STATISTICS DENMARK

Statistisk sentralbyrå
Statistics Norway

Statistiska centralbyrån
Statistics Sweden

MZ:2005:8

# Mission Report

## from a short-term mission on Data Modelling

## *31 January – 4 February  2005*

## TA for the Scandinavian Support Program to Strengthen the Institutional Capacity of the National Statistics, Mozambique

*Søren Netterstrøm*

Instituto Nacional de Estatística

Ref. no. 104.Moz.98

April 2005

*Søren Netterstrøm*
*Statistics Denmark*
*Sejrøgade 11*
*2100 København Ø*
*Denmark*
*Tel.: +45 39 17 35 84*
*sne@dst.dk*

# Table of contents

1 EXECUTIVE SUMMARY ........ 4

Appendix 1. Content of the course ........ 5
1.1.1 The micro data warehouse ........ 6
1.1.2 The macro data warehouse ........ 10
1.1.3 The dissemination database ........ 13
1.1.4 Metadata ........ 13

2 Appendix A. Creating a fact table from a questionnaire........ 15

3 Appendix B. Dimension tables. ........ 18

4 Appendix C. Data archiving ........ 20

Appendix D. Roadmap for implementation........ 21

Data Modelling........ 22

# 1 EXECUTIVE SUMMARY

**Scope of mission**
The main task of the mission was to carry out a workshop on data warehouse covering general data modelling and other issues related to creating a data ware house in a statistical office.

**The workshop**
The workshop was held over three days. The main points at the work shop was to look at the concept of a data warehouse based on Ralph Kimbals model and how this could be applied for a statistical office. A clear distinction was made between a data warehouse containing micro data (observation registers) and a data warehouse containing macro data (tables).
Details on the topics covered can be found in appendix 1.

**Main recommendations**
It is recommended that INE carries on with setting up the architecture of a data ware house system including both a micro data warehouse and a macro data warehouse and includes the dissemination database (PX-WEB) in the overall architecture.

It is recommended to see the data warehouse primarily as common library allowing any user within INE to get access to relevant data. From the data warehouse the user should be able to obtain a copy of the data in a relevant format of his choice for further processing.

INE should set clear goals for the data warehouse and the expected benefits for INE. IN the opinion of the consultants, the main benefit that can be obtained is an improved cooperation within INE by sharing micro and macro data across the organisation. By storing all survey data into a common data warehouse INE will benefit from a common standard format, improved security (backup) and improved availability.

The consultant strongly recommends building the data warehouse on standard components already in used within INE and to keep the system as simple as possible to meet to goals set. The system could be gradually developed.

**Concluding remarks**
The consultant would like to express their thanks to all officials and individuals met for the kind support and valuable information during stay in Mozambique, and which highly facilitated the work.

This report contains the views of the consultant, which do not necessarily correspond to the views of Danida or INE.

# Appendix 1. Content of the course

## Data warehouse at INE

*Scope*

The scope of the data warehouse at INE is to contain all survey data collected by INE, either through traditional surveys or from other sources (the central bank, line ministries etc.).
The data warehouse should also include data on aggregated levels ready for dissemination or for internal use, i.e. data for the System of National Accounts.

For reasons discussed below the actual data warehouse is split into a micro data holding and a macro data holding that together comprises the statistical data warehouse of INE.

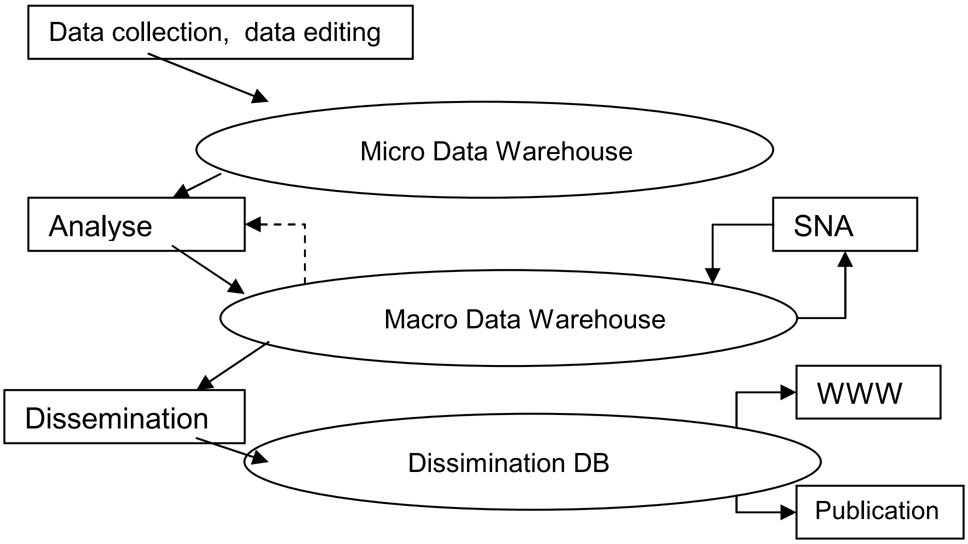
*A common structure*

The data warehouse offers a common structure for all micro data holdings and a common structure for all macro data holdings. This will in the long term lead to lower costs for maintenance within IT and minimize the number of products on which INE must keep sufficient knowledge.
The structures proposed will also lead to a more strict data discipline and ensure a basic level of quality.

*Not only data*

The data warehouse is not only holding data but must also hold metadata, that is, the data needed for proper understanding and usage of the data. For micro data this includes, but is not limited to, description of variables and, for classified variables, the classifications used (the value set).
For macro data (statistical tables) it is important to be able to handle footnotes on different levels.

*Suggested data flow*



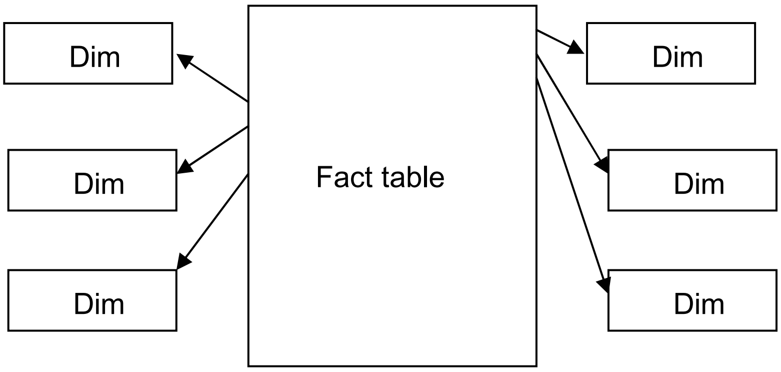
The above figure illustrates the main data flow with INE. Data is collected through surveys or from line ministries, the national bank and other external sources.
After being edited, to obtain data of a reasonable quality and to clean data (remove invalid codes etc), data can be loaded into the Micro Data Warehouse.

It should be noted, that some data from external sources are only used as input to the System of National Account or for dissemination purposes. Such data may be entered directly into the Macro Data Warehouse.
The primary source for analysis then becomes the Micro data warehouse, but data from the macro data base may be included as well, i.e. to calculate fixed prices the Consumer Prince index may be needed. Results from the analysis are stored in the macro database if it is suitable for dissemination or use within INE (primarily; National Account).
Not all data in the macro data base may be suitable for dissemination for confidentially reasons. Others are created with the purpose of being able to answer common questions, but where the level of detail is not suitable for general dissemination.

Data to be disseminated (through publications or on the Web) are transferred to the dissemination database.

### 1.1.1 The micro data warehouse

The micro data warehouse is using a modified star-model (Kimball).



The fact table stores the actual data collected. The Dimension tables holds information about classifications (or value sets) used for encoded variables.

The content of the fact table is divided into

- Dimension variables holding the encoded variables. A dimension variable is connected to a dimension table that holds the allowed value-set. Through referential integrity it is verified that each dimension variable only holds valid codes.
- Facts are numerical variables on which calculations can be performed, i.e. income in $10^6$MT, weight in kg, number of children in household etc. Facts are not associated with a dimension table.
- Identifiers are variables not used for analysis, but included to be able to make reference back to the original data source, i.e., the questionnaire.

*No updates*

Once a fact table has been loaded into the data warehouse, it should not be allowed to perform any changes to the content of the table.
The reason for this is to ensure that the statistics produced are consistent.
For some statistics where preliminary results are needed, a preliminary fact table may be loaded prior to the load of a fact table with final results. If (important) errors are found in

the statistics that requires corrections, i.e., when revised figures need to be published, a revised fact table may be loaded. It is however strongly recommended not to have frequent loads of such revisions and always in connection with a complete update of any prior published figures.
Likewise, the dimension tables may not be updated once loaded.

*Surveys and fact tables*

In general a fact table is related to an instance of a survey, i.e. each time a survey is conducted a new set of fact tables are loaded. Most surveys will produce one fact table, but larger surveys may produce 2 or more fact tables. The number of fact tables is dependent on the number of object type surveyed. There should be one fact table for each object.
In IFTRAB this results in 2 fact tables, one for households and one for persons as these are the object types. Even if there are special questionnaires for persons 7 years and above and for persons between 7 and 17 years, these does not lead to new objects because these would just be subtypes of persons.
In very special cases a fact table may cover more than one instance of a survey series, e.g., all prices collected during one year. However, this will complicate keeping track of preliminary and final data, so it is only recommended to do so if it is expected that data is loaded only once for each instance and no revisions takes place following this.

*Correspondence to questionnaires*

During the workshop it was demonstrated that there is a close correspondence between the questionnaire and the fact tables. Each question on the questionnaire in general turns into a dimension variable, a fact or an identifier. Appendix A contains a more detailed discussion of this issue.

It should be noted that, whenever a new questionnaire is prepared, it is possible to design the fact table and create the dimension tables so that these are ready before the actual data are ready for loading into the data warehouse. It is strongly recommended to do this concurrently with the questionnaire design because this operation may reveal weaknesses in the questionnaire design.

*Dimension tables*

The primary role of the dimension table is to contain the allowed codes for an encoded variable (a dimension variable) and explanatory text for the code.
This documents the content of the data warehouse and makes it possible to browse the data warehouse with access to the explanatory text rather than having to remember individual code values.
The dimension table also allows for introduction of building hierarchical recording schemes. If a dimension describes districts, then each district resides in a province.
The dimension table reflect this fact in the following manner

| District Code | District Name | Province Code | Province Name |
|---|---|---|---|
| 0101 | Lichinga | 01 | 01 Niassa |
| 0102 | Cuamda | 01 | 01 Niassa |
| 0201 | Changara | 02 | 02 Tete |
| … | | | |

On could further add region (code and name) or classify districts in any way, i.e. border district / non-border district, district having access to the sea or not etc.

This is a very powerful tool as it, in a simple manner, makes recoding directly available for analysis. When analysing the fact table, one could consider the fact table to have not only the variable district, but also the variables province, region, border district, access to sea etc.

Appendix B further discusses use and construction of dimension tables and their relation to classifications.

*Adding new regrouping*

It should be allowed to add new regroupings to a dimension table as this will not harm the consistency of the system. This in an exception from the rule stated above that dimension tables cannot be updated.

*Shared dimension tables*

There will be a number of dimensions that will be the same for a large number of fact tables across statistical domains. Examples are district, province, age, gender, and classification of economic activity.
Within a survey series (a repeated survey) this will be true as well. An effort should be made to share such dimension tables between fact tables. See also the discussion in appendix B.
One advantage of sharing dimension tables is that this would encourage users to reuse existing groupings of a classified item rather than inventing their own, in turn making the statistical system more coherent.

*Hierarchical fact tables*

Certain surveys, like IFTRAB, use a number of objects where the objects are in a hierarchy. In the case of IFTRAB we have household and person and each person is the member of one and only one household.

In this case all variables that apply to the household are variables about the person as well. I.e. if the household lives in a building made of cement blocks, then the person is a person living in a building made of cement blocks.

One way of describing this would be to duplicate all variables related to the household to the fact table for persons as well.

In such cases however, it is recommended rather to include the ID of the household in the fact table for persons to make a many to 1 relation between person and household. Then, rather than using the fact table, a (database) view could be constructed to reflect this hierarchy giving a full, virtual 'fact table' for person including all attributes for the person and the corresponding household.

It should be emphasised that this method should only be used, when there is a strict hierarchy as in the preceding example. In general, using relations between fact tables should be avoided. As an example, you could have a fact table about persons and a fact table about enterprises. Some of the persons are employed by an enterprise and some enterprises has persons employed, but not all persons are employed and not all enterprises has employees. In such a case, if we want to include information about the enterprise in the Person fact table we should duplicate this information in the person fact table. Fact tables are, in general, denormalised.

It should also be noted, that even if the relation between household and person would allow for constructs counting the number of persons in the household or the number of persons with certain characteristics in a household this should be avoided. The link should only be used to create an extended persons fact table, not to extend the household fact table. It may be used during data preparation to include such variables in the household record, i.e. number of persons, number of males, number of females, number of children etc.

*Using views*

Apart from using views to describe a hierarchy like above, views can be used to create clean fact tables, i.e., fact tables that only have dimensional variables and facts and thus leaving out the identifiers and other non-statistical data in the fact table.

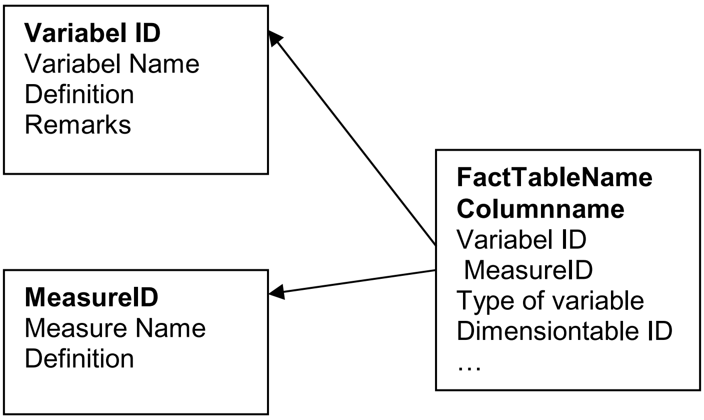
During analysis of data a subset of the data is often used, for instance only employed people or only people between 15 and 65 years of age etc.
There is often a small set of such subpopulations associated with a given fact table that are used repeatedly. In such cases one should consider creating a view establishing the subpopulation once and for all, ensuring the same constraint is always used.

### Metadata for micro data

Statistical metadata as been defined by Bo Sundgren as all the data needed to understand and make proper use of statistical data.
It is clear, that the star model on its own contains some important metadata. The dimension data can be seen as metadata giving the meaning of all encoded variables. Further it gives their relations to derived variables.
It is however clear that all the metadata needed are not present in the star model but must be supplied in a special extension or a dedicated metadata system.

### Variables

Each attribute of a fact table is indeed representing a statistical variable. There should be a place to give a more precise definition of each such variable. For each variable the actual question asked in the questionnaire should be used as part of the definition or a short description on how the variable is derived from other variables. The variable may include a more general description of the concept.
It makes sense to distinguish between the meaning of the variable and how the variable is actually measured. That is, the variable 'actual place of residence' is defined as the place where a person resides at a given point of time. This is true regardless of whether this is recorded as the province or as locality or even full address. For encoded variables, this information is already present in the dimension tables since they give the possible encodings of the variable. For facts, this needs to be recorded as the measurement unit (Count, Meticais, Million of Meticais, Kilos, Pounds, Minutes, Hours etc).
Since many fact tables may contain the same variable, either because they belong to the same survey series or because the variable is a commonly asked one, like gender, it is proposed to create metadata about variables in the fact tables using the following structure



All columns in a fact table refer to a variable. If the column represents a fact (a measurement) it should refer to a Measure to information about the unit used.
The number of variables across all surveys conducted by INE may be very large indeed. It may thus be useful to group variables by survey series to make it easier to locate an existing variable for reuse. A small number of variables may be declared as global

(associated indirectly with global dimension tables). It may also be useful to allow a variable to be associated with more than one survey, as it should be encouraged to reuse variables (whenever applicable) to make the total statistical output of INE coherent.
The number of measures will probably not be very large (less than 100). So it does not seem a problem to make a table containing all measures globally shared.

*The fact table*

Each fact table is related to a survey. There should be some formal descriptions of a fact table that may include a short description of the object type (person, household, enterprise, an event etc) and how the actual population is defined. For surveys the sample size and response rates may be noted. The time (date or period) covered should be indicated as well. Finally, through a link to the survey, there may be a link documents with more in depth information about the actual survey or the survey series.

### 1.1.2 The macro data warehouse

The purpose of the macro data warehouse is to contain the statistical estimates that are the result of analysis of micro data as well as other macro data that are useful within INE or used for dissemination purposes.

The content of the macro data warehouse can be seen as statistical tables. To allow statistical tables to be represented as fact table some extensions to the star model are needed. With such extensions it is possible to convert statistical tables into fact tables and associated dimension tables.

As an example assume a table with the following simple structure

Number of persons distributed by province, urbanity and gender

| | | Gender | | |
|---|---|---|---|---|
| Province | Rural/Urban | Male | Female | Total |
| 01 Niassa | Rural | | | |
| | Urban | | | |
| | Total | | | |
| 02 Cabo Delgado | Rural | | | |
| | Urban | | | |
| | Total | | | |
| … | | | | |
| Mozambique, total | Rural | | | |
| | Urban | | | |
| | Total | | | |

This could be expressed by the following fact table

| | |
|---|---|
| Province | Dimension |
| RuralUrban | Dimension |
| Gender | Dimension |
| NumberOfPersons | Fact |

At a first glance, the similarities between the micro data warehouse and the macro data warehouse might suggest, that the same structures could be useful for both types of warehouses. However, as demonstrated below, this is not entirely the case.

### Overlapping dimensions

As the preceding example shows, the values of the dimensions variables may include overlapping categories, in this example the totals. Contrary to this, in the micro data warehouse no dimension variable must use overlapping values because they are representing a single level of a classification.
Further, while a fact table (or cube) in the micro data set always can be collapsed over its dimensions, this is not true for a macro data set if it contains overlapping dimension variables.

### Indexes

A macro table may contain an index rather than absolute figures. An example would be the Consumer Price Index. In this case it does not make sense to perform further calculations on the table. The only useful operation would be to select parts of the table for display.

### Time series

A macro table often contains a time series, i.e., time is one of the dimensions of the table. Where each micro fact table refers to a specific point in time or a specific period, a macro table may cover a much broader set of time.
Time series are updatable in the sense that it should be allowed to add new elements to the time dimension as new data becomes available.

### Dimension tables

Dimension tables are in many ways similar to the dimension tables of micro data. However, for times series the problem of slowly changing dimensions may occur.
As an example consider external trade. According to the Harmonised System (HS) if time is 2001-2002 then 2001 will be according to HS 1996 while 2002 will be according to HS 2002. This could generate the following:

| HS 96/02 | 2001 | 2002 |
|---|---|---|
| **0101** Horses, asses, mules and hinnies, live | | |
| **010111** Horses, live, purebred breeding | | n/a |
| **010119** Horses, live, except purebred breeding animals | | n/a |
| **010120** Asses, mules and hinnies, live | | n/a |
| **010110** Pure-bred breeding animals | n/a | |
| **010190** Other | n/a | |

What happens is, that 0101 is the same between the HS 96 and HS 2002, but because of low traded of asses, mules and hinnies, the subdivision was changed in the 2002 version.

Where a dimension in a micro fact table will be either HS 1996 or HS 2002, a dimension in the Macro system must take changes over time into consideration. There is further a need to mark that a cell can logically have no content, that the value is “not applicable” (n/a).

### Footnotes

Data in a macro table may need to be annotated by footnotes.
Some footnotes may be used to indicated, that certain figures are provisional estimates or revised figures. Such footnotes could be created by adding a special section to the fact table where footnotes are treated much in the same way as dimensional variables. I.e., the fact table contains a code, and an auxiliary table contains the explanation for the code. (Even if these footnotes look like dimensions they are not to be used in any way similar to dimensions).
This mechanism could also be used for handling the situation described above with slowly changing dimensions by having an attribute to mark the value as not applicable. Another use could be to mark a value as confidential, not to be released.

The above technique is used in the GESMES/CB that is used for sharing statistical information between the central banks in Europe and when reporting from national statistical offices to Eurostat.

Another type of footnotes may be explanatory text connected to a specific value of a dimension, the intersection of such values from two or more dimensions, the whole table or a single cell.

Such footnotes could be expressed, using the above table as the base, in a structure like

| Province | Urbanisation | Gender | Note |
|---|---|---|---|
| 01 | | | The figures for Niassa are estimated based on the census 1997 |
| 02 | | 1 | Males in Cabo Delgado ….. |

This resembles the way PC-AXIS handles footnotes and a similar construct is available in the GESMES/CB mentioned above.

*Metadata for macrodata*

As for the micro data warehouse, the fact table, dimension tables, and the footnotes do not hold all the metadata needed. Further information that seems to be required for each fact table in the macro data warehouse is:

| | Example |
|---|---|
| Title | Number of persons distributed by province, urbanity and gender 1997-2001 |
| Source(s) | Census 1997, Population estimates |
| Released | 01/01/2002 (day the data was stored) |
| Created by | The person who created the information |
| Confidentiality | Mark if data can be used directly for publication |

Just like for the micro data a reference to variables may be desirable. However, if included it may be made optional.

The number of new fact tables in the micro data warehouse will be limited, it is estimated to be less than 100 fact tables (excluding revisions that do not require much metadata) will be created per year, and most of these will be new monthly or quarterly updates of existing fact tables. Due to the fact that these tables can be planned and (for the most part) documented well in advance.
Contrary to this, the number of new tables in the macro database will, by nature, be considerable higher and often may not be planned in advance.

Finally, if users feel it as a burden to put data in the macro data warehouse, experience from Denmark suggest that they will try to bypass the macro data warehouse using alternative ways for storing and retrieving the tables produced.

*PC-AXIS*

Since PC-AXIS has been chosen as a strategic tool for dissemination, care should be taken to ensure that it will be easy to create PC-AXIS files, including the metadata from data in the macro database.

*Quality*

The need to include data about quality (confidence levels) should be considered.

### 1.1.3 The dissemination database

The main purpose of the dissemination database is to hold figures to be disseminated regardless of the form of dissemination.
Where the macro data warehouse is for internal use only and may contain data that should not be disclosed due to confidentiality issues or because they are not released yet, data in the dissemination database has been released for publication.

The format of the dissemination at INE could be PX-WEB in order to make the dissemination database directly available through the internet portal.

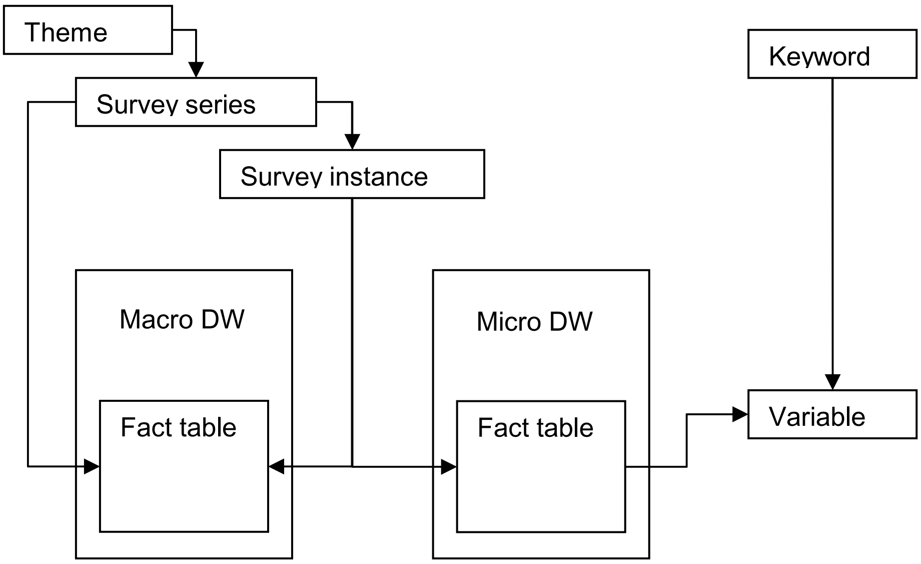
### 1.1.4 Metadata

The need for metadata in order to document the data of the system has been covered in relation to the micro and macro data warehouses.

Another use of metadata is to help locate a needed piece of information. As the number of fact tables in the data warehouse grows it will become increasingly difficult to find a certain piece of information without the help of such a system.

One way to locate a piece of information could be to drill down through a structured list until the information needed is located. Another method is search by keyword where a certain keyword would list fact tables related to the keyword. You could compare this to using either the list of content often place in the front of a book or to use the index normally placed at the end of a book. Depending on the kind of information you are looking for and how familiar you are with the structure each method has its advantages.

In order to establish such a system a structure like the following may be useful.



*Themes* A theme is a division of the universe covered by the statistics into a set of themes. Themes may be population, education, health, agriculture etc. They are broad groups, like the one used on the WEB portal and in the statistical yearbook.

*Survey series*

A survey series covers one or more surveys that used the same survey design and the same methodology - with no or only small changes between the surveys. Examples are the Consumer Price Index (CPI), QUIBB and perhaps the Population Census. It should be recognised that a survey may cover aspects of more than one theme.

*Survey instance*

A survey instance is a survey carried out at a specific point of time. Examples are CPI January 2004, CPI February 2004 .., IFTRAB 2004/2005 and Population Census 1997.

*Micro DW fact table*

As discussed above each survey instance produces one or more fact tables in the Micro DW and each fact table in the Micro DW comes from one survey.

*Macro DW fact table*

For the macro DW the situation is more complex. The Macro DW may be divided into two types of tables. Those that are the result of analysis on one fact table in the Micro DW, that is, they are related to one specific survey instance. The other group include time series, and may be related to one or more survey instances and even to one or more survey series.

*Variables*

Variables, as described under Micro Data Warehouse, are descriptions of the attributes of a fact table. Note that many facts (across a number of fact tables) may be associated with the same variable - even if they are using different value sets (i.e., different dimension tables).

*Keywords*

Keywords are a set of selected words to describe phenomena’s about which statistics exists. Examples are import, export, unemployment, malaria etc.

Keywords could be associated with variables and in this way give an indirect connection to fact tables in the Micro DW. How they are connected to the Macro DW if variables are not used is an open question.

Another possible source of keywords is dimensions. If we have the dimension gender, we might have the keywords male and female. However, where is the limit? It we have the dimension for type of goods according to the Harmonised System; we get a number of very detailed keywords. Are they needed? Are they relevant?

Bo Sundgren, one of the leading authorities on statistical metadata, once at a conference in Stockholm asked the question “What do we know about refrigerators”? The question has since been referred to as Sundgren's refrigerator.

We should probably not answer that question through keywords. At the same conference the Dutch delegate demonstrated how this could be solved using a tool similar to the one used when making queries on the web. It traversed all variable definition, all table titles and all dimension data looking for the word refrigerator.

*Balance*

The metadata system may very well be the most complex part of a Statistical Data Warehouse. The ideas outlined above may be used as a starting point for an internal discussion at INE on what metadata system should be established.

There is a need to find the balance between the resources needed to establish and maintain the system and the need for information and functionality, a balance that only INE can define.

*Planned tables*

Apart from giving information about actual fact tables in the data warehouse the metadata system might also contain information about planned tables and expected day of release. Such a feature might be useful for users like the editor of the statistical yearbook or the national account who would then be aware that new data might be available within a short time frame.

# 2 Appendix A. Creating a fact table from a questionnaire.

Using the questionnaire for IFTRAB 2004/2005 as an example it was demonstrated how a questionnaire can be turned into a fact table.

On the first page of the questionnaire, we find

Provincia ○ ○
Distrito ○ ○
Posto Administrativo ○ ○
Localidade ○ ○

In a traditional design, you would probably make 4 distinctive fields, one for each of the variables. However, a closer look to this structure reveals that this is a true hierarchy. The meaning of Localidade = 01 can not be known without knowing the preceding three levels and the meaning of Posto Administrativo and Distrito are in a similar way dependent on the previous levels.

So what we should have is one **dimension variable** (Localidade) that will be an 8-digit number. Associated to this variable there will be a Dimension table giving:

Localidade (8 digit code)
Name of Localidade (text)
Posto Administrativo (6 digit code)
Name of Posto Administrativo (text)
Distrito (4 digit code)
Name of Distrito (text)
Provincia (2 digit code)
Name of Provincia (text)

If Localidade can be classified otherwise, such classifications may of cause be added the dimension table.

The next item is
Urbano/Rural (1,2) ○

This is clearly a **Dimension variable**. The associated Dimension table is very simple having two columns

| **Code** | **Text** |
|---|---|
| 1 | Urbano |
| 2 | Rural |

The next two lines

Número de area de enumeração ○ ○ ○ ○
Número do agregado familiar ○ ○ ○

Do not seem to have any statistical meaning (after weights has been assigned).
However, to be able to reference back to the questionnaire they should be retained as **identifier variables**. They must also be kept to allow for a hierarchical structure between person and household as they identify the household uniquely. They should be concatenated into a single identifier variable in the fact table.

The remainder of the front page seems to be used only for administrative purposes during the conduct of the survey and the processing of the questionnaire and such information does not go into the fact table.

All the above information was related to a household, so it should go into the fact table **Household**.

From page 2, the questions are about individual persons in the household calling for another fact table, **Person**.

The Person table should include the values for enumeration of the district and the family number (taken together; the household ID) and add an extra ID for the person number.

The first question is related to head of household and is a coded value, so it turns into a dimension variable as does the next question specifying the gender of the person.

The third question is age (in whole years). This can be seen as a dimensional variable, allowing for a dimension table that creates new classifications of 5-years age group, 10-year age group or any other age grouping used within INE.
However, age could also be used as fact. It may not give much meaning to calculate the sum of ages, but average age may be useful.
For that reason, **age may be seen both as a dimension variable and as a fact**. This could be obtained either by duplicating the field (creating age_1year and age) or by allowing the metadata-system to allow a field to have this double function.

The next question seems only to be **a control question** (age over 7) and **should not be included in the fact table**.

The remaining questions on page 2 and page 3 are all simple dimensional variables, with the exception of the last question on page 3.
This is a **multiple response question**. **For each possible reply (01, 02 .. 06, 96) we must have a separate dimension variable**. All these dimension variables could share the same dimension table

| Code | Text |
|---|---|
| 1 | Yes |
| 2 | No |

Page 4 consist of a mixture of multiple response questions and simple dimension variables causing no comments.

Page 5-7 is just a repetition of page 2-4 allowing for another 10 persons.

Page 8 leads us back to questions related to the household, so everything on page 8 goes to the fact table for household.

Of interest here is question 30 that asked for the time in minutes (<997) used each day for a number of activities but which furthermore set aside two numbers (998, 999) for special responses ('not known' and 'not applicable', respectively)
By nature, the response to these questions is facts, holding a value in the range from 0 to 997. However, 998 and 999 are used to mark special responses and thus are more like dimension variables. One way to represent this dual usage is to create a dimension variable that has three values:

| Code | Text |
|---|---|
| 0 | Reply obtained |
| 998 | No reply |
| 999 | Not relevant |

and a fact variable (time), that will get the value NULL when the response is 998 or 999.

It may also be interesting to construct a new variable that classifies the response into some groups, i.e.
0 minutes
1-30 minutes
31-60 minutes
61-240 minutes
240+ minutes
or whatever is appropriate. This in turn would be a dimensional variable and could be used as such for cross tabulations.

It should be noted, that the rule that fact tables cannot be updated has an exception. It should be allowed to create such a new derived variable (based on the value of one or more existing variables) to the fact table, as this will not affect the consistency of the results produced from the fact table.

Such derived variable can be derived in many ways, like taking the sum or average of some variables, or it could be the result of a logical expression (if male and age > 17 and employed).

# 3 Appendix B. Dimension tables.

A dimension table should hold one row for each valid code of the dimension variable it is associated with.

The dimension tables should always include at least 2 columns, code and name, i.e., the dimension table for gender would be

| Gender | GenderText |
|---|---|
| 1 | Male |
| 2 | Female |

Most dimension tables will also be used for making one or more hierarchical groups. An example could be the dimension table for 1-year age group:

| Age | AgeText | Age5 | Age5Text | Age10 | Age10Text |
|---|---|---|---|---|---|
| 0 | 0 year | 1 | 0-4 years | 1 | 0-9 years |
| 1 | 1 year | 1 | 0-4 years | 1 | 0-9 years |
| 2 | 2 year | 1 | 0-4 years | 1 | 0-9 years |
| … | … | … | … | … | … |
| 5 | 5 year | 2 | 5-9 years | 1 | 0-9 years |
| … | … | … | … | … | … |
| 10 | 10 year | 3 | 10-14 years | 2 | 10-19 years |
| … | … | … | … | … | … |

*Don't snowflake*

The above table is deliberately of non-normalized form. It is clear, that a normalized version of this would be

| Age | AgeText | Age5 |
|---|---|---|
| 0 | 0 year | 1 |
| 1 | 1 year | 1 |
| 2 | 2 year | 1 |
| … | … | … |
| 5 | 5 year | 2 |
| … | … | … |
| 10 | 10 year | 3 |
| … | … | … |

| Age5 | Age5Text | Age10 |
|---|---|---|
| 1 | 0-4 years | 1 |
| 2 | 5-9 years | 1 |
| 3 | 10-14 years | 2 |
| .. | | |

| Age10 | Age10Text |
|---|---|
| 1 | 0-9 years |

| | |
|---|---|
| 2 | 10-19 years |
| … | |

Such a construct would be known as snowflaking the dimension table. But this should be avoided as it adds complexity to the data warehouse without adding value. Software trying to utilize the star schema becomes more complex to construct (and to use) in this case as the simplicity of the star schema is lost.

*Construct dimension tables from normalised tables*

In the example above, it is important that the text in Age5Text is exactly the same for all rows where the code is the same, just as it would be guaranteed in the snowflaked model. The result of

```
Select distinct Age5 from Dim_Age
```

And

```
Select distinct Age5, Age5Text from Dim_Age
```

And

```
Select distinct Age5Text from Dim_Age
```

should give exactly the same number of rows

One way of ensuring this might be to initially construct the dimension table using a normalized hierarchy and then simply construct the dimension table from that. In the above example this would be

```
Create table Dim_Age as
Select A.Age, A.AgeText, B.Age5, B.Age5text, C.Age10, C.Age10Text
From base_age A,  base_age5 B, base_age10 C
Where a.age5 = b.age5 and b.age10 = c.age10
```

This method may be useful when handling large dimension tables and makes it easier to maintain such tables, appending new groupings or making new versions where some codes are added or deleted.

*Classifications*

It should be noted, that dimension tables as described above are an excellent way to describe hierarchical classification.

# 4 Appendix C. Data archiving.

The content of the statistical data warehouse is one of the most valuable assets of INE. Even if it is true that the value of statistical data may decline over time, the data in the micro data warehouse may be useful for research in many years to come.

The actual data warehouse must be based on currently available technology and might even end up using proprietary formats. However, looking just 10 years ahead, such formats may be outdated and difficult to access. It may also at some point be suitable to change software (including operating system). It is therefore important to archive the contents of the data warehouse in a format that is likely to be able to survive for a long time. It is also important to have the option to recreate this data in the case of a disaster.

The Danish State Archive, facing enormous problems with electronic archives dating back from the 80's and before, has during the late 90's made a pioneering work in this area that has attracted attention worldwide. The solution they propose is to stick to a very simple format. All data are kept as pure ASCII-files and metadata are stored in XML according to a specific template.

As the data warehouse on its own imposes strict structures upon the data, a simple version of this system could be applied and the actual work involved with archiving would be minimal.

It is supposed, that each fact table and all dimension tables associated with the fact table (whether shared or not), are saved on CD-ROMS (or DVDs) as flat comma-separated files. Metadata is then stored using XML, also using pure ASCII files.

Each fact table from the micro data warehouse and its associated dimension tables and metadata-information is burned on one or more CDs (or DVDs). The first CD in the series should contain an index of the files actually stored and where in the set (each CD is given a unique volume ID). All CDs are created in two or more copies and stored in a safe place. At least one copy is stored outside the premises of INE (State Archive?).

The Danish system also allows for storing documents, like the questionnaire, as image files (using a simple TIFF-format) It should be considered to include this possibility in the archiving system to complete the metadata.

All data in the micro system should be archived immediately after they are loaded.

For the macro data warehouse similar procedures should be put in place, but in this case one should probably bundle many fact tables (from the same survey) on a single set of CDs, and one may accept do so with some time delay.

# Appendix D. Roadmap for implementation.

The role of the statistical data warehouse is to streamline the statistical production pipeline from the time the cleaned data arrive (and end up in the micro DW), through the analysis made in order to create statistically valid, and interesting, data (which end up in the macro DW), to publishing of chosen data (kept in the dissemination DB).

The streamlining is a result of making the data maximally available within INE and always using a common format so that data can more easily be retrieved -- the associated metadata system also play a crucial
role in this respect. Further, the metadata and annotation system can be used for documenting the quality of data, in turn increasing the credibility of INE.

Getting there is no simple feat as involvement of many different directorates is needed. This is particularly true of the work with the macro DW. Thus it is suggested to start by looking at the micro DW and the dissemination database.

The initial fact tables of the micro DW more or less will be given by the design of previous surveys. One can then make these fact tables and try to identify shared dimensions. For shared dimensions one in turn can attempt to use standardised versions from diverse international bodies, such as the UN. This latter work, and creating the associated metadata, is not a small task, but a worthwhile one because the knowledge inherent in such metadata can be used to inform survey design in the future: Using a relatively standardised set of metadata increases the possibility of comparing data and making time series while retaining a high quality level.

The dissemination database probably should be structured around the needs of users. For instance, one could structure the data base so that it was easy to find information related to a.o. the PARPA scheme or the millennium development goals.

The macro database should link up the micro DW and the dissemination DB in a way that supports the workflow during statistical analysis. Obviously, this requires a heavier involvement of statisticians. But to make sure that the system is well conceived one should at a relatively early time make a prototype macro DW containing aggregated/-analysed data of just a single survey. Since IFTRAB is the first survey more or less conforming to the new quality guidelines it may make a worthwhile example, allowing for experimenting with the metadata system.

**TERMS OF REFERENCE**

**for a short-term mission**
**on**

# Data Modelling

**January 31 – February 4, 2004**

within the Scandinavian Assistance to Strengthen the Institutional Capacity of INE/Mozambique

*Consultant: Soeren Netterstroem*
*Counterpart: Anastácia Honwana & DISI's applications group*

## Background

We have to start preparing for the data warehouse implementation. The critical part of this effort is to get the metadata right. We therefore have to make sure that our skill level is sufficient for execution of this part of the strategy. The DISI application group already has some experience in data modelling, but we both need to ensure that we have a common language and reference frame for the discussion of problems and that we have some understanding of the modelling techniques normally used for data warehousing. Reading the same books will help, but time for discussion is also needed. Preferably, some of this discussion should take place with the guidance of someone, such as Soeren Netterstroem, who is experienced in these issues.

### 4.1.1.1.1 Objectives

- To jumpstart the (DISI part of the) data warehousing project.
- To improve skills in data modelling and knowledge of data warehousing.
- To improve requirements analysis and implementation awareness.
- And to have lots of discussions.
- To make a preliminary design for the data warehouse architecture.
- To make a preliminary roadmap towards implementation.

### 4.1.1.1.2 Expected results

Improved knowledge of, and a common reference frame regarding:

- Conceptual versus physical modelling
- Dimensional modelling
- Object modelling
- Normal forms
- Data warehousing in general, including potential architectures.
- Requirements analysis.

And creation of:

- A data warehouse architecture
- An implementation roadmap

**Activities:**
The form will be that of a workshop with a mixture of presentation, group work and discussion. The subjects to be covered are those given in the objectives section as well as their relation to the current situation and discussion of how to proceed, i.e., to put it all into the framework of the task ahead of us.

The primary participants will be the DISI applications group, Anastácia Honwana, Clara Panguana and Karsten Bormann.

**Tasks to be done by INE to facilitate the mission**
The discussions will focus on the methodologies outlined by Kimbal & Ross for which reason it may be advantageous to buy a few copies of that book:

Ralph Kimbal, Margy Ross: The Data Warehouse Toolkit - The complete guide to dimensional modeling, 2nd ed Wiley 2002 416pp

**Preliminary program**
The workshop takes place from 8.00-12.00, with the possibility to use the afternoon for further discussions, if needed.
Monday: Data Warehouse versus Data Bases, Data warehouse in a Statistical Office
Micro data versus micro data. The needs of INE.
Tuesday: The star-model (Kimball). The role of metadata.
Thursday: Testing the Star-model on an example from INE
Friday: Implementation roadmap

Wednesday is a National holiday, so no meetings are expected.

**Consultant and Counterpart**
*Consultant:* Soeren Netterstroem
*Counterpart:* Anastácia Honwana

**Timing of the mission**
January 31 - February 4 (5 workdays).

**Report**
Will be a very short overview of the discussions and any recommendations that comes out of the discussion, particularly regarding the data warehouse architecture. The more specific analysis should be detailed by DISI staff and others in the coming following months.