Feasibility study into integrated End-to-End Statistical Solution for GCC-Stat

Prepared by

devstat

in consortium with

GLOBCOM
Bringing IT Together...

Acknowledgment

This report was prepared by **DevStat – Servicios de Consultoría Estadística S.L.** in partnership with **GLOBCOM.** The consortium was selected by GCC-Stat through a request of services through an open invitation to tender GCC-Stat/2016/03 of 31/03/2016.

The report was prepared based on the contributions from the following technical team: Mr José Cervera (Project Director), Mr Bouazza Bouckhar (Team Leader), Mr Karlis Zeila (IT Team Leader), Mr Norberts Talers (Senior IT expert), Mr Mohith Shakir (Senior local IT Experts) and Ms Florabela Carausu (Quality Assessment Expert). Project management and coordination was ensured by Ms Dovile Minkeviciute.

The authors are grateful for the helpful comments from GCC-Stat project coordination team and the GCC-Stat representants interviewed for the collection of information. They also express their gratitude to other entities that collaborated to the project by providing information on the aspects assessed.

The authors are responsible for any remaining errors or omissions.

The views expressed in this report are those of the project team and do not necessarily reflect those of GCC-Stat.

Table of contents

List of figures

List of tables

Acronyms

BMP6	The sixth edition of the Balance of Payments and International Investment Position Manual
CECT	GCC Common External Customs Tariff
COICOP	Classification of Individual Consumption by Purpose
CPI	Consumer Price Index
E2E	End-to-End
ESMS	Euro SDMX Metadata Structure
Eurostat	The Statistical Office of the European Union
FDES	Framework for the Development of Environment Statistics
GCC-Stat	Statistical Centre for the Cooperation Council for the Arab Countries of the Gulf
GSBPM	Generic Statistical Business Process Model
ILO	International Labour Organization
IRES	International Recommendations for Energy Statistics
IRWS	International Recommendations for Water Statistics
ISCED	International Standard Classification for Education
ISCO	International Standard Classification of Occupations
ISIC	International Standard Industrial Classification of All Economic Activities
ISO	International Organization for Standardization
MENA	Middle East and North Africa
OECD	Organization for Economic Cooperation and Development
SDMX	Statistical Data and Metadata Exchange
SNA	System of National Accounts
TSA	Tourism Satellite Account
UNCTAD	United Nations Conference for trade and Development
UNECE	United Nations Economic Commission for Europe
UNESCO	United Nations Educational, Scientific and Cultural Organization

Preface

This report is submitted by DevStat – Servicios de Consultoría Estadística (hereinafter **DevStat**) in cooperation with GlobCom in the frame of the contract for the implementation of the project **"Feasibility study into Integrated End to End Statistical Solution for GCC-Stat"** (tender reference n° GCC-Stat/2016/03), implemented for the Statistical Centre for the Cooperation Council for the Arab Countries of the Gulf (GCC-Stat).

The overall objective of the project is to support GCC-Stat in the continuous development process respecting strategic objectives and strategies, such as promoting and improving the quality and quantity of statistical products, developing and expanding statistics dissemination, aligning the national strategies in the GCC member countries, and especially establishing a culture of excellence in GCC-Stat building effective business systems and processes.

The specific objective of this project is to support and guide GCC-Stat in the assessment of the current arrangements that GCC-Stat has with six national statistical offices in member countries, aiming to streamline the statistical production process. More in detail, the specific objective of the project is two-fold:

- To assist GCC-Stat in assessing the statistical processes for data collection, management and dissemination and to formulate recommendations for improvement.
- To support GCC-Stat in developing a new, more efficient IT system, software tools and IT infrastructure, for an adequate support of the statistical processes from data collection to dissemination, followed by recommendations.

This report contributes to the achievement of the first specific objective, and provides the evidence that will be used for the formulation of recommendations. This report is drafted based on the analysis of a series of documents provided by GCC-Stat and on the information collected through the assessment missions.

This report is structured as follows:

- **Section 1** – presents the methodology and results of the assessment of the statistical processes for data collection, management and dissemination in GCC-Stat;
- **Section 2** – presents the assessment of the IT infrastructure of GCC-Stat;
- **Section 3** – presents options for the End-to-End solution;
- **Section 4** – presents gap analysis;
- **Section 5** – presents recommendations for the improvement of statistical processes and IT infrastructure in GCC-Stat as well as best scoring option for the End-to-End solution:
- **Section 6** – presents the time-bound plan.

Executive summary

Objectives

The overall objective of the project was to support GCC-Stat in the continuous development process as aiming at promoting and improving the quality and quantity of statistical products, developing and expanding statistics dissemination, aligning the national strategies in GCC member countries, and especially establishing a culture of excellence in GCC-Stat by building effective business systems and processes. More concretely, the project aimed **to support GCC-Stat in designing an efficient IT solution to create efficient business processes surrounding the statistical collections, data management and dissemination**.

The figure below presents the flow chart of all project activities and summarizes steps taken towards the implementation of the project.

Figure 1: Flowchart of project activities
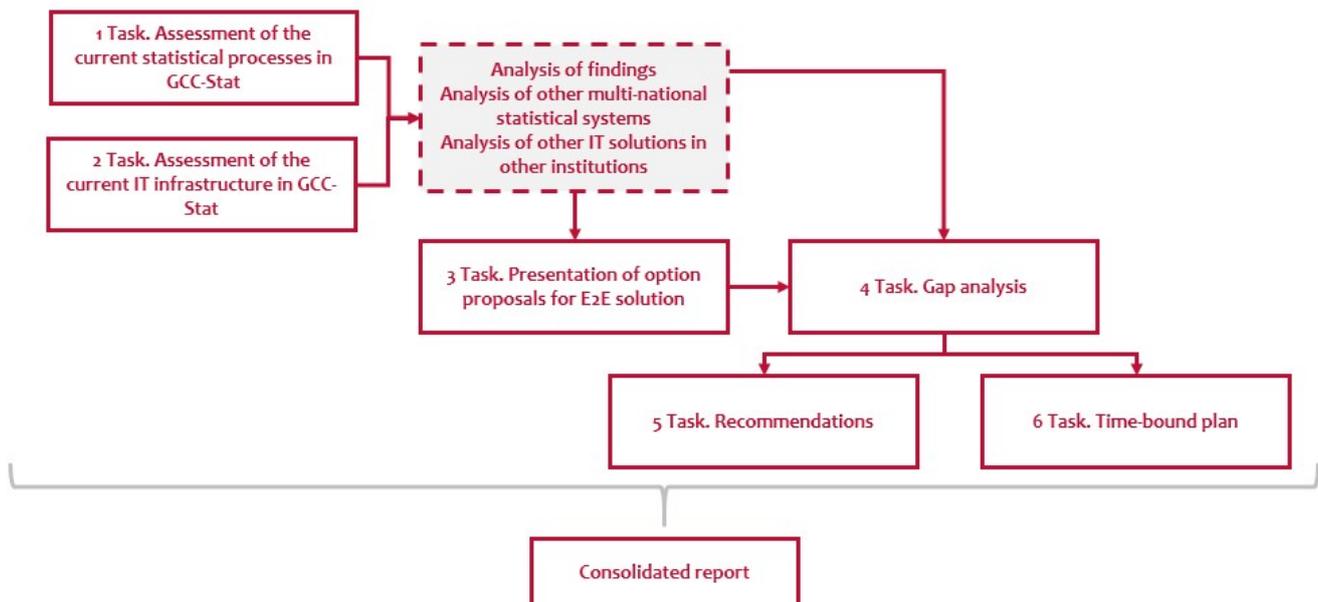
The two components of this project (i.e. improvement of statistical processes vs. improvement of the IT system for supporting the statistical processes) were subject of two complementary (but linked) analyses and recommendations reports.

Methodology

The methodology for the assessment of the statistical processes within GCC-Stat was largely based on international standards, using the Generic Statistical Business Process Model for Quality and Metadata Management (GSBPM) as a benchmarking framework, and of the Euro SDMX Metadata Structure (ESMS) as a guiding / structure tool for information collection and assessment of the production processes. The methodology of the assessment of statistical processes as well as of IT infrastructure proposed by DevStat included information collection

through questionnaires distributed to responsible GCC-Stat staff prior the assessment mission and face to face interviews

In order to obtain the information necessary for further development of the End-to-End (E2E) solution to the level required for stable functioning of proposed solution, an assessment of conformity of statistical processes to GSBPM was carried out. A complementary assessment of the IT infrastructure took place in parallel with the aim to determine its capacity, to record a series of different technical indicators characterizing the quality of the infrastructure and processed data security.

The key argument of linking GSBPM and SDMX / ESMS rested on the fact that the standardized metadata concepts can be applied across statistical domains, leading to a standardization of the related statistical processes.

The assessment of the statistical processes was done exclusively at GCC-Stat level, not at member country level, and referred to the set of statistical variables and indicators that GCC-Stat is aiming to collect from the six national statistical offices. The information was collected through missions carried out by subject-matter experts during October-November 2016. The information collected through interviews, using the template based on the ESMS, allowed for the formulation of findings.

Assessment results

The flowchart revealed a straightforward process, with the operations being almost exclusively done at GCC-Stat level. This revealed the serious work volume performed internally by GCC-Stat.

As key process operations were identified: (1) the data collection from the statistical centres of member countries, (2) the data checking and validation, done manually by GCC-Stat, and (3) the sensitiveness and time-consuming character of the data processing and analysis by GCC-Stat, as a consequence of the different standards and classifications used.

Overall, **the assessment of the statistics business case at GCC-Stat showed that it is characterized by the reduced documentation of the processes** and that it does not count with a communication strategy for statistics.

For the formulation of recommendations, the main findings were classified into aspects to strengthen and aspects to improve, based on their capability to impact in a favourable or unfavourable way GCC-Stat's plans to build an effective statistical business systems and process. Subsequently, a gap analysis of the statistical processes at GCC-Stat was performed, aimed at facilitating the formulation of recommendations for the improvement of some aspects. The gap analysis was not aimed to offer a concrete action plan, but alternatives to the current standing.

The recommendations formulated are grouped according to the following wide objectives:

- Strengthening the current **cooperation methods with member countries** in order to ensure the availability of harmonised timely statistics in specific statistical domains (allowing regional and international comparisons) and in line with expressed user needs;
- Improving the **dissemination and use of statistics** by the GCC-Stat, incorporating or developing additional elements such as communication strategy, visualization tools, quality

information about the data disseminated, etc., as well as focusing on the use of statistical data and facilitating the access and the use of statistical data bu all users (including dissemination and promotion of statistical literacy);

- Promotion of the culture of **statistical harmonization and quality** across GCC member countries and in all the statistical domains, by using common metadata and the implementation of adequate quality frameworks;

- Development of **automatic procedures and IT tools** to facilitate to GCC-Stat the automatic data correction routines and data processing;

- Building of **statistical capacity** in order to enable GCC-Stat and member countries staff the development of methodologies and common GCC standards, and the enhancement of their skills and competences.

The visual presentation of recommendations in the form of a Mind Map is presented in the next figure.

Figure 2: Mind map of recommendations
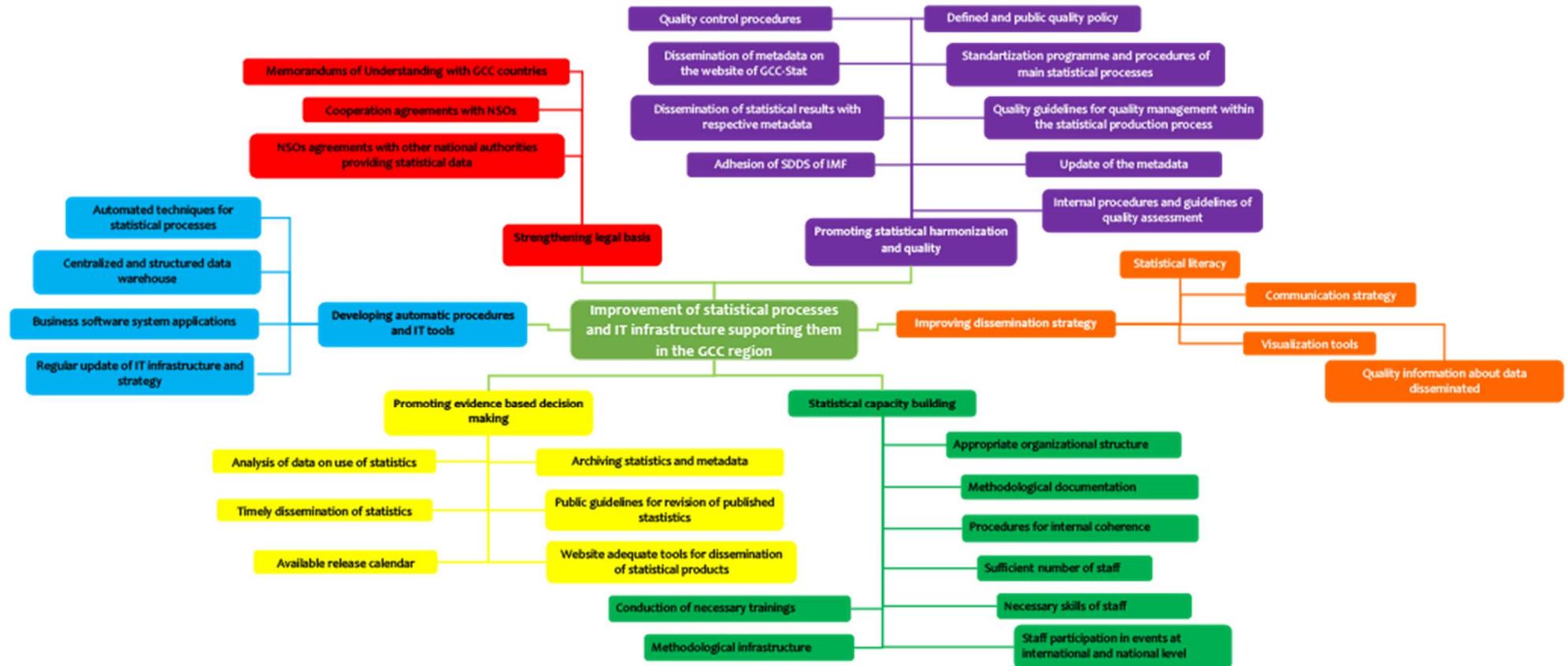
The recommendations concluded with the exemplifications of existing solutions in other advanced multi-national statistical systems, namely in the European Statistical System (ESS).

It was discovered that statistical data production processes in GCC-Stat are very close to GSBPM. Therefore, just minor changes were advised and a modified version of GSBPM was proposed to use as a basis for the elaboration of the chosen E2E solution.

The IT infrastructure from all points of technical assessment view was evaluated as good and well suited for the existing data processing environment in spite of lack of applications devoted directly for statistics production. All the statistics production processes are executed manually using MS Excel.

Based on findings of the assessment, three options for the E2E solution were elaborated according to the modified GSBPM version respecting GCC-Stat capacities and needs as well as international standards.

Alternative E2E solutions examined

The proposed options included the following: SAS based option, Microsoft based MD ISDMS option, Prognoz option.

- SAS based E2E option would ensure high level integrity and would provide access to large amount of statistical analysis. SAS software in different combinations is widely use in different organizations dealing with official statistics worldwide. At the same time SAS software does not ensure programming-free environment, it requires very deep and extensive training for statisticians and IT staff. It is most expensive software suite in the statistics area.

- Microsoft based Metadada Driven Integrated Statistical Data Management System (MD ISDMS) is a new generation statistical data production system. This system is fully integrated ensuring execution of all statistical processes in accordance with GSBPM within a programming-free environment. Metadata descriptions which can be prepared by any statistician, run the automated execution of processes. Efficient running of the metadata driven system has been proved in several NSOs.

- Prognoz platform based Statistical Data Management solution is built on the pillars of business intelligence. It covers all statistical processes but requires to be configured each of them. Strong IT support is necessary. Prognoz platform SDM provides wide range of tools for statistical analysis within the data production cycle as well as for data users on web. Training for statisticians as well as for IT staff plays a significant role in successful system exploitation. During the assessment phase it was clarified that Prognoz Company has suspended its activities and is currently transferring responsibilities to other companies and for the moment the Prognoz platform product is not available on the market.

As a result of evaluation of the options, **Microsoft based MD-ISDMS solution has been assessed as the most suitable option** to be introduced to and implemented in GCC-Stat. This option provides high level automation at all stages of statistics production and has user-friendly interface. The efficiency of the implementation of this option is proven by its use in several NSOs. The maintenance of the entire system is very easy and do not require expensive trainings.

Upon the request of GCC-Stat IT solutions used by the other international and national statistical organizations were assessed and are presented in Annex 9 of this report. In order to obtain more information on Prognoz platform use in AfDB, IMF, OECD, contact persons of these organizations were contacted. Unfortunately, no requested information was received. However, the information on correct and responsible person in these organizations was provided. Therefore, it is recommended for GCC-Stat to communicate with respective organizations at institutional level.

Furthermore, in order to learn about IT infrastructure in other international statistical systems such as the European Statistical System (ESS) represented by Eurostat (European Commission), DevStat upon discussions with GCC-Stat as well as with Eurostat, could support GCC-Stat in conducting a study visit to Eurostat premises in Luxembourg.

1. Assessment of statistical processes within GCC-Stat

Annexes of this section: Annex 1. Template for assessment of statistical processes. Annex 2. Assessment of statistical processes.

The assessment of statistical processes was built in a constructive way towards the formulation of recommendations for supporting an integrated End-to-End (E2E) solution harmonized among the six GCC countries and benefiting from the use of international statistical standards and classifications and from the data quality attributes following the GCC data quality framework. As stated in the terms of reference (ToR), it is particularly important that the integrated E2E supports comprehensive metadata, as well as data themselves.

The methodology proposed for the assessment of the statistical processes within GCC-Stat makes use of the following international standards in statistics:

- the **Generic Statistical Business Process Model for Quality and Metadata Management (GSBPM)** – established by UNECE, EUROSTAT and OECD – is used as a **benchmarking framework**;
- the **Euro SDMX Metadata Structure (ESMS)** – is used as a **guiding structure / tool** for the information collection (during the missions) and organization for the assessment of the production processes.

The original intention of the GSBPM was to provide a basis for statistical organizations to agree on standard terminology to aid their discussions on developing statistical metadata systems and processes. The GSBPM is therefore seen as a flexible tool to describe and define the set of business processes needed to produce statistics. The use of the model is also envisaged in separate, but often related contexts (as it is the case under the current project), such as: harmonizing statistical computing infrastructures, facilitating the sharing of software components, in the Statistical Data and Metadata Exchange (SDMX) for explaining the use of SDMX in a statistical organization, and providing a framework for process quality assessment and improvement[1].

For this benchmarking framework (i.e. the GSBPM), three main features are highlighted:

- it complies with the vision for quality of the main international statistics producers (e.g. Eurostat, OECD, etc.);
- is applicable to all the activities undertaken by the producers of statistics which result in data outputs;
- covers the whole production process, as implemented by all the entities / actors involved.

---

[1] GSBPM, version 5.0, December 2013
http://www1.unece.org/stat/platform/display/metis/The+Generic+Statistical+Business+Process+Model

The ESMS is used in the European Statistical System (ESS), being implemented both at Eurostat (supra-national level) and at national level. The ESMS contains the description and representation of statistical metadata concepts used for documenting statistical data and for providing summary information useful for assessing data quality and the production process in general. Although the template for the collection of information for the assessment of the statistical processes at GCC-Stat derives from the ESMS structure, it has been simplified and tailored to the specific needs of the project.

ESMS Metadata files are used for describing the statistics released by Eurostat. The concepts used in the ESMS files derive from the SDMX Content Oriented Guidelines (2009)[2].

It is highlighted, in this context, that the SDMX is not supported by the Prognoz web portal – currently in use in four of the six member countries. Nevertheless, the SDMX – at the same time, a technical and statistical standard – and corresponding ESMS are used here as a model to describe statistical data and metadata, rather than a standard for automated communication or a technology supporting standardized IT tools – i.e. it focuses on the content of the statistical data, and not on the structure. The assessment of using SDMX as an IT solution could be subject of the report proposing recommendations for the IT infrastructure – a separate deliverable produced in the frame of this project.

The key argument of the link between GSBPM and SDMX / ESMS rests on the fact that the standardized metadata concepts can be applied across statistical domains leading to a standardization of the related statistical processes. This efficiency add that could be obtained through combining these two standards is particularly relevant[3] in the context of the current assessment, considering that one of the main findings of the assessment missions is that there is a difference in the business processes developed and implemented for the priority statistical domains compared to the others.

Following the benchmarking framework, the statistical processes for the collection, management and dissemination of data and metadata in GCC-Stat are reviewed against the phases of the GSBPM, in order to guide the adaptation of the statistical processes at GCC-Stat towards an internationally accepted statistical business model. The benchmarking against the GSBPM will allow not only to review the statistical processes at GCC-Stat against the GSBPM recommended content for each sub-process, but also to elaborate a comprehensive processes flowchart for the identification of the sources of variation in the process which directly affect the output. In other words, it allows for describing the cause-effect pathway from activities (processes) to expected results (statistical outputs), or – in terms of quality – the cause-effect relation between 'quality of process' and 'quality of the output'.

The following components will be included in the process flowchart:

- The sequence of processes, indicating the decision points and their role (either as inputs or outputs);

---

[2] https://sdmx.org/?sdmx_standards=content-oriented-guidelines-cog
[3] At least theoretically, for the assessment of the statistical process and for the underlying recommendations for improvement

- The owners of the processes and their stakeholders;
- The key process variables.

Apart from allowing for identifying the key process variables, the processes flowchart can be used in the future as a tool for monitoring the results of the proposed recommendations.

The main evidence for this assessment report derives from the information collected during the sectoral assessment missions, following a template (see enclosed Annex 1) for the assessment prepared by the project team and agreed with GCC-Stat in advance.

The synthesis of the assessment missions that is presented in Section 1.2 will follow the structure proposed for the assessment template, organizing the information according to the following sections / activities:

- Access;
- Storage;
- Relevance;

- Metadata;
- Process analysis;
- Dissemination.

This structure reflects the compulsory link between the management of metadata and the management of quality. In fact, the GSBPM (see figure below) identifies as over-arching processes: Quality Management and Metadata Management. Quality management is closely identified with Phase 8 (Evaluation) – although presented throughout the entire GSBPM.

Figure 3: GSBPM phases and sub-processes
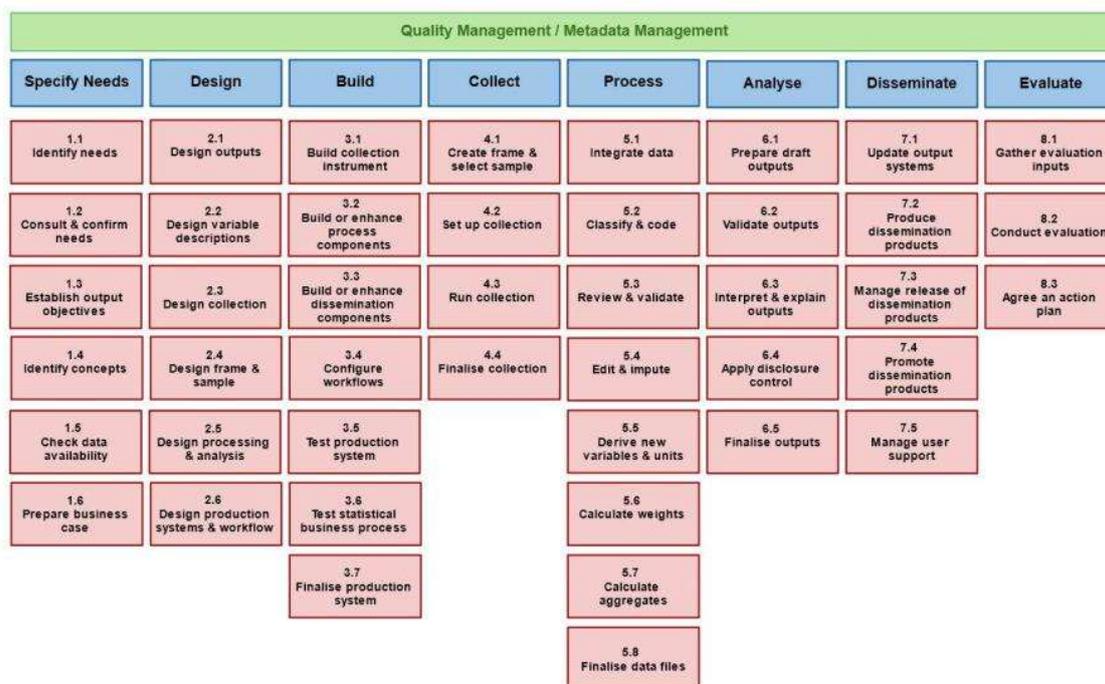
Source: Joint UNECE/Eurostat/OECD Work Session on Statistical Metadata (METIS), "Generic Statistical Business Process Model", version 5.0, December 2013

In the same way, although not represented through a specific phase, Metadata management must accompany the data throughout the entire statistical process, i.e. as an integral part of all steps in the statistical process.

The broad planning level represented through GSBPM phases Specify Needs and Design consist of defining all the relevant concepts related to the production of statistics, including details on:

- The data collected and from which institutions;
- The information needed in addition to purely numeric data;
- The frequency of collecting data;
- The quality controls that have to be performed after the collection of data;
- The processing that has to be performed for producing the statistical output;
- The users: who are they and how they will access the statistical data;
- The additional information that the user will need along with the statistical data (for an adequate interpretation and further analysis), etc.

This non-exhaustive listing of the content of some tasks that are undertaken in the planning of statistical activities shows that a data collection exercise envisages at the same time the collection of the data and the collection of the information about the data (i.e. metadata).

1.1. Overview of statistical system of GCC-Stat

The establishment of GCC-Stat was approved by the GCC Ministerial Council in September 2011 and the Charter of the Statistical Center was adopted by the GCC Supreme Council at its 33rd session held in December 2012. According to its charter the following tasks and responsibilities were assigned to GCC-Stat:

- Aggregation, classification, storage and analysis of data and statistics;
- Implementation of joint statistical projects in Gulf countries;
- Development of strategic plans for statistical work of GCC-Stat with the national statistical offices;
- Represent GCC in the statistical affairs;
- Contract international cooperation agreements for the development of the statistical system in the member countries, in coordination with the competent authorities;
- Application of international standards in all areas and phases of statistical work;
- Unifying and harmonizing methodologies, standards, definitions and classifications used in statistical work, in coordination with the national statistical of in the member countries;
- Policy development and identifying measures to ensure the quality of statistical work;
- Provide technical support for the improvement / modernization of the statistical work of member countries;
- Statistical capacity building in member countries;
- Contribute to building a culture of statistics and raising statistical awareness.

In 2014, GCC-Stat acquired the Prognoz data portal in order to improve the availability of statistics, as well as the dissemination methods. Along with the loading of the data to the portal, technical input is needed in parallel for ensuring quality statistical processes for an efficient processing, analysis and dissemination of statistics. The aim is to build a more streamlined process, which may include the direct transfer of data from countries, and a close integration of processing and dissemination using a common data storage, processing and dissemination portal.

Currently the data are **collected** in several ways, and namely through:

- Accessing statistics published by the member countries on their website, entered into Excel spreadsheets and then sent to countries for modification or confirmation;
- Spreadsheets in the form of transmission tables sent to the member countries for completion;
- Already formatted spreadsheets ready for dissemination as final publication tables.

As of February, 2015[4], in most cases, the **processing** of the data is a manual process, where staff visually check the spreadsheets for availability and validity. In few cases, formulas have been developed to validate the data. After an initial verification, the spreadsheets are re-checked by GCC-Stat and member countries for confirmation.

The **dissemination** is mainly in the form of:

- Electronic publications disseminated through the Data portal and the website;
- Paper publications delivered by post by GCC-Stat to a known list of regular users;
- Pdf files and spreadsheets published on the website;
- Press releases sent directly to the media;
- Social media dissemination (some).

A detailed overview of how the data collection and processing is done currently for each statistical domain is provided in Section 1.2. Still, it is hereby anticipated that major differences exist between the processes in place for priority statistical domains and the other domains.

GCC-Stat has identified the following fields as priority statistical fields:

- National Accounts;
- Monetary, Financial and Balance of Payments (BoP);
- Price and Short-Term Indicators;
- External Trade;
- Labour Market;
- Tourism Statistics;
- Energy and Environment;
- Progress and Development;
- Harmonized 2020 population census;

---

[4] See, mission report by M.A. Salmon: "Drafting – Collection, Management and Dissemination of data and metadata in GCC-Stat", February, 2015

- Business statistics, with a special focus on Business Registers.

The non-priority areas for GCC-Stat short-term planning work are:

- Agriculture and fishing;
- Culture and sport;
- Crime and justice;
- Technology;
- Health;
- Education;
- others.

Roughly, the processes implemented at GCC-Stat as per the GSBPM structure can be displayed as follows:

Table 1: GSBPM sub-processes implemented at GCC-Stat level

Specify needs	Design	Build	Collect	Process	Analyze	Disseminate	Evaluate
1.1. Identify needs	2.1. Design outputs	3.1. Build collection instrument	4.1. Create frame & select sample	5.1. Integrate data	6.1. Prepare draft outputs	7.1. Update output systems	8.1. Gather evaluation inputs
1.2. Consult & confirm needs	2.2. Design variable descriptions	3.2. Build or enhance process components	4.2. Set up collection	5.2. Classify & code	6.2. Validate outputs	7.2. Produce dissemination outputs	8.2. Conduct evaluation
1.3. Establish output objectives	2.3. Design collection	3.3. Build or enhance dissemination components	4.3. Run collection	5.3. Review & validate	6.3. Interpret & explain outputs	7.3. Manage release of dissemination products	8.3. Agree an action plan
1.4. Identify concepts	2.4. Design frame & simple	3.4. Configure workflows	4.4. Finalize collection	5.4. Edit & impute	6.4. Apply disclosure control	7.4. Promote dissemination products	
1.5. Check data availability	2.5. Design processing & analysis	3.5. Test production system		5.5. Derive new variables & units	6.5. Finalize outputs	7.5. Manage user support	
1.6. Prepare business case	2.6. Design production & workflows	3.6. Test statistical business process		5.6. Calculate weights			
		3.7. Finalize production system		5.7. Calculate aggregates			
				5.8. Finalize data files			
in place							
Partially							
under the current project (exclusively at recommendations level)							

Along with the work performed in parallel for each statistical domain, a set of tasks are generally undertaken horizontally in any statistical organization, which aim at standardizing and / or harmonizing concepts, classifications and methodologies.

Based on the information collected during the assessment missions, the following standards and classifications are used by GCC-Stat for different statistical domains.

Table 2: Statistical classification and standards used in GCC-Stat by statistical domains

Statistical domain	Priority (Y/N)	Classifications	Data collection
Foreign Direct Investment (FDI)	Yes	ISIC rev. 4 – at chapter level Country list and codes ISO Assets classification acc. to BMP6 Sectors classification acc. to BMP6	Transmission tables FDI data will be available through the forthcoming GCC-countries Coordinated Foreign Investment Survey (in collaboration with UNCTAD and MENA Centre for Investment).
Foreign Trade Statistics (FTS)	Yes	Country list and codes ISO GCC Common External Customs Tariff (CECT_ based on HS2012) – product nomenclature Mode of transport – IMTS 2010 nomenclature	Transmission tables complemented through manual downloads from the websites of other statistical providers (e.g. central banks)
Agriculture statistics	No	none	Email request to NSO's for selective indicators and Manual data collection from the websites of the statistical offices
Energy and environment	Yes	FDES (Framework for the Development of Environment Statistics) IRWS (International Recommendations for Water Statistics) Eurostat Waste Classification IRES (International Recommendations for Energy Statistics) Eurostat Energy Balance Standard ISIC	Transmission tables and manual data collection from national statistical offices websites
Industry and mining	No	ISIC rev. 4 – 2 digits	Email request to NSO's for selective indicators and Manual data collection from the websites of the statistical offices
National Accounts (NA)	Yes	SNA 93 and for KSA SNA 2008 ISIC Rev.3.1 and for KSA ISIC Rev.4 CPC 1.1 and for KSA CPC 2 COFOG, COPNI, COICOP HS (different versions) BOP (different versions)	Transmission tables and manual data collection from national statistical offices websites
Monetary and Financial	Yes	ISIC Rev. 3 IMF Monetary and Financial	Manual data collection from national statistical offices

Statistical domain	Priority (Y/N)	Classifications	Data collection
Statistics (MFS)		Statistics Manual, 2000 and Monetary and Financial Statistics: Compilation Guide, 2008	websites and other statistics providers (e.g. central banks)
Labour Statistics	Yes	ILO Standards 2008 (planned ILO 2013) ISCED Classification 1997 and 2011 ISCO 1988, ISCO 2008 and ISCO 2014 ISIC 3.1 and ISIC 4.0	Transmission tables and manual data collection from national statistical offices websites
Population	Yes	GCC 2020 Census Data Basket UN Standards on Vital Statistics (Rev. 2) – rev. 3 planned	Transmission tables and manual data collection from national statistical offices websites
Social Statistics	No	UNESCO Framework for Cultural Statistics 2009 UNESCO ISCED 2011 Classification WHO Core Health Indicators 2015	Transmission tables and manual data collection from national statistical centers websites
Sustainable Development Goals	Yes	ISCED 2011 UNESCO Framework for Cultural Statistics 2009 ISIC 4 ISCO 2008 UNESCO R&D manual	Transmission tables and manual data collection from national statistical offices websites
Price Statistics	Yes	COICOP 1999 – 4 digits	Transmission tables and manual data collection from national statistical offices websites (automatic extraction is under reflexion and test with the Sultanate of Oman)
Tourism statistics	Yes	Tourism Satellite Account: Recommended Methodological Framework 2008 (TSA: RMF 2008) International Recommendations for Tourism Statistics 2008 (IR TS 2008) International Standard Industrial Classification of all Economic Activities (ISIC)	Transmission tables and manual and automatic data collection from national statistical offices websites
Transport statistics	No	none	Email request to NSO's for selective indicators and manual data collection from the websites of the statistical offices

The standards and classifications used for the production and dissemination of statistics by member countries, and those used by GCC-Stat for the dissemination of the aggregate statistics produced, combined with the type of the data collection is important in the context of the processes done at GCC-Stat level; i.e. data harmonization and correspondence done by member countries or by GCC-Stat. Furthermore, it is also important in the context of GCC-Stat role / responsibilities to:

- Unifying and harmonizing methodologies, standards, definitions and classifications used in statistical work, in coordination with the national statistical offices of member countries;
- Application of international standards in all areas and phases of statistical work;
- Providing technical support to advance the statistical work of member countries;
- Statistical capacity building in member countries.

Figure 4: Process flowchart for the collection, management and dissemination of data by GCC-Stat
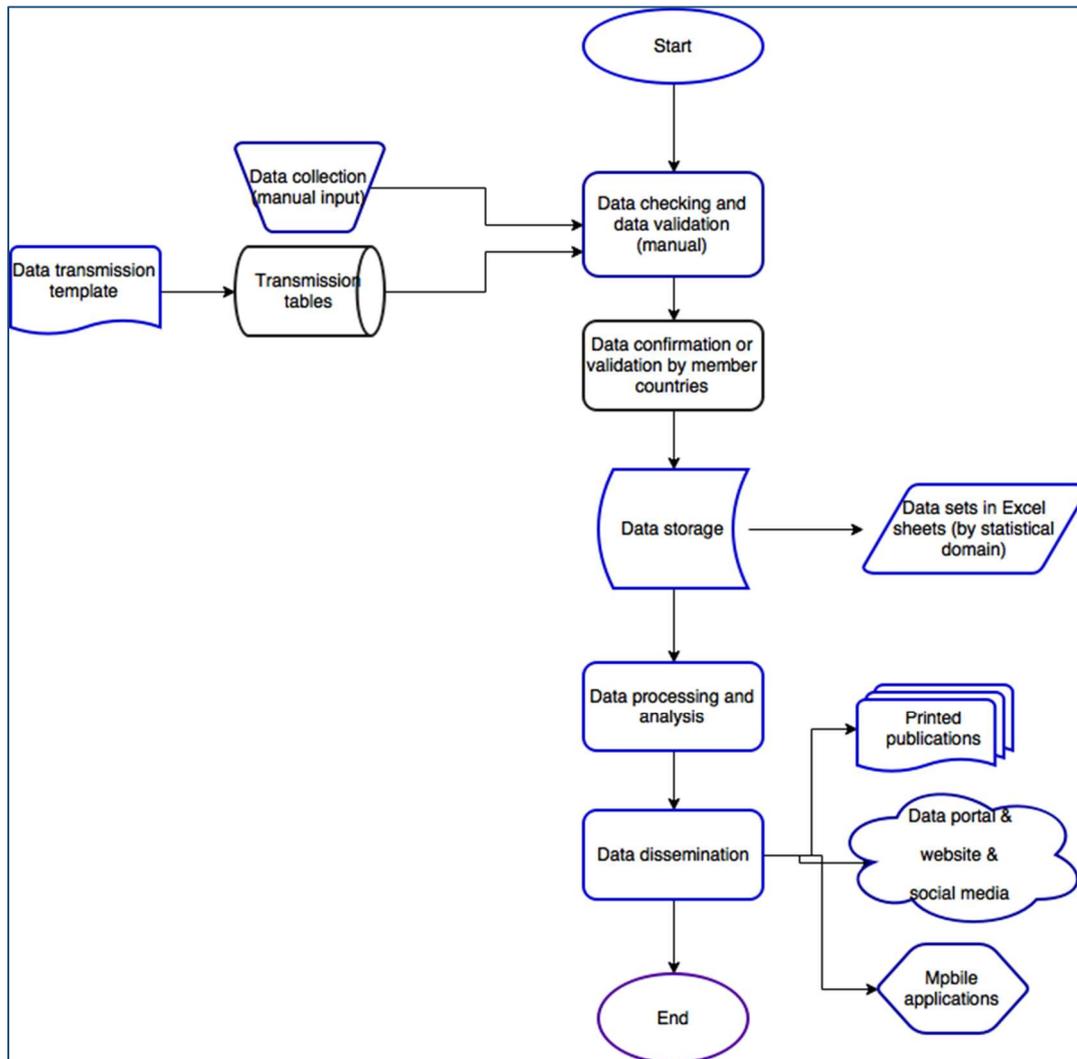
The current process for the collection, management and dissemination of data by GCC-Stat, as shown in the flowchart above is straightforward, with the operations being almost exclusively done at GCC-Stat level. This reveals the work volume that should be performed internally. Accordingly, key process operations are identified and are as follows:

- The data collection from the statistical offices of member countries. The data should be accompanied by detailed metadata, in order to ensure the proper processing and analysis by GCC-Stat;

- The data checking and validation, done manually by GCC-Stat (when, in general, manual operations for large volumes of data are error-prone);

- The data processing and analysis by GCC-Stat, as a consequence of the different standards and classifications used is considered as a time-consuming and sensitive task.

Overall, this statistics business case is also characterized by the reduced documentation of the processes and by the lack of a communication strategy for statistics. At the same time, it is also noted that the dissemination (the dissemination products) are not (necessarily) linked or derived from the data production platform (i.e. the data is only stored as Excel sheets).

1.2. Main findings of the assessment of the statistical system and processes within GCC-Stat

1.2.1. Institutional environment

- The general institutional frame refers to the Charter on the establishment of GCC-Stat and related roadmaps. There is no obligation for the member countries to provide statistical data to GCC-Stat. Meetings at the level of Standing Committee normally produce some recommendations to improve data exchange between member countries and GCC-Stat.

- The institutional environment is also characterized by the existence of many standing committees and working groups relating to the priority domains for GCC-Stat. These committees and WG are used to discuss the work programmes, the methodological issues and to coordinate the collection and dissemination of statistical data in these fields, following international standards, methodologies and classifications. Otherwise, there is no coordination mechanism (committees, working groups, task forces, etc.) enabling cross-domain discussions.

- It is recommended to put some efforts to strengthen the current cooperation agreements with member countries in order to ensure the availability of harmonized and updated statistics with regard to a number of topics and to allow regional and international comparisons.

- Member countries do not have legal obligation to provide the data and their consequent revisions to GCC-Stat. The data might be revised, but not immediately transmitted to GCC-Stat or are transmitted in different timings.

- Data confidentiality is sensitive as there are no developed standard rules on the treatment and transmission formats of confidential data. The confidential data mostly are not provided by member countries, and, if provided, then these are published only at GCC level.

- The Statistical Confidentiality is based on the 13th Primary Law of the creation of GCC-Stat. The law clearly states that individual data are to be treated as confidential and shall not to be used for other than statistical purposes. Otherwise, it should be noted that there are no specify legislations for statistical domains. Thus, the confidential data are not provided to GCC-Stat, they are suppressed (or not provided at all) before they are transmitted by member countries.

1.2.2. User needs

- User needs have been considered for main statistical areas, during the collection of variables / indicators, following the demands of GCC Secretary General or following discussions with the major GCC institutional users in regular meetings. Nevertheless, user needs for the other statistical sectors, are not considered yet.

- In some priority statistical domains, user needs (e.g., planners, policy makers, researchers) are regularly reflected when considerig indicators to be added to the transmission table. For example, some workshops and meetings are normally held in order to develop a culture of statistics and to improve the capacity building and enhance existing skills and competences.

- The organization of meetings and workshops, under the framework of the working groups and standing committees, showed a positive feedback as regards the satisfaction of user needs in sectors such as culture, education, etc.

- Not all user needs can be met in most of statistical domains due to the availability of data.

- In some statistical domains (i.e. mainly the priority domains) reflect their specific importance in the whole planning and policy decision-making process in the region. In this regard, some changes are planned by adding or deleting some variables based on national requirements.

- In some statistical domains (e.g. social statistics), relevant indicators are mainly collected for decision-making purposes. Regional comparison is desired but not yet possible due to the limited availability of data for a number of countries.

1.2.3. Use of international classifications and standards

- For the priority statistical domains, GCC-Stat and member countries use the international classifications and standards. However, in the other statistical domains, there is a lack in using such classifications, standards, methodologies and guidelines (e.g. agriculture, transport).

- In some statistical domains (e.g. population statistics, culture, …), the assessment revealed statistical data of member countries are in compliance with the international classifications (ISCO 2008, ISIC 4, ISCED 2011…).

- There is a lack of staff for the development of methodologies and common GCC standards.

1.2.4. Collection, processing and storage of data

- For the priority statistical domains, the decision for collecting data is taken during the annual regional working groups or standing committees meetings, or it is taken following the international recommendations. However, for the other statistical sectors, the decision is taken at internal level.

- The data are collected mainly from national statistical offices (NSOs), but in certain cases the data are downloaded from the website of other national institutions. These data are not included in consolidated databases, but used only for compiling and disseminating aggregates. Otherwise, the other sources at national level are not used even if they produce more detailed data.

- The data is collected by the Data Collection and Dissemination team andforwarded to the statisticians for validation.

- The collection of a set of variables by GCC-Stat through automatic accessing the national statistical offices website (e.g. through queries) is not available for any of the statistical areas.

- Some major constraints have been revealed in terms of data availability mainly due to: different and multiples sources (surveys, administrative data and census), data availability, gaps in terms of skills and competences and new topic for cultural statistics; different education systems in the region; and lack of detailed information for some health indicators. Time coverage for all sub-domains seems to be satisfying.

- In some statistical areas, not all the sub-domains are covered by GCC-Stat, even if data are available in most member countries for all the sub-domains of sectors (e.g. transport statistics, ICT, agriculture).

- Not all member countries are able to provide the entire set of statistics in the related fields. The situation varies from country to country.

- In general, the compiled and processed variables are mostly stored in Excel and are stored on the shared drive as well. In addition, there is no structured and centralized database.

- The data correction routines are not yet well established, there are no automated procedures, and all work is done manually. There is a substantial demand for development of automated procedures.

- The processing of the data is done manually. The risk of mistakes is therefore considerable. Next to that, the use of Excel files is always a risk for continuity of the work. For this reason, it needs to be documented in detail.

- For all stages of data validation, a regular feedback is ensured by GCC-Stat with the national statistical offices of member countries.

- The country treats the missing data. GCC-Stat has no rights to make an estimation. In this regard, no estimations are used to compensate missing data. Thus, if data are missing they are not included in aggregates and the data are not published.

1.2.5. Coverage and frequency

- In most of statistical domains, not all the variables / indicators collected are available for all the six member countries of GCC. There are many data missing (it depends on the country).

- In most of the statistical domains, the data are collected on a yearly basis, except for price statistics (monthly), external trade (monthly and quarterly) and national accounts (quarterly and yearly).

- Although concepts and definitions are the same, due to the data coverage problems and non-availability of some of data (because of confidentiality or other reasons), full harmonization and comparability of member countries data is difficult to achieve.

1.2.6. Use of metadata

- There is a reference metadata structure (based on OECD standard), which should be used by sectorial statistics. Only few priority statistical domains use it. In addition, for most of the statistical domains, the metadata for each member's country are not available.

- All aspects are taken into account when aggregating the data. However, for standards, concepts and definitions it is not possible to have an in-depth evaluation, as the metadata available from member countries do not contain the relevant data for some statistical sectors or there is no metadata produced by member countries.

- Standards, concepts and definitions are not possible to evaluate in depth, as there is no metadata or the metadata available from member countries is not sufficient.

1.2.7. Dissemination

- A data dissemination policy in GCC-Stat has been developed, which provide general framework for dissemination products. The data release calendar is available, where month of data release is mentioned. Strict implementation of this calendar not always possible, as it is subject to the availability of data.

- In principle, GCC-Stat cannot disseminate any statistics other than the ones already agreed with member countries.

- Furthermore, the data release calendar of GCC-Stat is not published on the website.

- GCC-Stat publishes statistical data in various ways (publications, social media, mobile application, website and data portal). Regarding the data portal, this tool enables GCC-Stat to disseminate data without metadata.

- Some general problems related to all statistical domains were highlighted: Data revision and dissemination policies are not harmonized not only between statistical domains but as well as between GCC-Stat and member countries.

- There are no restrictions to publish the data, everything what is received can be published. Harmonized dissemination of detailed data for all GCC countries currently is not possible either because of lack of the same details data or because of gaps in coverage.

1.2.8. Quality framework

- The existing quality framework produced by GCC-Stat is currently under review. The document referring to the technical guidelines ("Recommendations on Quality Reporting" specifically for dissemination), previously part of the document "Dissemination Policy", has become a separate document and is currently under finalization.

- The document "Quality Data Framework" for GCC-Statistics refers to all data produced by GCC-Stat.

- The appliance and usage of "Code of Practice" (data production, dissemination etc.) is currently under review.

- General quality framework was developed by GCC-Stat, which is based on the United Nations generic national quality framework. The framework is under review.

1.2.9. Other findings

- In general, the assessment revealed the continuing and appreciable efforts put in place by GCC-Stat in terms of capacity to comply with the metadata, process analysis and dissemination requirements.

- Some efforts are recommended to strengthen the current cooperation agreements with member countries in order to ensure the availability of harmonized and updated statistics with regard to a number of topics and allow regional and international comparisons.

- For most of statistical sectors assessed, there is the only one person available for each sector, which is certainly not enough for the task of providing harmonized aggregated data for the GCC countries.

1.3. Classification of main findings and construction of the argument

According to the methodology proposed, in this section the main findings formulated during the assessment phase are classified into 'aspects to strengthen' and 'aspects to improve', based on whether they are capable to impact in a favourable or unfavourable way GCC-Stat's plans to build an effective statistical business system and process. It should be noted that some aspects could be considered at the same time favourable and unfavourable; for example: the existence of many standing committees and working groups related to the priority domains for GCC-Stat, present at the same time an opportunity, but also a limitation; i.e. technical discussions and prioritization vs. fragmentation.

Table 3: Classification of main findings

Aspects to strengthen

- The institutional frame for the creation of GCC-Stat includes a roadmap to follow for the achievement GCC-Stat main objective (i.e. to build an effective statistical business system and process);
- The existence of many standing committees and working groups relating to the priority domains for GCC-Stat;
- Standing Committee meetings produce some recommendations to improve data exchange between member countries and GCC-Stat;
- The Statistical Confidentiality is based on the 13th Primary Law of the creation of GCC-Stat;
- For the main statistical areas, the user needs have been considered, during the collection of variables / indicators, following the demands of GCC Secretary General or following discussion with the major GCC institutional users in regular meetings;
- In some priority statistical domains, user needs (e.g. planners, policy makers, researchers) are regularly considered when deciding to prepare the indicators to be inserted in the transmission table;
- The user needs in some sectors are collected during the work progress of meetings and workshops, organized under the framework of the working groups and standing committees;
- The use of the international classifications and standards in the priority areas;
- For all stages of data validation, a regular feedback is ensured by GCC-Stat with the national statistical offices of the member countries;
- The missing data are treated by the corresponding National Statistical Office itself. GCC- Stat has no right to make an estimation;
- In most of the statistical domains, the data are collected on a yearly basis (in other words, a certain frequency is ensured);
- The existence of a reference metadata structure (based on OECD standard) is used in few priority statistical domains;
- The existence of a data dissemination policy in GCC- Stat with availability of a data release calendar;
- Diversification of the ways used by GCC-Stat for dissemination of statistical data (e.g. data portal, website, social networks, smart phone apps);
- GCC-Stat only disseminates statistics that are already agreed and disseminated with / by the member countries;
- The existing quality framework produced by GCC-Stat is currently under revision. In this regard, GCC-Stat is developing the Code of Practice and the general quality framework, which are based on international quality frameworks.

Aspects to improve

- Inexistence of any (legal) obligation for the member countries to provide GCC-Stat with statistics;
- Inexistence of developed standard rules on the treatment and transmission formats of confidential data;
- Inexistence of specific legislation for the different statistical domains;
- Difficulty to have regional comparison due to the unavailability of data for few countries;
- Lack of human resources for the development of methodologies and common GCC standards;
- Unavailability of harmonized procedures for collecting the data from member countries for all the statistical domains for which data is being disseminated by GCC-Stat;
- Existence of major constraints in terms of data availability mainly due to: different and multiples sources, data availability, gaps in terms of skills and competences, etc.;
- Non-coverage in some statistical areas / inability of some countries to provide data to GCC-Stat in some statistical areas;
- Inexistence of a structured and centralized database incorporating the data collected by GCC-Stat by

statistical domains;

- Inexistence of automated procedures for data correction routines and data processing;
- In the majority of statistical domains, the existence of many data missing for all the six member countries of GCC;
- Difficulty to have a full harmonization and comparability of member countries data;
- Unavailability of metadata in member countries for most of the statistical domains;
- Impossibility of strict implementation of the data release calendar;
- Inaccessibility to this data release calendar by the users via website of GCC-Stat;
- Non-harmonization of data revision and dissemination policies between statistical domains and between GCC-Stat and member countries;
- Insufficiency of human resources (statisticians) in charge of statistical sectors for the tasks of developing harmonized aggregated data for GCC region;
- The data / indicators disseminated by GCC-Stat through the data portal are disseminated without metadata

2. Assessment of IT infrastructure within GCC-Stat

Annexes of this section: *Annex 3. Information on the current IT infrastructure in GCC-Stat. Annex 9. Analysis of the other IT solutions in international and national statistical organizations.*

In this section the assessment of the existing IT infrastructure in GCC-Stat, statistical processes in comparison with the GSBPM, Business Software System currently in use as well as specificities of statistical domains related to the integrated E2E solution development is presented.

2.1. Overview of the assessment of IT infrastructure within GCC-Stat

Any organization or national authority aiming at producing high quality, internationally comparable official statistics is facing similar difficulties related to statistical processes as well as software supporting them. The ways how to tackle these issues differ from country to country and very often depend on national legislation, organizational specificities, available funding, etc. The most important issues assessed within GCC-Stat were its organizational structure, data production processes particularly data collection, data processing and dissemination, software used, the usage of metadata and its availability. All this information is of high importance and should be taken into consideration during the introduction of a new statistical data processing technology on the basis of integrated E2E solution. Special attention was given to the assessment of the IT infrastructure which ensures the statistical production.

The organizational structure with available human resources at GCC-Stat is presented in the table below. This part of organizational structure was chosen as a relevant to the scope of this project:

Department	Department's Section	Number of employees
Population and Social Statistics Department	n/a	5
Economic Statistics Department	n/a	7
Geographical, Environment, Energy and Tourism Statistics Department	n/a	3
Researches and Development Indicators Department	n/a	2
Technology and Information Department		
Technology and Information Department	Information Systems Section	4
Technology and Information Department	Data Collection and Dissemination Section	6
Technology and Information Department	Geospatial Information Systems Section	1

Data transfer from GCC member countries to GCC-Stat

Data collection from GCC member countries (MC) to GCC-Stat is mainly centralized and is carried out by Data Collection and Dissemination Section in Technology and Information Department. The exception is Trade statistics data in Economic Statistics Department. The data is collected directly from MS with Information technology specialists involved in problem solving stages but not directly involved in the data collection.

Data collection is organized in a way that data transmission tables are created in a standardized structure at GCC-Stat. Transmission tables have defined structure per statistical sub-domain and MS are expected to complete the tables and send them to GCC-Stat in a given time frame and periodicity which is defined at GCC-Stat level. The transmission tables are prepared in Microsoft Excel format and are sent as well as received per e-mail. In addition to that, the data is acquired from websites of NSOs of MS depending on the data availability as well as from hard copies of printed publications. It has been understood that in cases where data is acquired from websites or printed publications, GCC-Stat staff complete transmission tables manually.

Data validation algorithms are defined by statistical departments of GCC-Stat. They are given to Data Collection and Dissemination Section experts and are applied on received data sets in this section. Statistical departments deal only with clean data sets for further data analysis and compilation.

Data timeliness, integrity and availability

Even though, GCC-Stat has data production program on which basis workplans and deadlines were agreed, member countries are not always able to meet data reporting within particular timeframe due to the scarce availability of data as well as human resources constraints.

Quite often, data received in transmission tables do not contain all necessary values. Furthermore, data breakdowns do not always match the required structures, data from different member countries do not match methodologically and metadata is very often is of different level.

Data aggregation level

Most of the data received from MS are aggregated while in some cases micro-data is being received. GCC-Stat is eventually looking for possibilities of acquiring micro-data at a broader scope.

Data is further processed by analyzing and processing the data in order to create output tables representing the whole GCC region. The information is regularly published at the GCC-Stat dissemination portal.

Software used for data processing and dissemination

The main software used for data collection and processing is Microsoft Excel. There are two SPSS licenses. In addition, Prognoz data dissemination tool is used for data dissemination.

There are no data warehouses which consist of statistical data sets. Data is organized in Microsoft Excel files and stored in a structured system of folders on a network share. Folders access rights are set allowing access to a specific folder only from respective organizational unit of GCC-Stat. The exception is mentioned for Trade statistics section where Microsoft Access is used for data sets that fall under their responsibility due to the size which are too large to be regularly processed in Microsoft Excel.

Metadata usage and availability

International classifications, recommendations and other related information are being used in all statistical domains of GCC-Stat. Reference metadata structure is elaborated, but it does not appear yet in data dissemination portal. Reference metadata structure respect the structure used in OECD and IMF. Current version of the Prognoz data portal stores some text. However, is limited in its ability to provide a hierarchic structure or links to data being described by metadata. It is still not clear whether the Prognoz platform will be taken as a basis for improved / upgraded version of GCC-Stat data dissemination portal. The main disadvantages of the latter platform are a low performance, inconvenient interface, inability to store structured reference metadata, no assurance of linked datasets.

Structured metadata problems are similar to those related to data collection. Design and implementation of relation database for storing metadata and a separate classification database for storing and maintaining classifications would be necessary. In order to ensure an easy access to national statistics offices of member countries would be a significant step to improve data / metadata quality in the entire region.

IT Infrastructure details

High Level Infrastructure details are presented in the figure below.

Figure 5: GCC-Stat high level network diagram
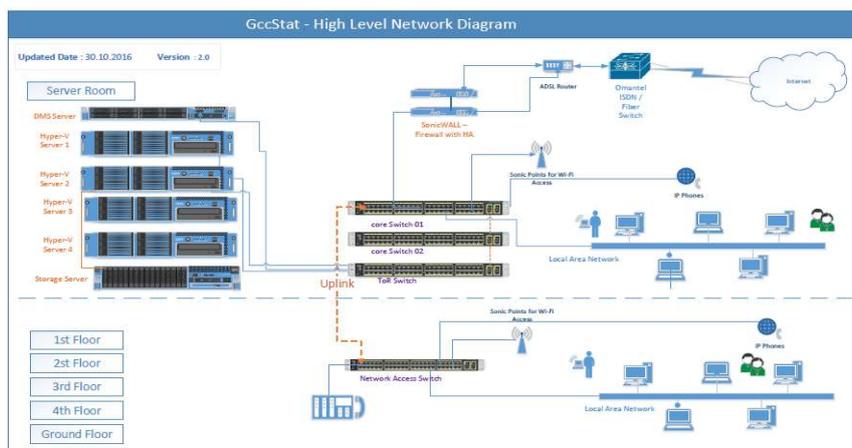
GCC-Stat is using a high-speed internet connection with the bandwidth of 64mb provided by a local telecommunication company Omantel. GCC-Stat has Data Center (Tier 2) on the 2nd Floor, Server and Network Infrastructure is hosted in the Data Center.

Server infrastructure details

#	Server Type	OS	Year of Production	No of Units	No of Processors	CPU Utilization
1	Rack Server	Server 2012 R2 DataCentre	2013.06.25	1	24(logical)	5%
2	Rack Servers	Server 2012 R2 DataCentre	2015.10.16	1	40(logical)	7%
3	Rack Servers	Server 2012 R2 DataCentre	2015.10.16	1	40(logical)	50%
4	Rack Servers	Server 2012 R2 DataCentre	2013.02.16	1	32(logical)	5%
5	Rack Servers	Server 2012 R2 DataCentre	2013.02.16	1	32(logical)	5%
6	Storage	-	2014.01.27	1	-	-
7	Virtual	Server 2012 R2 DataCentre	-	30	2(logical)	5%

Servers functionality details are given below:

1. Document Management Server - Production - Used to store soft copies of documents

2. Database Server - Production - Hosts the database of COSEC system

3. Financial Server - Production - Hosts GCC-Stat's Financial System and Database

4. Business Process Management Server - Test - To evaluate a BPM solution

5. CCTC Server - Production - Hosts the CCTV media Sever and files

6. Application Server 1 - Production - Hosts WDS (Windows Deployment Services) and WSUS

7. Network File Server - Production - Hosts all working files of GCC Stat users.

8. Geographical Information System Server - Production - Hosts the ARC GIS Sever

9. iGCC-Stat Server - Production - Hosts the internal Development Server.

10. HR Server - Production - Hosts the COSEC (Access Control, Leave Management and Time & Attendance System)

11. Application Server 2 - Production - Hosts the Print Server

12. Datacenter Access Control Server - Production - Hosts the application for the access control for the Datacenter

13. Domain Controller 2 Server - Production - Working as a secondary domain controller

14. Digital Post Server1 - Production - Working as a server for digital poster 1

15. Digital Post Server2 - Production - Working as a server for digital poster 2

16. Test Server 1 - Test - for testing purposes of server configurations and updates

17. Test Server 2 - Test - for testing purposes of server configurations and updates

18. Test Server 3 - Test - for testing purposes of server configurations and updates (Linux)

19. Data Portal Server - Test - Acting as a server for new data portal website

20. Data Portal Server - Test - Acting as a server for new data portal database

21. Backup Server - Production - Use as the server for backups with Veeam Application

22. Data Portal Test Client 1 - Test - Acting as a demo client for new data portal

23. Data Portal Test Client 2 - Test - Acting as a demo client for new data portal

24. Data Portal Test Client 3 - Test - Acting as a demo client for new data portal

25. Data Portal Test Client 4 - Test - Acting as a demo client for new data portal

26. Test Client - Test - Acting as a test client for server configurations and updates

27. Juno Space Server -Production - Host Juno Space application for management Juniper switches \ and routers

28. SonicWall Analyser Server - Production - Use for SonicWall Analyser tool for firewall monitoring

29. Monitoring Server - Production - Use spice works application for network monitoring

30. System Centre Server 1(CM) - Test - System Centre.

31. System Centre Server 2(DB) - Test - System Centre.

32. System Centre Server 3(DPM) - Test - System Centre.

33. System Centre Server 4(OM) - Test - System Centre.

34. System Centre Server 5(ORCH) - Test - System Centre.

35. System Centre Server 6(SM) - Test - System Centre.

36. System Centre Server 7(CM) - Test - System Centre.

37. System Centre Server 7(VMM) - Test - System Centre.

Desktop Environment

#	Hardware type	Characteristics	Manu	Memory	Disk space	OS	Number of units
1	Desktop Tower	ESPRIMO P510 E85+, Intel Core i7-3770 3.4GHz	Fujitsu	8 GB	500 GB	Win 8.1 64bit, Win 10 64bit	30
2	Desktop Tower	ESPRIMO P420 E85+, Intel Core i7-4790s 3.60GHz	Fujitsu	8 GB	1 TB	Win 8.1 64bit, Win 10 64bit	25
3	Desktop Mini Tower	Think Center E73, Intel Core i7-4770s 3.10GHz	Lenovo	8 GB	500 GB	Win 8.1 64bit, Win 10 64bit	30
4	Notebook + Tablet	HP X2 Split 13*2	HP	4 GB	500 GB	Win 8.1 64bit	8
5	Notebook	Thinkpad E431	Lenovo	8 GB	1 TB	Win 8.1 64 bit	5
6	Notebook	Thinkpad T540P	Lenovo	8 GB	500 GB	Win 8.1 64 bit	5
7	Notebook	Thinkpad T550P	Lenovo	8 GB	1 TB	Win 8.1 64bit, Win 10 64bit	5
8	iPAD	Apple iPAD AIR 2(Wifi+4G) Triple-core 1.5 GHz Typhoon	Apple	2 GB	128 GB	iOS 10	15
9	iMAC	iMAC Intel Core i5 3.2GHz	Apple	8 GB	1 TB	iOS X 10	3

High availability and redundancy

GCC-Stat is mainly using a Failover environment for the critical Network Services as a high availability solution. Any Disaster Recovery (DR) or Business Continuity Planning (BCP) solutions is not used. At this stage, it is not recommended to have DR. However, after analyzing data growth and use of new E2E System usage for 3 years, analysis might suggest the need to have DR and BCP.

Data and network security details

- Access control systems:
  - o All the employees are using the Access Control System within the organization.
  - o For the Access Control Systems, GCC-Stat is using Employee ID as a RF ID and the finger print.
- Application security:
  - o Antivirus software
    - ▪ F-secure
- Authentication:
  - o All the users are using domain username and the password for the PCs and Email accounts.
- Authorization:
  - o In the files server, the employee has given permission to access the relevant department and their personal folder.
  - o With the Access Control System the employees has restricted to access to the other departments and areas physically which are not related within the organization premises.
- Firewall (computing):
  - o With the firewall, users have been restricted and limited to access unwanted websites.
- Intrusion Detection and Prevention System:
  - o IDS and IPS systems are enabled in the firewall.
- Mobile secure gateway:
  - o Currently this feature is not available in GCC Stat network.

Applications hosted in DC

- Document Management System – ELO
- COSEC system
- Financial System (custom made)
- CCTV Monitor
- WDS (Windows Deployment Services)
- WSUS (Windows Update Server)

- Geographical Information System - ARC GIS
- Veeam Backup and Replication
- Juno Space
- Sonic Wall Analyser
- Spice Works Application
- System Centre Configuration Manager.

GCC-Stat future plan

Based on the discussion with GCC-Stat team, a plan to procure few applications which will help GCC-Stat in day to day operations was mentioned. Existing Data Dissemination portal is hosted outside GCC region, and GCC-Stat is willing to host it within the region. For this an application for Data Collection, Data Processing and Data Dissemination is needed. A list below presents Applications / Tools required for GCC-Stat and are planned to be procured in near future.

- Data Management Tool
- New website
- Mobile Application – Like a dash board
- Hosting of Data Portal (within GCC)
- ERP
- Data Management and Data Collection Tool.

2.2. Assessment of the Prognoz platform use by other organizations

Upon the request of GCC-Stat IT solutions used by the other international and national statistical organizations were assessed and are presented in Annex 9 of this report. In order to obtain more information on Prognoz platform use in AfDB, IMF, OECD, contact persons of these organizations were contacted. Unfortunately, no requested information was received. However, the information on correct and responsible person within these organization was provided and is presented in the table below.

Organization	Contacted person and contact details
African Development Bank	Mr Charles Lufumpa, c.lufumpa@afdb.org (responded and redirected a message to the following persons)
	Mr Louis Kouakou, k.l.kouakou@afdb.org (not responded)
	Mr Rafik Mahjoubi, r.mahjoubi@afdb.org (not responded)
OECD	Mr Arnaud Atoch, Arnaud.ATOCH@oecd.org (responded and redirected a message to the following persons)
	Mr Jonathan Challener, Jonathan.CHALLENER@oecd.org
International Monetary Fund	Ms Patrizia Tumbarello (not responded) Mr John St. Hilaire, jsthilaire@imf.org (not responded)

3. Presentation of End-to-End options

Annexes of this section: Annex 4. Presentation of the SAS based option. Annex 5. Presentation of the MS MD ISDMS based option. Annex 6. Presentation of the Prognoz platform based option. Annex 8. Full description of SDMX implementation. Annex 9. Analysis of the other IT solutions in international and national statistical organizations.

3.1. Main findings of the assessment of IT infrastructure

In this section the main findings of the assessment of IT infrastructure are presented. We have grouped the findings following statistical data management processes and covering the IT infrastructure:

Data and metadata collection:

- Several methods of data collection are in use. Most data are collected using data transmission tables elaborated by GCC-Stat, filled-in by respondents and submitted by e-mail.
- Some data are collected directly from webpages of MS data providers.
- Minor data are collected from printed publications.
- Most data from MS statistical offices or other data providers are collected in aggregated form.
- Microdata are requested and collected rarely.
- Data are collected together with structural metadata. Not all reference metadata from MS are available.
- Data collection processes are manual and most of them are performed by Data collection unit.

Data processing and analysis:

- Data processing and analysis are performed by statisticians of statistics subject mater domains.
- Data validation is performed by the Data collection unit running validation programs coded on the basis of validation algorithms elaborated by statisticians.
- Validations are carried out manually at GCC-Stat. In case of data discrepancies respective MS is contacted.
- Data analysis is performed and reports are prepared using Excel.
- There is only one Foreign trade statistics database used.
- Data is organized in Excel files and stored in a structured system of folders on a network share.
- Data security is ensured via granting access rights to related experts to specific data folders.

- Two SPSS software licenses have been acquired for data processing and analysis needs. Trainings for users are being organized and performed.

Data dissemination:

- There is a data publication calendar which is available on the GCC-Stat website.
- Prognoz data dissemination software have been acquired for data dissemination purposes.
- On the basis of Prognoz data dissemination software platform data dissemination portal has been elaborated and implemented.
- Data dissemination portal is well organized and provides data presentation and analysis services to users. Data sets geographical interpretation is ensured.
- Due to weak structural metadata collection and management this part of metadata is not available on the data dissemination portal.

IT Infrastructure:

- GCC-Stat has high speed internet connection provided by a local telecommunication company Omantel.
- GCC-Stat has a Data Center, where also a Server and Network Infrastructure is hosted.
- Server Environment is hosted on Hyper V Environment.
- Data and Network Security:
  o All GCC-Stat staff are using the Access Control System.
  o For the Access Control Systems, GCC-Stat is using Employee ID as a RF ID and the finger print system.
  o All users are using domain username and password for PCs and Email accounts.
  o In the files server, the staff have given permission to access the relevant department and their personal folder.
  o With the Access Control System the staff have restricted access to the other departments and areas physically which are not related within the organization premises.
  o With the firewall, users have been restricted and limited to access unwanted websites.
  o IDS and IPS systems are enabled in the firewall.
- Backup and replication is done using Veeam Backup.
- Desktop environment is secured with F-Secure Antivirus System
- Applications are hosted on DC and managed by IT Support team
  o Document Management System – ELO
  o COSEC system
  o Financial System (custom made)
  o CCTV Monitor
  o WDS (Windows Deployment Services)
  o WSUS (Windows Update Server)

- Geographical Information System – ARC GIS
- Veeam Backup and Replication
- Juno Space
- Sonic Wall Analyzer
- Spice Works Application
- System Centre Configuration Manager

- Existing IT Infrastructure is utilized to its maximum.
- GCC-Stat is mainly using a Failover environment for the critical Network Services as a high availability solution. DR and BCP solutions are recommended based on Business need.
- New separate IT Infrastructure recommended for the Implementation of E2E Integration System at GCC-Stat.

Aspects to be taken into account during the elaboration of the E2E solution

This section presents aspects to be taken into account from findings described above.

Data collection is a phase consuming most of the time and resources in statistical data life cycle. Statistics produced depends a lot on the quality of the collected data. Therefore, several aspects listed below should be taken into account when elaborating E2E solution.

The process of data collection has to be based on advanced methods used for data collection in official statistics:

- Web based data collection using predefined web forms:
  - The continuation of the use of currently used data transmission tables as web forms.
- Newly elaborated web scraping tools to be used for web scraping data collection.
- Implementation of the SDMX based data collection using SDMX RI in GCC-Stat and in MS:
  - Necessity of significant improvement of cooperation between GCC-Stat and MS. Organization of the cooperation based on adopted legal acts.
- Organization of data collection as fully automated process.
- The primary validation of the collected data should be a part of a data collection process thus ensuring a storage of true data in the GCC-Stat Data warehouse.

Data processing, analysis and reports preparation have to be organized as:

- An integrated data management on the basis of GCC-Stat corporative data warehouse supported by necessary application set;
- Excel, which will continue being used as a widely-spread format for data import and export as well as statisticians' desktop analytical tool.

For data dissemination, the Prognoz data portal is considered to be integrated as data dissemination component into the E2E solution and the whole E2E solution based system should be run on internal GCC-Stat IT infrastructure.

It is important to underline that workplans for legal cooperation agreements between GCC-Stat and member countries have to be elaborated and implemented in order to ensure necessary improvements of timeliness and accuracy of collected data and metadata.

The current IT Infrastructure is supporting current requirement of business and operations, in order to have high performance, efficiency and stability of the proposed solution, we would recommend to have new IT Infrastructure.

3.2. Short presentation of End-to-End options

Three different options for E2E solution were elaborated based on the existing advanced IT Technology and available software solutions:

1. Statistical Analysis Software (SAS) option;
2. **Metadata driven option**;
3. Oil and Gas industry option.

In agreement with GCC-Stat 1 and 3 options were discarded due to high licence costs. And instead the **Prognoz integrated platform option** already used within GCC-Stat was elaborated.

Options presented are elaborated based on the assumption that main organizational and legal issues related to data and metadata collection from GCC member countries are implemented and member countries provide GCC-Stat timley comparable high quality data.

The two remaining options (**Metadata driven and Prognoz integrated platform options**) were presented and discussed with the GCC-Stat leading staff. All GCC-Stat requirements have been taken into account and elaborated in the final version of both options. Elaboration of the E2E solution options was based on:

- The assessment results formulated as findings and aspects to be taken into account within the further steps of the feasibility study (Section 2.2);
- The modified GSBPM version proposed for the use within GCC-Stat (detailed description presented in Section 5.3.1);
- The detailed analysis based on available information of the IT usage in the area of Official Statistics – National Statistical Offices. Analysis of results on other IT solutions is presented in Annex 9.

- **Metadata driven option**

This solution is characterized as a new generation statistical data processing system which integrates several subsystems, and is Metadata driven. Innovations which have been elaborated within the process of the system design and development was a standardization of the statistical data production processes (full compliance with GSBPM) and usage of the statistical structured metadata as the driving force for automatic applications generation to run above-mentioned processes.

The new system is developed as centralized system, where all data is stored in corporate data warehouse. This in practice means a movement from classic stovepipe to process oriented data processing approach.

Metadata driven data processing system is based on systematic description of all metadata on surveys, indicators and classifications used, their storage in the centralised metadata base, thus ensuring preconditions for definition of entire system functionality and automatic applications generation. Metadata becomes as a key element of the entire system.

The system is built on principles of client server architecture and has a modular structure in MS.Net environment. The system handles MS SQL Server databases running on OS MS Server platform 2003 and higher.

There are no specific requirements for IT infrastructure to run the solution and a standard software is sufficient. It is necessary to have a Windows Server standard edition starting from version of 2003 and MS SQl Server Standard edition starting from 2005. Capacity of the IT infrastructure depends on data volumes to be processed. Detailed description of the option is presented in Annex 5.

- **Prognoz Statistical Data Management platform option**

Prognoz Statistical Data Management (SDM) is designed on Prognoz software platform to harness large amount of statistical data. It enables to fasten and simplify data collection, validation, and analysis. It also provides a powerful toolkit to manage a statistical data warehouse, monitor and analyze desired information, generate reports, and deliver data to a wide range of users.

The proposed solution combines a full-featured desktop data management application; a Web-based module for statistical data collection, population surveys, and census; and a Web-based data visualization and analytics application available to external users.

Prognoz SDM provides powerful designer of indicator structure, including dictionaries and such attributes as indicator mnemonics, full descriptor, labels, etc. You can add desired metadata for further data analysis, transformation, validation, and report generation. A global metadata environment enables to build a common metadata dictionary across data sets from different sources.

A list below contains the necessary standard software:

1. Microsoft Windows Server 2003 R2/2008/2008 R2 Enterprise Edition
2. Microsoft SQL Server 2008/2008 R2 Standard Edition (64-bit)

As for the Metadata driven option, capacity of the IT infrastructure for Prognoz SDM also depends on data volumes to be processed. Detailed description of the option is presented in Annex 6.

It is important to mention that all elaborated options have a number of common features:

- Generic IT architecture is able to ensure E2E solution functioning regardless **the software platform that will be used for E2E solution implementation** (IT architecture is presented for each option and added to their detailed descriptions in Annexes 3, 4, 5). The chosen software platforms and approach for the design of E2E solution make options different from each other.
- Data collection tools like web forms, web scraper and SDMX are the same for all options and comply with the GCC-Stat requirements.

SDMX is an international standard for statistical data and metadata exchange. SDMX-ML is the data structure used in SDMX. This machine-readable structure, therefore usage of that structure, throughout the whole chain of data production processes (E2E solution) would lead to designing the entire system from scratch. SDMX is a very complex system that requires accurate fulfilment of all cooperation requirements. The full description of SDMX implementation is provided in Annex 8.

- Foreign trade statistics is a specific domain in official statistics that substantially differs from all others due to high volumes of data. Structurally Foreign trade statistics is treated as an integral part of the E2E data processing system. For data collection, the same mechanisms as for data transmission tables can be used by operating with data import facilities. Technically data should be stored in a separate database. This would allow direct interaction with data in database for specialists without any risk of disrupting data of other domains. Furthermore, due to high volumes of data, the processing of Foreign trade data would not disrupt the performance of other statistical domains data processing if common data warehouse would be used. Also, specific data processing requirements could be the case for the domain and usage of a separate database would allow easier adjustment of metadata functionality as well as data processing / aggregation algorithms without affecting other statistical domains. However, foreign trade statistics data processing subsystem should be treated as an integral part of the system and should not raise the implementation costs in any proposed option.

- Assumptions stated are relevant for all three options.

- Specific data that differs from option to option is reflected in the options descriptions in Annex 4, 5, 6. Exception is SAS option:

   o During the option evaluation process, the prices of SAS modules licenses, implementation and training costs for Middle East region were not provided by a local branch located in UAE. The prices obtained are of Eastern Europe region and therefore they are just showing a range of price.

   o Another reason is a decision taken to exclude SAS option from detailed evaluation described above.

- Training for IT specialists as well as for statistical staff is foreseen for both evaluated options (Metadata driven and Prognoz platform options).

- In both evaluated options data and metadata storage is statistical data warehouse. The warehouse is built up on MS SQL Server database management platform. MS licenses types and a number of them are the same.

- Structure and costs of IT infrastructure that should ensure stable functioning of the solution implemented for five years has been reflected at the end of Section 5.3.5. The capacity of required IT infrastructure mainly depends on data volumes to be processed and not so much on software tools used. Moreover, taking into account that both options are using the same database management system, it is possible to affirm that necessary IT infrastructure is the same for both options. Equal costs were excluded from evaluation without affecting the results of the evaluation.

- In order to ensure an efficient functioning of IT department it is necessary to expand it by three additional IT specialists:
  - For Metadata driven option the specialists will maintain the system SQL Server based data warehouse and assist statisticians in creating metadata descriptions. It is important that specialists are familiar with MS SQL databases and have an understanding of metadata role in official statistics,
  - For Prognoz integrated platform option specialists will maintain the system SQL Server based data warehouse and the overall Prognoz platform. The specialists have to be familiar with MS SQL databases and have knowledge of Prognoz platform structure, are able to maintain it and to configure new data sets production environment. Knowledge and understanding of official statistics would be an advantage.
- Explanations of ratings for calculating the benefits criteria is reflected in Annexes 4, 5. Explanations of ratings for other criteria are reflected in the following chapter.

3.3. Options evaluation criteria

Options evaluation criteria were elaborated presented, discussed and approved as follows:

- **Benefits**

Tangible and intangible benefits, which GCC-Stat expects to achieve upon implementation of the solution. This criterion was calculated in the pre-evaluation table for each option and results were added to the options evaluation table (see Section 3.3). Rating frame is 1 to 10 with weighting coefficient 1.

- **Applied to official statistics (best practices)**

Comparison on how solution of each option meets the best practices in the area of official statistics worldwide. Rating frame is 1 to 10 with weighting coefficient 0,9.

- **Necessary investments**

Comparison and rating of investments necessary for acquisition of software and additional hardware[5]. Rating frame is 10 to 1 with weighting coefficient 1.

- **Training provision complexity**

Comparison and rating of necessity for trainings and their complexity. Rating frame is 10 to 1 with weighting coefficient 0,8.

- **Qualified IT professionals available in the market**

Comparison and rating of the availability of necessary additional staff in the employment market. Rating frame is 1 to 10 with weighting coefficient 0,7.

- **Application development complexity**

---

[5] In case necessary hardware is equivalent for comparing options then the cost of it would be excluded from comparison and evaluation.

Comparison and rating of application development complexity with the aim to expand functionality of the solution. Rating frame is 10 to 1 with weighting coefficient 0,8.

- **Maintenance complexity**

Comparison and rating of solution options from the maintenance complexity point of view. Rating frame is 10 to 1 with weighting coefficient 0,8.

- **Product ownership**

Comparison and rating of the software system source code ownership. Rating frame is 1 to 10 with weighting coefficient 0,5.

- **Impact on member countries**

Comparison and rating of the impact to data providers of GCC member countries. Rating frame is 10 to 1 with weighting coefficient 0,7.

- **Risks evaluation**

Comparison and rating of risks[6]. Rating frame is 10 to 1 with weighting coefficient 0,5.

Explanation of rates used for evaluation

Criteria No	Ratings explanation
Benefits	Detailed rating explanation is provided in benefits table in Annex 5, 6.
Applied to official statistics (best practices)	Integrated metadata driven data management solution has been implemented in Latvia, Ireland, Croatia and Serbia. Solution has been presented in several International conferences, UNECE Work sessions and Eurostat. There are no data available on Prognoz SDM implementing as E2E solution in official statistics
Necessary investments	n/a
Training provision complexity	Only a general and brief training for Metadata driven option is necessary. Statisticians work in a dialogue with the system. For Prognoz SDM platform option statisticians must be trained in order to have sufficient knowledge to understand system's architecture.
Qualified IT professionals available in the market	For Metadata driven option IT specialists must be familiar with MS SQL Server database development and administration. For Prognoz SDM platform option IT specialists in addition must have skills to run the Prognoz platform environment.
Application development complexity	To add new functionality to the system requires programming for both options
Maintenance complexity	For Metadata driven option IT specialists must be familiar with MS SQL Server database development and administration. For Prognoz SDM platform option IT specialists in addition must have skills to run the Prognoz platform environment.

---

[6] Risks identified for the comparing options should be equal but rating of them could differ.

Criteria No	Ratings explanation
Product (source code) ownership	To secure Metadata driven option implementation the system source code is provided. This is a substantial benefit in cases when product developer stops vendoring the product, client is capable to continue running the system by maintaining and developing a system on his own or by handling the source code to other developer company. Prognoz does not provide a source code.
Impact on NSOs of MC	Impact is minor in both cases
Risks evaluation	Risks listed in Section 6.2 are mostly the same and only likelihood differs for the reasons already described.

3.4. Options evaluation table

Criteria No.	Criteria	Rating frame	Weight	Option 1 Metadata driven		Option 2 Prognoz SDM	
				Initial Rating R	Final rating RxW	Initial Rating R	Final Rating RxW
1	Benefits	From benefits table	1	7,2	7,2	6,2	6,2
2	Applied to official statistics (best practices)	1 to 10	0,9	6	7,2	5	3,6
3	Necessary investments	10 to 1 ($LI^\wedge$ = 10, $HI^{\wedge\wedge}$ =LI*10/HI)	1	4,4	4,4	10	10
4	Training provision complexity	10 to 1	0,8	7	4,8	6	5,6
5	Qualified IT professionals available in the market	1 to 10	0,7	8	7	6	2,8
6	Application development complexity	10 to 1	0,8	4	5,6	4	3,2
7	Maintenance complexity	10 to 1	0,8	8	6,4	6	4,8
8	Product (source code) ownership	1 to 10	0,5	10	9	1	0,5
9	Impact on NSOs of MC	10 to 1	0,7	8	5,6	8	5,6
10	Risks evaluation	From risks table	0,5	6,9	1,6	6,2	1,7

Criteria No.	Criteria	Rating frame	Weight	Option 1 Metadata driven		Option 2 Prognoz SDM	
				Initial Rating R	Final rating RxW	Initial Rating R	Final Rating RxW
	TOTAL			n/a	54,2	n/a	49,4

LI^ – Lower Investment

HI^^ – Higher Investment

4. Gap analysis

This section contains a gap analysis of the statistical processes at GCC-Stat – which, in fact, is a guide to facilitate the formulation of specific recommendations – should be read as complementary to the gap analysis prepared for the IT system gap analysis, which is more precise in what regards the identified gaps and the action plan proposed.

4.1. Gap analysis for the improved statistical processes

The aim of this section is to reduce the list of divergent aspects – prepared in Section 1.3 – into few narratives / wide objectives, transcending the delimitations of the areas of intervention used in the questionnaire for the collection of information that will support the formulation of straightforward recommendations.

This section can be read as a gap analysis exercise – as an analytical process used to help an institution to understand and quantify the gaps that exist between its ideal future state and its present state. By analysing these gaps, the management can create specific action plans to move the institution toward its goals and close the gaps identified in the exercise.

The gap analysis of the statistical processes facilitates the formulation of recommendations to GCC-Stat for the improvement of some aspects considered as lagging behind in the continuous development process of GCC-Stat. Therefore, it does not offer a concrete action plan, but alternatives to the current standing.

The main narrative / wide objectives are:

- Strengthen the current cooperation agreements with member countries in order to ensure the availability of harmonized timely statistics in specific statistical domains (allowing regional and international comparisons) and in line with expressed user needs.

- Continue the improvement of the dissemination strategy by GCC-Stat, incorporating or developing additional elements such as: a communication strategy, visualization tools, quality information about the data disseminated, etc.

- Promotion of the culture of statistical harmonization and quality across GCC member countries and in all the statistical domains, by the use of common metadata and the implementation of adequate quality frameworks.

- Development of automatic procedures and IT tools to facilitate to GCC-Stat the automatic data correction routines and data processing.

- Building of statistical capacity in order to enable GCC-Stat and member countries staff the development of methodologies and common GCC standards, and the enhancement of their skills and competences.

- Promotion of evidence-based decision making focusing on the use of statistical data and and facilitating the access and the use of statistical data by all the users (including as well the dissemination / promotion of statistical literacy).

4.2. Gap analysis for IT infrastructure

The development of IT infrastructure solution is a core stone for all national statistical offices which ensures and increases the efficiency of work, quality and timeliness of the statistics produced as well as reduces the burden of statisticians. In the case when Integrated E2E solution has to be implemented it is necessary to discuss a business case change. It is a great step forward to move from manual statistics production using MS Excel and File server as data storage towards an Integrated E2E solution where most of processes are automated and data is stored in corporative data warehouse. It is important to point out that the strategic objective of institutions in official statistics is not a creation of IT systems but production of timely and high quality statistics that is possible with the exploitation of advanced IT systems and tools. This section shows the main steps to be considered in order to achieve strategic objectives that will support the implementation of the chosen integrated E2E solution.

Strategic objective of the winning solution	Current standing (including operation resources)	Deficiency (gaps / lack)	Action plan (including needed optimal operational capacities)
To raise quality and timeliness of produced statistics	Statistical data suppliers, GCC member countries, provide GCC-Stat with data / metadata they produce on a voluntary basis. Thus, in many cases it may lead to delayed data provision, processing and publishing at national and regional levels	Lack of legal acts (bilateral agreements) on collaboration between GCC-Stat and member countries.	1. GCC-Stat elaborates bilateral agreements for collaboration with member countries where data collection methodology, formats, time frames, etc. are defined. 2. GCC-Stat undersign bilateral agreements with member countries, thus defining methods, formats and time frames for data collection on a legal level mandatory for all parties involved
Automate as much as possible data collection process by introducing web forms, web data scraping technology and SDMX standard. Combine data validation with data collection.	Data are collected using data transmission tables (not for all statistical domains), searching web pages, manually entering from printed publications and using customs declarations for Foreign trade statistics	Lack of advanced technologies and knowledge	1. Data collection subsystem of the new Integrated E2E solution will support the transition of data transmission tables to data collection web forms. 1.1 Organize trainings for statisticians on statistical metadata 1.2 In parallel with the software system adaptation and implementation, organize introductive seminars for statisticians 2. Create a new position for a specialist and provide a training on work with web data scraping module. 3. Create a group of specialists who will be responsible for SDMX standard implementation for data collection and dissemination. 4. Organize seminars for relevant specialists of NSOs of member countries to harmonize methodology and rules supporting the implementation of the

Strategic objective of the winning solution	Current standing (including operation resources)	Deficiency (gaps / lack)	Action plan (including needed optimal operational capacities)
			SDMX Standard.
Fasten production, ensure statistical, confidentiality, use advanced statistical analysis tools	MS Excel is the only tool for statistics production. Few SPSS licenses are purchased. Training on SPSS is organized. Data files are stored on file server which determines a manual search.	Lack of technologies and knowledge	1. The implementation of the new Integrated E2E solution will transfer data production from file server environment to Data Warehouse environment. Data sets will be easy to find by navigating through Metadata base. Import export procedures will allow exporting data sets to special analytical tools like SPSS, work with data and afterwards importing them back to Micro or Macro databases. Confidentiality checks will be elaborated. 2. In order to ensure successful implementation and use of all production features of the system, a substantial training will be provided to statisticians as well as to IT specialists.
Disseminate produced statistics in user friendly way. Provide users with data analysis and visualization tools.	Prognoz Data dissemination portal is in use within GCC-Stat. Portal ensures all advanced requirements for statistical data dissemination. Knoema, a company located in US provides a technical support. Portal is installed on US based Cloud environment. Portal does not support Reference metadata Dissemination.	Lack of the necessary link between the portal and production environment that would ensure reference metadata dissemination together with the related data. Production is organized in a stove pipe mode.	1. The new Integrated E2E solution establishes a link between reference metadata storage and data dissemination portal. Statistical indicator would help as an identifier. 2. The instalment of the new system and the existing data portal at the same environment is necessary. The system can be installed in cloud environment as well as in the GCC-Stat internal IT environment. In case of cloud based environment, the cloud should be located in the GCC region. 3. In order to ensure an independent run of the system the training to GCC-Stat IT specialists should be provided.
Create an advanced in-house IT infrastructure to ensure stable and sustainable running of E2E solution	Existing infrastructure described in the assessment report is not foreseen fotr running a new E2E solution.	Decision of GCC-Stat management	New IT infrastructure will ensure full support for running of E2E solution in an advanced IT environment that will allow: 1. The implementation of Virtualization Solution, best utilization of Infrastructure at optimized cost. 2. The operation of main servers on high availability (cluster) solution, thus significantly ensuring redundancy. 3. The common storage system which will ensure the efficient and safe data storage from statistical data warehouse

Strategic objective of the winning solution	Current standing (including operation resources)	Deficiency (gaps / lack)	Action plan (including needed optimal operational capacities)
			and from supporting servers. This will ease data back-up. 4. The use of supporting servers (virtual servers) for development and adaptation of business software system as well as web servers.
In-house E2E solution implementation. Flexibility of enhancement.	GCC-Stat is using cloud-based solution, hosted in cloud environment. The current solution supports only Data Dissemination. Data is collected manually.	Lack of technology and knowledge	Proposed integrated E2E solution – MD (Metadata driven) SDMX 1. Set up of a new infrastructure without disruption with the current infrastructure 2. Implementation of the Virtualization Solution, best utilization of the infrastructure at optimized cost 3. Secured environment 4. High availability solution 5. Dedicated specialist for managing a new infrastructure 6. Leverage of the current back-up solution for taking back-up. 7. Enhancement and customization based on changing Statistical Environment / requirement.

5. Recommendations

Annexes of this section: *Annex 7. Presentation of multi-national statistical systems.*

5.1. Recommendations for the improvement of statistical processes within GCC-Stat

This section builds on the main findings identified at the end of the assessment of the statistical processes phase (see Section 1.2). Accordingly, the recommendations will directly respond to the key findings identified during the information collection and analysis phase.

As a first step of the approach followed for the formulation of recommendations, the main findings have been classified into 'aspects to strengthen' and 'aspects to improve', based on whether they are capable to impact in a favourable or unfavourable way GCC-Stat's plans to build an effective statistical business system and process (see Section 1.3).

It is highlighted that for the classifications of the main findings, the separation between the areas of intervention identified during the collection of information (i.e. institutional environment, user needs, use of international classifications and standards, etc.) was not be kept. This helps constructing a narrative argument supporting recommendations for transversal actions and integrated measures, but also helps to see how some 'aspects to strengthen' can be balanced by specific 'aspects to improve'.

Based on the comprehensive view of the two divergent categories, consolidated recommendations are formulated in this section. The recommendations could be prioritized, following the feedback received from GCC-Stat and in collaboration with the project team. An important criterion that will be considered for the formulation of recommendations is their feasibility, in terms of both operational feasibility and political / organizational feasibility or will.

For the prioritization of the recommendations – if requested by GCC-Stat – the following considerations will be assessed:

- Is it operational feasible?
- Is there political / organizational will or any potential opposition?
- Does it fill an existing gap or need?
- What is the likely impact?
- Are resources available (human, IT, financial)?
- Does it require a policy reform?
- What is the estimated time frame to implement it?

Finally, Annex 7 concludes with the exemplifications of existing solutions in other multi-national statistical systems. These should not be understood as a benchmarking framework, for GCC-Stat system to be measured against, but as best practices in similar environments. The future

implementation of any of these would be in any case subject to a feasibility assessment and tailoring to GCC-Stat needs and specificities.

5.1.1. Consolidated recommendations

For each narrative / wide objective the following recommendations are proposed:

Strengthen the current cooperation agreements with member countries in order to ensure the availability of harmonized timely statistics in specific statistical domains (allowing regional and international comparisons) and in line with expressed user needs.

- Enhance the revisions of the law and/or the agreements for collaboration.
- Implement standardized and harmonized statistics laws that govern all statistical work practice in the GCC region.
- Elaborate strategic goals and plans and publish these regularly.
- Prepare a multi-annual programme for the production of statistics at GCC-Stat level in collaboration with member countries;
- Carry out regular user surveys for all the statistical domains GCC-Stat is disseminating statistics;
- Plan the support member countries need to transmit the data to GCC-Stat;
- Support member countries in providing the statistics to GCC-Stat by designing specific transmission tables for all the statistical domains GCC-Stat is disseminating statistics;
- Develop the processes of harmonization of definitions, concepts and methods used in GCC region and develop standard classifications, statistical methodologies for uses in the GCC region by the expansion of Standing Committees activities and establishment of new Working Groups to cover the other statistical domains.
- Ensure cooperation and interaction between Working Groups through the development of crosscutting activities.
- Organize interim meetings (i.e. semestrial or quarterly) of the Working Groups, in order to improve the organization and the follow-up of activities.
- Take into consideration possible synergies with other regional/international projects and initiatives when establishing the Working Group action plans.
- Implement procedures to prioritize between different users' needs in the work programme.
- Establish service level agreements or similar arrangements with most important users.
- Carry out a periodic evaluation of the work programme to identify negative priorities and emerging needs.
- Define improvement actions arising from the user satisfaction surveys and schedule it for implementation.
- Implement specific measures to assess satisfaction of the key users with particular or new products (e.g. specific user satisfaction survey / indicators on product level or organizing meetings / seminars / workshops with these specific users).

Continue the improvement of the dissemination strategy by the GCC-Stat, incorporating or developing additional elements such as: a communication strategy, visualization tools, quality information about the data disseminated, etc.

- Present data comprehensively and improve the consistency of data obtained from various sources or statistical offices, their detail and temporal and geographical comparability.

- Include quality information for the data disseminated by GCC-Stat (such as Quality Indicator Profiles).

- Present the data in various release formats, offering the users the possibility to tabulate variables and time points according to their needs.

- Make available the use of interactive tools and graphical / mapping presentations.

- Use different dissemination and communication channels for the effective dissemination of statistical products and services.

- Develop the understanding and use of statistical products by taking into account elements of statistical literacy.

- Prepare a communication strategy for statistics.

Promotion of the culture of statistical harmonization and quality across GCC member countries and in all the statistical domains, by the use of common metadata and the implementation of adequate quality frameworks.

- Review and develop the current quality control procedures to cover all steps of the statistical production process starting from data collection to data dissemination in GCC-Stat.

- Encourage the member countries to take the necessary steps towards adhesion to Special Data Disseminations Standards (SDDS) of IMF.

- Disseminate the metadata on the website of the GCC-Stat for all the statistical domains with more details on the process of aggregation. In addition, it is very important to encourage member countries to develop and disseminate the metadata relating to all the statistical domains on their websites.

- Define a quality policy and made it available to the public. In addition, it is recommended to establish an organizational structure and to develop tools to deal with quality management.

- Develop quality guidelines on how to implement quality management within the statistical production process. The guidelines should comprise:

  o A description of the statistical production process and the identification of documentation for each stage, following the GSBPM for statistics or any other equivalent process representation;

  o A description of the methods to monitor the quality of each stage of the statistical production process.

- Made the quality guidelines, as defined above, available to all users at least in a summary version.

- Define and implement standardization programmes and procedures in the main stages of statistical production areas, for example data collection, data processing and data exchange, according to the business process model.

- Develop internal procedures and guidelines for data quality assessment and address accuracy and reliability issues.

- Disseminate all statistical results together with the respective metadata allowing for a better understanding of the results.

- Update metadata regularly and establish procedures to ensure the updating are available.

Development of automatic procedures and IT tools to facilitate to GCC-Stat the automatic data correction routines and data processing.

- Develop IT architecture and strategy and upgrade and update them regularly. In this regard, automation might be a strong incentive for harmonization and standardization. However, for an E2E solution of automation, it is necessary to coordinate, standardize and harmonize the statistical input, throughput and output. It is especially relevant to use the same classifications, concepts, reporting standards and timeliness of the data.

- Develop business software system applications to ensure all stages of the statistical process of all the statistical domains from data collection to data dissemination. This business software system will ensure whole cycle of statistical production.

- Develop a centralized and structured data warehouse.

- Establish policies, procedures and tools to promote automatic techniques for data collection, validation and processing.

Building of statistical capacity in order to enable GCC-Stat and member countries staff the development of methodologies and common GCC standards, and the enhancement of their skills and competences

- Strengthen the skills of staff through continuous training and presence in the regional and international events in this area.

- Implement a methodological infrastructure (e.g. units, committees) which defines statistical methods, monitors their implementation and validates the results. In particular, it defines and makes available standard tools for every stage of the business process model (e.g. collecting, processing data and disseminating.).

- Elaborate methodological documentation for each statistical domain containing all pertinent information on metadata, namely concepts, methods, classifications, and made it public at least in a summary form.

- Enhance the participation of GCC-Stat staff in conferences, seminars and workshops at national or international level on the application of standards, methods, classifications, etc. (e.g. The NTTS conference, the quality conference, etc.).

- Update the staff skills concerning new tools, methods and fields of study.

- Implement an appropriate organizational structure, which will provide guidelines, recommend appropriate methodologies and periodically examine the methods used for all statistical domains.

- Develop and carry out procedures and guidelines to monitor internal coherence in a systematic way. Where appropriate they should deal with consistency between preliminary and final data (i.e. continuity), between individual data and aggregated data, between annual, quarterly and monthly data, between statistics and National Accounts and also with non-deterministic consistency (e.g. consistency between economic growth and employment, also called plausibility).

- Establish procedures and guidelines to ensure combination of outputs from complementary sources and to assure internal coherence and consistency.

- Develop a common repository of concepts.

- Provide training courses for staff on metadata.

- Develop and conduct training courses for writing interpretations and press releases.

Promotion of evidence-based decision making focusing on the use of statistical data and and facilitating the access and the use of statistical data by all the users (including as well the dissemination / promotion of statistical literacy).

- Ensure that all users have equal access to the statistics disseminated by GCC-Stat in the same time.

- Make the release calendar available and accessible to users through the GCC-Stat website.

- Carry out a major development of the GCC-Stat website to become an adequate tool for dissemination of GCC-Stat products and to enable the users the ability to find and access to various data, which they are looking for. Such development should include:
  - Establishment of a link to metadata of the six counties.
  - Development of efficient and interactive search engine.
  - Facilitating self-tabulation in the most appropriate formats (e.g. XLS, HTML).

- Apply routinely guidelines and principles relating to the revision of published statistics and made it known to users. In addition, the revisions have to be accompanied by all necessary explanations and made available to users.

- Analyze the data on the use of statistics (e.g. evaluation of downloads, subscribers of reports) to support priority setting and user consultation.

- Publish regularly the release calendar covering all statistics.

- Develop and implement a policy for archiving statistics and metadata.

5.2. Recommendations for the improvement of the IT infrastructure

Three options for E2E solution were elaborated and presented by DevStat IT experts team:

1. Option 1 – SAS based option
2. Option 2 –Metadata driven approach option
3. Option 3 – Option based on software tools used in oil and gas sector in the region

After the presentation of Options 1, 2 and 3 it was agreed that Option 1 and 3 will not be directed towards further analysis and evaluation due to high costs necessary for their impelemtantion. Therefore, it has been agreed to elaborate one more option (Prognoz Statistical Data Management Solution). Thus, the evaluation of usage of Prognoz platform for Integrated E2E solution was elaborated taking into account that GCC-Stat is currently exploiting Prognoz Data Portal. Thus, two options were evaluated in detail:

- Option 2 – Metadata driven approach option
- Option 3 – Prognoz Statistical data management option

Evaluation process and results of options are reflected in Section 3.1 of this report and detailed descriptions are added to Annex 4, 5 and 6.

Feasibility of each option has been analysed separately and the analysis results were achieved in accordance with the criteria elaborated, rated and compared at the final step of the entire work. Results of the evaluation have showed that:

- Option 2 – Metadata driven approach option was rated with 54,2 points;
- Option 3 – Prognoz Statistical Data Management (SDM) option was rated with 49,4 points.

The applied rating system shows that Metadata driven approach option from many points of view (except implementation costs) is more suitable solution for further development of IT infrastructure in GCC-Stat.

Analysis of the current situation in the area of Official Statistics in both national and international statistical institutes shows that there is no standard software product available for Integrated E2E solution. Each organization develop their own software system for statistics production. In most cases, these systems are not integrated and support only a stovepipe approach.

The proposed solution based on the Metadata driven statistical data management is integrated, very advanced and has a proven ability to work in real statistical production environment. This system has been launched for statistics production in Central Statistical Bureau (CSB) of Latvia in 2002. Since then four major upgrades of the system functionality have been implemented and it is a tangible presentation of the system working capacity.

At the same time, the Prognoz SDM option is elaborated and well-advertised but not implemented yet in any national statistical office. Prognoz SDM solution was implemented for institutions and organizations such as African Development Bank (AfDB), International Monetary

Fund (IMF), OECD, to support them in harnessing large amount of statistical data and to manage data warehouse, monitor and analyse desired information, generate reports and disseminate data to a wide range of users. However, no information was found on the implementation of the solution in national statistical offices. In IMF a specific version of Prognoz is in use and is known as EcOS. It covers various aspects of the World Economic Outlook process, including data collection, validation, transformation, visualization and data collaboration.

Prognoz SDM tool is good for data analysis, reporting and visualization but there is a scarce information as regards the implementation of the Integrated E2E solution except the ones proposed by Prognoz themselves. The Prognoz Open Data Portal is significantly more widespread as a data distribution solution while SDM implementations are just few. There is quite a big difference between Prognoz data dissemination portal that is in use in GCC-Stat and Prognoz SDM. Experience gained by exploiting Prognoz data dissemination portal will not provide a knowledge necessary to operate with Prognoz SDM. Additional training will be required.

Both proposed options are possible to implement within GCC-Stat and both of them will ensure statistical production needs of GCC-Stat. The only differences of these two options is user friendliness and the level of maintenance. In both cases the development of the solution will not require the development from scratch. In case of Metadata driven option the adaptation of already elaborated software system will be needed. And in case of Prognoz option the configuration of the Prognoz SDM tools for GCC-Stat requirements will be required. Numeric evaluation presented in Section 3.3 is in favor of Metadata driven solution.

The information on implementation and maintenance costs for both options was obtained from vendors of the hardware and software. For Metadata driven option costs were obtained from Information Technology Services company iSoft Solutions http://www.isoft.lv/index_en.html, and for Prognoz SDM option – part of information was provided by IT company Knoema https://knoema.com/ and additional information was provided by Prognoz Russia located in Moscow.

The obstacle faced during the feasibility study, was the difficulty to obtain necessary information. Information provided by Knoema was very scarce and inadequate. Moreover, Knoema did not respond neither to our additional request for more detailed information, nor to our reminder and request to provide us the continuous support.

Therefore, we directly approached Prognoz Russia for clarifications and the outcomes are listed below:

- Prognoz Russia, as mentioned by Ms Margaret Ann Salmon in her email of 25 January 2017, are in process of transferring responsibilities of their products to other companies. Furthermore, we were informed that Prognoz is not operating as a vendor and Prognoz Russia did not sign the partnership agreement with Knoema yet.

- The rights to resell, modify, etc. Prognoz platform are still under negotiation with Knoema and ITG Holding.

- As regards Prognoz SDM products group[7] it is still not clear which company will have the rights to sell these products.

All the information obtained from Prognoz Russia indicate that Prognoz SDM tool will not be available for potential users for an uncertain period of time until the vendor will be clarified.

The recommended activities to be considered for the implementation after the finalization of the project are the following:

- On the basis of the results of entire project, to elaborate Technical and Organizational requirements (Terms of reference) for tendering procedure on implementation of the winning option solution;

- To launch tendering procedure for the winning Integrated E2E Solution, it's adaptation and implementation in GCC-Stat;

- To consider the expansion of GCC-Stat IT staff. In order ensure the successful implementation of the new Integrated E2E solution it is recommended to expand GCC-Stat IT team by three specialists with higher educational background in IT, mathematics and / or statistics. They should to have skills in area of SQL databases administration and development, should have a knowledge of official statistics production particularly statistical metadata, should have experience to operate with different data formats particularly with XML. The additional staff should ensure the creation and maintenance of statistical metadata descriptions, structural and reference metadata maintenance as well as the implementation of SDMX-RI. Nevertheless, additional training will be necessary.

- To ensure training of at least two specialists of the current IT staff on MS SQL Server Standard edition administration and development.

- The proposed IT infrastructure shall ensure a stable and sustainable running of the new system, increase its performance and security. The expenses for configuration and for the necessary IT infrastructure upgrade is similar for both options. IT infrastructure capacity depends mainly on data volumes to be processed and not from software tools used for processing. The implementation of the new IT infrastructure from the scratch or the upgrade of existing infrastructure will be a decision of GCC-Stat after internal discussions. We strongly recommend an upgrade of existing IT infrastructure for efficiency reasons. The current servers at GCC-Stat are powerful enough to ensure the running of both, existing working environment and a new software system in parallel.

- To continue working on legislative acts for creating and implementing data exchange among GCC member countries.

- To increase a number of GCC-Stat statistical staff. GCC-Stat is a developing institution and the assessment of the current situation showed that the statistical production is at the development stage. The increase of the data volumes depends on the organization of data collection from respondents in a legally organized manner. The capacity of the current IT infrastructure significantly exceeds (appr. 90%) the level necessary for statistical production

---

[7] Product Statistical Data Management is one that fits the best for GCC-Stat needs. This product forms the backbone for Option 3.

within existing production environment with existing volumes of data. It can be carefully stated that GCC-Stat will not be able to decrease a number of existing statistical staff, and that the increase of it should be considered. A number of necessary additional staff will depend a lot on factors such as the success of the implementation of E2E solution, the quality of trainings for both IT specialists and statisticians, the planning of implementation of statistical domains, etc. Metadata driven solution implementation in CSB of Latvia showed that in a period of 3 years since the system was launched in statistical production environment, the system covered 25 statistical surveys. This was one of the factors to decrease a number of statistical staff by 100 employees.

5.3. Detailed description of best-scoring option

Microsoft based MD-ISDM option

The quality of statistics produced by NSOs depends on methodologies applied for data collection, processing, safeguard and dissemination of statistical information. The management of statistical information is not possible without a support of adequate tools of modern information technologies integrated in statistical information systems (SIS).

The latest activities in the field of the development of SIS witnessed an important move towards SIS that standardize and integrate the tasks throughout the statistical activity. Leading NSOs integrate the data and metadata repositories and processing systems, harmonize the tools used with a view of having efficient, transparent and easy to maintain systems. Nowadas it is a general trend in NSOs.

However, the integration of SIS requires the consideration of various specific needs of individual phases of data processing. Therefore, as a first step in the development of SIS, a creation of a model describing the functions for NSOs should be mentioned.

Using modern information technologies, the system which ensures statistical processing trough the particular model processes performing functions as defined in the relevant sub-processes, have to be created. Thus, the implementation of the standardization and process-oriented approach in the statistical production is common for all kinds of statistics produced.

The model describing functions for NSOs will also contain terminology of data types in order to ease the following of data lifecycle. The term "microdata" will refer to the data that comes from a member country in any form and "macrodata" will refer to the data processed by GCC-Stat.

5.3.1. Generic Statistical Business Process Model (GSBPM)

Specifying needs

Taking an example of any statistical survey, it responds to a concrete demand for statistical information. Therefore, at the beginning of the cycle, it is necessary to specify the need for the statistics. At this stage communication with the potential users of data should take place.

In order to determine the need for statistical information to define the output, the business case and communication software with users is sufficient. However, it is necessary to access the

existing database(s), in order to check which current data are availabile. The standard office software for preparing the business case is sufficient.

Design

Design is the phase when statisticians begin their own work on the future survey, census or other activities related to data sets, transmission tables, etc. Based on the results of the Specifying needs phase it is necessary to design outputs.

The design phase is the first step towards formalizing the requirements. In case of more complex design, it is possible to use a modelling approach such as UML (Unified Modelling Language) which is supported by commercial software like Rational Rose, etc.

Build

The extensive use of the information technology in the statistical business process cycle requires an attention to the information systems and their components. This should be completed well in the early stage of the statistical data lifecycle, and a due attention should be paid to testing. This will ensure that the actual statistical activity will be implemented smoothly, and will help to avoid any delays due to technological issues.

A particular focus is on building and tailoring the tools that are used in data collection and data processing phases. When configuring tools, it is important to take into account the specifities of subject matter domains. Therefore, it is important to have a clear data model. The data model comprises information obtained during phases "Specify needs" and "Design" as well as details of the multidimensional structure of variables and indicators, metadata, code lists, relationships between those components, etc.

Another aspect, when building the information system components, is not only to focus on the current survey, but at the same time to consider the reusability of components for other surveys. *Vice versa*, it might not be necessary to build all components from the beginning, because they may already exist. In simple words, it is preferable to have in mind consistent information architecture of NSO rather than taking ad-hoc solutions. While the latter may give faster results from the immediate perspective, consistent information architecture will pay back in the long term.

Therefore, the components that will be built and / or configured during this phase include data collection instruments and process components. These will be used immediately at the beginning of the survey. Further, the workflows (between components, units, etc.) will have to be configured, and all components and workflows will be tested in order to avoid errors during the survey. After the production systems are finalized the core statistical operation (survey, census, etc.) will start.

Collect

Data collection tools, built in the previous phase, will depend on the sources of data. In case of GCC-Stat where data transmission tables are in use, electronic data collection software is proposed as a solution. Web scraping tool could be in use for getting additional data from the websites of the respondents such as GCC member countries, National Banks, Ministries, etc.

In case of collecting data from administrative registers and records, it would be necessary an off-line module for import and pre-processing of administrative data or a link to the administrative source itself.

The basic data checks will be performed already in data collection phase. The data checks have to be built into the data collection software that would flag questionable records. When using web scraping tool, it is necessary first to check the quality of the website data, and subsequently proceed with the basic statistical checks for outliers or other inconsistent data.

In case of internet based data collection, when respondents themselves enter data into electronic online questionnaires (transmission tables), the questionnaire forms need to be equipped with the basic consistency checks that would flag the inconsistent data and would prevent the respondent from proceeding further. In this respect, it is necessary to stress that the built-in checks in online questionnaires should not appear improper. It is important to focus only on the most important checks. When checking minor discrepancies, they may be considered legitimate from the data quality viewpoint, but the respondent might be discouraged, and might abandon the questionnaire what can lead to a non-response.

Process

This phase combines a large number of tasks that are aimed at protecting the confidentiality, ensuring quality, integrating data with data already existing in databases and performing all necessary calculations and estimations. Therefore, the complexity of IT tools used is much higher at that stage than in previous stages. It is possible to combine the existing statistical packages like SAS and SPSS or Stata with specific modules designed by a NSO using MS SQL Server BI facilities. In all cases, it is important to use an integrated data repository and an integrated metadata system, and provide linking and data sharing between all tools employed in this phase.

The processes listed in the GSBPM are all possible processes in an organization dealing with official statistics. A particular survey or census may not necessarily need to perform all of them.

With respect to the coherence of statistical information system the calculations and estimations should not take place off-line on a statistician's desktop. They should be supported by the system. It is also important to integrate all data into a database, to have a record of the raw data, all intermediary aggregates as well as the final aggregates and estimators. The integration of data also implies the integration of underlying metadata, coding and classifying.

It has been mentioned that basic checks have been done already at the data collection phase. This allows to get back to the respondent to verify inconsistent and / or missing data. In the data processing phase, it should be necessary to use data editing tools that are more profound. This will help to perform more in-depth evaluation of data quality. Unlike in the data collection phase, it will not be necessary to go back to the respondent, but rather replace outliers and missing data with estimates. The errors and edits should be fully recorded, and should be analysed against systematic errors. The goal is to make adjustments to the design of data collection and avoid the same systematic errors in the next instance of the survey.

Finally, calculations and estimations will be needed to define outputs that are deriving from new variables, calculating weights and aggregates.

Analyze

Analyze phase aims at creating desired outputs from the collected and processed data. The use of statistical data analysis tools mentioned above would be preferable than the custom designed ones. The set-up of the technology should take into account that this phase is performed by statisticians from specific subject matter domain. It is suggested to consider the user-friendly technologies as statisticians are more skilled in statistics than in IT. Their substantive knowledge should be reflected when designing the IT tools.

A variety of standard and tailored statistical packages would be integrated into a coherent SIS. This will help to perform statistical analysis and prepare draft outputs.

Interpretation and explanation of the results is based on the personal expertise of statisticians. However, they should have a possibility to access metadata system and record their findings, so that these can be used when disseminating and communicating data to users.

Before the final aggregated data and results of analysis may be presented to users, it should be ensured that the private and confidential information is protected. Appropriate tool for data disclosure control elaborated in the system could ensure that neither direct identification, nor statistical matching can be used to obtain confidential information.

Once the subject matter specialists finalize the outputs for dissemination, these can be passed on to dissemination and communication specialists.

Disseminate

The statistics is significant only when it reaches a user. Official statisticians usually think about the statutory users, who already expect the data to be presented to them. However, public, media, researchers, etc. should not me omitted. Therefore, this phase includes also an outreach to new users through marketing and maintaining public relations. When designing metadata descriptions structures as well as level of detail of data sets, it is important to take into account different types of users (from student to data scientist) to who the data will be disseminated. This phase has a potential to use already existing IT tools for publishing printed or electronic data. However, there is still a need for specific tools that enable statistical capabilities. Prognoz data dissemination portal would be sufficient.

Evaluate

As mentioned above, statistical surveys in general are periodic conducted on a certain time frame basis. However, it is important to ensure that issues and problems tackled in the first round of the survey do not repeat again in the next rounds. Therefore, as mentioned in the process phase, all edited outliers and missing data should be recorded. These records should be then carefully analysed for a pattern of systematic errors. It is essential to ensure data quality and avoid the continuous editing and imputation with respect to the same systematic errors in the next survey rounds.

Similarly, as for the initial phase, the standard office and communication software will be used mostly. However, evaluation phase should not be treated as a separate sequential process,

instead it should be an essential part of all other process groups in GSBPM. The evaluation phase concludes the phases of the GSBPM.

5.3.2. Summary scheme of the GSBPM adapted to GCC-Stat needs

The following scheme summarizes the model presented in the previous section and is modified respecting GCC-Stat needs. It is important to keep in mind that it is a generic model of the statistical business process cycle. Each data set will comprise only of those phases and tasks that are vital for the specific survey. It means that some of the phases and tasks might be skipped for some surveys.

Figure 6: Generic Statistical Business Process Model (GSBPM) adapted to GCC-Stat needs
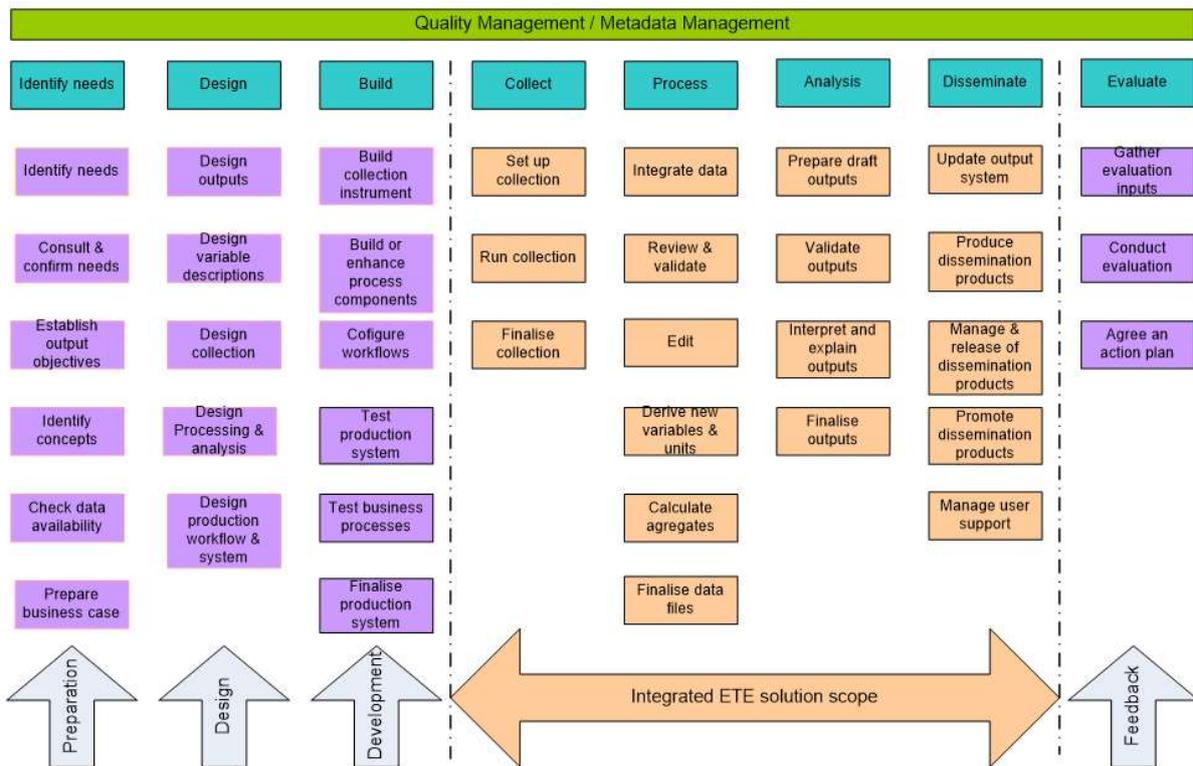
5.3.3. High level architecture of Metadata Driven Integrated Statistical Data Management System

Metadata Driven Integrated Statistical Data Management System (MD-SDMS) system would be characterized as a new generation of statistical data processing system which integrates several subsystems, and is Metadata driven. Innovations which have been elaborated within the process of the system design and development was a standardization of the statistical data production processes and usage of the statistical structured metadata as the driving force for automatic applications generation to run above-mentioned processes.

The new system is developed as a centralized system, where all data are stored in corporate data warehouse.

This in practice means movement from classic stovepipe to process oriented data processing approach.

Metadata driven data processing system is based on systematic description of all metadata of surveys, indicators and classifications used, storage in the centralized metadata base, thus ensuring preconditions for definition of the entire system functionality and the generation of automatic applications. Metadata becomes as a key element of the entire system.

The system is built on principles of the client server architecture and has a modular structure in MS .Net environment. The system handles MS SQL Server databases running on OS MS Server platform 2003 and higher.

The following business application software modules cover and support all phases of the statistical data processing:

- Core metadata base module – the key part of the system ensures metadata collection and storage, defines all entire system processes (starting from data collection and ending with output reports preparation). All system software modules are linked with the Core Metadata module.

- Web based data collection module – ensures electronic data collection via web.

- Wed data scraping module – ensures data collection from predefined web pages.

- SDMX reference infrastructure module – ensures data collection from respondents and data interchange with international organizations.

- Microdata validation and analysis module – generates date entry and validation applications, executes validation and data editing processes, and stores clean data sets in Micro Database.

- Data analysis module – via microdata exports to MS Excel and / or Access, ensures data analysis processes by MS OLAP tools and MS SQL Server Enterprise edition BI analysis tools, ensures a wide range of analysis at micro and macro data level.

- Data aggregation module – ensures data aggregation on different conditions and storage of the aggregated data sets in Macro Database.

- Data import export module – ensures data interchange on both micro and macro data levels

- Data Dissemination module – Open data portal that allows users to access necessary data and related metadata as well as provides a wide range of data visualization tools.

- Users administration module – administrate users' roles and rights.

The MD-SDMS high level architecture has been elaborated respecting the GSBPM described in previous sections and is presented in Figure 5.

Figure 7: MD-SDMS High Level Architecture on the basis of MS Business Intelligence and Metadata Driven Approach
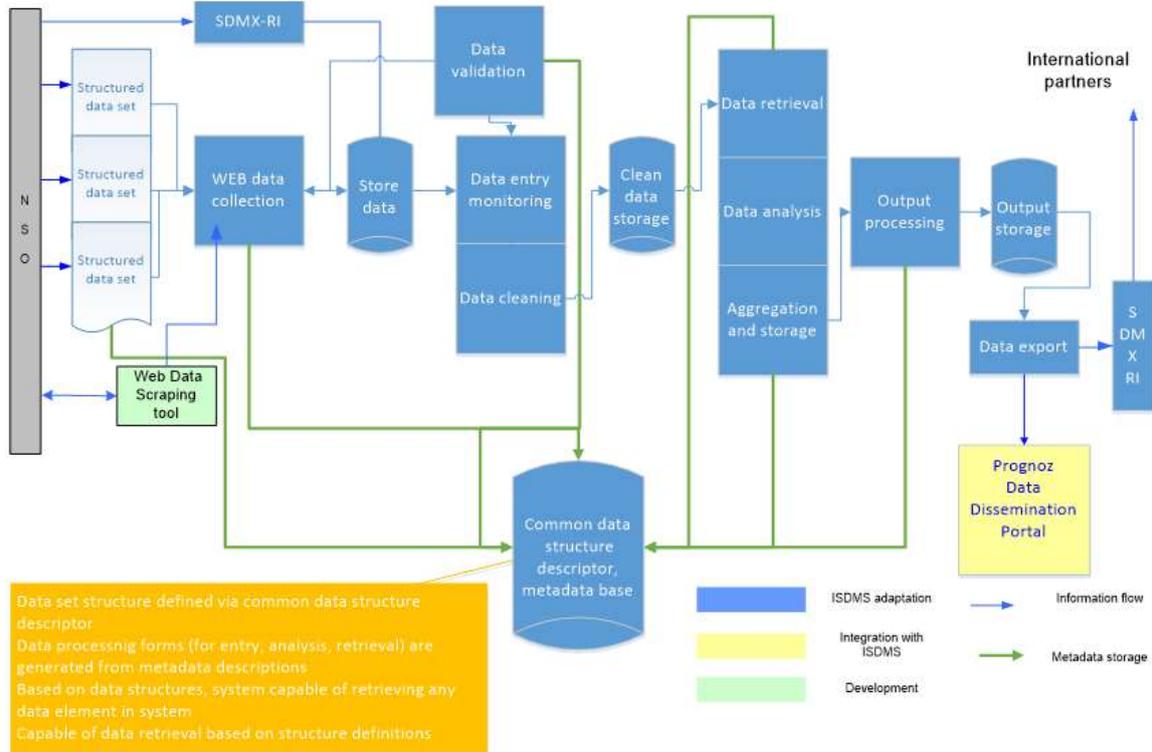
MD-SDMS functionality is reflected in Figure 6.

Figure 8: MD-SDMS structure – data warehouse links with application subsystems (modules)
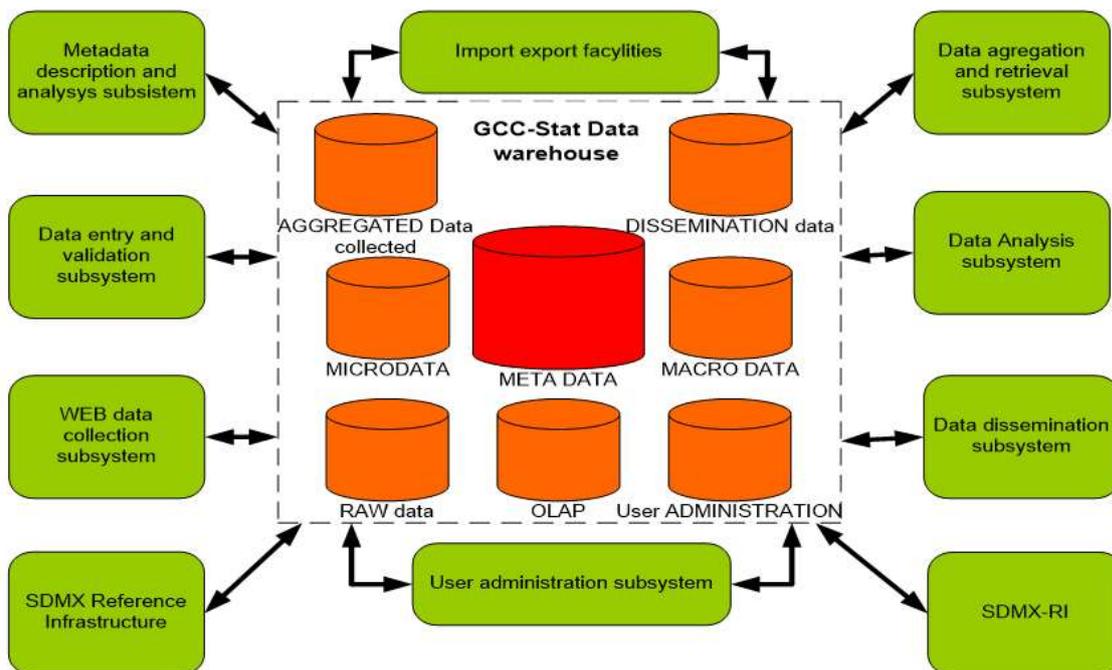
Further detailed description of the MD-SDMS is presented in the breakdown of three main processes: Data Collection, Data processing and Data dissemination. These processes will incorporate related sub-processes.

Data collection

To ensure the highest quality of data collection three different technologies will be implemented for parallel use:

- Online data collection using web forms (data transmission tables);
- Data scraping from pre-defined websites;
- Implementation of SDMX -RI[8].

Online data collection using web based data transmission tables

To ensure programming free environment using metadata driven approach, any activity have to start with the process metadata descriptions entered to the metadata base. Use of structured metadata descriptions will provide a possibility to ensure automation of data collection and primary data validation processes using online web forms (data transmission tables in case of GCC-Stat). Figure 7 illustrates the main metadata description steps and structure within the data collection process.

Figure 9: Metada description

---
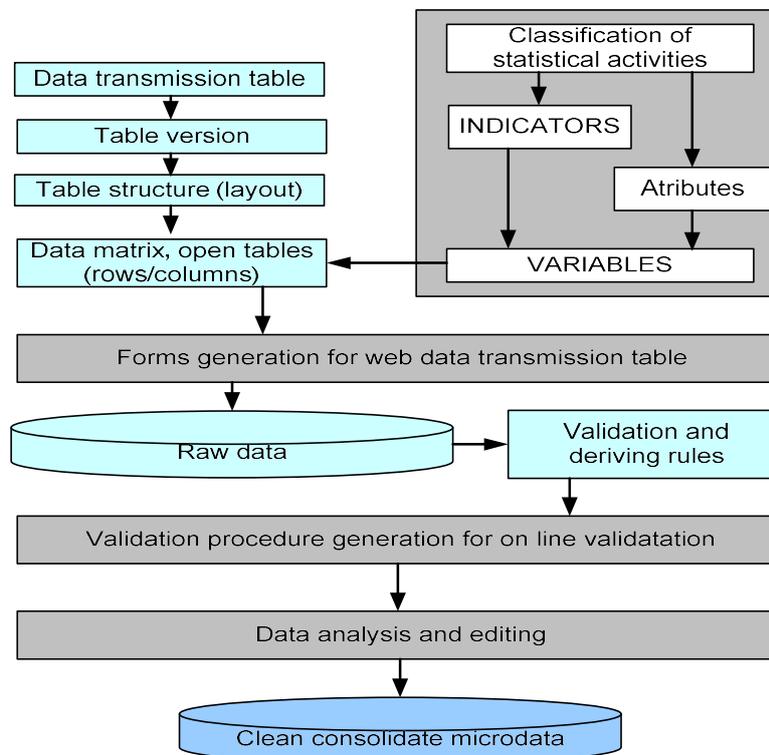
[8] SDMX-RI should be implemented on both the respondent side and the date receiver side to ensure data exchange.

The figure illustrates metadata functions. These functions ensure necessary metadata descriptions entry, which are used for the generation of automatic data entry forms. Data processing steps, which are driven by metadata rules, are illustrated from data transmission tables entry process to clean microdata preparation process. The hierarchical data selection approach will be used for data querying and analysis. On the top level of this hierarchy there is a list of all data transmission tables that are registered in the system. The next hierarchy level represents years for which existing survey data are available. The next detailed level is made up of a list of chapters (if exists) for each registered data transmission table. Further, under each chapter there is a list of available cells (smallest data entry units) with coordinates (row and column code) and variable names that are ready for data selection. Thanks to the hierarchical approach, it is possible to combine data from any survey and from any time period stored in the data warehouse and create complex data queries.

Electronic online data collection module is based on web data entry applications, including survey design and preparation, special data validation algorithms, automatic request sending, response check and management.

The core element for Electronic Data Collection Module is web based data entry and validation forms that are used for different data transmission tables. Responses (completed forms) are transferred to GCC-Stat through internet. Certain features of electronic surveys that contribute to the increase of data quality can be checked immediately.

The business and design goals for Electronic data collection module are:

- Design and preparation of electronic tables in automated mode;
- Collection of necessary data in electronic form from respondents;
- The improvement of the quality of collected data by using online validation rules that were missing in forms sent by email;
- Automatic request to respondents and automatic response control, implementation of reminder system to respondents;
- The removal of unnecessary stage of recopying of information by GCC-Stat personnel.

The Electronic data collection software module key features are:

- Use of Metadata provides universal approach for generation of the web based data entry forms, which can be done without participation of IT specialists;
- Management system for web forms is created, including version control;
- Response control application allows defining automatic reminders / requests, sending timetable and checking responses;
- Software module ensures registration of respondents and defines detailed access rights for them;
- Web forms offer the following new features for respondents:
  - Pre-loaded data: respondent or survey specific data (e.g. respondent's name and address);
  - Feedback data: historical data is available;

- o Automatic validation: Web forms should include validation rules;
- o For periodical surveys in Web based applications respondent has a possibility to access previous years' data;
- o During data entry process in-form validation should be provided;
- o Where necessary respondents can search, and use classifications such as NACE, PRODCOM, etc.;
- o The respondent is able to print questionnaires for internal use;
- o The respondent is able to see a list of data transmission tables to be filled with the data as well as the information on the submission deadline;
- o Filling in of the web questionnaire is possible in both ways, manually and importing data from XML file;
- o Help facilities.
- • System provides means of security by using user access rights control and data transfer via HTTPS protocol;
- • GCC-Stat collects all data in the raw database, which ensures a transfer of the data to the Micro database to continue data processing in the system.

Detail functional diagram of the Web Data Collection System (WDCS) is provided in Figure 8.

Figure 10: Web Data Collection System
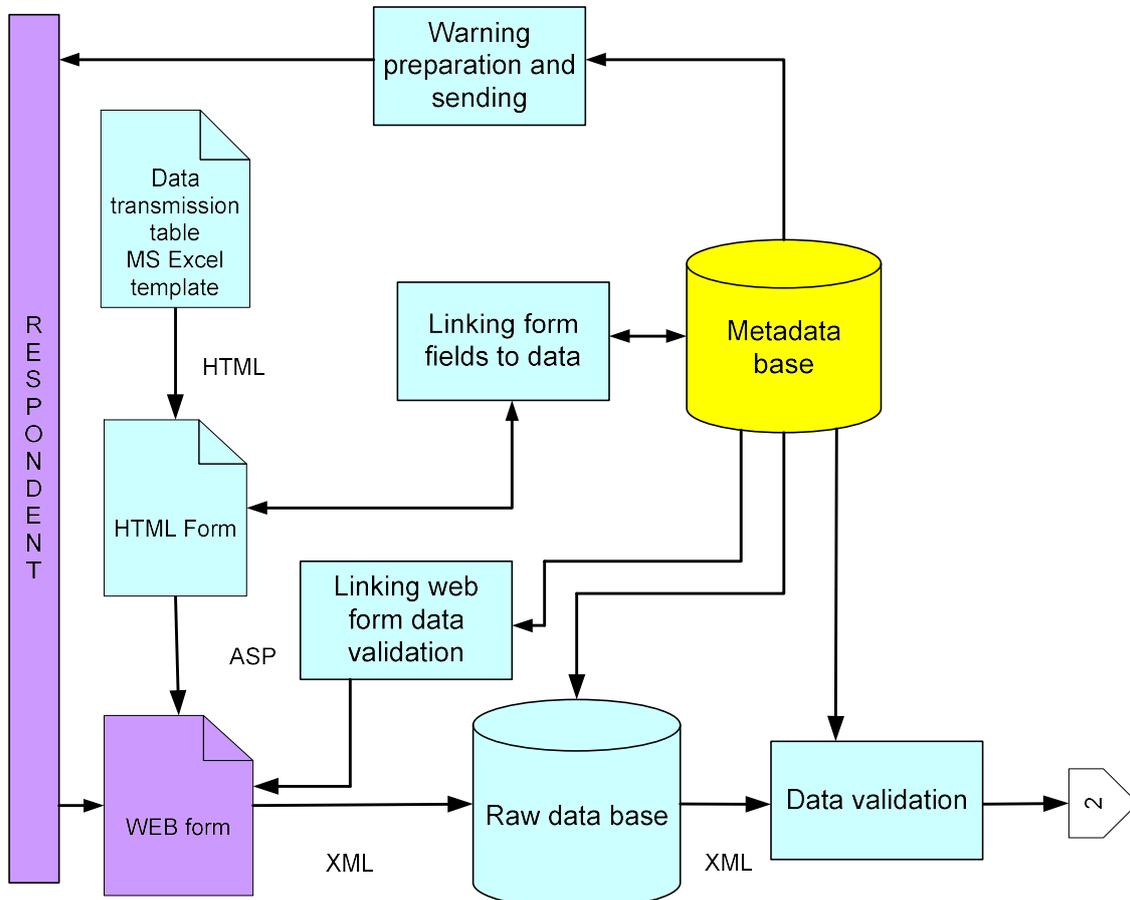
Data scraping from defined websites – data scraping tool

Information from predefined websites will be obtained using Web Scraping Tool.

The purpose of this tool is to provide a semi-automatic way to import data from respondent web pages, where data is published, into data processing system database. The process consists of several steps which are presented in Figure 9.

Each step of this process is responsible for certain functionality, which is described below. Also, users' intervention is necessary during some steps.

WEB page – Some respondents (country) may disseminate their data on their web pages. These web pages become as a source, where survey data might be located. The whole data grabbing begins with the provision of data source, where potential survey data are stored.

WEB page grabbing – is a process when user provided webpage is downloaded as temporal data file for further data processing.

Figure 11: Web scraping tool
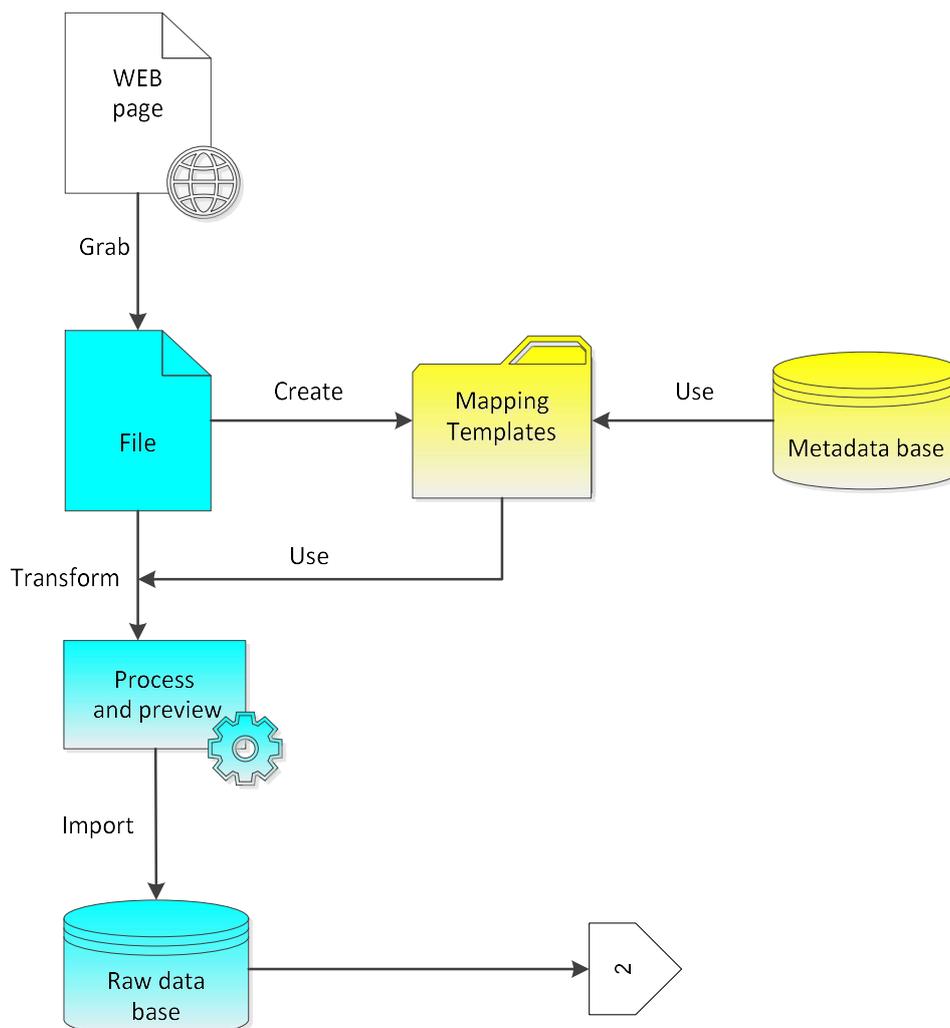
Mapping template – When data are downloaded from certain webpage, users' main task is to describe, which survey data are on this web page and where these data must reside in the survey questionnaire form. This means that users for each numeric value in the source web page must point the destination cell in the questionnaire, where data will be located. This is the mapping between each source value and destination cell. The metadata base information about each survey layout and content will be used to specify each destination cell. All the mapping must be done for each respondent and survey before real data import can be started. The mapping will be saved, for recurrent usage next time, when data grabbing for certain respondent survey will be processed. Such mapping should be created for each respondent and for each survey, because each respondent has its own website, with different data structures and differently structured reports where survey data potentially can be stored.

Transform – is a process, when system automatically finds respondent source data, transforms and saves them into survey form. This process will use user's created data mappings in previous steps. Thus, system will know which values from source webpage must be taken and where these values must be stored in destination survey form.

Process and preview – After data transformation, results will be presented to the user. User's task is to evaluate these results and to take a decision to store data into database or to discard them. The reason of data discard can be wrong data source webpage, structural changes in respondent's webpage, incomplete data mapping. Depending on certain case, user must provide different source – webpage, where to find certain survey data, or readjust data mapping if data are transformed incorrectly and placed in wrong destination survey cells.

Data import – After data evaluation, the accepted data set can be stored in raw database, as original data for a certain respondent survey.

Validate and refine – is a part of standard survey data entry procedure. Data is already imported from webpage, and user performs data validation and corrections to meet logical and arithmetical rules. Final corrected data is stored in Microdata base, where all survey data is located for further data analysis and dissemination.

Validated primary database (Microdata base) – is a central data warehouse, where all survey data for all respondents and all time periods is stored. This database is a basis for data analysis, data aggregation and dissemination.

Implementing SDMX reference infrastructure

Information consists of data and related metadata. The SDMX standards are designed for exchange or sharing of statistical information between organizations at national and international level. SDMX consists of technical and statistical standards and guidelines, together with IT architecture and IT tools, to be used for the efficient exchange and sharing of statistical data and metadata.

In general, SDMX is built around the SDMX Information Model, which was designed to describe aggregated statistics transmitted from NSOs to Eurostat, ECB and other international organizations; the SDMX Information Model also supports various domains of statistics which are disseminated at national or international level.

For supporting the development of SDMX in different statistical domains, Eurostat promotes and maintains the SDMX Reference Infrastructure (SDMX-RI) as a foundation for a rapid and standardized implementation.

SDMX-RI is a generalized service infrastructure that can be re-used partially or completely by any organization interested in using SDMX for data exchange.

An organization can decide to use the SDMX Reference Infrastructure as a whole, can extend the infrastructure adding new modules, can modify some modules, or integrate some building block within its existing dissemination environment.

The most common SDMX Reference Infrastructure modules and supporting tools are described below.

- **SDMX Query Parser**, is an XML parsing API implementation for incoming SDMX-ML Query messages. It validates the received SDMX Query with the SDMX-ML Query schema and in the next step translates this Query to the internal SDMX Data Model, returning it to the Web Service Provider.

- **Data Retriever** is another tool of the RI, dedicated to retrieve respective data from dissemination databases. It operates by translating the internal SDMX Data Model Query to an SQL statement which helps requested data to be taken from the dissemination database, using in the process the mapping information from the Mapping Store. It results with an SDMX dataset represented in the internal SDMX Data Model.

- **Structure Retriever** in its operation is similar to the module mentioned above, which retrieves the SDMX Data Structures based on an SDMX Structure Query. It translates the query to an SQL statement and takes the SDMX Structural Metadata from the Mapping Store, delivering at the end an SDMX-ML structure message.

- **SDMX Data Generator** is used for creating responses to be sent to clients. It translates the (internal Data Model) Data Message to an SDMX-ML Dataset in requested data format. The Data Structure Definition and the SDMX-ML Dataset message format are the input in this case.

There are two web components of an SDMX Reference Infrastructure: Web Client and Web Service Provider.

- **Web Client** acts as a GUI to interact with the Web Service for the display of Structural Metadata and also for the selection, display and exporting of data. It works as a web interface for creation of basic SDMX queries to expose structural metadata from a Mapping Store and data from dissemination databases. All data and metadata are retrieved using the (SDMX-RI) Web Service. User is able to customize the presentation of the tabular data and download it in various formats (e.g. SDMX-ML, XLS and CSV).

- **Web Service Provider** receives from a Client SDMX (Data or Structure) Query Messages and responds with an SDMX-ML (Data or Structure) message related to the input SDMX Query. It performs "XML Validation" of the received SDMX Query and includes a "Soap Error Handler". It also operates and controls the information exchange with the other modules.

- **Mapping Assistant**, is a tool developed by Eurostat and plays an important role in the SDMX Reference Infrastructure. The purpose of this desktop application is to translate "Legacy" Dissemination systems to "SDMX-World". Comparing to mentioned above modules, operating "under the hood", the Mapping Assistant requires user's interaction.

The implementation of SDMX in a statistical domain is made progressively and is generally divided into four main steps. The duration of each of these phases varies according to the number of data sets, the involvement of other international organizations and the expected date of use of SDMX into production.

Figure 12: Main Steps of SDMX implementation
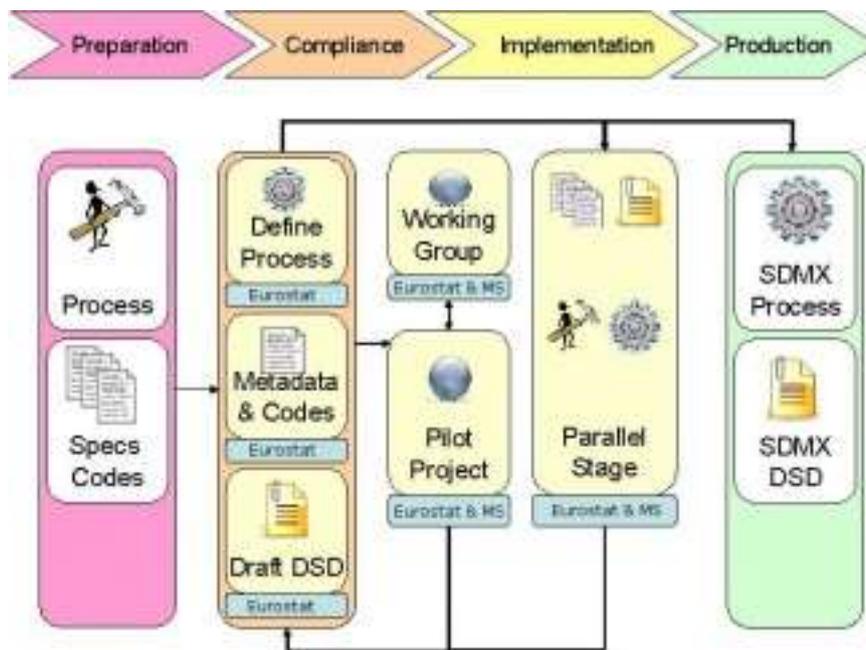
Preparation

The initial phase is called "preparation" and starts with first meetings organized between the concerned GCC-Stat production units and the unit responsible for the implementation of SDMX. The objective of these meetings is to have an overview of the current data collection / treatment / dissemination system in place and to analyse if and how SDMX could be used to facilitate the processes. Therefore, during this step, sample files are analysed and depicted in order to understand the structure of the data collected as well as the code lists which are used for exchanging those data. In case other institutions are involved, it is also important to see if the data are collected in the same manner or differently and whether the data collection is regulated by any legal acts. In this phase, the way reference metadata are collected / treated is also analysed especially in case the Production Unit is collecting national information from NSOs and / or exchanging metadata with other national / international organizations or institutions. This phase ends with a set of decisions on the future work / tasks to be undertaken: continuation or abandonment of this project; which datasets to be concerned, etc.

Compliance

Once the decision of further implementation of SDMX has been taken, the second phase concerns the "Compliance" with the SDMX standards as defined in the SDMX Information Model, the SDMX technical standards and the content-oriented guidelines. During this phase, the first new structures for data exchange are drafted and the code lists used are analyzed in-depth and possibly re-adapted in order to comply with the SDMX standards. This phase ends with a proposed Data Structure Definition (DSD). Regarding the analysis of reference metadata, the compliance phase consists of alignment to the existing standards, for example the ESMS (Euro SDMX Metadata Structure) for the collection of reference Metadata and to the ESQRS (ESS Standard for Quality Reports Structure) for quality related metadata.

Implementation

The implementation phase encompasses the practical creation of developed SDMX elements to be delivered to the production unit. The created package of SDMX objects (DSD, code lists, etc.) is then generally used in a pilot project where several NSOs are involved. This pilot project consists of installation processes in member countries as well as a series of tests. These actions might result in changes of the structure of the DSDs or in certain re-adaptations which are also part of the Implementation Phase. A series of validation tests of generated SDMX-ML sample files created are also performed. The possible re-adaptation of the GCC-Stat production systems is also made during this phase and could imply significant changes. Concerning the implementation in terms of reference metadata, this phase consists in pilot projects where member countries will make use of the European standard structures (ESMS, ESQRS) and the dedicated IT application (National Reference Metadata Editor) to generate their first test files.

Production

This phase marks the concrete application of SDMX for data and metadata exchange. All countries are using a final version of the DSD and the transmission chain is fully based on SDMX. From that point onwards, normal maintenance tasks are required to keep code lists and validation mechanisms up-to-date. National metadata files are created and sent by member countries to GCC-Stat through the National Reference Metadata Editor. These files are checked and validated by GCC-Stat domain specialists. The produced files may also be disseminated on the GCC-Stat website.

Distribution of duties among GCC-Stat and member countries

To make the whole SDMX system effective and efficient a clear division of roles, responsibilities and duties among GCC-Stat and member countries should be defined.

GCC-Stat can use available tools to generate the necessary DSDs and make them publicly available to member countries. The DSDs then will be used to map existing data structures against DSD to generate SDMX files. Separate DSD must be created for each data file or a group of data files if a decision of combining some data sets is taken. Over time the content of the required data sets can change, thus GCC-Stat must implement a proper versioning of DSDs to be able to keep a trace of changes. For countries to be able to map their data structures to DSD, a publicly available metadata repository must be set up where data matrices are explained with all respective code

lists and variable lists. This will allow to understand how to link existing data sets against DSD. On the receival part of SDMX data flow, proper mechanisms must be set up in order to interpret SDMX files and extract data which will be imported into corporate data warehouse for normal data processing.

Such requirements foresee a well prepared structural metadata before starting data exchange based on SDMX approach. Besides metadata need to be agreed and harmonized among member countries to avoid any misinterpretation of data. All metadata must be available for countries and versioned properly. This is necessary to keep historical versions of metadata to be able to reproduce and read SDMX files with historical information where metadata does not correlate with a current version.

Member countries can download DSDs and access the metadata repository. By using SDMX-RI infrastructure data mapping against DSDs is carried out and SDMX files are generated. Proper data validation mechanisms must be set up to be sure that generation was successful and content of SDMX file is valid. Then they are sent to GCC-Stat for further processing.

Data processing

Data processing combines a number of tasks that are aimed at protecting the confidentiality, ensuring quality, integrating data with data already existing in databases and performing all necessary calculations and estimations. Data analysis at micro and macro data level is a significant part of data processing phase. Specific applications designed using MS SQL Server BI facilities will ensure the implementation of all mentioned tasks. In all cases, it is important to use an integrated data repository and an integrated metadata system, and provide a linking and data sharing between all tools employed in this phase.

The data processing functional diagram is reflected in Figure 11.

Figure 13: Data processing phase functional diagram
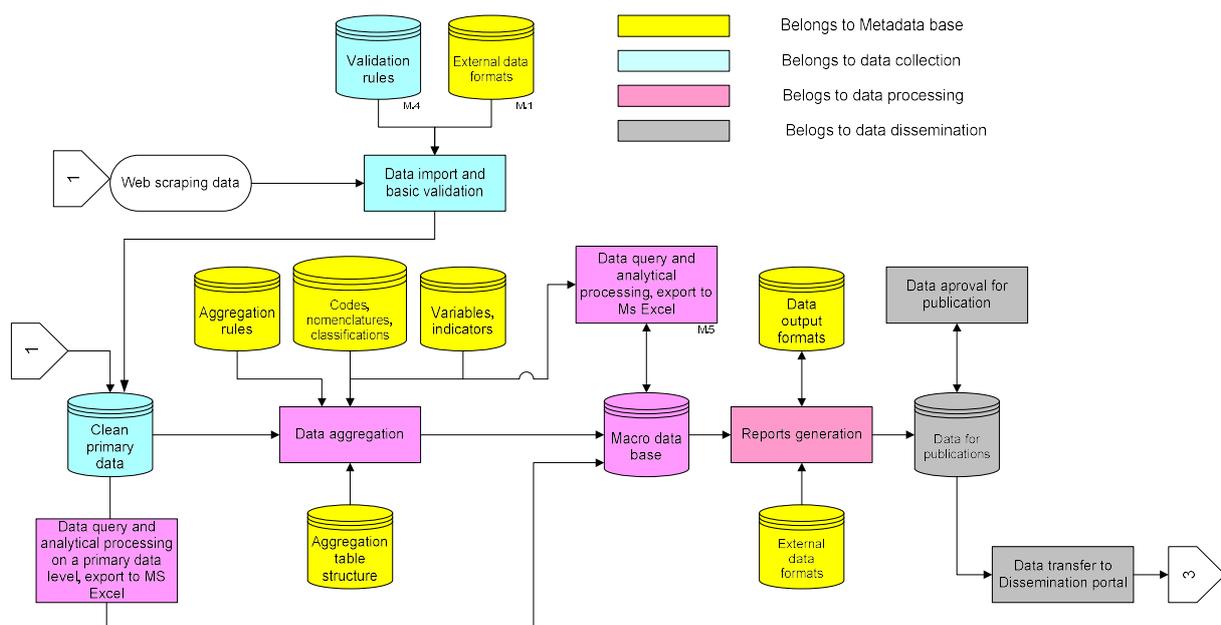
The core Metadata base module

The Core Metadata base module is one of the main parts of the system and can be considered as the core stone of the system. Metadata base data handled by this module are used by all other modules of the system.

The data in the Metadata base is the information on micro and macro data, i.e. description of the numerical data within the statistical production process and the real meaning of this numerical data. Also, the Metadata base system contain descriptions of statistical data transmission tables, data files obtained by Web scraping subsystem or collected using SDMX tools, their content and layout, description of validation, aggregation and reports preparation rules.

The system ensure that Metadata base is used as a key element for creating universal, common, programming-free environment for different statistical data sets collected and for data processing. System users can easy query necessary data form Microdata / Macrodata bases navigating through Metadata base. Metadata are widely used for data analysis and dissemination.

Metadata base is linked with database structure model level with Microdata base and Macrodata base thus creating a system data model. The system data model is reflected in Figure 12.

Figure 14: Metadata base linked with Microdata / Macrodata bases – system data model
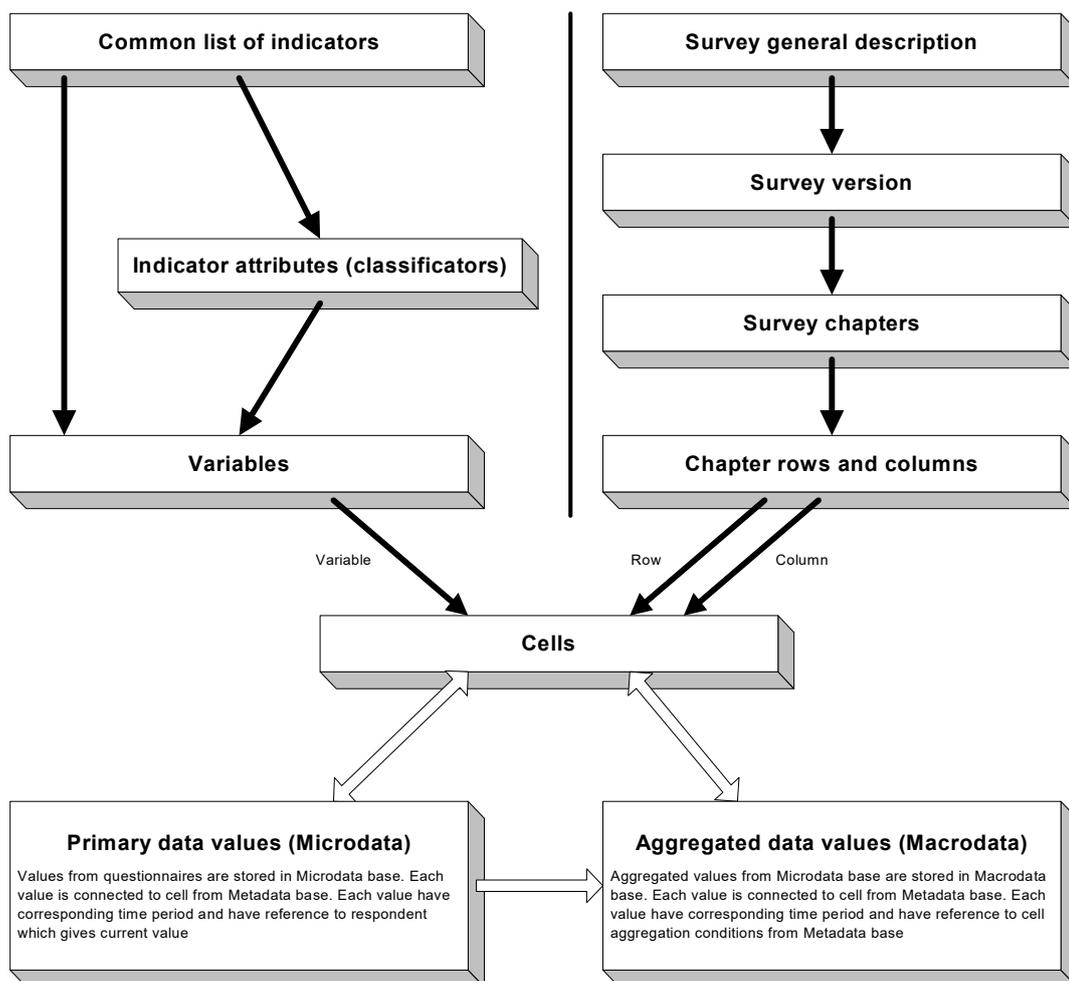
Data processing through statistical data transmission tables begins with the data transmission table metadata entry into Metadata base. Each new data transmission table should be registered and described in the system with the support of applications for metadata description. For each data transmission table, it is necessary to create a table version, which is valid for at least one year. It is important not to change table structures frequently in order to have possibility to retrieve data in time series manner. Many table structure versions may raise complexity of data processing and can cause data inconsistency problems. If data transmission table content and / or layout does not change, then current data transmission table version and its description in Metadata base will be used for the next year.

Each statistical data transmission table could contain one or more data tables called chapters. In Metadata base for each chapter it is necessary to describe table type. System recognizes three types of tables:

1. Constant tables with fixed number of rows and columns;
2. Tables with variable rows number and constant number of columns,
3. Tables with variable columns number and constant number of rows.

For each data, transmission table version chapter describes rows and columns with their codes and names in the Metadata base. The information about data transmission table version is necessary for automatic data entry application generation and further data processing.

During the description of the data transmission table version, the information about indicators is entered into Metadata base. This information is stored in Metadata base in the common list of indicators. In cases, where the use of newly created indicator in data transmission table version is foreseen, it is necessary to add its description to the common list of indicators. Indicators themselves are independent from data transmission tables. That gives a possibility to use one indicator in several data transmission tables and to get information about one indicator from several data transmission tables as well.

For each indicator, it is possible to define attributes. Attributes in the system are classifications, which give opportunity to describe and store values of indicators in breakdown by classification items. Indicators could be presented without attributes.

The meaning of Variable used in the entire system is a value of an indicator for the exact item of classification. Therefore, the value of the variable is equal to the value of the indicator only in cases when attributes are not in use and variable represents the total value of the indicator for a measuring object.

When indicators and attributes are defined, it is necessary to define variables. As it has been described above, variables are combination of indicators and corresponding attributes. Created variables are connected to the cells of data transmission table.

Last step in the data transmission table content and layout description is cells formation. Cells are created as combination of row and column from data transmission table version side and a variable from indicators and attributes side.

All data transmission table values from questionnaires are stored in Microdata base and each value is linked with a cell (from Metadata base), which describes value meaning. Also, each value

in Microdata base has additional information about respondent, which gives current value and reporting period.

The metadata structures described above allows to easy navigate, retrieve and analyse data by browsing or querying the system. User can browse through a structured list of indicators and attributes and retrieve respective information from any data table where particular indicator or classification is used. Another option is to browse through a structured list of data tables and to drill-down through particular indicators, attribute and variable tied to particular cell. This gives possibility to visually browse data warehouse without specific knowledge of data retrieval from a relational database only knowing data transmission tables' structures and the information they consist of.

The same situation is in Macrodata base, where aggregated values are stored. Each aggregated value has reference to cell (from Metadata base), reference to each value aggregation conditions (from Metadata base) and correspondent reporting period.

Such cells definition – combination from two sides – has several advantages. For statisticians, who are familiar with data transmission tables and work with them every day, can easily work with values using row and column language (row and column codes). All validation rules for data transmission table are described in Metadata base using row and column coordinates.

Described data transmission table version is valid for at least one year. If it is necessary to introduce any changes in the data transmission table, then a new data transmission table version has to be created in the Metadata base. For data analysis purposes, users can use approach, where a search of cells' values from variables side is possible (list of variables stored in the Metadata base). The definition of variable cell location is not important.

The content and functionality described through application packages of Metadata base module, Microdata base module and Macrodata base module are reflected sequentially in Figures 13, 14 and 15.

Figure 15: Metadata base module content and functionality
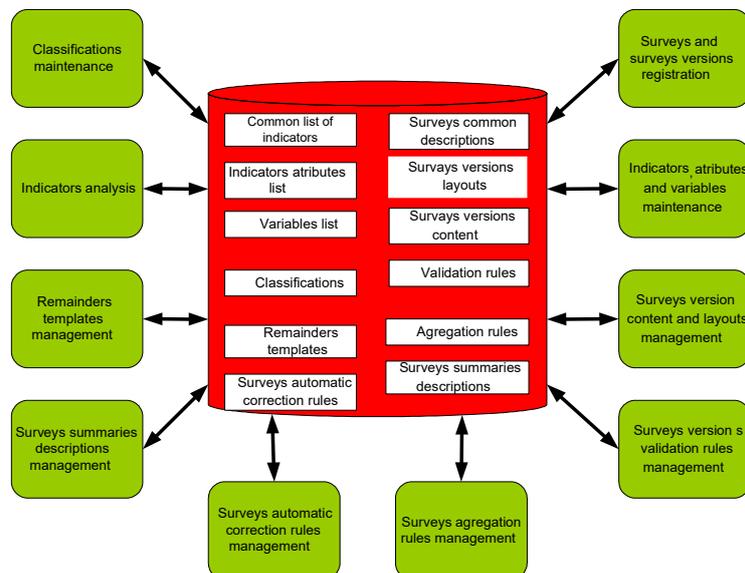
Figure 16: Microdata base module content and functionality
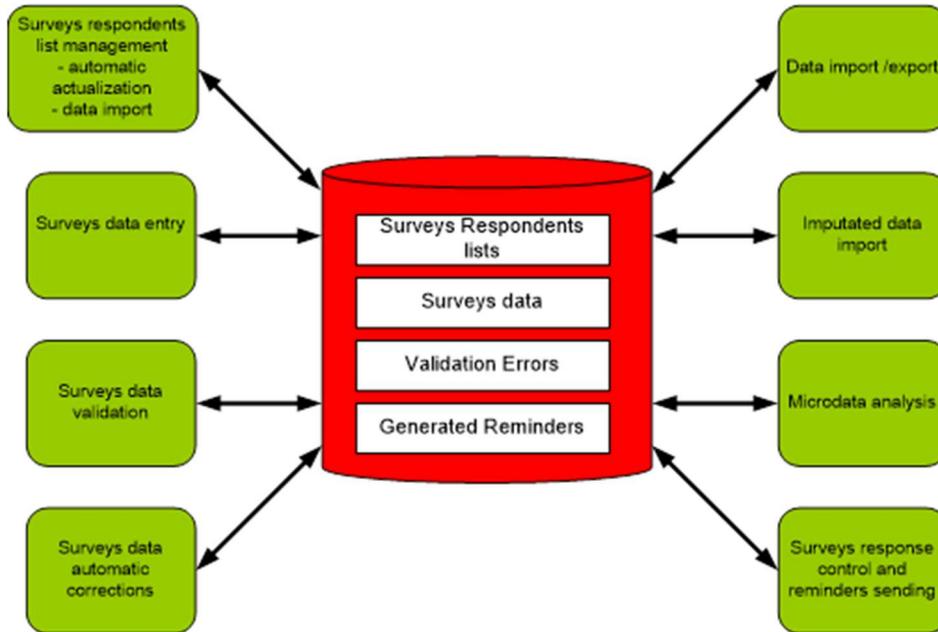
Figure 17: Macrodata base content and functionality
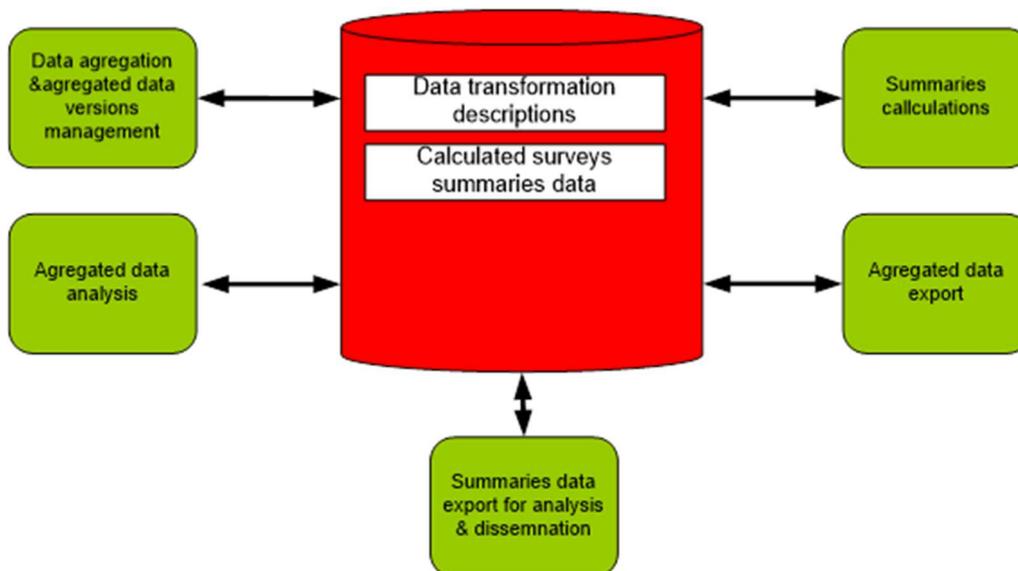
Data aggregation module

As mentioned before, users can run application from data collection and validation module. This application is used to maintain aggregated data. For each data transmission table version for selected period several aggregated data versions in the system can be stored. Using mentioned application, users maintain aggregated data versions, i.e. create new or delete unnecessary versions of aggregated data. When users create a new version of aggregated data, application analyze Metadata information on the current survey version aggregation conditions and create data aggregation procedure. After running the aggregated data in Macrodata base are stored. If an exact data set for selected time period has several data aggregation versions, then one of

them must be selected as active. Active aggregated data version is used in all applications, which works with the aggregated data (reports creation, aggregated data analysis).

From Data collection and validation module users can also run application for common Macrodata base analysis. Using this application, users can extract from Macrodata base any data they need. Data selection starts from classifications selection. Then application display whole data sets list, which have aggregated data for selected classifications. For each survey, it is possible to see years when aggregation has been done as well as to see cells' list for current survey version. Using this application, users can select and combine aggregated data from different surveys for different time periods and to add additional selection conditions. All selected data will be exported to Microsoft Excel or Microsoft Access. Selected data can be converted to PC-AXIS file format for dissemination purposes.

Special solution for Foreign trade statistics

Structurally Foreign trade statistics is treated as an integral part of the E2E data processing system. For data collection, the same mechanisms as for data transmission tables can be used by operating with data import facilities. Data tables used for data storage and further processing must be described in metadata which would allow easier data validation at data receival stage as well as definition of data aggregation procedures through metadata. Technically data should be stored in a separate database. This would allow direct interaction with data in database for specialists without any risk of disrupting data of other domains. Furthermore, due to high volumes of data, the processing of Foreign trade data would not disrupt the performance of other statistical domains data processing if common data warehouse would be used. Also, specific data processing requirements could be the case for the domain and usage of a separate database would allow easier adjustment of metadata functionality as well as data processing / aggregation algorithms without affecting other statistical domains.

Data analysis module

Data analysis module is based on On-Line Analytical Processing (OLAP) techniques. OLAP is a set of technologies that accumulates data in data warehouse and transforms the data into multi-dimensional structures, called cubes. Cubes allow a better response to complex queries.

Data analyses module is implemented using Microsoft SQL server component – Analysis Service for multi-dimensional data cube formation and storing and include facility for multi-dimensional statistical data analysis reflected in Figure 16.

A work with Data analysis module can be divided into two main parts:

- Statistical survey data preparation for analysis;
- Browsing and analysis of survey data.

The first part is performed by data administrators, who are familiar with existing surveys and well introduced to Metadata base, Macrodata base data structures. This part is divided into two steps:

- Data transformation from Metadata base and Macrodata base, and Microdata base in special cases using Microsoft Data transformation service to special data structure format, which is required for Microsoft SQL Analyses service;

- Formations of multi-dimensional cubes, which administrators create with Microsoft Analysis service manager.

Figure 18: Data analysis module functionality
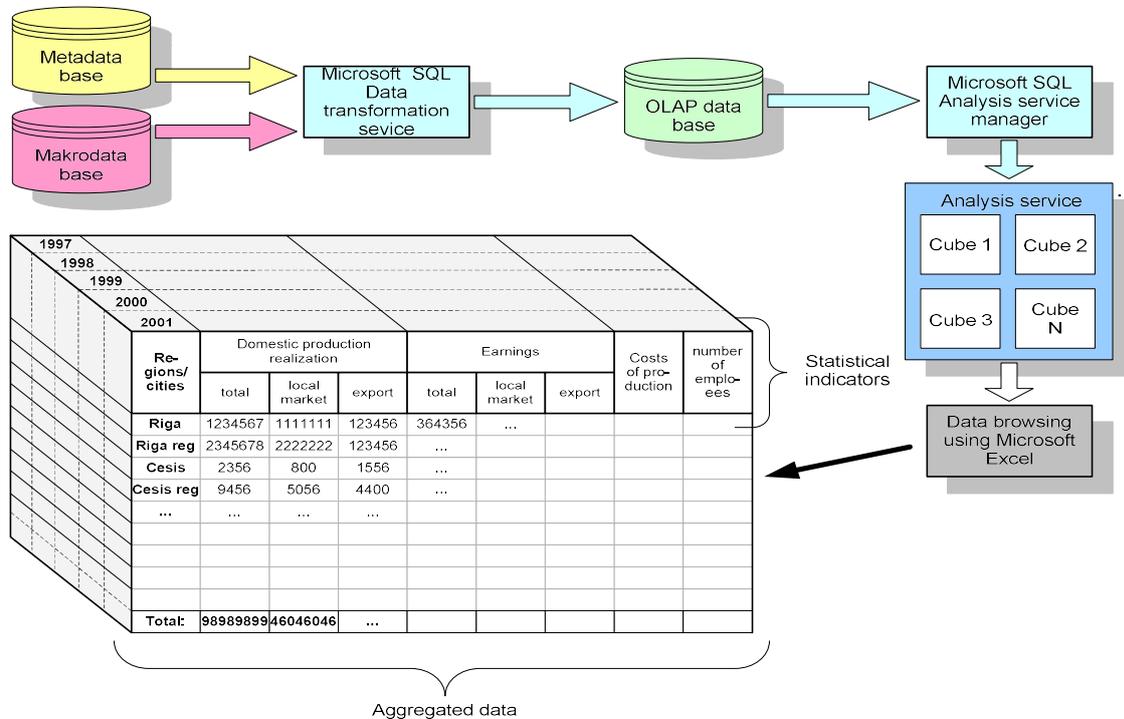
The second part of Data analysis module will be done by system's end-users using Microsoft Excel component – PivotTable, which allows to easy view, rearrange, regroup data in different ways. It is also possible to use other Microsoft Excel component – PivotChart, which allows creating diagrams from PivotTable data.

In special cases, specific analysis data could be transferred to SPSS or another special analysis software package.

User administration module

To ensure data security and whole system maintenance, to provide easiest as possible system users administration The Users Administration module has been designed.

System in general consists of corporative data warehouse where data and metadata of different processing stages are stored. It also contains a large amount of different applications that ensures necessary activities such as data validation import / export, analysis, etc. with data and metadata. The main element of that module is user's roles database.

User's roles must be predefined. Roles' granularity depends on duty separation of system users. The more granular are the roles the easier is to separate duties by assigning proper roles and giving access only to a particular user who will carry out job duties.

Another important aspect is user's activity monitoring and auditing. It is a system which records user's activities for later troubleshooting of possible user errors. This will also help to identify a

possible information misuse. System users must be internal ones, not DBMS. Latter are a basis for administration ease and future migrations of the system to newer platforms. The system then interprets user's roles to give proper access to system functionality and data warehouse parts that user needs to access.

System application administrators can grant rights to users. Application administrators can create new users in the system. They can create user groups and add users to these groups. The next applications administrators' duty is to prescribe application functions list for each user group and define with which exact user group each group can work.

For example, it is possible to assign roles to user groups and to view system journal of applications. Moreover, the information on the latest survey data changes, author and timing of those changes is available.

Data dissemination

Prognoz data dissemination portal and data dissemination software packages PX-Win and PX-Web are proposed to be used for data dissemination.

The PX-Win family consists of a number of programs for Windows and Internet environment that are used to present, facilitate and disseminate statistical information. These are mostly used by NSOs in different countries and give users possibilities to retrieve statistics.

PX-Web is a solution for the web, which is used to establish dynamic tables on Internet from PX-Win files. PX-Web can also be a part of solution with connection to SQL databases.

The PX-Web applications are generated automatically from a PC-Axis database i.e. a catalogue structure containing PX-Win files. By using style sheets, the application will fit into most website layouts.

Prognoz Data Dissemination portal integration to E2E solution determines its existence and usage in GCC-Stat. This portal provides a user with a sufficient choice on the necessary data visualization option.

Taking into account that Prognoz Data Dissemination Portal is a commercial product and its internal structure description is not published, the integration detail will be elaborated during the software system adaptation and implementation process.

SDMX-RI will ensure data sets transfer to international organizations such as UN, WB, IMF, etc. Detail description is provided in the data collection section of this section.

IT Technical architecture

The figure below presents Hardware Technical Architecture of MD-SDMS.

Infrastructure for proposed solution is based on the standard approach for Application Hosting and Publishing. Architecture is designed with 3–tier having presentation / application layer, business logic layer and data layer. Our recommendation is to have a separate infrastructure for E2E Integrated Solution (MD-SDMS) ensuring that load is not distributed to existing network infrastructure. The other recommendation is to have separate Network link with ISP Server

provider (2MB). Consolidated High Level Network diagram is presented below (including new Infrastructure for MD-SDMS).

Figure 19: Architecture of proposed IT infrastructure
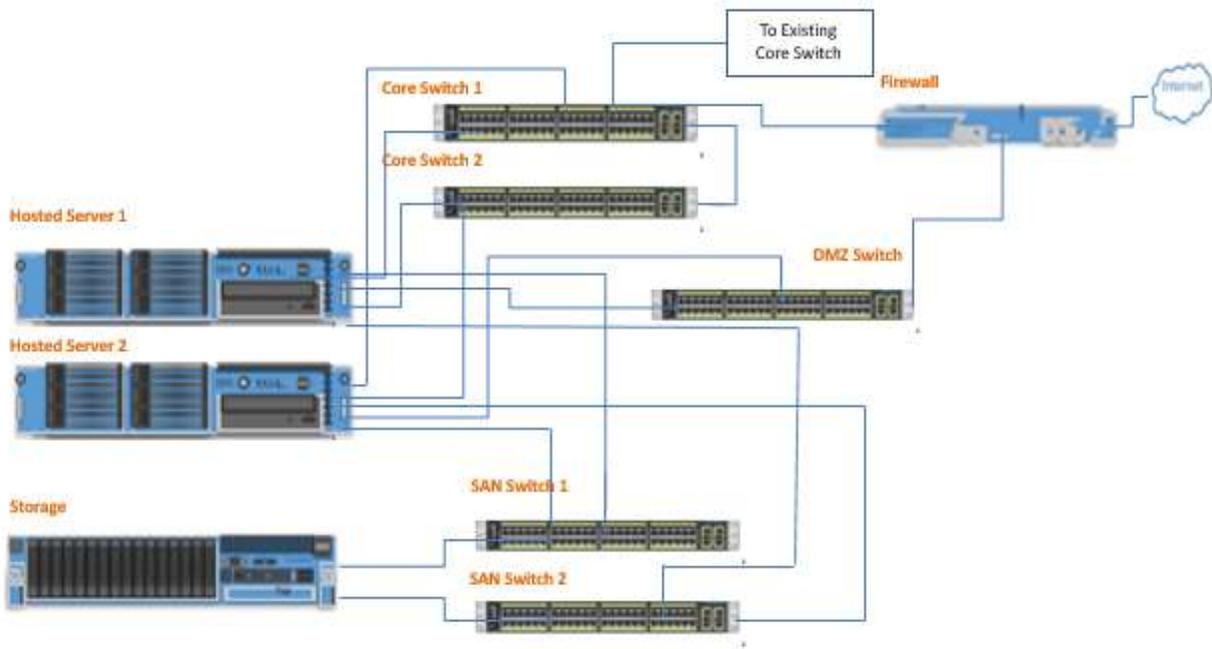
Proposed Infrastructure consists of Server Infrastructure (Implementing Virtualization Solution – preferably Microsoft).

Hosted Servers

a. Specification:

Latest Generation Servers deliver the latest performance and expandability. The server ensures reliability, serviceability and continuous availability with minimum configuration of 128 GB RAM which can be expanded up to 256 GB RAM, 12 Core. Servers are configured for High Availability.

b. Functionality:

Main functionalities of the proposed infrastructure will be:

*   Database Management System (MS SQL Server 2016 Enterprise) – For Data WareHouse.
*   Integrated Statistical Tool MD-SDMS (Metadata Driven Statistical Data Management system).
*   Host Virtual Servers for Database, MD-SDMS, development and testing.
*   Development server will be utilized for application development and testing.

c. Requirement:

Having in mind the scope of overall system requirements, specified system infrastructure will be necessary to obtain long term sustainability and support according to projected system growth and technology change.

d. Advantages:

Proposed architecture solution will be implemented in single hosted solution with high availability. Best utilization of infrastructure to optimize the price.

e. Limitations:

One server will not be fully utilized until it failover.

f. Licenses:

Proposed standard system software to be run on such an infrastructure will be:

- Microsoft Windows 2012 Enterprise – 16 Nos.
- MS SQL Server 2016 Enterprise – 2 Nos.

Storage

a. Specification:

SAN Storage. Latest Hybrid Storage System with configuration 900 GB 12 SAS 10k.

b. Functionalities:

- System that stores data to be utilized by different users that are connected by the computer network, and that are in demand for various data working in heterogeneous applications.
- Will have separate LUN's for proposed solution (Database, Application, Development and Testing)

c. Advantages:

Easy deployment and management, affordable price, low utilization costs, high performance, scalability, very high reliability and availability, robustness, with remote replication capability.

SAN Switch

The 8/8 Base SAN Switch and 8/8 SAN Switch are targeted for small to medium sized enterprises and are ideal for entry-level SANs, and as an edge switch for core-to-edge SAN environments. The 8/8 Base SAN Switch and 8/8 SAN Switch provide SAN connectivity that simplifies IT management infrastructures, improves system performance, maximizes the value of virtual server deployments, and reduces overall storage costs.

Core Switches

24 Port Data IP Base – Latest Switch should support full IEEE 802.3at Power over Ethernet Plus (PoE+), modular and field-replaceable network modules, RJ45 and fiber-based downlink interfaces, and redundant fans and power supplies. Should enable uniform wired-wireless policy enforcement, application visibility, flexibility, and application optimization.

DMZ Switch

Latest Switch which should enable intelligent services such as application visibility, Smart Operations, Layer 3 routing and delivering best-in-class energy efficiency with up to 80% energy saving. Designed for operational simplicity to lower TCO.

Firewall

Any Latest Firewall Solution delivering end-to-end network security with one platform, one network security operating system and unified policy management with a single pane of glass – for the industry's best protection against the most advanced security threats and targeted attacks.

Notes:

1. There is a possibility to use existing Firewall – SonicWall

2. Back Solution is not proposed, recommend to use existing VEEAM data back-up solution.

3. Proposed architecture can be optimized by using exiting Core Switches.

Hardware Specification

Description	Number	Price	Total
Hosted Servers with accessaries	2	40,100	80,200
SAN Storage	1	23,727	23,727
SAN Switch	2	7,800	15,600
Core Switch	2	8,200	16,400
DMZ Switch	1	2,600	2,600
Firewall	1	4,100	4,100
		Total	142,627

Product Number	Unit Price	Qty	Product Description	Extended Price
HP DL380 Gen9 8SFF Server – Hosted Servers				
719064-B21	$2,175.00	2	HPE DL380 Gen9 8SFF CTO Server	$4,350.00
719064-B21 B19	$0.00	2	Europe – Multilingual Localization	$0.00
817943-L21	$1,725.00	2	HPE DL380 Gen9 E5-2650v4 FIO Kit	$3,450.00
817943-B21	$1,725.00	2	HPE DL380 Gen9 E5-2650v4 Kit	$3,450.00
817943-B21 0D1	$0.00	2	Factory integrated	$0.00
805351-B21	$675.00	4	HPE 32GB 2Rx4 PC4-2400T-R Kit	$2,700.00
805351-B21 0D1	$0.00	4	Factory integrated	$0.00
724865-B21	$150.00	2	HP DL380 Gen9 Universal Media Bay Kit	$300.00
724865-B21 0D1	$0.00	2	Factory integrated	$0.00
652564-B21	$300.00	4	HP 300GB 6G SAS 10K 2.5in SC ENT HDD	$1,200.00
652564-B21 0D1	$0.00	4	Factory integrated	$0.00
726537-B21	$150.00	2	HP 9.5mm SATA DVD-RW Jb Gen9 Kit	$300.00
726537-B21 0D1	$0.00	2	Factory integrated	$0.00
665243-B21	$810.00	2	HPE Ethernet 10Gb 2P 560FLR-SFP+ Adptr	$1,620.00
665243-B21 0D1	$0.00	2	Factory integrated	$0.00
AF573A	$30.00	4	HP Rdnt 2m,10A, C13-C14 Jmpr Cord	$120.00
AF573A 0D1	$0.00	4	Factory integrated	$0.00
749974-B21	$720.00	2	HP Smart Array P440ar/2G FIO Controller	$1,440.00

Product Number	Unit Price	Qty	Product Description	Extended Price
733660-B21	$110.00	2	HP 2U SFF Easy Install Rail Kit	$220.00
733660-B21 0D1	$0.00	2	Factory integrated	$0.00
AJ763B	$1,844.00	2	HPE 82E 8Gb Dual-port PCI-e FC HBA	$3,688.00
AJ763B 0D1	$0.00	2	Factory integrated	$0.00
720478-B21	$325.00	4	HPE 500W FS Plat Ht Plg Pwr Supply Kit	$1,300.00
720478-B21 0D1	$0.00	4	Factory integrated	$0.00
455883-B21	$995.00	4	HPE BLc 10G SFP+ SR Transceiver	$3,980.00
455883-B21 0D1	$0.00	4	Factory integrated	$0.00
H7J32A3	$0.00	1	HPE 3Y Foundation Care NBD Service	$0.00
H7J32A3 TT3	$248.00	2	HPE ProLiant DL380 Gen9 Support	$496.00
HA114A1	$0.00	1	HP Installation and Startup Service	$0.00
HA114A1 5A6	$1,753.00	2	HPE Startup 300 Series OS SVC	$3,506.00
748922-421	$667.35	16	WinSvrDCCore 2016 SNGL OLP 2Lic NL CoreLic Qlfd	$10,677.60
7JQ-00253	$17885.51	2	SQLSvrEntCore SNGL LicSAPk OLP 2Lic NL CoreLic Qlfd	$35,771.02
R39-01181	$1751.14	1	WinSvrExtConn 2016 SNGL OLP NL Qlfd	$1,751.14
			Subtotal	**$80,319.76**
HPE MSA 2042 SAN DC SFF Storage (Q0F72A)				
Q0F72A	$11,255.00	1	HPE MSA 2042 SAN DC ME SFF Stor	$11,255.00
H7J32A3	$0.00	1	HPE 3Y Foundation Care NBD Service	$0.00
H7J32A3 YYY	$3,257.00	1	HPE MSA 2042 Support	$3,257.00
HA114A1	$0.00	1	HP Installation and Startup Service	$0.00
HA114A1 5J0	$2,608.00	1	HPE MSA Family Startup SVC	$2,608.00
J9F47A	$567.00	11	HPE MSA 900GB 12G SAS 10K 2.5in ENT HDD	$6,237.00
C8R23A	$370.00	1	HPE MSA 2040 8Gb SW FC SFP 4 Pk	$370.00
			Subtotal	**$23,727.00**
HPE 8/8 Base 8-port Enabled SAN Switch (AM867C)				
AM867C	$4,388.00	2	HPE 8/8 Base 8-port Enabled SAN Switch	$8,776.00
AM867C 05Y	$0.00	2	2.4m Jumper (IEC320 C13/C14, M/F CEE 22)	$0.00
H7J32A3	$0.00	1	HPE 3Y Foundation Care NBD Service	$0.00
H7J32A3 9LJ	$576.00	2	HPE B-Series 8/8 and 8/24 Switch Support	$1,152.00
HA113A1	$0.00	1	HPE Installation Service	$0.00
HA113A1 5GA	$512.00	2	HPE LowEnd SAN/Edge Switch/HAFM Inst SVC	$1,024.00
QK734A	$114.00	16	HPE Premier Flex LC/LC OM4 2f 5m Cbl	$1,824.00
AJ716B	$179.00	16	HPE 8Gb Short Wave B-Series SFP+ 1 Pack	$2,864.00
			Subtotal	**$15,640.00**

Core Switch				
WS-C3850-24T-S	$6,500.00	2	Cisco Catalyst 3850 24 Port Data IP Base	$13,000.00
CON-SNT-WSC384TS	$563.75	2	SNTC-8X5XNBD Cisco Catalyst 3850 24 Port Data IP Base	$1,127.50
S3850UK9-36E	0.00	2	CAT3850 Universal k9 image	0.00
PWR-C1-350WAC/2	$650.00	2	350W AC Config 1 SecondaryPower Supply	$1,300.00
C3850-NM-4-1G	$500.00	2	Cisco Catalyst 3850 4 x 1GE Network Module	$1,000.00
CAB-TA-UK	0.00	4	United Kingdom AC Type A Power Cable	0.00
STACK-T1-50CM	0.00	2	50CM Type 1 Stacking Cable	0.00
CAB-SPWR-30CM	0.00	2	Catalyst 3750X and 3850 Stack Power Cable 30 CM	0.00
PWR-C1-350WAC	0.00	2	350W AC Config 1 Power Supply	0.00
			Subtotal	**$16,427.50**
DMZ Switch				
WS-C2960X-24TS-L	$2,395.00	1	Catalyst 2960-X 24 GigE, 4 x 1G SFP, LAN Base	$2,395.00
CON-SNT-WSC296XT	$198.00	1	SNTC-8X5XNBD Catalyst 2960-X 24 GigE, 4 x 1G SFP, LAN	$198.00
CAB-ACU-RA	0.00	1	Power Cord UK, Right Angle	0.00
			Subtotal	**$2,593.00**
Firewall				
FG-100D	$1,564.00	1	Fortinet FortiGate-100D Firewall	$1,564.00
Fortinet 100D - BDL	$2,500.00	1	FortiGate-100D Hardware plus 8x5 Forticare and FortiGuard UTM Bundle, 1 Yea	$2,500.00
			Subtotal	**$4,064.00**
			Total	**$142,770.50**

6. Time-bound plan

Activity	Duration in months																		Responsible executives
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	
1. Establishment of the legal bases for data interchange and cooperation between GCC-Stat and GCC member countries with particular attention to data collection issues like methodology, data formats and time frame. Thus, improving data / metadata collection processes and collected data / metadata quality and timeliness.	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	GCC-Stat & GCC MC
1.1 Discussion all the issues relating to the new co-operation agreement (data collection, methodology, metadata)	■	■																	GCC-Stat working groups and Standing Committees
1.2 Elaborating draft cooperation agreement			■	■	■	■													GCC-Stat Standing Committees
1.3 Negotiating elaborated draft with member countries					■	■	■	■	■	■	■	■							GCC-Stat & GCC MC
1.4 Undersign cooperation agreements													■	■	■	■	■		Board of Directors & GCC Ministerial Council
2. Acquisition of IT infrastructure elements, installation, checking	■	■	■	■	■	■	■	■	■	■									GCC-Stat IT Department & Supplier
2.1 Tender preparation and launching	■	■																	GCC-Stat IT Department
2.2 Tender evaluation			■																GCC-Stat IT Department
2.3 Equipment supply							■	■											GCC-Stat IT Department
2.4 Installing the proposed IT infrastructure								■	■										Supplier under supervision IT department
2.5 Checking new IT infrastructure										■									Supplier & IT Department
3. Tender preparation and launching for E2E solution adaptation and installation	■	■	■																GCC-Stat IT Department
4. Establish project implementation unit. Prepare a planning of the solutions implementation in statistical domains			■																GCC-Stat Management
5. Tender Evaluation and agreement signing							■	■											GCC-Stat Evaluation team

Activity	Duration in months																		Responsible executives
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	
6. Organize a project Kick-off meeting								■											GCC-Stat Project Board
7. Project on Integrated E2E Solution adaptation and implementation								■	■	■	■	■	■	■	■				Supplier under supervision of GCC-Stat
7.1 Analysis of adaptation detail according to description attached (2 months)									■	■									Supplier with Assistance of GCC Statistical Domains Managers
7.2 Seminar on project status and findings											■								Supplier & GCC-Stat responsible experts
7.3 Adaptation and installation of the system modules part A according to description attached (2 months)												■	■						Supplier
7.4 Video conference on project status and findings																			Supplier & Project board & Steering committee
7.5 Adaptation and installation, part B according to description attached (2 months)														■	■				Supplier
7.6 Video conference on project status and findings																			Supplier & Project board & Steering committee
7.7 Pilot testing according to description attached (1 month)																■			Supplier & GCC-Stat responsible experts
7.8 Video conference on project status and findings and recommendations																			Supplier & Project board & Steering committee
8. Implementing system by Introducing subject matter data sets into system with training in parallel according to description attached (1 month)																	■		Supplier & GCC-Stat statisticians from subject matter domains & subdomains
9. Project closing seminar																	■		Supplier & GCC-Stat responsible experts
10. Continue implementing solution in all statistical domains according to planning prepared (see e)																			

6.1. Project implementation unit

To ensure a smooth project implementation a Project implementation unit containing the assigned project board, project manager, and project team should be created. Despite of three mentioned roles there are many supplemental roles that help ensure requirements and standards are met, and that project runs smoothly.

Main roles and responsibilities

The Project Board, includes multiple people: the Senior Executive (the customer), the Senior User covering all statistical domains (the end user), and the Senior Supplier (the supplier). **The Project Board** checks for project assurance from three perspectives:

1. The Senior Executive makes sure the project is still viable from a business sense (cost vs. benefit).
2. The Senior User ensures user needs are being met.
3. The Senior Supplier checks whether the project is working towards a realistic, practical solution.

The Project Board is accountable to a corporate or programme management on the implementation of the project, and has the authority to manage the project within the remit set by corporate or programme management as documented in the project mandate.

Furthermore, the Project Board is responsible for the communication between the project management team and external stakeholders involved (e.g. corporate and / or programme management).

Depending on the complexity and importance of the project, the Project Board members may assign Project Assurance to obtain an objective opinion from Independent Experts. The Project Board may also delegate decisions regarding changes to a Project Change Authority.

The Executive – is business-oriented person who is responsible for the successful and realistic implementation of the project. The Executive, in this case, is a Senior Executive, who has an outstanding knowledge of all statistical processes and IT infrastructure, is appointed by the GCC-Stat management. The Executive is supported by the Senior User and the Senior Supplier.

The Executive is responsible for designing and appointing Project Management Team, including the rest of the Project Board and the Project Manager.

Senior User – one or more people who present the final users' requirements to the Project Board.

The Senior User has the following responsibilities:

- To specify the needs (requirements) of Users that will use project products.
- To liaise between the Project Management Team and Users.
- To make sure that the solution will meet the needs of Users, especially in terms of quality and ease of use, and against requirements.
- To supply the benefits information for the Benefits Review Plan.

The Senior Supplier – one or more people who present the interests of suppliers.

The Senior Supplier presents the interests of those designing, developing, facilitating and implementing the project's products. They provide supplier resources to the project and ensure that the right people, tools, equipment and knowledge are in place, and that the products will meet the expected criteria, including quality criteria.

The Project Assurance – assures the interests of primary stakeholder. The Project Assurance must be independent from the Project Manager and the Project Team and is responsible for supporting the Project Manager by giving advice and guidance.

The Project manager – responsible for the day-to-day management of the project on behalf of the Project Board.

The Project Manager runs the project on behalf of the Project Board within specified constraints and liaises throughout the project with the Project Board and the Project Assurance. The Project Manager usually is assigned by the Executive and is responsible for all project management processes except for the directing process. The latter is a responsibility of the Project Board that reviews and evaluates project's execution.

The Change Authority – The Project Board may assign a person or group who consider requests for change or off-specifications. The Project Manager could be assigned as the Change Authority for some aspects of the project (e.g. changing baselined Work Packages, changing budget, changing control).

Figure 20: Project implementation structure proposed according to PRINCE2
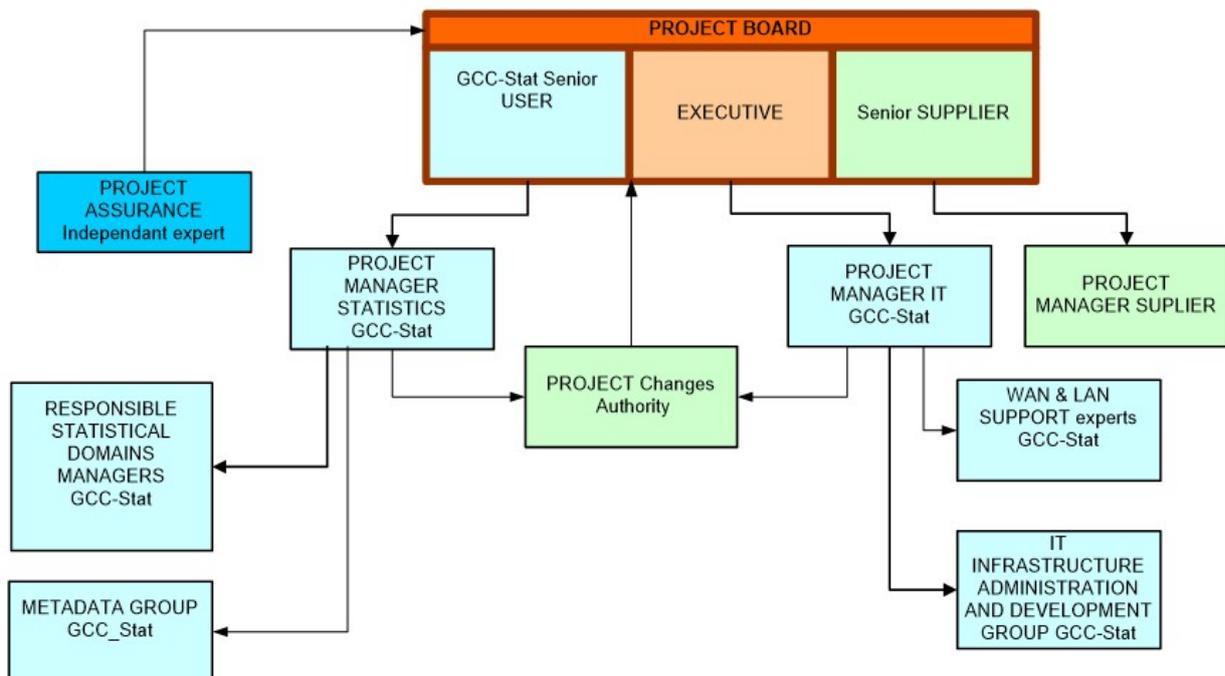
Resources to be provided:

By supplier: Senior Supplier representative to Project Board, Project Manager, IT Specialists conducting analysis, installation and testing. Specialists participating in project change authority group, Training providers (at least two experts).

By GCC-Stat: Senior User who has an outstanding knowledge of statistical processes, Executive who has an outstanding knowledge of IT infrastructure, IT project manager who will work under Executive's supervision, Project Manager-Statistician who will work under Senior User's supervision, IT experts, specialists of subject matter domains.

6.2. Possible risks and risks' mitigation

Risk management has the primary goal of identifying and responding to potential problems with sufficient lead-time to avoid crises, so that is possible for the project management to achieve its goal. Risk management needs a systematic planning at the beginning of the project and constant review and readjustment during the evolution of the project.

The table below summarizes the identified risks. Each risk is assessed through a likelihood (based on prior information) and an evaluation of impact. Both likelihood and impact are measured on a qualitative scale (L = Low, M = Medium, H = High).

Risk identification	Likelihood	Impact	Preventive actions
Member countries do not support cooperation and data exchange agreements	H	H	Clear presentation of the necessity of harmonised statistical data exchange, presentation of best practises.
Delay with hardware and software supply	M	M	Ensure timely equipment supply and monitor of the evaluation process of the related tender
Some functionality of the system under implementation do not meet GCC-Stat business requirements	L	H	Only those solution providers should be considered who can adopt the necessary functionality
Implementation time schedule will exceed the expected	M	M	Careful planning and organization of work
Staff allocated to work with solution do not have necessary skills	M	H	Revise training policy and increase the number of trainings to be provided
Staff will resist to work with solution	L	H	Clearly present staff advantages and functionality of the solution
Lack of staff necessary to maintain the solution	L	H	Increase the number of staff
Lack of statistical staff to work with solution implemented	M	H	Perform regular staff workload analysis, increase the number of staff
All tender proposals are of higher cost than expected	M	H	Check the tender technical requirements, make changes, negotiate costs with the provider. Relaunch tender
Solution implemented, expected results not achieved	L	H	Check the work organization and the knowledge level of staff. Do analytical audit inviting solution providers. Implement the results from audit.

7. List of Annexes

Annex 1. Template for assessment of statistical processes

Annex 2. Assessment of statistical processes

Annex 3. Information on the current IT infrastructure in GCC-Stat

Annex 4. Presentation of the SAS based option

Annex 5. Presentation of the MS MD ISDMS based option

Annex 6. Presentation of the Prognoz platform based option

Annex 7. Presentation of multi-national statistical systems

Annex 8. Full description of SDMX implementation

Annex 9. Analysis of the other IT solutions in international and national statistical organizations