Conceptual Architecture of the

Statistical Information System

at the Swiss Federal Statistical Office

Version history:

Version	Date	Comment	Responsible
01	July 5, 2006	Initial version	D. Tombros
02	July 10, 2006		D. Tombros
03	July 12, 2006	New structure as discussed in today's workshop	D. Tombros
04	July 17, 2006	New structure, chapters 4.1.2, 4.1.3, 4.2. 4.3 added	B. Loison
05	July 18, 2006	Chapters 4.1.1, 4.1.5, 4.4 reviewed, Data Quality Management added	D. Tombros F. Maurer
06	July 26, 2006	Review of chapters 4.1, 4.2, and 4.4	D. Tombros F. Maurer
06.5	July 31, 2006	Review of all chapters	B. Loison
07	August 18, 2006	Added 3.4, modified 4.1.1, 4.1.3, 4.4.1, 4.4.2, 4.4.9, small form corrections in all chapters	D. Tombros F. Maurer W. Bausch
08	August 30, 2006	Review of chapter 3.4, 4.3.4, 4.4; Added chapter 4.4.1; small form corrections in all chapters	W. Bausch F. Maurer
09	November 21, 2006	Changing "Data Storage Area" to "Data Working Area" Review of chapter 4.1.2, 4.4.5, 4.4.6.1, 4.4.9.3	W. Bausch F. Maurer
091	February 16, 2007	Technical topics moved to Appendix Additions to application type descriptions in Chapter 4.3 Corrections in all chapters Management summary	D. Tombros
092	March 5, 2007	Updated overview graph	D. Tombros
1	March 5, 2007	Review of all chapters	B. Loison

Contents

1. Management Summary

This document defines an overall conceptual architecture for the Statistical Information System (SIS) of the Swiss Federal Statistical Office (SFSO). In its final form the SIS will support all process groups in the SFSO Statistical Value Chain (SVC). The SVC comprises data collection processes, data transformation processes, data analysis and interpretation processes, the processing of statistical information, and the distribution of statistical products. It describes the SIS vision and as such it can be positioned beyond past and ongoing studies, projects, or existing systems like BS-SYSTEM, CODAM, or G-SOA@BFS.

The SIS conceptual architecture considers Swiss federal and international standards and is based on the architecture model proposed by the ISB. The conceptual architecture is divided into four layers of abstraction: the business layer which covers business processes, the application layer which covers the application systems which implement the business processes, the data layer which covers the data used by the application layer, and the technical layer which covers the software systems and technical infrastructure required for the operation of the SIS.

While all abstraction layers are briefly described, the focus of this document is on the application and data layers. Specifically its purpose is to provide certain high-level principles and guidelines for these layers to which the architecture of the SIS should adhere. Furthermore it defines certain basic concepts and gives a generic classification of various elements of the SIS including application types, data types as well as functional components of the data layer. The technical abstraction layer is mostly beyond the scope of this document and is described in the documentation of specific projects and systems. Certain technical information on aspects such as data mastership and data quality management is included in the Appendix.

2. Scope

The aim of this document is to define a conceptual architecture for the Statistical Information System (SIS) of the Swiss Federal Statistical Office (SFSO) in compliance to the IT strategy adopted on January 1, 2004. Based on the Statistical Information Value Chain, the whole process is considered starting by the collection of the data from the different suppliers and ending by the distribution of the statistical products to the customers.

This document is based on the architecture model proposed by the federal information strategy organization (ISB) in [18], covering the business, application and data layers. It can be positioned beyond past and ongoing studies, projects, or existing systems like BS-SYSTEM, CODAM, or G-SOA@BFS. It is the aim of this document to give an integrated view of the SIS at SFSO to be prepared for the challenges that SFSO will be faced within the next years. SFSO maintains for example a statistics portal[1] [19] that builds an information platform for all official statistics producers. The aim of this platform is to provide all statistical information independent from the producer at a central place.

The top-down and integrated approach applied here enables to plan and implement processes, responsibilities and competencies to build the described architecture.

Once the concept described in this document has been accepted, the details and technical implementation of the architecture of the business, application and data layers, will be described in detail.

3. Standards and Reference Architectures

3.1. International Organizations

Several international organizations develop policies and publish guidelines and recommendations to help national statistical offices to implement statistical information systems. This facilitates the exchange of data and metadata with other countries and international organizations. Eurostat, OECD, UNECE, and the World Bank are such key players. In Table 1, some important standards of international organizations relevant to this document are listed.

[1] http://www.statistik.admin.ch

STANDARDS OF INTERNATIONAL ORGANIZATIONS	
1)	Information Systems Architecture for national and international Statistical Office – Guidelines and Recommendations, United Nation Statistical Commission and Economic Commission for Europe, Conference of European Statisticians – No 51, 1999.
2)	Dublin Core Metadata Initiative (DCMI) is a well-known standard providing a simple but effective set of elements (consisting of 15 unstructured elements) for describing a wide range of data. (http://dublincore.org/)
3)	Data Documentation Initiative (DDI) is an international effort to establish a standard for technical documentation describing social science data. A membership-based Alliance is developing the DDI specification, which is written in XML (http://www.icpsr.umich.edu/DDI/).
4)	Statistical Data and Metadata Exchange (SDMX) is an initiative sponsored by BIS, ECB, Eurostat, IMF, OECD, UN and the World Bank to foster standards for the exchange of statistical information (http://www.sdmx.org).
5)	ISO/IEC 11179 describes the standardizing and registering of data elements to make data understandable and shareable. Data element standardization and registration as described in ISO/IEC 11179 allow the creation of a shared data environment in much less time and with much less effort than it takes for conventional data management methodologies (http://www.iso.org).

Table 1 Standards of international organizations related to statistical data, metadata and the architecture of statistical information systems

3.2. Swiss Federal Government

The Swiss Federal Strategy Unit for Information Technology (FSUIT) is responsible to enact and maintain policies, guidelines and standards for the Swiss Federal Administration. The most significant relevant policies, guidelines, and standards are listed in Table 2.

STANDARDS OF SWISS FEDERAL GOVERNMENT	
6)	Information Technology Policy of the Swiss Federal Administration, 2000
7)	R001 - Referenzmodell für die Informatikarchitektur Bund (RIAB), Version 1.3, ISB, 2003.
8)	R009 – Architekturvorgaben Bund, Version 3.0.001, ISB, 2006
9)	KOGIS, GM03 – Metadatenmodell, FD (Final Draft) Version 1.4, Juni 2004

Table 2 Related standards of the Swiss Federal Government

3.3. Swiss Federal Statistical Office

The Swiss Federal Statistical Office (SFSO) is responsible for designing, developing and implementing a Statistical Information System (SIS) that enables to satisfy the requirements of national data suppliers and customers but also to exchange data and metadata with other countries and international organizations. To follow this goal the SFSO enacts architectural guidelines and concepts listed in Table 3.

STANDARDS OF SWISS FEDERAL STATISTICAL OFFICE	
10)	System Architektur „CODAM“, BFS, R. Liviero, 2004
11)	Strategische Informatikplanung, BFS, 2004
12)	Architekturstudie „90-Grad“: Generische Architektur für Erhebung, Erfassung und Verarbeitung, B. Loison , A. Marzetta, M. Schröder, M. Moret, M. Körsgen, 2005.
13)	Systemziele G-SOA@BFS, B. Loison, V. Scholl, 2006
14)	Systemanforderungen G-SOA@BFS, B. Loison, V. Scholl, 2006
15)	StatiX „XML-Konzept für das BFS“, B. Loison, D. Profos, 2006

Table 3 Related standards of the Swiss Federal Statistical Office

3.4. Conceptual Architecture Framework

The present document defines a conceptual architecture framework for the SIS based on the architecture model proposed by the ISB (see [18]). Accordingly, architecture definition and evolution is not an isolated activity and has to be realized considering operational processes and the business environment. Data, business objectives, organization, technology, financial resources and last but not least people play an important role in this context. Therefore, a wide and integrating view is required that defines basic constraints for the enterprise and IT landscape.

To avoid a fragmented and heterogeneous IT landscape with redundant and monolithic systems and applications, it is important that the architecture defines all aspects of the framework as shown in Figure 1. It thus has to consider the strategic organization targets, must provide adequate guidelines for different implementation projects, and take into consideration the technical and business requirements as well as technical feasibility and other constraints.

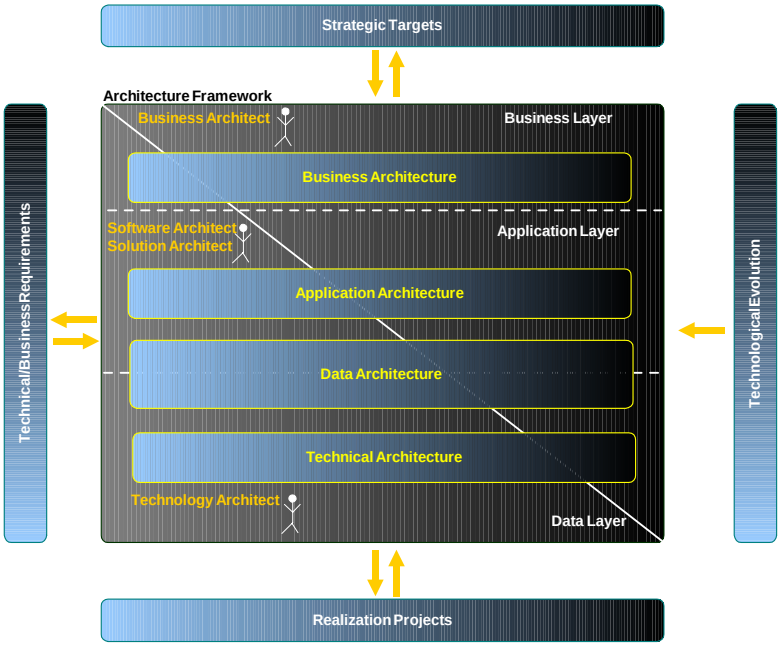


Figure 1 A conceptual architecture framework for the SFSO Statistical Information System

The conceptual architecture framework itself contains four layers of abstraction that must be addressed:

- Business architecture layer:
  The business architecture layer contains an overview of all business activities and their objectives including the relations between each other from a business point of view. It defines the necessary roles and organizational units that are involved in these activities. The business architecture has to ensure that process-related, organizational and strategic decisions are based on current visions and business objectives. In this document, business architecture is assumed as predetermined.
- Application architecture layer:
  The application architecture consists of a set of applications and their interactions. The tasks of different business processes are mapped to several functions provided by the applications. The main focus of the application architecture lies on understanding all the functions and their interrelations that help to construct and maintain the architecture. Beside any business constraints (like defined roles or locations) also technical constraints like standards and trends (XML, Web Services) should be considered. The application architecture aspect is in scope of this document.
- Data architecture layer:
  The data architecture is driven by the business objectives of the business architecture. It defines what data is needed to meet business user needs and how this data is conceptually structured. It examines the completeness and correctness of source systems that are needed to obtain data and identifies the data facts and dimensions. It also defines data models, establishes a generic aggregation plan and to set up the metadata infrastructure. The data architecture defines the applicable standards concerning the data management, distribution and access. The data architecture aspect is in scope of this document.
- Technical architecture layer:
  Here, focus is on the description of required infrastructure elements like hardware and software and their relationships, as well as possibilities for standardization and sharing between different systems. Moreover, the technical architecture defines technologies that support applications and data management, services and protocols, as well as development methods and tools. The technical architecture must take into consideration the architecture aspects mentioned above, implying that these aspects need to be defined beforehand. This is why technical architecture is not in scope of this document.

In Figure 1, these four aspects are separated in three layers: business, application and data layer (see also chapter 3.4 of [18]) . Additionally, responsibilities and roles for the architecture management are shown. The business architect is responsible for the business architecture and is involved to a certain degree in application, data and technical architecture work. The software or solution architect has to consider on one hand the application architecture and its application landscape, on the other hand he has also a data view considering the data architecture from the point of view of the applications. The technology architect is responsible for the technical architecture and the data architecture from a technical point of view.

In what follows, these architectures will be further specified, assuming that a bus infrastructure will be used. Also, guidelines for the application-, data-, and technical architecture will be defined.

In general, an architecture framework is a complex system whose components evolve frequently and independently of each other. For example, some data sources may disappear while others will be added. The enterprise data model can evolve with the enterprise objectives and strategies. The technical environment changes with product evolution and updates. Design choices at the implementation level can also evolve to follow changing user and administration requirements. Therefore, the maintenance of a conceptual architecture of the SIS is a permanent process. In Table 4, the relevant aspects that need to be considered and maintained in this are shown. It is evident, that most of these aspects are closely dependent on each other.

EVOLUTION ASPECTS	
SYSTEM-TECHNICAL ASPECTS	Allocation and maintenance of the hardware and software infrastructure. Guarantee a transparency related to both the available system heterogeneity and the inherent dataflow complexity for the interaction between the different components.
PERFORMANCE MANAGEMENT	Controls whether the system technical architecture fulfills the performance requirements and optimizes it, if required. E.g., aggregation queries require a permanent adaptation of the system support (index structures, partitioning, or sum tables) and raises the question of optimal load balances.
QUALITY SURVEILLANCE AND ASSURANCE	Controls and manages the overall service quality of the architecture framework, defines actions for the input and output of the data flow and it addresses the aspects of system oriented monitoring (→ quality of operations) as well as data oriented monitoring (→ quality of data content).
CAPACITY PLANNING	Capacity planning is a continuous process that adapts the current capacity situation according the current requirements and known bottlenecks (considering hard- and software components as well as manpower).
USER SUPPORT	The goal of IT architecture is to have satisfied users and customers, respectively. Therefore this aspect is one of the most significant and tries to adapt to customer requirements on different levels (e.g., from improvements of the GUI down to changes in the data schemas).
PROTECTION- AND SECURITY MANAGEMENT	In general, an overall protection- and security management needs to be established that prevents the abuse of the architecture (authorization and authentication model for the users) and that establish a concept to minimize the negative impacts of a system failure.
EVOLUTION CONTROL	The purpose of the evolution control is to design the changes of the permanent evolution process of an architecture in that way that the operation of the systems is not disturbed (or at least kept minimal) and to guarantee that the new changes follow the new guidelines.
STRATEGY AND PLATFORM	This aspect contains all tasks that are related to the project planning of subsequent systems. It can be further divided into aspects that are related to technical issues and aspects related to data issues. Its goal is a continuous improvement of the whole Architecture framework.

Table 4 Overview of aspects that need to be considered when planning and managing the evolution of IT architecture

A way to control architecture evolution is to define and maintain complementary metadata, enabling to track change history and providing the basis for active enforcement of consistency rules.

4. SIS Conceptual Architecture

4.1. Overview

An overview of the SIS Conceptual Architecture is shown in Figure 2. The architecture consists of three layers: “Statistical Information Value Chain” (corresponding to the “Business Layer” from the previous section), “Application Layer”, and “Data Layer” that are bounded by a “Suppliers” and “Customers” area. In this section, there is a brief description of these layers.

4.1.1. Suppliers

Data suppliers are the sources and input channels for data and in some cases metadata in the SIS [8]. It can be seen from Figure 2 that the suppliers feed the application layer via different channels available at the portal representing the different possibilities to collect data. A short description of the different channels as defined in [8] is outlined in Table 5. It must be differentiated between different technologies to collect data by the suppliers and technologies used to import the data into the SIS architecture.

SIS CHANNELS	
TECHNOLOGIES TO COLLECT DATA BY THE SUPPLIERS	
ELECTRONIC	There exist several electronic technologies to collect data: • CATI: Computer Assisted Telephone Interviewing • CASI: Computer Aided Self-Aided Completion Interviewing • CAPI: Computer Assisted Personal Interviewing • CAWI: Computer Aided Web Interviewing • WATI: Web Assisted Telephone Interviewing • EMS: Electronic Mail Survey • External Databases: Data exchange with SFSO-external and SFSO-internal registers and databases.
PAPER	Scanning and OCR technologies are used to converst paper questionnaires into XML files.
IMPORT INTO THE SIS (SFSO) FROM COLLECTED DATA	
FILE[2]	Data may be imported in the SIS architecture only via files (e.g., using XML or CSV format). The files can be transmitted by FTP, SFTP, Web Services or by using file transfer with Online Services Computer Interface (OSCI).

Table 5 Channels for data input into the SIS

It should be noted that for each statistical data collection provided by a supplier, a responsible contact person on the supplier side is defined independently of the channel used. Additionally, for each supplied (vital) data collection the appointment of data management-specific roles is required (e.g. data steward, data owner). This way, responsibilities for data circulating inside SFSO are clearly defined, enabling the implementation of data management activities.

[2] A collection of data or information that has a name, called the *filename*.

4.1.2. Statistical Information Value Chain

The Statistical Information Value Chain provides a high level view of the purpose of the SIS which is to collect data from several data suppliers, to analyze and process these data and finally to distribute statistical products to different customers. The following business process groups are distinguished as described in Table 6.

BUSINESS PROCESS GROUPS	
DATA COLLECTION	Collect the data from the different data suppliers via different portals
TRANSFORMATION	The data needs to be prepared for the use in the analysis step. This includes amongst others a validation of the data. Validation rules may be installed and applied automatically within the transformation functionality of the different areas of the Data Layer (chapter 4.4). Manual validation of the data is done directly in the corresponding applications of the Application Layer (chapter 4.3), or as part of the process of creating a statistic.
ANALYSIS / INTERPRETATION	Analyze and interpret the data
INFORMATION PROCESSING	Process the information from the analysis step, create the products
DISTRIBUTION	Distribute the products via different channels to the customers

Table 6 High-level business process groups of the SFSO supported by the SIS

4.1.3. Application Layer

The architecture of the Application Layer is based on the reference architecture [1] and includes the component types listed in Table 7. It provides the required applications and tools to support the Statistical Information Value Chain layer (functionalities).

APPLICATION LAYER	
COMPOSITE APPLICATIONS	An application built by combining multiple services. A composite application consists of functionality drawn from several different sources within a service-oriented architecture (SOA). The components may be individual web services, selected functions from other applications, or entire systems those outputs have been packaged as web services (often legacy systems).
MONOLITHIC APPLICATIONS	In such applications, there isn't any clear separation across functionality boundaries (modules), implying that functionality cannot be easily encapsulated for reuse. At this time, most statistical applications are monolithic applications.
COMMERCIAL APPLICATIONS	A commercial off-the-shelf product is one that is used "as-it-is." Those products are designed to be easily installed and to interoperate with existing system components. Almost all software bought by the average computer user fits into the commercial off-the-shelf product category: operating systems, office product suites, word processing, e-mail programs, SAS, and Superstar are among the myriad of examples.
ENTERPRISE **S**ERVICES **A**RCHITECTURE **P**LATFORM	Enterprise Services Architecture is the lean manufacturing model applied to IT architecture. An Enterprise Services Architecture platform assures the collaboration and the process management of the three different types of applications as described above and manages the access to the Data Layer.

Table 7 The different types of applications encountered in the SIS application layer are distinguished based on their integration characteristics. The ESAP provides common services.

4.1.4. Data Layer

The architecture of the Data Layer is based on the reference architecture [1] and is divided in the areas listed in Table 8. Its purpose is to provide the functionality to support the Application and Statistical Information Value Chain. This layer provides a service interface. The different areas and components listed below will be explained in more detail in section 4.4.

AREAS	
DATA COLLECTION AND STAGING AREA	The Data Collection and Staging Area collects data from the data sources and it extracts, transforms, and loads the collected data to the Data Working Area.
DATA WORKING AREA	The Data Working Area contains all collected data from the Data Collection and Staging Area in a consolidated form for creation of statistical products and loads the data into the Data Analysis Area.
DATA ANALYSIS AREA	The Data Analysis Area hosts the data loaded by the Data Working Area optimized for analysis operations. From there, the data are loaded into the Data Dissemination Area.
DATA DISSEMINATION AREA	Based on the loaded data from the Analysis Area, the Data Dissemination Area provides the required functionality to provide the end products to customers.
DATA MANAGEMENT AREA	Controls all processes concerned with the data processing workflow in the Data Layer.
METADATA AREA	Manages and stores metadata about the hosted data in the Data Layer.

Table 8 Elements of the SIS Data Layer

4.1.5. Customers

Customers are SFSO-external users that are interested in statistical information. All information that is distributed to the customers is offered as different products. To provide the appropriate products for the different customer needs, there are different product profiles and several products that can be combined in product portfolios.

The product profiles are oriented towards to the different customer groups. Customer groups are classified as shown in Table 9. Note that there are also products that are relevant for several or for all customer groups like information on current events.

CUSTOMER GROUPS	
OBSERVERS (“BEACHTER”)	Interested customers, which are occasionally looking for information for private or professional purposes. Occasional usage of information by the general public.
USERS (“BENUTZER”)	Customers, which require information for professional purposes (often decision-makers). Goal-oriented usage by politics or industry representatives.
PROCESSORS (“BEARBEITER”)	Experts that regularly and professionally search, collect, and prepare data for their business needs. Subsequent processing usage of information by specialists and researchers.

Table 9 SIS Customer Groups

The products of the SIS can be distributed via different media to the customers. The available media are shown in Table 10.

DISTRIBUTION MEDIA	
ONLINE	Electronic reports like public web reports in the Internet, web reports for members, Extranet, or e-mail
OFFLINE	Electronic reports distributed via CD-ROM, DVD, or other data media.
PRINT	Printed products like brochures, reports, or books.

Table 10 – Distribution media to the customers

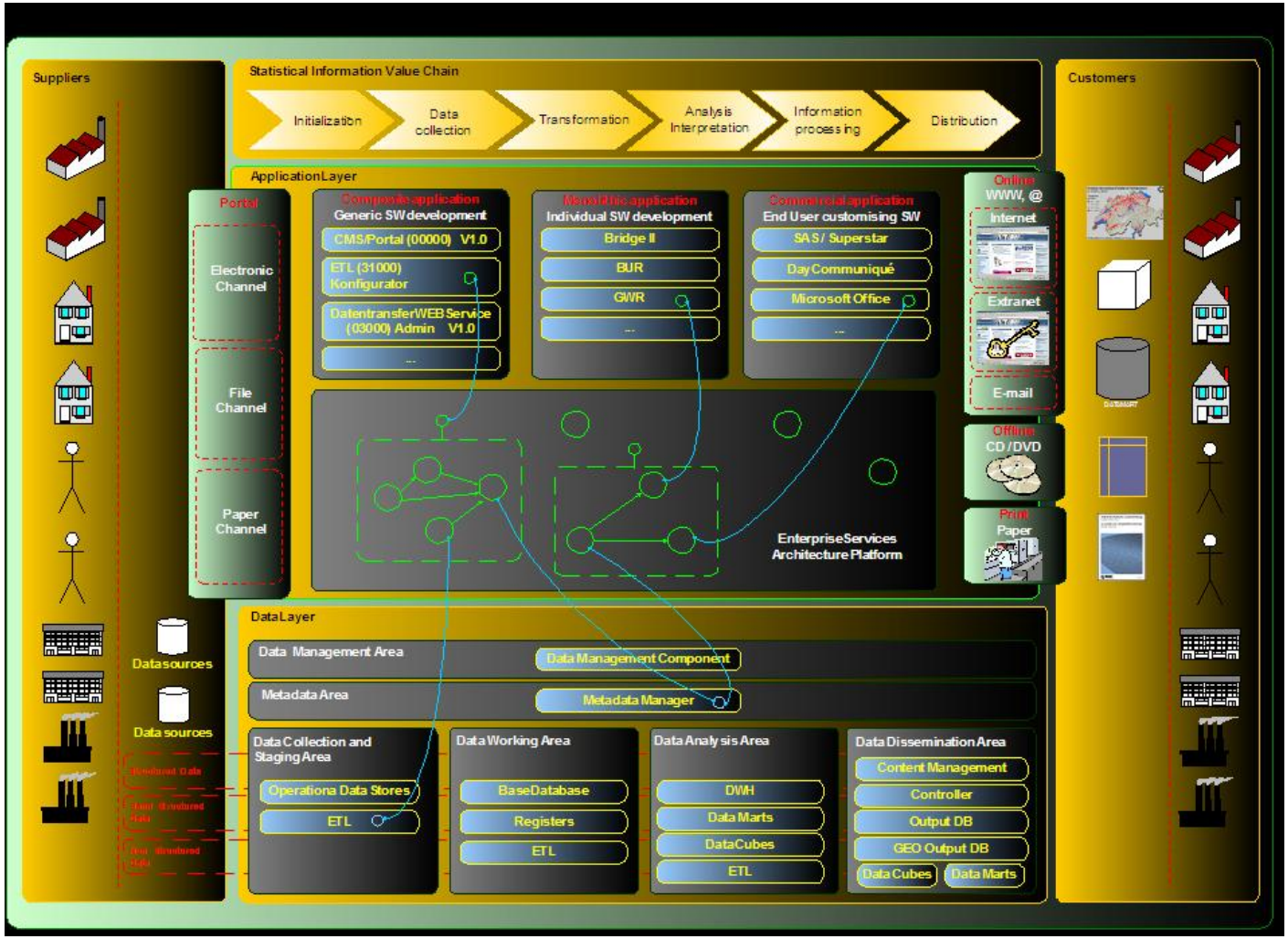


Figure 2 – Overview of the SIS conceptual architecture with examples *of applications implemented with services*

4.2. Statistical Information Value Chain

The *Statistical Information Value Chain* (SIVC) provides a complete description of all primary and support activities that the SFSO performs to turn inputs into value-added outputs for its external customers. The *Statistical Information System Value Chain* (SISVC) can be considered to be the IT-specific subset of these activities. The SISVC describes an operational IT infrastructure aimed at both directly adding value for external customers and indirectly adding value by supporting other enterprise operations. Ideally, an optimal alignment between the SIVC and SISVC should be maintained. The SIVC of the SFSO is given in Figure 2. The shown business process groups correspond to the defined steps of the core processes at SFSO as described in [17] and [16]. The responsibility for the maintenance of the value chain lies within the section P+P (Processes and Products) of the SFSO.

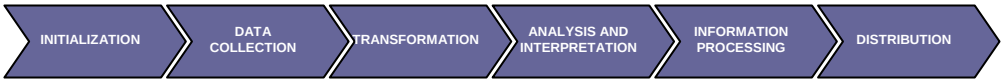


Figure 3 – Statistical Information Value Chain (business process groups)

The most important issue of the process modeling is to distinguish the difference between the value chain of the SFSO and the processes that support the creation of this value. The SFSO modeled processes will be implemented in the Enterprise Services Architectures Platform (see section 4.3.4). This implementation will ensure that the SIS supports all statistical processes mentioned in section 4.4.

All processes that support the creation of the values of the Statistical Information Value Chain must be implemented in a reliable Enterprise Services Architecture Platform to ensure that the business processes will be really applied at the SFSO. Also, the notion of data quality and the ability to trace data quality are intrinsically linked to the definition of these business processes.

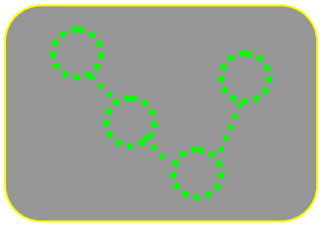
4.3. Application Layer

The Application Layer contains the required applications and tools to implement the Statistical Information Value Chain. The Application Layer provides the interface to internal users (internal use of the SIS) and to external users (data suppliers and customers). It is important to note that internal users, suppliers, and customers will not be authorized to access directly the data layer. The access to the data layer will be allowed only through the application layer. In other words, the application layer manages by whom, when, and how users access to the data of the SFSO.

4.3.1. Composite Applications

The first type of applications in this layer is called composite applications. The most important aspects of composite applications are the new kind of automation that can be enabled and made affordable (e.g., orchestration). Because a composite application is a composition of existing applications that have been split into components, it can now

ignore the boundaries between these underlying applications (e.g. used by ETL tools) and components (reusability of software components). This cross-functional automation of processes spans easily application boundaries. Composite applications consist of orchestrated services (as depicted below).



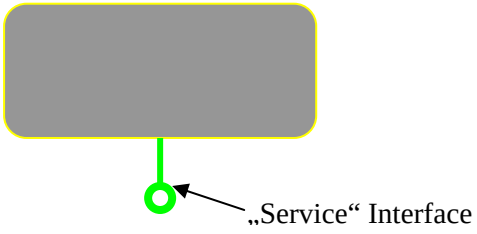
One of the goals of the project G-SOA@BFS is to develop and implement composite applications based on a Service Oriented Architecture (SOA) at the SFSO. The SOA defines an application architecture in which all functions or services are defined using a description language and they have interfaces that can be called to perform business processes (see section 4.1.2). In a SOA each interaction is independent of every other interaction and the interconnection protocols of the communication system do not affect service interfaces. Because interfaces are platform-independent, a client from any device using any operating system in any language can use the service. Note that SOA is not a technology but a kind of IT architecture (similar to client/server architectures) and can use different technologies like Web services, SOAP, or XML.

A Service Oriented Architecture (SOA) enables to develop customized application based on reusable software components.

Examples: eSurvey and SFTP are some functional composite applications used at SFSO. The project G-SOA@BFS is propagating the use of this philosophy in software development.

4.3.2. Monolithic Applications

The second type of applications built on this layer is called monolithic applications. In other words, there exists no clear separation across the functionality boundaries (modules) that comprise this type of application. The application provides a single “interface” and its functionality can only be used as a whole block (see below).



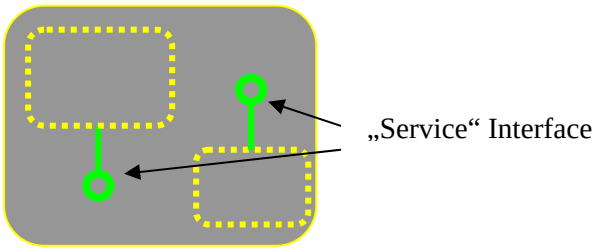
At this time, all statistical applications from the SFSO are monolithic applications. One of the goals of the project G-SOA@BFS is to reduce the number of monolithic applications by using composite applications. Because it is not possible and also too risky to eliminate all monolithic applications in one step, this replacement will take place within the next five years. Note that it may not be possible to eliminate all of the monolithic applications. Some applications are too specialized to be modularized to support reuse (e.g. register based applications). In that case, the aim is to encapsulate

such monolithic applications by a service interface before integrating them into the enterprise services platform.

Examples of monolithic applications are PAUL, PRESTA 1, or BIS.

4.3.3. Commercial Applications

The third type of applications built on this layer are commercial applications. A commercial off-the-shelf product is one that is used "as-it-is." Those products are designed to be easily installed and to interoperate with existing system components. As such they may provide “open” interfaces which be used in a modular way (see below).



An important difference with the current situation will be that such commercial applications will interact in the future only by the use of services. A further difference will be that the access of these applications to the Data Layer will only be possible by the use of predefined data layer services. Finally, the interfaces of these applications are beyond the control of the SFSO so that appropriate encapsulation mechanisms must be used for their integration in a service-oriented architecture.

Examples of commercial applications used at SFSO are SAS, Superstar, SQL Navigator, or Microsoft Office.

4.3.4. Enterprise Services Architecture Platform

An Enterprise Services Architecture Platform (ESAP) is designed to assure the collaboration and the orchestration (process management) of the three different types of applications as described in the previous sections and to manage and to track the access to the Data Layer (see section 4.4). The main components of an ESAP are described in Table 11.

FUNCIONALITY OF THE SIS ESAP		
CUSTOMER/SFSO INTEGRATION	The customer/SFSO integration brings functionality and information to the end users. It is geared towards providing a seamless user experience by hiding the complex and heterogeneous IT infrastructure.	
	MULTI – CHANNEL ACCESS	Multi-channel access delivers pervasive user access to applications or data. End users are able to access information from anywhere using any device (e.g. desktop, laptop, handheld, smartphone,…).
	PORTAL	A portal delivers personalized information to the end user in a unified view (CI/CD).
INFORMATION INTEGRATION	Information Integration makes both structured and unstructured information available in a consistent and accessible manner. The integrity of that information will have to be guaranteed.	
	DATA ANALYSIS	Data analysis enables to integrate, analyze, and disseminate relevant and timely information. It can be

		delivered through Business Intelligence Tools (e.g. SAS, Superstar,…) that provide functionality to create and deploy customized, interactive reports and applications, supporting decisions at every level.
	KNOWLEDGE MANAGEMENT	Knowledge management with user-centric services provides a single access point to content management system and third-party repositories (with integrated search, taxonomy, classification, content management, publishing, and related workflow processes). At the moment, this capability is not a primary focus for the SIS architecture.
	INFORMATION LIFECYCLE MANAGEMENT	Information Lifecycle Management promotes information integrity across a business network in a heterogeneous IT environment. It enables the sharing of harmonized data formerly trapped in multiple systems and ensures cross-system data consistency.
PROCESS INTEGRATION	Process Integration enables business processes to run seamlessly across heterogeneous IT landscapes: the business processes that span systems and organizations have to be well-orchestrated and offer high performance.	
	INTEGRATION BROKER	An integration broker for instance enables XML/SOAP-based communication between application components and allows to model software components, interfaces, data mappings, and content-based routing rules.
	BUSINESS PROCESS MANAGEMENT AND AUTOMATION	Business process management and automation enables to model and drive processes in a dynamic IT environment and to combine underlying applications into adaptive, end-to-end processes spanning the entire value chain.
APPLICATION PLATFORM INDEPENDENCE	The application platform provides a complete infrastructure to develop, deploy and run platform-independent, robust and scalable business applications. To allow this flexibility, different technologies have been established (e.g., J2EE).	
	SOLUTION LIFE CYCLE MANAGEMENT	An ESAP must be able to provide core technology for all stages of the software life-cycle, including design, development, deployment, implementation, versioning, testing, operations, and end-of-life phases.
	COMPOSITE APPLICATION FRAMEWORK	A Composite Application Framework provides a robust environment for the design and use of monolithic, commercial and composite applications that comply with Enterprise Services Architecture. A Composite Application Framework comprises design tools, methodologies, services and processes, an abstraction layer for objects, and user interface and process pattern libraries.
DATA SECURITY	Data security is a prerequisite to be able to provide end users with access to data. An ESAP must offer a comprehensive security solution that protects data and ensure the confidentiality of business transactions (internally and externally).	

Table 11 Functional components of an ESAP

One of the goals of the project “G-SOA@BFS” is to implement ESAP with the help of a commercial platform. This will build the “backbone” application for the SIS. The main difference to the current situation is the design of the composite application architecture. The key idea is to implement the process logic of the SIS within the ESAP.

One important goal of this platform is to build a logical layer and therefore to eliminate all the procedures stored currently in the databases (stored procedures) and the hard coded statistical methods (black boxes).

An Enterprise Services Architecture Platform (ESAP) enables to implement at the operational level the software logic of the SIS (rules engine, tracking, data access, business processes, supporting processes, statistical methods, …). This platform will be in charge of the orchestration of the SIS.

Examples: At this time there are four commercial applications in evaluation from SAP, Oracle, IBM, and Microsoft.

4.3.4.1. *Adapter Objects and Services*

The aim of the adapter objects and services is to manage the interfaces between the three types of applications (composite, monolithic, and commercial applications as described in the previous sections) and can be seen as the interfaces between the ESAP and the other components of the Application Layer.

4.3.4.2. *Data and Service Adapters*

The aim of the Data and Services adapters is to manage the access (read/write access) to the objects of the SIS in order to maintain the consistency of all parts of the SIS. Data and Services Adapters can be seen as the interfaces between the ESAP and the Data Layer.

4.4. Data Layer

In Figure 4, a detailed illustration of the Data Layer of the SIS and its components is shown. In the following sections, the different areas and its components will be further explained.

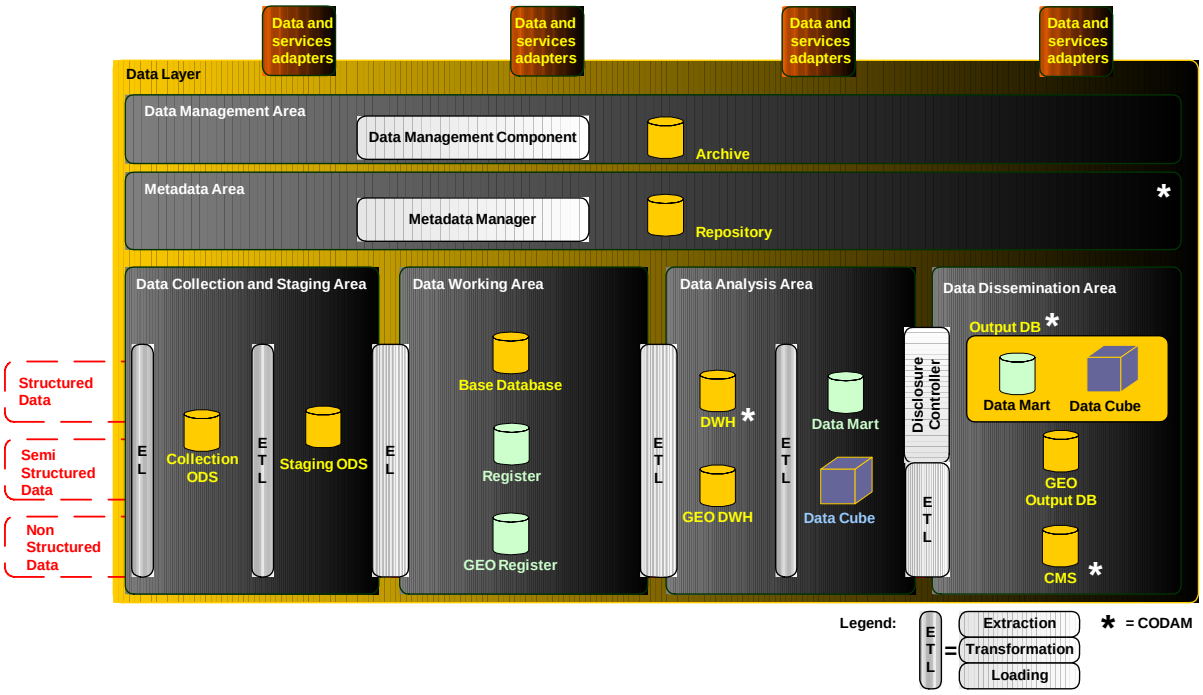


Figure 4 – Data Layer of the SIS Architecture

In Table 12, the main processes of the Data Layer are shown. In addition, there exist processes for the support of the main processes as described in Table 13. A description of these supporting processes is given in section 4.4.10. Note that the supporting processes need to be implemented in all of the four main areas of Figure 4 (Data Collection and Staging Area, Data Working Area, Data Analysis Area, and Data Dissemination Area), because data quality management, data lineage tracking and data security are aspects that have to be treated in all of the areas.

DATA MANAGEMENT CORE PROCESS GROUPS					
		INVOLVED DATA AREAS			
		COLLECTION AND STAGING	WORKING	ANALYSIS	DISSEMINATION
COLLECTING STATISTICAL DATA	Collect data from the different external data suppliers via different portals.	X			
PREPROCESSING STATISTICAL DATA	Prepare data for the analysis step (e.g., validate, anonymize). External data is prepared in the staging area. Internal data may be prepared in the Working Area (registers) and in the Analysis Area (statistical end products), if routing it by the Staging Area is too cumbersome. However, preprocessing capabilities may be limited.	X	(X)	(X)	
STORING STATISTICAL DATA	Store data in the Working Area for statistics in the process of being created. Store statistical end products in the Analysis Area (incl. historical and lineage tracking data). Store statistical end products to be externally disseminated in the Dissemination Area.		X	X	X
RETRIEVING STATISTICAL DATA	Extract data from the Data Working Area or the Analysis Area for the use in the analysis and interpretation step (happens in the application layer).		X	X	
ANALYSING STATISTICAL DATA	Analysis and interpretation of data. For statistics in the process of being created, this mainly involves data in the Working Area. For statistical end-products, data in the Storage Area may be accessed.		X	X	
INFORMATION PROCESSING	Process the information from the analysis and interpretation step and create the different products. Such products may be created in the Working Area before being transferred to the Storage Area, or they may be created in the Storage Area before being transferred to the Dissemination area.		X	X	
DISSEMINATING STATISTICAL DATA	Distribute the products via different channels to the customers.				X

Table 12 Statistical data management processes groups

DATA MANAGEMENT SUPPORT PROCESS GROUPS					
		INVOLVED DATA AREAS			
		COLLECTION AND STAGING	STORAGE	ANALYSIS	DISSEMINATION
DATA QUALITY MANAGEMENT	Enables a sustainable assurance and improvement of the data quality.	X	X	X	X
DATA LINEAGE TRACKING	Enables to trace data items back to the origins of source data items that produced this item.	X	X	X	X
DATA SECURITY	Enables data security assurance concerning confidentiality, integrity, and availability.	X	X	X	X

Table 13 Supporting data management process groups

4.4.1. Roles and Responsibilities

As listed in Table 14, there are three different roles that can be defined within the data responsibilities.

DEFINITION OF ROLES	
DATA OWNER	A data owner is defined as an organizational role for defining relevant properties of a data collection, with respect to data processing, but also with respect to data suppliers and users. Data owners have the authority to create organizational rules and policies for business data. They appoint data stewards in order to implement these rules and policies.
DATA STEWARD	A data steward is defined as organizational role for monitoring the defined properties of a data collection and to react accordingly if discrepancies to these properties are detected. Data stewards account to the data owner concerning the degree of conformance with the relevant properties.
DATA USER	A data user is defined as organizational role that actively maintains data in a collection or reads data from a collection. It has the authority to access, create and modify the contents of a data collection according to the rules set forth by the data owner. Also it is responsible for implementing monitoring mechanisms that support the data steward in assessing data properties.

Table 14 Data responsibility roles

It is important not to confuse data owners with data users. The term of “usage” refers to the right to access, create, and modify data whereas data ownership means the right to allocate these privileges. The field of data ownership covers yet a wider area than just granting and revoking privileges on data access. It includes the responsibility for defining quality, availability, security and access performance levels and for justifying

trade-offs in these areas. Depending on the complexity of this task, the data ownership concept may be implemented as hierarchy, where “top level data owners” are concerned with overall policy establishment, whereas “bottom level data owners” are concerned with implementing these policies in the context of individual data collections. Data ownership should not be confused with the technical term “data mastership”. For an exact description of data mastership see the Appendix.

Data ownership cannot be defined only for externally supplied data only. Statistical products generated as part of the Statistical Value Chain also need to have a defined ownership. For instance, consider the production of a statistical database combining several external registers. The external registers are seen to by a (number of) data owner(s). Before this new statistical database goes into production, a new data owner needs to be appointed. He/she is responsible for providing sound data out of this statistical database. Therefore he/she may have to coordinate with the register data owners in order to establish the required data properties. Using this approach, any data collection created an maintained within the SIS must have a defined data owner. All data users working on this data collection need to abide to the rules and regulations set forth by the data owner. It is possible to relax this requirement by allowing individual users or groups to create data that is not owned by anyone. However, such data must not be incorporated into the value chain and be used by other parties or groups. It should be considered as “private data” used within a project team or organizational unit.

Data Stewards are responsible for monitoring data properties and for enforcing the policies and rules defined by data owners. Together with IT operations and security, a data steward oversees data acquisition, maintenance, interpretation, application of business rules, integrity and security. Note that it is not the responsibility of the data owner or steward to actually implement these rules and policies. This needs to be taken over by SFSO staff using the applications and systems. Accordingly, a data steward must have the authority to initiate activities with involved staff. Thus ownership and stewardship for the data of a given product must be well-aligned with the business that is creating the corresponding products. In analogy to the data owner, data stewards may have to cooperate in order to achieve a certain overall quality goal.

The data user term covers a wide ground. Basically it subsumes everyone that has access to a data collection, but is not bearing the responsibilities of a data owner or steward. Data users may be staff that maintains a register database, but it may also be “read only” users of a statistical db.

4.4.2. Data and Services Adapters

As shown in Figure 2, the Data Layer provides service-oriented interfaces to the Application Layer illustrated by the “Data and services adapters”. It is crucial to closely match the services required by the application layer (high-level, aggregated services) to the services providing access to the data (rather low-level, fine granular) The data layer services should to the service catalogue defined by the GSOA@BFS project in order to efficiently integrate both service abstraction layers.

The data layer services enable the different applications to get data from the different areas of the Data Layer (the problem of data access control is discussed in section 4.4.10.3). It is also important that the processed data of the different applications can be stored back in the Data Layer. For this, different approaches can be chosen:

- All provided service interfaces allow to modify data. This means that applications may load data directly into the Collection and Staging Area, Data Working Area, or Analysis Area. It is important that the service interfaces guarantee that data coming from the Application Layer fulfill appropriate quality standards. If an application directly wants to load data into the Analysis Area, the service interface may need logic that checks this data with respect to the requirements imposed by the Analysis Area. Such an approach has the most degrees of freedom concerning the mode of work of statistical groups. However, it prevents global rules (e.g. pertaining to data quality or formats) to be defined and enforced at a single point in the data layer.
- A more purist approach is to concentrate “modification services” to the Collection and Staging Area. This means that the only way to inject external data or data from the Analysis Areas into the Working Area, where new statistics are created, is via the Collection and Staging Area. This guarantees that all data undergoes the same procedures (e.g. quality checks, anonymization, etc.). While this approach allows more rules to be centrally enforced in the Staging Area, it bears the risk of being overly restrictive for the mode of work of statistical groups, which would need to route all new input data through the Staging Area to be able to use it in the Working Area to create statistics.
- Depending on the type of modifications to support and the desired behavior of the SIS (e.g. when applying modifications), a mix of both approaches may be best. For instance, data heavily shared by different statistics may need to go through the Data Collection and Staging Area to ensure an equal degree of consistency for all users of the data. Data that is rather application specific may be allowed to directly go into the Working Area. Note that such an approach needs appropriate governance to be in place.

4.4.3. ETL functionality

It can be seen from Figure 4, that the areas of the Data Layer are connected via Extraction-, Transformation- and Loading (ETL) or just Extraction-, and Loading (EL) components. Within these areas, EL or ETL functionality may also be used. Whenever data has to be loaded from one data storage to another data storage, ETL functionality is required.

4.4.3.1. Extraction functionality

In general, the extraction component controls the choice of the data sources and the extracts of data that will be imported into the target data storage from the particular sources.

For choosing the right time to import and transfer the data to the next storage, several strategies can be used:

- Periodical extraction, whereas the time period needs to be based on the minimum required actuality.
- Extraction on demand
- Event-driven extraction, e.g., if a certain amount of changes is reached.
- Immediate extraction by detected changes

Note that these strategies may be in conflict with each other. For instance, combining a data source with periodical extraction with an on-demand extracted source may generate undesired effects in subsequent processing. This is why the update strategy choice for each data collection needs to be coordinated to fit the overall picture.

4.4.3.2. Transformation functionality

Before the extracted data can be loaded into the target data store, data transformation may be required for the integration and standardization of the data from heterogeneous sources. It may be required that this transformation is performed manually by qualified personnel that interactively transform the data, or automatically based on information stored in the Metadata Repository.

The following actions can be performed on the extracted data:

- Data validation
- Data quality analysis
- Adjustment of data types
- Conversions of encoding and units of measurements
- Simplification of strings and dates
- Aggregation respectively separation of attribute values
- Anonymization

In addition to the data transformation, further data processing may be performed:

- Plausibility checks
  - Removing of redundancies and duplicate data entries
  - Cleaning up of “null” or missing values
  - Etc.

Note that complex transformations may require application layer components to be involved (e.g. transformations requiring some complex mathematical function to be applied). If just a copy of the data shall be transferred from one data storage to the other, the transformation functionality is generally not needed.

4.4.3.3. Loading functionality

After completing the data transformation, the integrated and cleaned data are ready to be stored in the target data storage. The forwarding of the data into the data storage is performed by the loading component according to pre-defined schedules and rules.

4.4.4. Data Collection and Staging Area

The Data Collection and Staging Area can be considered as a main data injection point for the Data Working Area, where statistics are built. Data externally and – depending on the approach chosen (see 4.4.2) -- internally supplied for building a new statistic *may* be imported into the Data Working Area through the Data Collection and Staging Area.

Data injected can be classified in non-structured, semi-structured, and structured data. A definition of the structure classification is given in Table 15. In addition to the structure classification, there exists also a statistical classification as shown in Table 16.

STRUCTURE CLASSIFICATION OF DATA	
NON-STRUCTURED DATA	Non-structured data does not follow a specific format or sequence. The data is not predictable and does not comply with specific rules. There is no metadata available about the data itself. Examples of non-structured data include video, sound and imagery but also unformatted text data.
SEMI-STRUCTURED DATA	Semi-structured data can also be called “self describing” data because the information about the data is contained within the data itself and does not need to be predefined, e.g. the number and order of attributes may dynamically change. A typical example for semi-structured data is an HTML web page.
STRUCTURED DATA	Structured data is organized within a predefined schema. All data needs to conform to this schema, e.g., the order and format of attribute cannot be arbitrary, and values are required for all attributes[3]. An example of structured data is a time series produced by sensors.

Table 15 Structure Classification of Data

STATISTICAL CLASSIFICATION OF DATA	
MICRODATA	Data about individual objects (persons, companies, events, transactions, etc). Objects have properties, which are often expressed as values of variables of the objects. For example, a ”person” object may have variables such as ”name”, ”address”, ”age”, and ”income”. Microdata represent observed (e.g., birthday) or derived (e.g., calculated age based on birthday information) values of certain variables for certain objects.
MACRODATA	Macro data, also called "statistics", are estimated values of statistical characteristics concerning sets of objects, "populations". A statistical characteristic is a measure that aggregates the values of a certain variable of the objects in a population. ”The average age of persons living in OECD countries” is an example of a statistical characteristic. Some statistical characteristics, e.g. correlations, summarize the values of more than one variable. Macro data represent estimated values of statistical characteristics. Estimated values deviate from true values because of different imperfections (errors and uncertainties) in the underlying observation (measurement) and derivation processes. The difference between ”estimated” and ”true” values is an issue not only on the macro level, but also on the micro level, since the observed (measured) values deviate from the true values because of measurement errors.

Table 16 Statistical classification of data

The incoming data from the different suppliers are first filled in an Operational Data Store (ODS) called Collection ODS in Figure 4. Unlike a Data Warehouse, which contains historical data, the contents of the ODS is a snapshot of current data sources.

[3] Consider „undefined“ as a value as well.

It is designed to support data consolidation. In this area, the Collection ODS is used as interim database for incoming data via the different channels (see section 4.1.1) before it gets further processed by the ETL components and loaded into the Staging ODS.

The quality and properties of the data (and in some cases metadata) provided by the sources has an impact on the analysis results. Therefore, the source of externally supplied data becomes relevant. The adequacy of a given source can be assessed along the following dimensions:

- **Aim of the SIS**
  It is evident that the choice of the data sources has to be based on the aim of the SIS in order to deliver the required analysis results as output.
- **Data quality**
  Insufficient quality of the data can cause considerable costs and falsify statistical analysis results. Identified cost factors caused by insufficient quality are:
  – Additional required effort for the correction of the data
  – Misinterpreted data analysis resulting in wrong tactical and strategic decisions
  – Increasing dissatisfaction of the end users (e.g. of Data Warehouse data)

  Dimensions for requirements to the quality of data are given in Table 17. It is the task of the Data Quality Management (see section 4.4.10.1) to define and adjust the levels of quality such that data is “fit for use” by the end user.

DATA QUALITY REQUIREMENTS		
CORRECTNESS	COMPLETENESS	UNIFORMITY
CONSISTENCY	CURRENCY	CLEARNESS
RELIABILITY	NO REDUNDANCY	COMPREHENSIBILITY
ACCURACY	RELEVANT	UNAMBIGUOUSNESS

Table 17 Data quality dimensions

- **Availability** (legal, social, organizational, technical)
  Make sure that the organizational and technical preconditions are given for using the data sources in the Data Warehouse. For instance, this means that there are no legal constraints for using the data and that access to the data is technically possible.
- **Costs for getting the data**
  The cost for getting the data for external and internal data sources has to be considered by choosing the data sources, as well.

The SFSO distinguishes between internal and external registers, depending on whether it controls data definition, entry and maintenance in a register or not. It is important to see that the SFSO does not have the responsibility nor the authority to enforce quality of external register data. However, the SFSO is responsible for

- managing the internal metadata describing these registers and
- defining SIS-internal quality requirements with respect to external register data in order to support sound statistical analysis

This requires specific data management processes to be in place at the interface between the Data Supplier and Data Collection and Staging areas. These processes need to aim at adjusting the form and quality of externally supplied register data:

- Harmonize statistical metadata (nomenclatures, formats,…) so that subsequent analysis may abstract from data source details
- Identify data quality issues (e.g. missing or incorrect contents)
- Implement appropriate countermeasures (e.g. enrich data using results from questionnaire campaigns…)
- Document modifications to the original data as part of cleaning activity

The overall aim of these activities must be to provide “good enough” data to subsequent statistical analysis procedures, where “good enough” needs to be quantifiable in order to assess the impact of data quality on the analysis process. Efforts are currently undertaken to define such procedures on a conceptual level (SFSO register strategy and concept, GSOA@BFS).

After collecting and storing of the received data from the different data sources in the Collection ODS, the data need to be prepared for the use in the subsequent areas. Therefore, ETL components are required. To support these functionalities, the Staging ODS is used. It can be considered to be a working area for the ETL logic. From the Data Collection and Staging Area the data will be loaded into the Data Working Area.

4.4.4.1. Guidelines

- Data collection needs to avoid information loss by design:
  - Data from external suppliers must not be aggregated when loaded into the Collection area (no loss of information).
  - Data from internal suppliers may be stored in aggregated from, if this aggregation is part of the statistical processing logic. The SIS infrastructure must not aggregate this data.
  - Aggregated data needs to support drill down in order to trace the aggregated figure back to its atomic components.
- No OLAP database schemas are implemented in the data staging area.
- No data analysis services are provided in this area.
- No direct data access is provided to the end user (e.g., via SQL*PLUS).

4.4.5. Data Working Area

The Data Working Area consists of the Base Database (BDB) containing statistics in the process of being created and register databases (Register and GeoRegister) containing data shared by several of these statistics.

As shown in Figure 5, the BDB consists of a limited set of database instances grouped by a common theme, like person-centric statistics (shape) or company-centric statistics (GUS). These groups can be chosen to fit the organizational structure, such that data maintenance in one BDB instance can be attributed to a defined organizational unit, and ideally such that dependencies between BDB instances can be minimized. A BDB

instance contains one or more individual schemas, which can be seen as processing areas for statistical groups when creating a statistics.

Most of the statistical applications found in the application layer will create, modify or access data in such a schema, and should be free to do so. Data in a specific BDB schema should not be generally visible to other statistical groups, since it may consist of half-products not intended for general use.

The BDB is characterized by the following properties:

- Each one of the BDB schemas integrates external, register and DWH data that form the basis for creating a new statistic. It may contain non-anonymized data. It may contain temporary statistical products. It may contain register-like data sets, if not covered by the two register databases.
- Neither the modeling nor the optimization of the BDB is focused on a specific analysis, but rather it can be considered as application neutral.
- For a given statistic, the BDB contains an integrated view on all data and metadata required for its creation. This view is evolving as part of the analysis, interpretation and processing activities. This may require data to be appropriately prepared before entering the BDB.
- The BDB contains enough data to guarantee lineage traceability from the statistical end product back to the original data. The extent of historical data to keep depends on the nature of the statistics and needs to be determined when designing the statistic.
- The BDB should be considered volatile. Once a statistic is published to the Analysis Area, it should be possible to clean up and eventually release related BDB schemas. This requires that all data needed for history preservation and lineage tracking have been published to the Analysis Area.

In addition to the BDB, there are two registers available in the Data Working Area. The first register contains general master data provided by governmental departments (“Register”) and the second register provides geographical master data (“GeoRegister”) that are imported via the Data Collection and Staging Area.

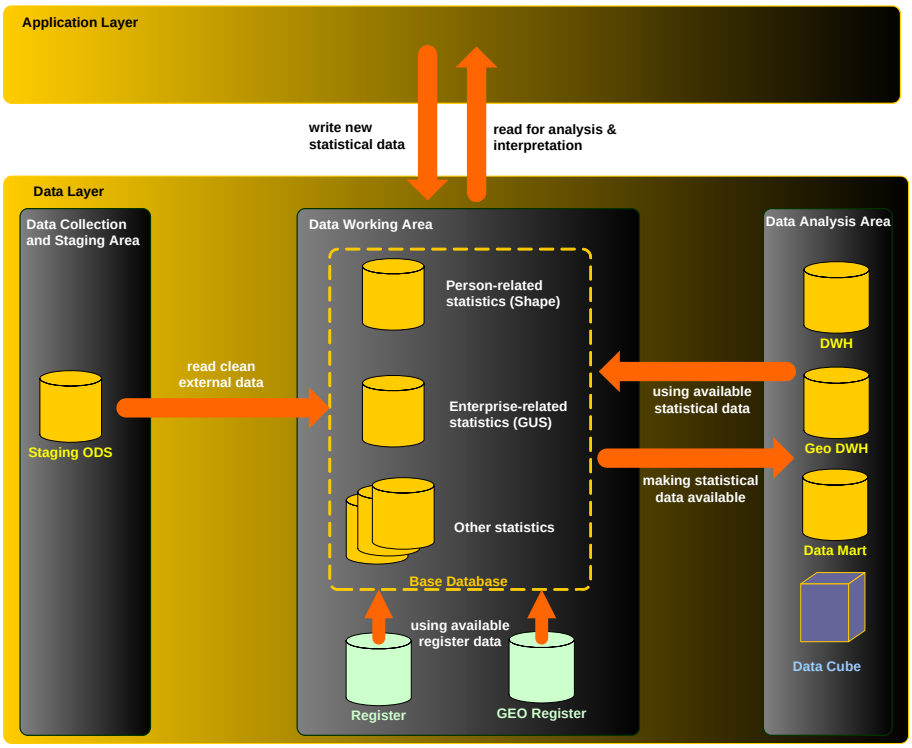


Figure 5 – Data flow in the data working area

Figure 5 shows possible data flow paths between data areas, assuming that it is allowed to directly load data into the Data Working Area without passing through the Data Staging Area (see 4.4.2). The orange arrows indicate the data flow in and out of the BDBs of the Data Working Area. If a new statistic is produced, the input data can come from different sources:

- New external data from the Collection and Staging Area
- Available register data previously imported into the Data Working Area
- Data on existing statistics from the Data Analysis Area. This backward data flow of the Data Analysis Area to the Data Working Area is important to reuse previously achieved statistical results.
- Newly created data produced by a statistical application in the application layer.

Once a statistic is ready for publishing, it needs to be decorated with the required metadata and may be published to the DWH in the Data Analysis Area. Note that within the DWH of the Data Analysis Area only finished statistics will be stored, as opposed to the Data Working Area, where unfinished statistics may be stored.

4.4.5.1. Guidelines

- The group using it to create the statistic content defines to a large extent the structure of the working area. The data structures should generally match the structure required by the Data Analysis Area, to where it will eventually be published.
- Schemas in the BDB should be independent from each other and thus autonomous. This guarantees freedom of action for individual statistical groups when creating statistics while minimizing unintended side effects between

working areas. Objects physically shared by several working areas (register-like) should be read-only.

- A schema should only have access to external, register or DWH data that it needs to serve its intended purpose. Its data interfaces (to Staging and/or Analysis Areas) should be documented upon creation and kept up-to-date.
- Once a statistic has been published, it should not be dependent schema where it has been created. This requires publishing all relevant data to the DWH (e.g. microdata or metadata related to lineage tracking), not only aggregated end products.
- Schemas in the BDB should be created on request of statistical groups.

4.4.6. Data Analysis Area

The Data Analysis Area contains two data warehouses, the CODAM DWH and the GEO DWH used especially for the geo-referenced data. The databases in this area are optimized towards online analytical processing (OLAP)[4]. They get periodically updates of anonymous data from the Data Working Area and in conjunction with the Metadata Repository contain all the information needed for the desired analysis operations.

Even if it is the goal of a DWH to provide a logical integration of the different available data, it is not required to physically integrate the data in one database. Therefore, it is possible to distribute the different data of a DWH into several Data Marts that are managed by the Data Management Component. This provides different advantages:

- Reduction of the data model complexity
- Reduction of the data volume
- Improved load balancing
- Optimized performance tuning for certain operations

To support the required OLAP operations, data cubes (also known as hypercubes) may be used. They consist of dimensions and measures and allow data to be viewed and analyzed in multiple hierarchical dimensions.

Note that the customers of the Data Analysis Area are still internal customers. This is different to the Dissemination Area, where the data will be prepared for external customers and more stringent requirements regarding data privacy protection have to be considered.

The user acceptance is the key success factor for every Data Warehouse and is closely coupled to a high quality Data Analysis Area. All technical or modeling decisions should be taken under this perspective and it is important to provide user-friendly and powerful Analytical and Business Intelligence Tools. Business Intelligence Tools are the interface of the Data Warehouse to the end user and provide fundamental report formats (tables, figures, text, and multimedia elements), various functional areas (as shown in Table 18), and multiple platforms (fat or thin clients, active warehousing).

[4] This as opposed to Online Transaction Processing (OLTP), where frequent modifications by concurrent users and applications of the data are performed. Depending on the type of processing database structure has to be optimized in different ways to ensure adequate performance.

FUNCTIONALITY OF BUSINESS INTELLIGENCE TOOLS	
DATA ACCESS	Read only access to the data used often for reporting tools.
ONLINE ANALYTICAL PROCESSING (OLAP)	Interactive data analysis that provides the possibility to browse through different tables and figures.
DATA MINING	Enhanced data analysis methods that enable to use statistical analysis methods, for instance.

Table 18 Differentiated Functionality of Business Intelligence Tools

4.4.6.1. Guidelines

- Only dimensional models using star- or snowflake schemas are allowed. A star schema has a central fact table, connected to a set of dimension tables. The snowflake schema is an adaptation of the star schema in which the dimension tables build a hierarchy. Dimensional models have to be chosen because they have the better performance than for example models using normal forms or object oriented models and are focused on using relational databases.
- It has to be possible to generate data cubes automatically based on metadata information. If these generated data cubes will be stored in the DWH, they have of course to be provided also with appropriate metadata information.

4.4.7. Data Dissemination Area

The data dissemination area provides all the required functionality to provide the end products of the Data Analysis Area to external users. In the Output Database the passed data from the Disclosure Controller component will be stored and provides the base for the products of the Data Dissemination Area. Similar to the Analysis Area, also Data Marts and Data Cubes are available. The Content Management System manages the different available products.

As shown in Figure 4, a Disclosure Controller component connects the Data Analysis to Data Dissemination area where stringent requirements regarding data privacy protection are required. In the Dissemination Area, the informational content of collected and processed data should be preserved as much as possible whilst guaranteeing that particular individuals cannot be re-identified. This is known as the statistical disclosure control (SDC) problem[5].

[5] The SDC problem is related to individual (microdata) and aggregated data. Re-identification happens when data on the same individual but from different products or data files can be successfully linked. Record linkage is a re-identification mechanism used to link records that correspond to the same individual. There exist different approaches and mechanisms to protect released data like distortion, aggregation, or suppression ([14],[15]), before they get into the Dissemination Area. If the files share a set of common variables, re-identification procedures based on probability distributions (probabilistic record linkage) and based on similarity functions (distance based record linkage) have been developed. But recent developments in the field of data mining consider as well re-identification for non-common variables across different data files. This is of interest when considering data files sharing a set of individuals that describe similar information (e.g., correlated variables). Re-identification for non-common variables is based on the existence of some relationships between individuals that are kept across files. These relationships imply some underlying structures in both files that can be obtained through the manipulation of the data and can be, afterwards, related through a re-identification mechanism.

4.4.7.1. Guidelines

- Only data that have been cleared and appropriately transformed for fulfilling data privacy and protection requirements are transferred into the Data Dissemination area.
- It is in the responsibility of the Disclosure Controller first to measure the disclosure risk according the known mechanisms, and then to apply the appropriate statistical disclosure control techniques. Within SFSO, section METH (statistical methods) is responsible for developing and establishing an appropriate SDC methodology.

4.4.8. Data Management Area

The Data Warehouse Management Component is responsible for the initiation, control, monitoring and logging of the processes in the SIS Data Layer.

Starting by the extraction of the data from the different data sources and ending by the analysis and data dissemination processes. For the control of the processes, the Data Management Component uses the information stored in the repository of the Metadata Management.

One of the most important tasks of the Data Management Component is the initiation of the data collection. Therefore, the Data Management Component contains different monitor components that observe the different data sources and report relevant changes to the Data Management Component. The job logs of the monitor components are stored in the archive database.

4.4.8.1. Guidelines

- All accesses to data in the SIS by services and applications are monitored and logged.
- All data management-related processes in SIS are configured using technical metadata. No hard-coded process logic is used.

4.4.9. Metadata Management Area

Metadata are the information about the contents and uses of the Data in SIS (also known as data about data).

4.4.9.1. Classification of Metadata in SIS

Metadata are created by several components of the SIS and can be divided into two general categories:

- *Statistical Metadata*
  Statistical metadata provides mainly information for the end user of the SIS. Most national statistical institutes and international organizations use the following definition:

  “Statistical Metadata is descriptive information or documentation about statistical data, i.e. microdata and macro data. Statistical Metadata facilitates sharing, querying, and understanding of statistical data over the lifetime of the data” [6].

Statistical metadata describe different quality aspects of statistical data, e.g.

- Content aspects: definitions of objects, populations, variables, etc.
- Accuracy aspects: different kinds of deviations between observed or estimated and true values of variables and statistical characteristics.
- Availability aspects: which statistical data are available, where they are located, and how they can be accessed.

With statistical metadata the analysis tools can be used more effectively, it is easier to find the relevant data and to interpret the results of the analysis.

- *Technical Metadata*
  Technical metadata contain technical information like logical and physical data schema, integrity conditions, or implementation information used by tools for automated processing.

Statistical and technical metadata, the metadata can be further classified as shown in Figure 6.

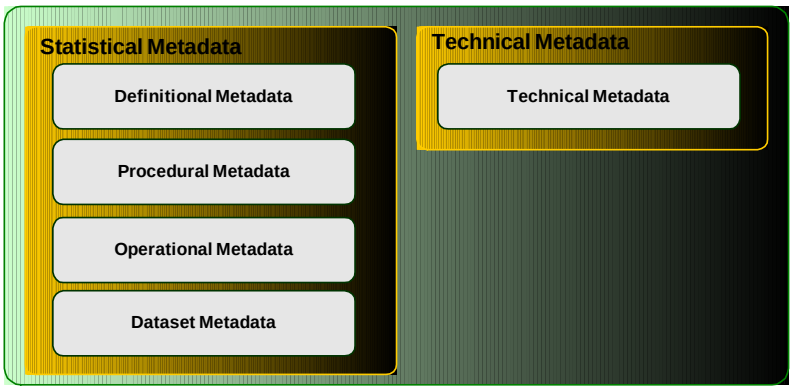


Figure 6 – Overview of metadata types

Definitional Metadata

The Definitional Metadata contain international, national and SFSO-internal standards and classifications. Some examples of stored metadata within the Definitional Metadata are shown in Table 19.

DEFINITIONAL METADATA		
ABBREVIATIONS	CLASSIFICATIONS	LANGUAGES
DEFINITIONS	VARIABLES	MEASUREMENT UNITS
DESCRIPTION	VALIDITY DATES	FOOTNOTES
INTRODUCTION	CONTACT PERSONS	
PERIODICITY	RESPONSIBLE ORGANIZATION	

Table 19 Examples of definitional metadata

Procedural Metadata

The Procedural Metadata contain information about general processes and procedures that are related to statistical activities like surveys or statistics.

PROCEDURAL METADATA	
QUESTIONNAIRE LAYOUTS	RULES FOR TAKING SAMPLES
CONVERTING RULES	SURVEY CONTENTS
WORKFLOW DEFINITIONS	STATISTICAL GROUPS
PLAUSIBILITY RULES	STATISTICAL ACTIVITIES

Table 20 Examples of procedural metadata

Operational Metadata

The Operational Metadata contain information about specific procedures of a statistical activity.

OPERATIONAL METADATA	
USED QUESTIONNAIRE	DATA SUPPLIERS
QUALITY OF RESPONSES	CUSTOMERS
RATIOS OF RESPONSES	
QUALITY RULES	

Table 21 Examples of operational metadata

Dataset Metadata

The Dataset Metadata is the actual data about data and contains information about the stored data from a conceptual and a physical view.

DATASET METADATA	
ATTRIBUTE CATALOGUE	DATASETS
VARIABLE LISTS	PUBLICATIONS
VALUE RANGES	DOCUMENTS

Table 22 Examples of dataset metadata

Technical Metadata

The Technical Metadata contains all relevant information from a technical view to store and manage the statistical data within the information systems.

TECHNICAL METADATA	
LOGICAL DATASCHEMA	ADMINISTRATION INFORMATION
PHYSICAL DATASCHEMA	IMPLEMENTATION INFORMATION
INTEGRITY CONDITIONS	

Table 23 – Examples of technical metadata

4.4.9.2. Management of Metadata

Metadata are created and used in all the different areas of the SIS and therefore Metadata Management can be considered as a support layer for all other areas. As

shown in Figure 2, this area contains two components: the Metadata Repository and the Metadata Manager:

- *Metadata Repository*
  In the Metadata Repository all metadata of the SIS are stored.
- *Metadata Manager*
  The Metadata Manager is the interface to the Metadata Repository. It controls the access of the different architecture components to the repository and enables also the exchange of metadata between different components. The Metadata Manager manages all data flows from and to the repository.

Despite the classification in Figure 6, the different metadata classes may be linked to each other. It is possible, for example, that a variable described under Dataset Metadata is part of a classification within Definitional Metadata or that a statistical activity described in the Procedural Metadata produces a table with variables in Dataset Metadata. Therefore, it must be possible to create references between the different classes of metadata.

To simplify the communication between the producers and customers of statistical data especially including metadata, several international standards have been defined that should be considered by the implementation:

- *ISO/IEC 11179* [3]
  This standard is focused particularly to metadata and specifies the kind and required quality to describe data. It specified also the management and administration of metadata in a metadata registry.
- *Dublin Core* [4]
  The Dublin Core metadata standard is a well-known standard providing a simple but effective set of elements (consisting of 15 unstructured elements) for describing a wide range of data.
- *Data Documentation Initiative (DDI)* [5]
  DDI is an effort to establish an international XML-based standard for the documentation respectively creating metadata of content, presentation, transport, and preservation of datasets. There are over 300 defined tags that can be used in the metadata. The tags are categorized in document description, study description, files description, data/variables description, and other related material.
- *Statistical Data and Metadata Exchange (SDMX)*
  ”The BIS, ECB, EUROSTAT, IMF, OECD, UN, and the World Bank have joined together to focus on business practices in the field of statistical information that would allow more efficient processes for exchange and sharing of data and metadata within the current scope of our collective activities. The goal is to explore common e-standards and ongoing standardization activities that could allow us to gain efficiency and avoid duplication of effort in our own work and possibly for the work of others in the field of statistical information.“ (mission statement from www.sdmx.org)

4.4.9.3. Guidelines

- The definition of metadata structure and handling policies is performed centrally by one organizational unit (CODAM), for which it is responsible. These definitions must be in accordance with international metadata standards (e.g. SDMX).

- The definition of metadata semantics, is centralized as much as possible (e.g. in a similar way as the BDB instance groups similar statistical subject data), in an effort to reduce the overhead of defining similar concepts over and over again. Responsibility for this definition is with the SFSO sections.

- Structure and content of metadata must be checked for conformance with defined rules and policies before it is published. These checks are performed / supervised by the unit defining the structure.

- Up-to-date metadata must be made available to a broad audience in a timely and transparent manner. This may be facilitated by using and enforcing a central metadata repository.

4.4.10. Supporting Processes

4.4.10.1. Data Quality Management

Simple corrections of the data to improve the data quality are just fighting the symptoms. For a sustainable improvement of data quality, the data cleaning and correction needs to be embedded into a Data Quality Management (DQM). For this management, the following potential “traps” need to be considered:

- Neglecting of the cultural aspects:
  - Data quality problems often arise at the human-machine interface.
- Big picture is missing
  - The data quality problem is handled only on the system layer without considering the business processes that are still producing data with insufficient quality.
  - The term of data quality is interpreted locally without considering the interfaces to the external systems.
- Missing governance
  - Data quality improvements are performed “blind” without knowing the effects of the different actions that have been taken.
  - The responsibilities are not traceable. Therefore, it is difficult to provide appropriate incentives.
- Complex effects of data combination
  - Data Quality properties are non-additive. In environments where data from diverse sources is combined and processed using complex statistical functions, it is not obvious what the resulting quality properties are.
  - There is no reference point for quality measurements. In environments where new data products are continuously created, it is difficult to define reference

values for data quality assessment. In a first step, such values have often best guess character.

In Figure 2, a DQM according the Deming-Circle (best practice for any quality management processes) that contains four steps (plan, do, check, and act) in a continuous optimization cycle is shown. The steps of the Deming-Circle are described in detail in the Appendix.

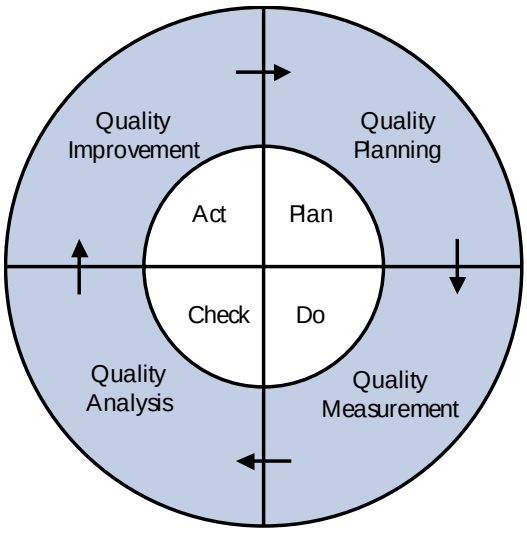


Figure 7 – Data Quality Management according to Deming-Circle ([2], p.221 f.)

It is important to see that these four steps are part of an iterative optimization cycle that assures a continuous improvement of both data quality and quality management processes. For a successful realization of a DQM, it is important to establish a quality management system and to provide the required methods and tools for all of the different phases.

In the Data Layer of the SIS architecture, the data quality issue arises in every area where data are hosted. It is required that every area has its own Data Quality Management process because due to the processing of the data by the ETL components or the Controller component (between Data Analysis Area and Data Dissemination Area) it is not guaranteed that if the data quality is fulfilled in one area that this is also the case for the subsequent area. In addition, it is also possible that different areas have different requirements to the data quality that can be adjusted by the own DQM process. But it is evident, that for example the data quality requirements of the Data Dissemination Area influence the requirements of the Data Collection and Staging Area. Therefore, it is required as shown in Figure 8, that all DQM processes are derived from a high level DQM strategy that defines the business objectives and high-level data quality requirements.[6] In this high level view, the DQM should be aligned and improved according the business processes. The responsibility for the high level DQM strategy is at the business level in collaboration with the suppliers that are delivering data to the SIS. Therefore, as already mentioned in chapter 4.2, the defined business objectives and requirements are strongly influencing the DQM processes of the whole SIS architecture. On the Data Layer, the different sections within SFSO have to maintain and guarantee the corresponding DQM processes of the different areas.

[6] This concept of hierarchy relates to the data owenrship hierarchy concept discussed in 4.4.1

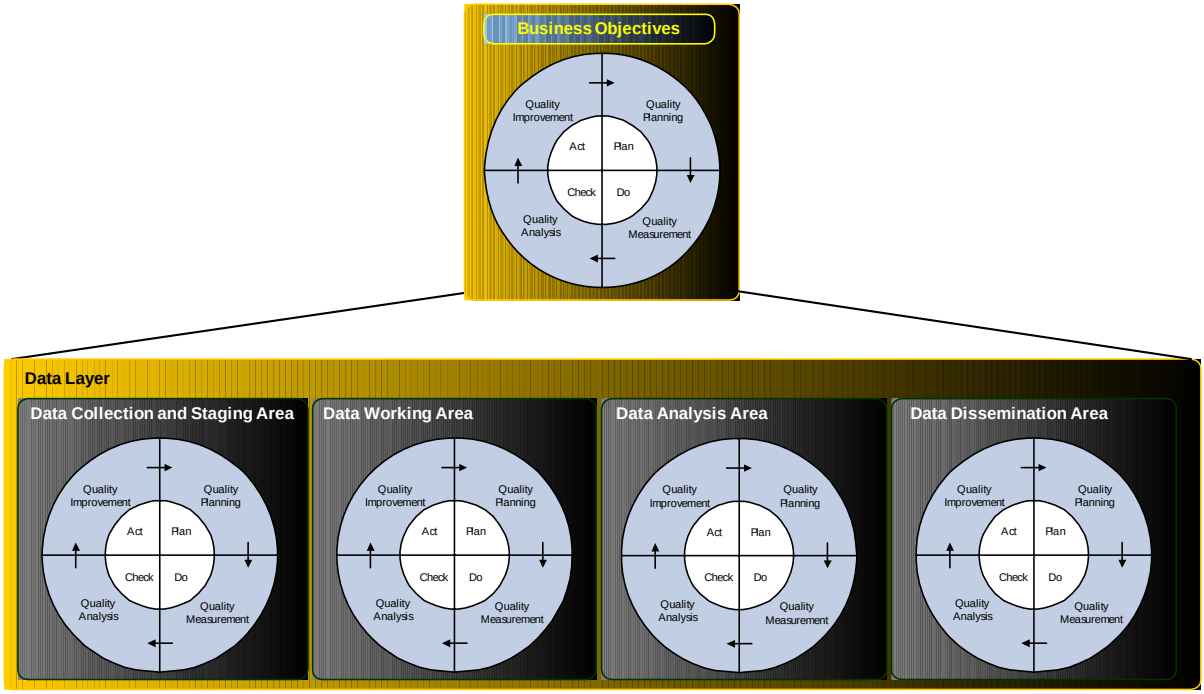


Figure 8 – Data Quality Management applied to Data Layer

4.4.10.2. Data Lineage Tracking

In general, in a Data Warehousing environment, the data lineage-tracking problem can be divided into two use cases:

- If you have an aggregated data item (e.g., summed value), data lineage tracking provides the possibility to drill-down this value to the source values (e.g., the values that contributed to the sum).
- As it can be seen in the Data Layer (see Figure 4) of the SIS architecture, the data coming from the data sources of the different suppliers travel through the different areas before it gets to the customers. In these areas several transformations (containing for example, data cleansing, anonymization, or aggregations) are performed. Task of the Data Lineage Tracking is also to be able to determine for any given data item to identify the exact set and the origins of the source data items that produced this item (e.g. which data items from which data sources in the Collection and Staging Area contributed to an aggregated data item in the Data Analysis Area).

The first use case is mostly relevant for users that try to get the bottom of any given data item. The second use case is relevant both for users to feel confident about their statistics results and for developers to assess impact of change within the SIS architecture. Also, it can be used for error analysis. If there are any incorrect values, potential faulty data sources could be identified. Within the scope of change management, Data Lineage Tracking provides the possibility to get information about relations between data over different areas of the Data Layer.

There exist several approaches and algorithms how to guarantee a consistent Data Lineage Tracking (compare [12], [13]). Similar to all approaches is that the tracking is based on appropriate metadata that are produced by any data transformation or aggregation. Hence, it is important to identify the needs and requirements imposed by Data Lineage Tracking. Some aspects that should be considered are mentioned in Table 24.

EVOLUTION ASPECTS	
GRANULARITY	The level of detail at which the drill-down of an aggregated data item should be possible.
DEPTH OF LINEAGE-TRACKING	The level of depth at which the backtracking should be possible.

Table 24 Data lineage tracking aspects

4.4.10.3. Data Security

The stored data in the different areas of the Data Layer of the SIS Architecture are valuable resources that need to be protected. Unauthorized access, breakdowns, manipulations, or destruction can have disastrous impacts to the SFSO. Data Access Controls and Security deals with secure storage, processing, and transmission of data. This includes the undisturbed operating of the IT systems and applications as well as the control that data and information can only be read, changed, or deleted by authorized users. Both internal and external threats need to be addressed.

In general, security requirements can be divided into 3 aspects: confidentiality, integrity, and availability.

Confidentiality, Secrecy

All resources, data, and information have to be accessible only by authorized people or organizations and only the minimal required authorizations should be granted in each case (based on the “need-to-know” principle). There exist three approaches for enabling data access control as listed in Table 25.

DATA ACCESS CONTROL APPROACHES	
DISCRETIONARY ACCESS CONTROL (DAC)	User related access control that restricts access to objects based on the identity and need-to-know of users and/or groups to which the object belongs. Controls are discretionary in the sense that a subject with certain access permission is capable of passing that permission (directly or indirectly) to any other subject.
MANDATORY ACCESS CONTROL (MAC)	System related access control where the system security policy (as set by the administrator) entirely determines the access rights granted, and a user may not grant less restrictive access to their resources than the administrator specifies.
ROLE-BASED ACCESS CONTROL (RBAC)	Users are not assigned permissions directly, but only acquire them through their role (or roles). Management of individual user rights becomes a matter of assigning the appropriate roles to the users.

Table 25 Data access control approaches

Basic requirements and functionalities of a integrated data access control is shown in Table 26

BASIC FUNCTIONALITIES OF DATA ACCESS CONTROL	
IDENTIFICATION AND AUTHENTIFICATION	All subjects (that will get access control) and objects (to which access control is granted) need to have unique, checkable, and unforgeable identification attributes. Authentication means the validation of a pretended identity (this includes validation of users and resources).
ACCESS CONTROL MANAGEMENT AND AUTHORIZATION	Access Control Management includes allocating, canceling, and administering access rights. Before the access to an object is granted, it has to be authorized.
PRESERVATION OF EVIDENCE	The preservation of evidence increases the traceability, liability, and integrity. All granted access and access attempts are logged. The log files need to be checked regularly concerning any irregularities.
TRANSMISSION SECURITY	It has to be assured, that the transmission of data and information between different system components is secure. Especially, if users access the system from outside via public Internet, the transmission of the access data need to be secure.

Table 26 Basic functionalities of data access control

Integrity, Accuracy

Aim of integrity is the correctness of the data that are available in the SIS architecture. Loss of data integrity can have several causes:

- Human factors (intended or unintended, e.g., caused by greenness or stress)
- Software problems (bugs, capacity problems like tablespace overflow)

- Hardware breakdowns (defective harddisks, network problems, power breaks)
- External causes (e.g., fire, sabotage, terrorism)

It can be distinguished between preemptive and corrective actions that can be taken for data integrity assurance. Preemptive actions try to avoid integrity problems (e.g., disk mirroring) in contrast corrective actions try to recover the data integrity after an incident (e.g., backup and recovery).

It is important that data integrity can be detected as early as possible. For example, if there is a software problem in the ETL component of the Collection and Staging Area that loads corruptive data into the Storage Area, this should be detected before the data are delivered to the Analysis or even Dissemination Area. Once an integrity problem is detected, a backup and recovery strategy is needed that enables to go back to a consistent state for the concerned systems as fast as possible. Metadata play also here an important role because they reflect not only the objects that need to be stored but contain as well information about the backup system, schedules and log files.

Availability

It needs to be distinguished between availability of data and availability of the system and its components. Whereas data availability is covered in the previous section about data integrity, this section covers the availability of the SIS architecture and its system components. In general, if a system component fails, it can be replaced. Important is that this replacement can be quickly performed (e.g., by using “hot swappable” that enables exchangeable modules during runtime). Alternatively, a redundant design of the architecture can be used meaning that a second failover-system takes over, if the first one fails.

Because the SIS Architecture is a business relevant component of the SFSO and can have disastrous impacts if it is not available, the strategy of the availability assurance and how to react in breakdown has to be addressed within the setup of a Business Continuity and Contingency Management. This is an ongoing process of risk assessment and management with the purpose of ensuring that the business can continue in the case of a breakdown. These risks could be from the external environment or from within your organization, such as deliberate or accidental damage to systems. Business continuity is not just concerned with disaster recovery; it addresses anything that could affect the continuity of service over the long term, such as staff shortages in specialist areas. In the following some tasks and goals of the Business Continuity Management are listed:

- Avoid loss of confidence and minimize disaster damage.
- Identify operational risks and weak points within the SIS architecture.
- Evaluate appropriate risks and develop a continuity strategy.
- Develop and realize actions for risk preventions.
- Take precautions for the replacement of processes and resources.
- Elaborate and provide plans for the re-establishment of processes.
- Assure appropriate economics for the precautions and re-establishment actions.

References

[1] A. Bauer, H. Günzer, „*Data Warehouse Systeme*“, dpunkt.verlag, 2. Auflage, ISBN 3-89864-251-8, 2004

[2] BFS, Die Erstellung des Artikelportfolios: Erläuterungen und Hilfsmittel. ANHANG 3 der Publikationsleitlinien BFS. Version 2.0 07.07.2006.

[3] International Organization for Standardization, „*ISO/IEC 11179 Information technology – metadata registries (MDR)*“, http://www.iso.ch, 2004

[4] Dublin Core Metadata Initiative, http://dublincore.org, July 2006

[5] Data Documentation Initiative, http://www.icpsr.umich.edu/DDI/codebook/index.html, July 2006

[6] C. Dippo, D. Gillman, „*The Role of Metadata in Statistics*“, UN/ECE Work Session on statistical Metadata, Geneva, September 22-24, 1999

[7] KOGIS, GM03 – Metadatenmodell, FD (Final Draft) Version 1.4, Juni 2004.

[8] A. Marzetta, M. Moret, B. Loison, Architekturstudie „90-Grad“: Generische Architektur für Erhebung, Erfassung und Verarbeitung. 19.2.2005.

[9] E. von Maur, R. Winter, „*Data Warehouse Management*“, Springer Verlag, ISBN 3-540-00585-4, 2003

[10] M. Berberov, R. Eder, P. Gerstbach, „*Data/Information Quality Managem*ent“, Seminararbeit, University Vienna, 2002 http://www.gerstbach.at/2002/DataQualityManagement/

[11] P. Vassiliadis, M. Bouzeghoub, C. Quix, „*Towards Quality-Oriented Data Warehouse Usage and Evolution*“, Proceedings of the 11th International Connference on Advanced Information Systems Engineering, p.164-179, 1999

[12] Y. Cui, J. Widom, „*Lineage Tracing in a Data Warehousing System*“, 16th International Conference on Data Engineering, IEEE Computer Society, February 2002

[13] H. Fan, A. Poulovassilis, „*Tracing data lineage using schema transformation pathways*“, In Knowledge Transformation for the Semantic Web, p. 64-79, IOS Press, 2003

[14] J. Domingo-Ferrer, V. Torra, „*On the connections between statistical disclosure control for microdata and some artificial intelligence tools*“, Information Sciences 151, p. 153-170, 2003

[15] P. Doyle, J. Lane, J. Theeuwes, L. Zayatz, „*Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies*“, ELSEVIER, ISBN 0-44-50761-2, 2002

[16] BFS, “Strategische Informatik Planung 2003”, January 1, 2004

[17] BFS, “*PRAXIS Teilprojekt Prozessgestaltung, Schlussbericht*”, Version 1.1, May 21, 2002

[18] ISB, “*R001 Referenzmodell für Informatikarchitektur Bund (RIAB)*”, Version 1.3, March 31, 2003

[19] SFSO, “*Statistikportal Schweiz*”, AG/BFS Info,9/03, September 2003

[20] SFSO, “Umgang mit Metadaten im BFS”, 10/06.

Appendix

Data Ownership vs. Data Mastership

This section is devoted to clarifying the role of Data Mastership in the context of this document. Whereas the roles described above are focused on data management activities in the organization, the data mastership concept addresses data placement on a system-based level. It will help to identify redundancies and thus provide a tool to analyze potential data quality issues. It can be distinguished between a centralized or decentralized approach for the Data Mastership. In a centralized mastership, a data item of a certain type (e.g., name and prename of a residents) has a dedicated system that is responsible for storing values of that data item type. In contrast, in a decentralized mastership several systems are responsible for data items of the same type. Decentralized mastership is thus more complex to handle since it has inherent redundancy issues that need to be addressed. The benefit is that it is less restrictive onto the IT architecture and business processes. Criteria for deciding in favor of a centralized or decentralized data mastership are shown in **Error! Reference source not found.**.

CRITERIA FOR USING CENTRALIZED OR DECENTRALIZED DATA MASTERSHIP	
CENTRALIZED MASTERSHIP	• High consolidation potential (many similar data entries). • Critical (external or internal) data. • A lot of systems are involved and interested in the data, respectively.
DECENTRALIZED MASTERSHIP	• Disjunctive data entries (e.g., if there are disjunctive customer profiles, the data about the customers can be stored in two different systems). • Low usage and relevance of the data. • Few systems are involved and interested in the data, respectively.

Table 27 – Criteria for using centralized or decentralized Data Mastership

Data Quality Management: The Deming Circle

The Deming-Circle defines pest practice with respect to data quality management implementing a feedback loop consisting of the definition of data quality and ist metrics in a specific context (quality planning), measurement of the data quality, analysis of the collected quality data, and implementation of improvement measures.

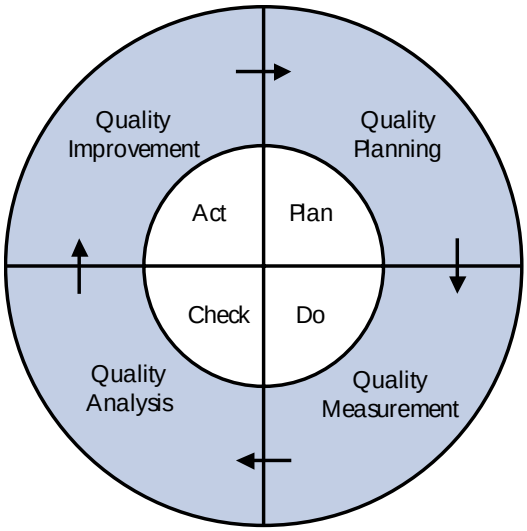


Figure 9 – Data Quality Management according to Deming-Circle ([2], p.221 f.)

- *Quality Planning* (Plan):
  There are different expectations towards the requirements of the data quality. In general, “data A is better quality than data B if it meets customer’s demands better” (according Thomas C. Redman, [10]). The aim of Quality Planning is to transform the expectations concerning the data quality into requirements that can be measured. This can be achieved by the actions described in **Error! Reference source not found.**.

ACTIVITY	TASKS	ROLES		
		OWNER	STEWARD	USER
DEFINE DATA QUALITY REQUIREMENTS	Define required data quality dimensions (accuracy, completeness, consistency, reliability, relevance, comprehensibility, and so on).	L, D	R	
	Profile data to determine potentially recurring quality defects (Data-, Dataflow-, and Metadata analysis).	R	L	D
KPI DEFINITION	Define quality characteristics (attribute values, structure, integrity conditions, value frequencies) and how it will be measured.	R	L, D	
				L=lead, D=do, R=review

Table 28 – Tasks of Quality Planning

- *Quality Measurement* (Do):
  The aim of the Quality Measurement is to measure the data quality according the defined requirements of the Quality Planning step (see **Error! Reference source not found.**).

ACTIVITY	TASKS	ROLES		
		OWNER	STEWARD	USER
MEASURING DATA QUALITY	Measure the data quality according to defines KPIs.		L	D
CONSOLIDATION OF DATA QUALITY LACKS	Collect facts about data quality lacks.		L, D	D
	Identify recurring patterns		L, D	
	Identify data quality problems	R	L, D	R
				L=lead, D=do, R=review

Table 29 – Tasks of Quality Measurement

- *Quality Analysis* (Check):
  The Quality Analysis can be considered as support for the Quality Planning and Measurement and its task is to pinpoint existing quality problems based on the data collected in the previous step and to define appropriate countermeasures (see **Error! Reference source not found.**).

ACTIVITY	TASKS	ROLES		
		OWNER	STEWARD	USER
ACTION AND REASON ANALYSIS	Locate the business relevant impacts of the insufficient data quality.	L, D	R	
	Analyze the reasons for insufficient data quality, e.g., data input, data movement, or data transformation.		L, D	D
DEFINITION OF COUNTERMEASURES	Data cleansing activities, monitoring and logging improvements, quality assurance improvements, extended use of metrics	R	L, D	
				L=lead, D=do, R=review

Table 30 – Tasks of Quality Analysis

- *Quality Improvement* (Act):
  The Quality Improvement facilitates the dynamic increase and continuous improvement of the data quality (see **Error! Reference source not found.**).

ACTIVITY	TASKS	ROLES		
		OWNER	STEWARD	USER
REALIZATION OF DEFINED COUNTERMEASURES	Adaptation of data architecture	L, D	R	
	Issuing new guidelines	L, D	R	
	Enhancement of metadata	L, D	R	
	Improve metrics	R	L, D	
	Clean data		L R	D
	Improve logging		L	D
	Improve quality assurance	R	L, D	
				L=lead, D=do, R=review

Table 31 – Tasks of Quality Improvement