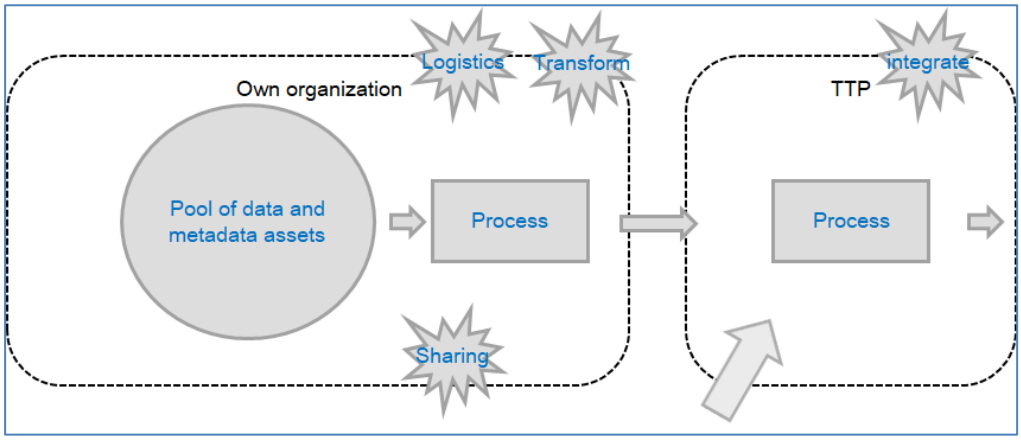

Figure 18: Example Stats Netherlands UseCase

XIII. Annex: Old vs New Data Sources

Data Architecture for Old and New data sources

189. We tried to analyse a number of examples of new and traditional data sources, trying to generalise from the actual examples and investigating the general nature of newer data sources as compared to the more traditional ones. Our conclusion is that the newer data sources tend to provide data that is ever closer to the actual phenomena that the NSO is trying to describe. In the traditional cases, we still needed "agents", usually of the human kind, to provide the data. Talking about registers, i.e. data already collected for other purposes by other parties (usually organisations), we can find ever more data sources that others have set up for an ever growing variety of reasons, containing or delivering data describing an ever increasing number of phenomena.

190. The interesting thing is, that this data seems to be ever more "raw", "primitive", ever more elementary. Instead of using data that is already "pre-processed", filtered, coded, etc., we tend to tap into unfiltered, unprocessed, raw data, that is more detailed by nature. Sometimes, it is "unfocused", as for example satellite imagery, where a certain image may be analysed for determining land usage, harvest estimation or counting the number of solar cells on roof tops. In other cases, the nature of the data is very focused, such as the output of traffic loops. These newer types of data very often are directly related to individual transactions or events ("happenings"), such as buying in a shop, money transfers, vehicles passing a certain spot on a road, making a phone call, goods passing the border, etc. etc. rather than the aggregated data that we used to collect in the past.

191. Where traditional data was purely alphanumerical, the new data is increasingly also of other types. Also, most of the new data is still of the alphanumerical type. But increasingly, we see data of other types, for now primarily imagery (still or video), but potentially also audio.

192. The general consequence of the use of such newer data is, that the statistical organisation itself has to do the processing. Or, at least, determine what processing is required to extract the desired "pure" information out of the "raw" data. Execution of this processing may be outsourced to others, such as the data provider or a third party. But determining the processing required involves a preliminary stage of investigation, analysis and specification, maybe even building prototypes.

193. Due to the more detailed nature of the new data, the volume is increasing rapidly. So, not only do we have to process more in order to extract the true value from the detailed data, there is more of it as well. Whether we "hit" more units of the population under study, still remains to be seen.

194. Below we try to analyse the consequences for the architecture derived from the different characteristics of current and new data sources.

Table 3: Old vs New Datasources

Current	New	Impact on Architecture
Structured	Un(semi) structured	Although the general perception might be that traditional data is structured, that impression is not necessarily correct. We should not forget that the data that an NSI receives is not the output of the collection phase, but the data that "is out there in the real world". The NSI uses an instrument (the interview, structured through a questionnaire) to select, filter and format the data from the raw source, the interviewee. So, indeed the data received from the collection phase is pre-filtered (by the NSI itself). Because of the long experience, NSI's have a lot of in-house knowledge about how to do this. In the case of new data sources, that the NSI is (as yet) unfamiliar with, it needs to study (a sample of) the raw data and decide how to filter, format and structure it. Example: web scraping, exploration phase for Big Data. The impact is that we need capabilities for this exploration and for designing and building the necessary instruments.
High Latency (infrequent)	Low Latency (frequent) Streaming	Receiving the information (much) soon(er) after the event happened in the real world means generally better quality. It also means an opportunity to improve the actuality of the published statistics. In order for this to happen, the internal processing must also be speeded up, until the point of "straight through processing". However, the nature of statistics, aggregation, puts restraints on this: per cell a minimum amount of relevant data must be received before the cell summary value can be calculated. It will become a local trade-off whether to process directly or to store data awaiting more data.
Large Batch Size	Small Batch Size	Receiving small but frequent amounts of data opens the possibility to spread the load of processing over time,

Current	New	Impact on Architecture
		reducing the need for high peak capacity in the infrastructure that stands idle for the rest of the period. However, just like with latency, there are limits to the gain that can be reached.
Pre-filtered	Raw / Detailed	As explained above, the pre-filtering involved in Primary data collection and most likely also in the collection from Administrative Data sources, is almost "sub-conscious" in most NSI's. Due to the amount of detail available in new data sources, also the volume of data tends to explode (hence the Big Data). All this data needs to be processed, so the requirement for the capabilities involved is clearly to be able to handle (much) larger volumes of data. Also, as explained before, the NSI must develop new methods and instruments to select/filter, format and potentially aggregate the detailed data. In certain cases, this work can then be off-loaded to parties (data providers) upstream in the value chain.
Controlled Quality (Less Noise)	Uncontrolled Quality (Noise Bias)	Like explained before, in primary data collection and to a certain extend also in administrative data sources, the NSI can control or at least influence the processes involved in the data collection, with positive effects on the quality of the data. In the case of Primary data collection, this includes the sample selection. Newer data sources in most cases do not allow for this up-front control, the data is what it is and the NSI needs to take measures after the actual ingestion to manage the quality, maybe discarding part of the data received as irrelevant or unusable due to bad quality.
In-House Processing	External Processing	Due to the large volume of data, organisations increasingly feel the need to off-load some or all of the processing involved in checking, selecting, reformatting, aggregating, etc. to stations upstream in the value chain. This involves mechanisms to express the algorithms to be applied, and sometimes even distributing actual code, as well as mechanisms to validate and manage

Current	New	Impact on Architecture
		compliance. Big impact on partner collaboration. Technically, this may or may not involve cloud technologies.
Less Complex Supplier Management (Internal Skills sufficient)	More Complex Supplier Management (External Expertise needed)	Both for understanding the meaning and usefulness of the data as well as for the collaboration on external processing, there is a need for closer cooperation between the NSI, its data providers and possibly intermediate parties involved. The complexity can arise in a number of ways: (a) legal contracts could be more difficult to agree, (b) some suppliers will want to work with NSI as partners (e.g. supplier data for free in the expectation that there will be benefits to them from the statistical output, such as obtaining more detailed information or obtaining the information sooner).
Less Work for Standardizing / Cleansing	More Work for Standardizing / Cleansing	Due to the fact that the NSI has less influence on the upstream processes, and wants to be able to use data from more and different sources, it needs to invest more in improving the quality of data for its own purposes. As explained before, in the traditional cases, most of this work is done in the preparation of the data collection instruments (the questionnaires). This however is not only a quality issue. It also requires a mapping exercise or a transformation process to ensure the source data is made to comply with the standards which are needed for statistical processing.
Mastering Reference Data Internally	Linking to External Reference Data	External data providers of digital data most likely have their own reference data. Instead of creating and maintaining an internal copy of that reference data, NSI's may choose to re-use the external reference data by linking to it, as an (external) live repository. This involves collaboration on organisational as well as technical levels.

Current	New	Impact on Architecture
Storing Processed Data	Storing Raw Data (Volume and Retention Period)	
Traditional Legal Framework	New Legal Framework	
Accessibility, Stability / Continuity	Accessibility, Stability / Continuity	
Digitized	Digital by nature	Most NSI's these days rely on digital data for their internal (automated) processing. Any data not in digital form must be digitized as part of the Ingestion phase. Most current CAPI, CATI and CAWI data collection practices includes this. Due to the nature of the new data sources, the data from these sources is already digital from the outset, eliminating the need for digitizing.
Human sources (mediated)	Machine Generated (Robotic sources)	There may be a need to understand the logic applied by the intelligent end-points and intermediate devices upstream as part of understanding the meaning an value of the data provided. Processes that provide Machine Generated data will sometimes need training data in order to build the model / logic for processing the data. This process needs to be understood and reviewed periodically to arrange for the model to be re-trained where necessary.
Often No Partnership needed	Adaptive Design Partnership	The only partnership involved in primary data collection is the (willing or forced) collaboration of the interviewee (the ultimate data provider). For new data sources, as explained before, closer collaboration is needed in many cases, extending into the preparatory stages of design and build. Fortunately, these data providers provide data

Current	New	Impact on Architecture
		for more or all units, which means much less providers involved per survey.

195. Dimensions that can explain the differences:

- Vs: Volume, Velocity, Variety/Variability, Volatility
- External non-government (private) origin
- Machine-generated Internet of Things
- No possibility to influence data-sources: sometimes unknown structure
- RealTime/Streaming phenomena
- Quality (Veracity) aspects
- Local/global sources, supranational data
- Legal framework? Availability/accessibility? Ethical aspects
- Need partnership with external users to validate/understand/give meaning to the data

XIV. Annex: Gartner DM model vs CSDA

196. The latest Gartner Information Capabilities Framework (ICF) contains a number of capabilities that are important for new or challenging use cases, such as for example new (big) data sources, data sharing via data hubs and wider data ecosystems, IoT/streaming mode etc. The table below shows that CSDA also covers those capabilities, albeit in a slightly different grouping, under different names.

Table 4: CSDA vs Gartner ICF

Gartner name	Gartner description	CSDA
Describe	Collect knowledge about data assets including where they are, what format they are in, what level of quality they represent and their potential value to the enterprise.	Source Discovery & Profiling
Organize	Align and structure data assets so that they can be readily found and easily consumed by other use cases. Decide if data should be structured in a way that conforms to the organization's standards of syntax (format), semantics (meaning) and terminology (use of common terms), or whether the use case allows for local standards. Opting to organize data locally may affect the ability to integrate with other sources or support other use cases.	Data Description & Organization
Integrate	Support accessing and ingesting diverse data types, performing transformations (changing formats and semantics, or combining data, for example) and allow independently designed data structures to be used together toward a common objective.	Information Logistics
Share	Make data available to consumption points. This can mean a single use case or a variety of use cases depending on the trade-offs made for organizing and integrating data.	Information Logistics
Govern	Provide for risk assessment, control and compliance as it relates to data quality, security, privacy and retention. Data governance will need to take a trust-based approach that is no longer a one-size-fits-all, top-down approach, but an approach that adapts to the situation and the level of central governance required.	Information Governance, Security & Information Assurance
Implement	Support the deployment and execution of the other five capability types. The decision of collecting versus connecting to data only needs to be resolved at implementation. Changes in implementation can also occur over time as the level of usage (or the use case) evolves.	GAMSO: Capability Development

197. Notes:

1. **Integrate:** Mapping Gartner's Integrate to CSDA's Information Logistics is based on the assumption that Gartner puts emphasis on "support for" performing transformation and integration, much like CSDA has separate capabilities for the actual Data Transformation and Data Integration.
2. **Implement**: Just like CSDA, Gartner recognizes two modes of operation: COLLECT and CONNECT, where COLLECT represents the (more traditional) way of ingesting data into the own systems before processing (centralized data and processing), whereas CONNECT represents the way where processing happens by directly connecting to the (external) data sources (distributed data and processing). CSDA supports the view that decisions about "collect" vs "connect" should be made on a case-by-case basis, preferably during Source Discovery & Profiling. Implementing such decisions is "business as usual" and should be part of Execution ("Production" in GAMSO terminology). Information Logistics therefore should include the ability to implement (configure) solutions for both modes of operation.

XV. Annex: Examples of Exchange Channels

198. In many organizations, channels are implemented as generic capabilities, like CATI, CAPI or CAWI, that can be configured or organized to execute specific surveys. Statistical organizations are known to organize themselves in such a way that Data Collection is a separate organizational unit, of which the channels are a part. Often, they provide a business service to other parts of the organization.

199. Exchange channels deliver data (datasets) and metadata, mostly process meta and para-data. Exchange channels are configured through metadata (designs), often embedded in instruments, tools and questionnaires.

200. The following examples will clarify some of this. For clarity, we'll start with some examples that should be familiar, examples taken from traditional Data Collection.

A. Example 1: The CAPI (person-to-person interviewing) channel.

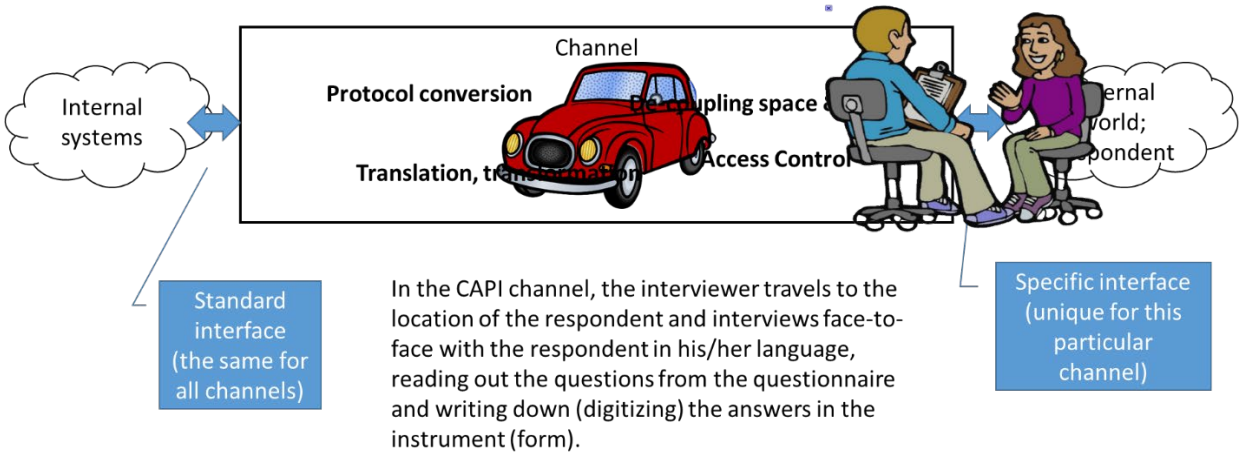

Figure 19; CAPI channel

201. In this example, the interface presented on the outside, is the oral and body language interaction between interviewer and respondent. The interviewer may have to translate the questions and/or the answers, but as a minimum, he will have to digitize the answers, using some means (such as a laptop computer) that is also part of the channel, just as the interviewer himself is part of the channel.

B. Example 2: The CATI (telephone interviewing) channel.

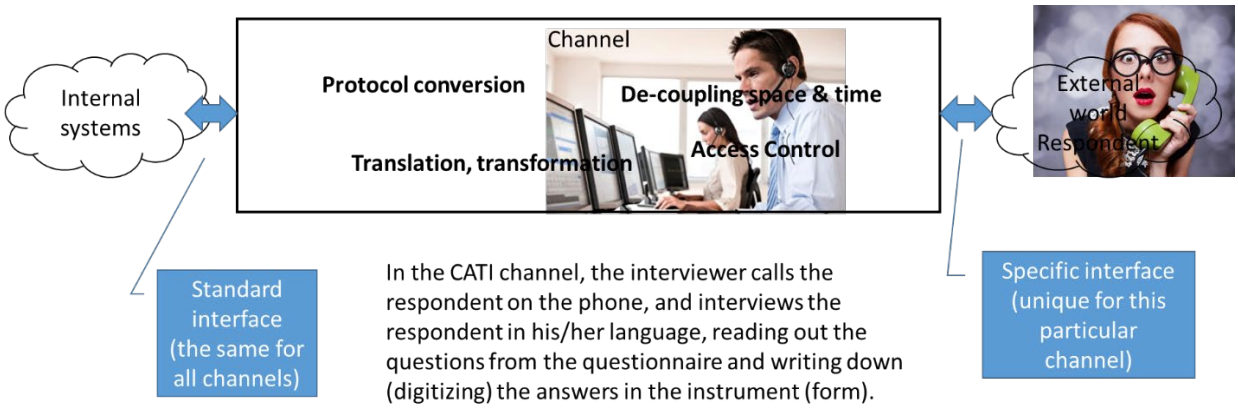

Figure 20: CATI channel

202. The CATI channel presents a different interface on the outside, namely the sound waves the are transmitted from/to the telephone. Note, that in this case the channel uses equipment (the telephone network and the respondent's extension) that is not owned by the statistical organisation. Again, the interviewer will have the task to translate (digitize) the spoken word into digital information.

C. Example 3: The CAWI (web form self-interviewing) channel.

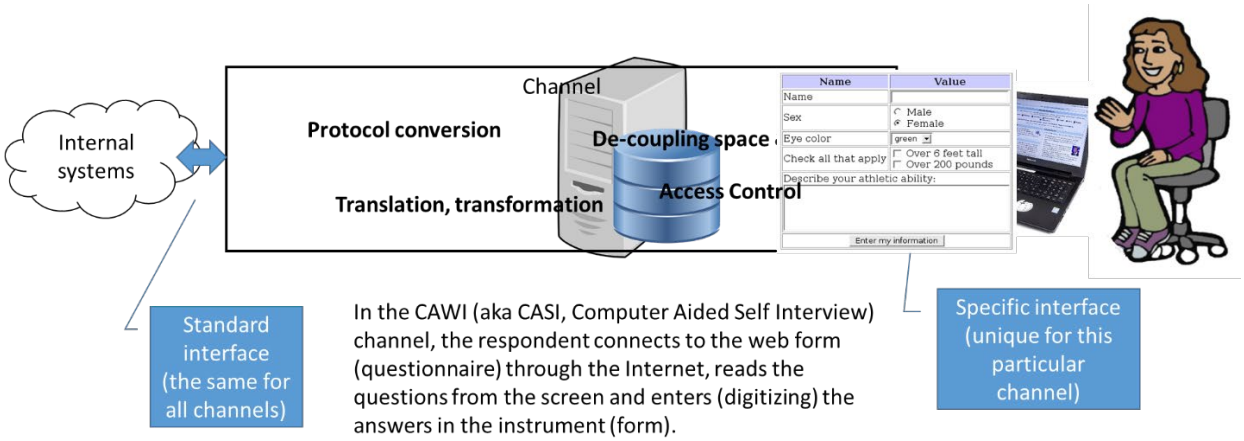

Figure 21: CAWI channel

203. In this example, again the channel uses equipment that is not owned by the statistical organisation: the internet and the respondent's computer. The interface presented to the respondent (the information provider) is the image (the questionnaire) on his/her computer, and the mouse and keyboard for entering the answers. The act of digitizing is "outsourced" to the respondent.

D. Example 4: A web scraping channel

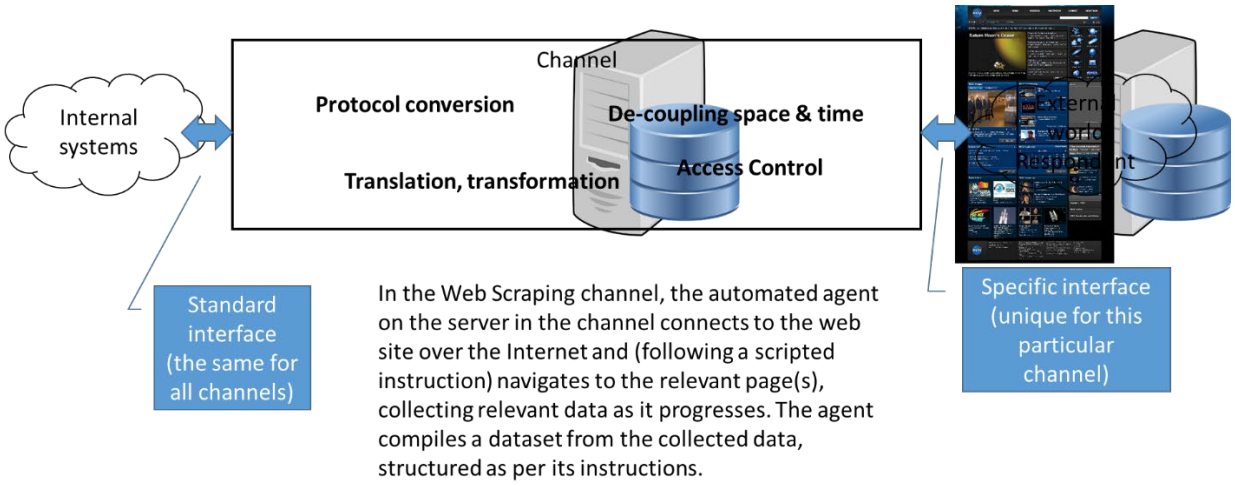

Figure 22: Web scraping channel

204. In web scraping, the channel uses the Internet to connect to some relevant web site and uses an internal script to navigate that web site and collect from it the relevant information. The script, executed by some automated agent, is the equivalent of the human interviewer, the external web site is the representation of the data provider (the respondent). Again, the channel uses some technology (such as the Internet) that is not owned by the statistical organisation.

E. Example 5: connecting to a public administrative data source

205. This example is very similar to the previous one in terms of technology used. The relevant difference is, that the provider is not just providing information about a single entity (unit), but about a large collection of units, such as a whole population. And of course, there may be technological differences, too, such as directly connecting to a database or using some web service provided by the data provider.