

## TWINNING CONTRACT

### **Institutional Capacity Building for the Central Agency for Public Mobilisation and Statistics (CAPMAS) and Developing the Legal Framework for Statistics in Egypt**

**EG/07/AA/F106**



## MISSION REPORT


**on**

### ***Activity 4.2***

### ***Establishing the framework for a common data warehouse***

Mission carried out by  
Mr Søren Netterstrøm and  
Mr Povl Valeur, Statistics Denmark  
Cairo, 15-19 February 2009

Final version

EE2009:09		 <b>STATISTICS DENMARK</b>
Central Agency for Public Mobilisation and Statistics		Statistics Denmark

***PHARE 2005***

*Author's name, address, e-mail*

*Søren Netterstrøm  
Sejrøgade 11  
DK-2100 Copenhagen Ø  
Denmark  
Tel. +45 39173917  
sne@dst.dk*

*Povl Valeur  
Sejrøgade 11  
DK-2100 Copenhagen Ø  
Denmark  
Tel. +45 39173917  
pov@dst.dk*

## Table of contents

Executive Summary .....	5
Main conclusions and highlights. ....	6
1. General comments.....	6
2. Results.....	6
Data Mining.....	7
Metadata .....	8
Data warehouse(s) .....	8
DPA level .....	8
Micro data warehouse .....	9
Macro data warehouse .....	9
3. Action plan for establishing a common data warehouse .....	9
Annex 1. Terms of Reference .....	11
Background .....	11
Purpose of the mission .....	11
Expected Results .....	11
Activities .....	11
Tasks to be done by CAPMAS to facilitate the mission .....	12
Consultant and counterpart .....	12
The mission will be carried out jointly by:.....	12
Mr. Søren Netterstrøm, Statistics Denmark, and.....	12
Mr. Povl Valeur, Statistics Denmark .....	12
The beneficiary's counterpart will be Ms. Zeinab Gharib. ....	12
Timing .....	12
Report.....	12
Annex 2. Programme for the mission.....	13
Annex 3. Persons met.....	14
Annex 4. Modulus 11 .....	16
Annex 5. References .....	18

## List of Abbreviations

CAPMAS	Central Agency for Public Mobilisation and Statistics
ToR	Terms of Reference

## **Executive Summary**

**A model/framework for a common data warehouse has been discussed and it is the opinion of the consultants that the model will provide a good starting point for the further work towards establishing a common data warehouse in accordance to the action plan outlined in chapter 3. The consultants recommends that CAPMAS further investigate the model and adapt it to the specific needs and conditions at CAPMAS in order to take a strategic decision on this issue that may be integrated into the overall IT-strategy.**

**As an intermediate result it was acknowledged that there is a need for making two different data warehouses depending on the characteristics and usage of the data. There should be one data warehouse for microdata and another data warehouse for macrodata. In missions 4.8 and 4.10 the macro data warehouse will be further explored.**

During the mission the consultants met a dedicated and skilled staff working with high motivation and willingness to discuss detailed IT matters.

## Main conclusions and highlights.

### 1. General comments

This mission report was prepared within the Egyptian-Danish Twinning Project „Institutional capacity building for the central agency for public mobilisation and statistics”. This activity is the second activity within component 4, *Improved IT Function*. The objectives for this component are to give recommendations for an integrated IT function for central and regional offices with the MS operating system and MS office. Also for upgrading of statistical databases, including metadata and rules for statistical production and publishing are further elaborated. Dynamic and user friendly website with output database is implemented.

This activity will contribute to this objective and especially to the benchmarks set out in the contract: *By the 6<sup>th</sup> month, the action plan for establishing a common data warehouse.*

The concrete objectives of the mission were:

- Establishing the framework for a common data warehouse
- An action plan for establishing a common data warehouse

The consultants would like to express their thanks to all officials and individuals met for the kind support and valuable information which they received during the stay in Egypt, and which highly facilitated the work of the consultants.

The views and observations stated in this report are those of the consultants and do not necessarily correspond to the views of EU, CAPMAS or Statistics Denmark/Finland/Sweden/ Latvia/ Czech Republic.

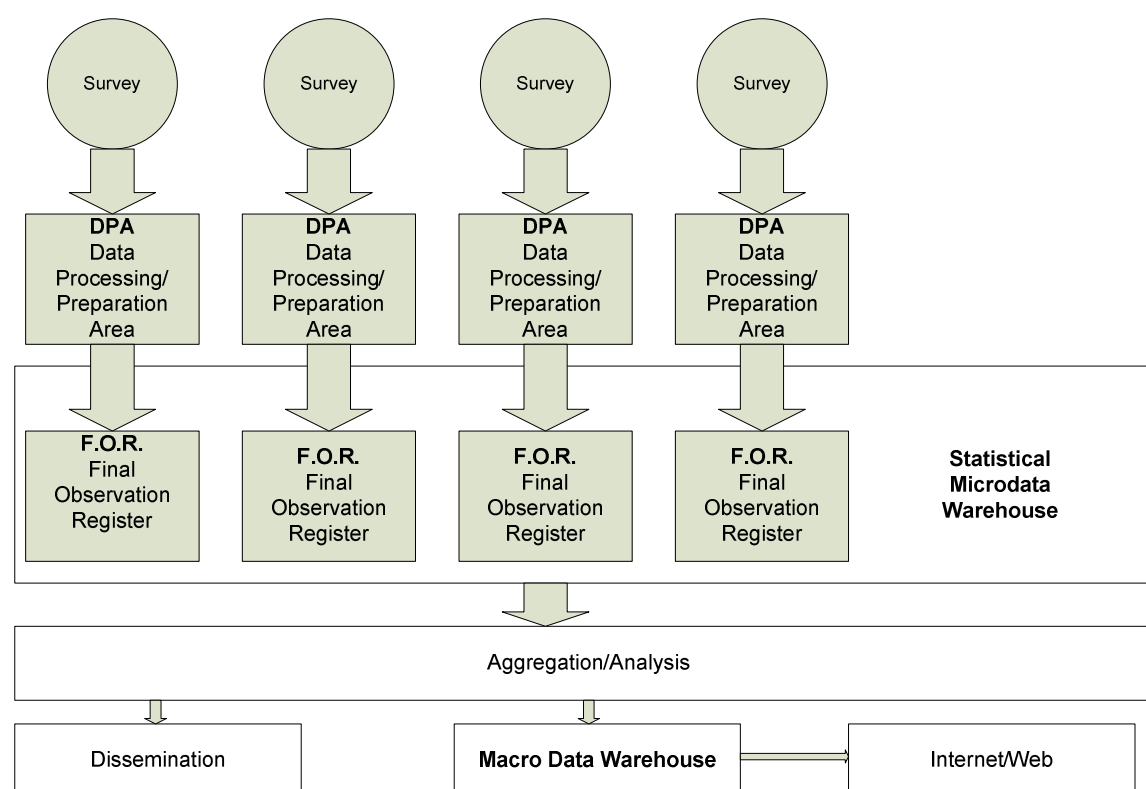
### 2. Results

The recommendations in this report are based on discussions and workshops with managers, employees and supervisors at CAPMAS.

All in all the consultants find that the IT function in CAPMAS is working at a professional level achieving good results. During the mission the consultants met a very dedicated and skilled staff working with high motivation and willingness to discuss detailed IT matters.

Some of the observations mentioned in this report should be analysed and discussed in more detail in the coming activities. In action 4.8 and 4.10 the macro data warehouse as a dissemination tool will be further examined.

One of the objectives for the mission was establishing a model/framework for a common data warehouse and to illustrate it the following diagram can be useful. It is based upon the works of Bo Sundgren (Annex 5) which again has been the underlying theory of statistics production in the Nordic countries and elsewhere during the last decades.



Data collected from a survey (or rather an instance of a survey) is loaded into a database in the Data Preparation Area (DPA) from where it will be processed in regard to Data Cleaning etc. This stage may be further subdivided into data entry, primary data cleaning on micro level and macro editing as appropriate.

When the processing of data has taken place in the DPA, the data will be loaded into a Final Observation Register (FOR). At this stage the data as a minimum should be formally correct and all errors that may influence the statistics in an unacceptable way should have been detected and corrected. The Final Observation Register is the primary result of the survey and should be the only source for producing tables and other materials for dissemination purposes.

The FOR has some distinct characteristics:

Data is static in the sense that it *cannot be updated*.

All attributes that are coded (i.e. sex, activity code) are linked to the appropriate classification.

For most surveys, there will be only one Final Observation Register. In some cases however, where there is a need for rapid results there might be a preliminary and final version of the FOR. They should be clearly separated and any figures disseminated should clearly identify if they are preliminary or final.

From the FOR the statisticians can make all kind of aggregations and statistical analysis eventually leading to dissemination.

This model and the terms used are referred to in the remainder of this report.

## **Data Mining**

It is the opinion of the consultants that data mining in general is **not** part of the core business of CAPMAS (or other statistical bureaus) at present and as such should be avoided unless a strategic decision is made to follow this path.

Data mining should be done by researchers on the level of FOR (eg. Statistical data warehouse) and it should be placed in a special research unit if decided.

On the DPA-level some techniques from data-mining can be utilized as part of the data cleaning process, where drill-down techniques could be used to determine the nature of “potential errors/suspicious values” (macro editing).

## **Metadata**

The question of the importance of Metadata was especially raised in the mission report from Activity 4.1 and was discussed further in this activity with focus on metadata in combination with data warehouse.

A model with two types of metadata was discussed:

- Metadata that can be used by statistical software
- Descriptive metadata for each survey

The first type of metadata is normally embedded with the data, i.e. resides in the same data base as the data or even in the data file itself. To meet the needs of different packages more than one format may be utilized. However, this type of metadata, once established is rather static. It could be considered to establish a single source (database) from where such metadata is drawn and could also be used as a point of reference for internal planning. It is however strongly recommended not to engage into the building of one single meta data base but rather to take a pragmatic approach of having a repository that makes it easy to locate and identify commonly used classifications and other types of metadata.

Also the second type of metadata (eg. Documents describing a survey or instance of a survey) should be put into a common repository (database). It will be essential for the end users understanding of the specific survey, but also for internal users not familiar with the specific survey.

The metadata for the macrodata warehouse can be word, excel etc file(s) – for the descriptive metadata part. These can be very useful for external as well as internal users. This kind of metadata should be updated whenever the survey is changed.

The structure of the metadata database should be kept as simple as possible.

## **Data warehouse(s)**

The discussions made it clear that it is necessary to distinguish between the different layers in the model. The need for microdata in the production of statistics has some distinct differences compared to the needs of a macrodata warehouse targeted at end users.

Today CAPMAS uses four different database tools: Oracle, Sybase, Terradata and MS SQL Server.

In the opinion of the consultants a single tool should be selected for each horizontal layer. Using a single tool opens for creation of reusable components across surveys and for use of common tools to access and analyse data.

## **DPA level**

For example the RDBMS used at the DPA-level should be limited to one tool (Oracle seems to be most widely used in CAPMAS today at this level). At the same time it should be ensured to use standard interfaces as far as possible since this will provide the highest degree of independence of distinct vendors. For example only standard SQL should be used to create tables, to insert and update etc and on the extraction side standard interfaces as ODBC should be used. This will ensure that the RDBMS can be replaced (with as little cost as possible) with a similar product from another vendor.



## Micro data warehouse

On the FOR-level CAPMAS has some experiences using the Teradata tool. This tool was not known beforehand by the consultants, but from the observations done it seems to fulfil the requirements for a DB at the FOR-level. It seems to have sufficient linkage between data and metadata and also seems to be working on static data (used in the way CAPMAS demonstrated) and because of the speed it is likely that it is internally working with some data structures suitable for a data warehouse with microdata. (The STAR and Snowflake model was discussed during the mission).

Teradata does support both a relational view on data and a datawarehouse view using unrelated fact tables using a star or snowflake schema. As discussed during the mission, the relational model should be kept to be able to explore the full richness of the survey material if this has more than one observation unit, i.e. census data about buildings, family and persons. However to facilitate the use of the data, fact tables should be created for each type of observation unit. This makes access to the data simpler and opens for a broad range of analytical tools to be applied.

It was demonstrated, that using Teradata it would take around 4 minutes to create a table on a fact table covering the whole census population of Egypt. This seems to imply that even the largest fact tables produced by CAPMAS can be processed using ROLAP (that is having a fact table with one entry for each observed unit). MOLAP, where some pre-aggregated cubes are used to speed up common request are used, seems not to be interesting, taken the speed of ROLAP and the expected rate of usage into account.

The micro data warehouse should contain data from all surveys conducted by CAPMAS including the census, surveys on economic unit, socio-demographic surveys etc. To what extent historic data for a survey should be included, where the survey consists of a series of survey instances, must be decided for each survey taken the cost of inclusion and expected further use of the data into consideration.

## Macro data warehouse

The macro database is intended primarily for dissemination purposes. It should consist of a set of ready made cubes and embedded metadata to ensure proper processing of the data and that information to correctly interpret the content of the cubes. As a tool for the data warehouse for macrodata it should be considered to use either PC-AXIS or PX-WEB since these tools are directed at making statistical metadata available to a broader audience by internet publishing. PC-AXIS is currently being used by more than 40 countries (including all the Nordic countries) for this purpose. Activity 4.8. and 4.10 will cover this in more detail.

It should be noted, that the macro data warehouse is likely to contain not only data that has its origin in the micro data warehouse described above. Data collected and processed by external partners may well be disseminated by CAPMAS. Currently the National Account is not produced within CAPMAS as an example. In the macro data warehouse, many cubes should have a time dimension (time series) and this may include historic data not covered by the micro data warehouse.

The strategic goals of using a data warehouse should be addressed in activity 4.9.

## 3. Action plan for establishing a common data warehouse

A number of actions are needed in order to establish a common data warehouse corresponding to the framework described in this report:

- **Make a timeplan**

- **Consider model as discussed**
  - Does it apply to CAPMAS needs on a strategic-, business- and technical level
- **Decide on model**
  - Go through model and adapt to local needs if needed
    - Eg CAPMAS might need an extra layer with MS-SQL that is used for data-entry before loading into DPA
  - Should the micro datawarehouse contain FOR both in a relational version and a version with fact table.
- **Choose one set of software for each horizontal layer (DPA, FOR, Analysis, Dissemination)**
  - Microdata warehouse (Teradata already in use)
  - Macrodata warehouse (Teradata/PC-AXIS?)
    - Activity 4.8 and 4.10 are related to this issue (dissemination from output database)
  - Select a single tool for development/redesign of surveys at the DPA level. If DPA is split into several layers different tools could be used for each layer. Existing applications at this layer may not need to be immediately converted if they meet the basic needs of the model.
- **Implementation**
  - micro data warehouse (FOR-level) is currently being investigated by the Teradata project. It is likely that some changes might be needed if the recommended model is used
    - consider to what extent historic data should be included into the micro data warehouse.
  - Macrodata warehouse
    - Await the results of mission 4.8 and 4.10

**It should be noted, that each of the layers identified could be implemented in a separate process as the boundaries between the layers and the purpose of each layer is well defined in the overall model.**

**When making an actual implementation plan, it should be considered to start with surveys where data is readily available in a suitable format for smooth inclusion in either the micro or macro data warehouse respectively.**

## Annex 1. Terms of Reference

EG/07/AA/F106

Statistics Denmark, International Consulting

8. juli 2009

POT/-

### Terms of Reference

*for a short-term mission to the Central Agency for Mobilisation and Statistics  
on*

#### *Activity 4.2*

*Establishing the framework for a common data warehouse*

## Background

CAPMAS and Statistics Denmark with partners have established a fruitful cooperation in the framework of Twinning. This twinning project is EG/07/AA/F106.

This activity is the second activity within component 4, *Improved IT-Function*. The objectives for this component are to give recommendations for an integrated IT function for central and regional offices with MS operating system and MS office. Also for upgrading of statistical databases, including metadata and rules for statistical production and publishing are further elaborated. Dynamic and user friendly website with output database is implemented

This activity will contribute to this objective and especially to the benchmarks set out in the contract: *By the 6<sup>th</sup> month, the action plan for establishing a common data warehouse is set.*

## Purpose of the mission

Based on the assessment report from activity 4.1, the task of the mission is to give recommendations for upgrading of statistical databases spread across different RDBMS, including metadata, data warehouse and data mining.

Also recommendations on a coherent database structure combining data input/collection, data processing and output

Identifying training needs for IT personnel in the area of database management.

## Expected Results

- Establishing the framework for a common data warehouse
- An action plan for establishing a common data warehouse (Benchmark)

## Activities

A tentative schedule for the mission is:

*Sunday 15 February*

Introduction to CAPMAS and overall discussion on the activity – RTA and BC project management

Meeting with component leader and the relevant staff

Determining the agenda for the mission

Purpose and goals for a common Data Warehouse – the vision of the IT structure in CAPMAS

Definitions and clarifications of terms, e.g. data mining

*Monday 16 February to Wednesday 18 February*

Discussions on the challenges of reaching for a common data warehouse, e.g.:

- Data availability
- Technical issues
- Staff requirements
- Training
- Input from different sectors
- organisation of the work
- Linking data from various sources
- Metadata
- Data mining

The beneficiary has especially requested input on

- Metadata
- Database structure in statistics Denmark
- Data mining

*Thursday 19 February*

Finalise the action plan for establishing a common data warehouse

Final discussions and clarifications with CAPMAS

Presentation of preliminary results and findings with BC project management

## **Tasks to be done by CAPMAS to facilitate the mission**

The beneficiary will arrange meetings with the relevant staff in CAPMAS.

## **Consultant and counterpart**

The mission will be carried out jointly by:

Mr. Søren Netterstrøm, Statistics Denmark, and

Mr. Povl Valeur, Statistics Denmark

The beneficiary's counterpart will be Ms. Zeinab Gharib.

## **Timing**

The mission will be carried out during 15-19 January 2009 in Cairo.

## **Report**

A final report from the mission should be made available not later than two weeks after the termination of the mission.

## **Annex 2. Programme for the mission**

### *Sunday 15 February*

Introduction to CAPMAS and overall discussion on the activity

Meeting with component leader and the relevant staff

Determining the agenda for the mission

Purpose and goals for a common Data Warehouse – the vision of the IT structure in CAPMAS

### *Monday 16. February*

Discussions on the challenges of reaching for a common data warehouse – focus on

- Data availability
- Technical issues

### *Tuesday 17. February*

Discussions on the challenges of reaching for a common data warehouse – focus on

- Data Mining
- Metadata

### *Wednesday 18. February*

Discussions on the challenges of reaching for a common data warehouse – focus on

- Establishing Action Plan for a common data warehouse

### *Thursday 19. February*

Finalising the action plan for establishing a common data warehouse

Final discussions and clarifications with CAPMAS

Presentation of preliminary results and findings with BC project management

## Annex 3. Persons met

### Consultants from Statistics Denmark:

<b>Name</b>	<b>Title</b>
1- Mr. Søren Netterstrøm	TODO Head of Division (Central IT Function)
2- Mr. Povl Valeur	Head of IT (Central IT Development)

### CAPMAS: <TODO – Peter sørger for at sende opdateret liste>

<b>Name</b>	<b>Title</b>
1- Mr. Yasser El Sayed	Head of Central Department for IT Training
2- Mr. Nabil El Hotey	Head of Central Department of Information system
3- Ms. Zeinab Gharib	General Director of Technical Affairs, Planning and Follow-up
4- Ms. Madeha Rashad	General Director of Data Software for Transportation & Communications
5- Ms. Azza Taha Taher	General Director of Data Software for Indicators & Trade
6- Ms. Abla Salah	General Director of Projects Development
7- Ms. Neveen Ali	General Director of Computer and Software Engineering
8- Ms. Amal Ali	General Director of Data Preparation
9- Ms. Wafaa Mohammed	General Director of Data Software for Industry & Construction
10- Mr. Ali El Kashef	General Department for Computer Engineering
11- Mr. Basel Monir	General Department for Preparing and Auditing

12- Ms. Salwa El Shazli	General Department for Data Bank
13- Ms. Fadia Hosney	General Department for Data Bank
14- Mr. Mohammed Ismail	General Department for Technical Affairs
15- Mr. Monin Abdel Fadeel	General Department for Technical Affairs
16- Mr. Abo Fadl Hammad	Advisor of Head of CAPMAS
17- Ms. Nadia Farid	Statistical Sector
18- Ms. Dorreyya Abbas	Statistical Sector
19- Mr. Tarek Rashad	The CAPMAS Presidency Office
20- Mr. Mohammed Zeid	Central Department for Training

## Annex 4. Modulus 11

During the mission it was discussed how to make unique numbers for businesses from the census as an intermediate solution until the business register (Activity 5.5) is ready. The consultants recommends that modulus 11 type of numbers are introduced.

Modulus 11 detects single digit errors, single transpositions, and double transpositions. Unlike other check digit systems, it is based on a weighted checking factor for each digit in the basic number. The modulus 11 check digit is obtained as follows:

Each digit position of the basic number is assigned a weighted checking factor . The following factors are assigned, starting with the units digit and progressing toward the high-order digit:

2 3 4 5 6 7 2 3 4 5 6 7 2 3 4 . . .

Each digit in the basic number is multiplied by its checking factor.

The products are summed and then divided by 11. The remainder is subtracted from 11. The result is the check digit.

Example:

Assume a basic number

5 1 6 1 9 2 8 7 2

Checking factors

4 3 2 7 6 5 4 3 2

Add the products

$20+3+12+7+54+10+32+21+4=163$

Subtract remainder from 11

$11-9=2$

Check digit =

2

Self-checking number

5161928722

If a check digit is generated using modulus 11 calculations and the result is 10, the check digit cannot be used and an error is returned. Modulus 11 check digits are the remainder from dividing the product of the calculations by 11 (see example above). Thus, if check digits are being generated for a continuous series of numbers, every eleventh number must be skipped to avoid this error.



If the product generated through the modulus 11 calculations is evenly divisible by 11 (no remainder), the resulting check digit is 11. In this case, the digit 0 is appended to the basic number.

## Annex 5. References

*Bo Sundgren: Guidelines for the Modelling of Statistical Data and Metadata*  
United nations, Geneva 1994

<http://www.unece.org/stats/publications/metadatamodeling.pdf>

The paper is rather theoretical and CAPMAS shouldn't try to implement it in total, but it gives some ideas for implementing a more pragmatic model and describes the theory behind the model proposed by the consultants.