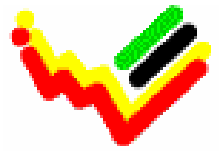
MZ:2006:8

PC-Axis and the Macro Data Warehouse

6 - 16 June 2006

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

Jesper Ellemose Jensen



Instituto Nacional de Estatística

Ref. no. 104.Moz.98 July, 2006

Jesper Ellemose Skou Jensen
Statisitcs Denmark
Sejrøgade 11, 2100 Copenhagen Oe, Denmark
jej@dst.dk
+ 45 39 17 30 56

Table of contents

List of abbreviations

CO	Scanstat Coordination Office in Statistics Denmark
Danida	Danish International Development Assistance
DKK	Danish Kroner
DSt	Statistics Denmark
EUR	European Euro
INE	Instituto Nacional de Estatística, Mozambique
INE-P	Instituto Nacional de Estatística, Portugal
IIS	Internet Information Server (A Microsoft product)
LTA	Long Term Advisors
STA	Short Term Advisors
MZM	Mozambique Meticais
NOK	Norwegian Kroner
PX	Family of Software produced by Statistics Sweden
Scanstat	Consortium between Statistics Denmark, Statistics Norway and Statistics Sweden
SCB	Statistics Sweden
SEK	Swedish Kronor
SSB	Statistics Norway
USD	US Dollars
ZAR	South African Rand

1 EXECUTIVE SUMMARY

The mission should be seen as a logical follow up on the mission on Internet Database pilot, also conducted by Ellemose Jensen in June 2004. Then the main task was to train DISI staff in the use of PC-AXIS / PX-Web, and to install a PX-Web version for test and development purposes.

Since then INE's has launched a new version of their web portal www.ine.gov.mz and the original PX-web solution has been a part of that portal. As the PX-Web installation was originally intended as a test and development environment, and not as a production solution, the layout and structure was not as compliant with the design and layout of the INE Portal, as desirable.

Upgrading PX-Web

The main undertaking of the mission was the installation and putting to work the latest version of PX-Web. The new version integrates better with the INE portal regarding look and feel and also has an integrated search function which will make it easier for end-users to find and extract data from the portal.

The other activities of the mission focused on discussions on the Data Warehouse project and a small number of pilot transformations of data.

Continue with the Data Warehouse

It's recommended that INE continues to work with the Data Warehouse implementation, in which PX-Web is an important part.

Use a dedicated server

As PX-Web is in production on the INE portal, it is very important that PX-Web is moved to a dedicated web server. This will strengthen the production environment and make the data much more accessible.

Train additional staff members

As the staff situation at the IT department improves it is strongly recommended that additional staff members are given training in the PC-AXIS / PX-MAKE / PX-WEB software. At present INE is highly dependent on one or two staff members.

Future support

Some working time was spend on discussions with the Danish Embassy, INE management, and the local team leader regarding the future financing and organisation of the Scanstat project. To some degree this was a diversion from the main objectives of the mission. However, during the mission a firm commitment was received from the Scandinavian donors for continued and extended support for the period 2006 – 2007.

2 INTRODUTION

The mission was a follow up to the mission conducted in June 2004 by the same consultant. The purpose of the 2004 mission was to introduce the PC-Axis / PX-Make / PX-Web software at INE. Two years ago a pilot installation was completed and later set to work as part of the INE portal (www.ine.gov.mz). So in 2006 it was due time for an update of the original pilot installation and retraining of INE staff, as a significant number of the staff members who received training in 2004 has now left INE.

PX-Web will play an important part in the coming data warehouse at INE, as the primary output resource on the internet. As INE prepares to conduct the 2007 Census, proper procedures for long term storage and management of data at both the micro and the macro level through the data warehouse will be even more critical for the efficient day to day running of the statistical production. It is hoped that the PX-family will play an important part in this work.

I would like to express my thanks to all officials and individuals meet during the mission. They all provided me with the necessary information in a kind and open atmosphere which greatly facilitated my work in Mozambique. But specially, I would like to thank Mr. Lars Carlsson for being an excellent host and for a very constructive sharing of his thoughts on the project.

Finally it should be noted that this report contains the views of me, as the consultant, and that they therefore do not necessarily correspond to the views of Statistics Denmark, Danida or INE.

3 ACTIVITIES DURING THE MISSION

The main objective of the mission was to install and putting to work the latest version of PX-web. The latest version of PX-Web was - although it is called 2005 – released in the middle of May 2006, so INE is now totally up to date regarding the PX-Web database software.

3.1 PX-Web

Installing PX-Web

The latest version of PX-web was installed on an empty Windows 2003 server for test at development purposes. During the process, knowledge on setting up and installing PX-Web was transferred to the local counterpart, Mr. Anselmo Nhane.

The interaction between the IIS (Internet Information Server) and the PX-Software was demonstrated to the local counterpart.

Windows 2003 reinstallation

Do to unforeseen circumstances it was necessary to reinstall Windows 2003 on the development server twice. Although the time spend doing this could have been spent on more in depth training in other areas of the software family, it gave the counterpart an excellent opportunity to improve his skills and practice the steps involved in installing PX-Web.

However the unforeseen circumstances also strongly highlighted the need for continued work and focus on the security aspects of the IT infrastructure at INE.

Revised design

The earlier version of PX-Web was launched on the internet together with the present version of the INE homepage www.ine.gov.mz in 2004. This first version of PX-Web was originally intended as a prototype / pilot installation for INE to become familiar with the software family. The “look and feel” was therefore not exactly paying any kind of justice to the design of the INE portal. The latest version was therefore adjusted by the consultant and Mr. Nhane to better reflect the overall design of the INE portal.

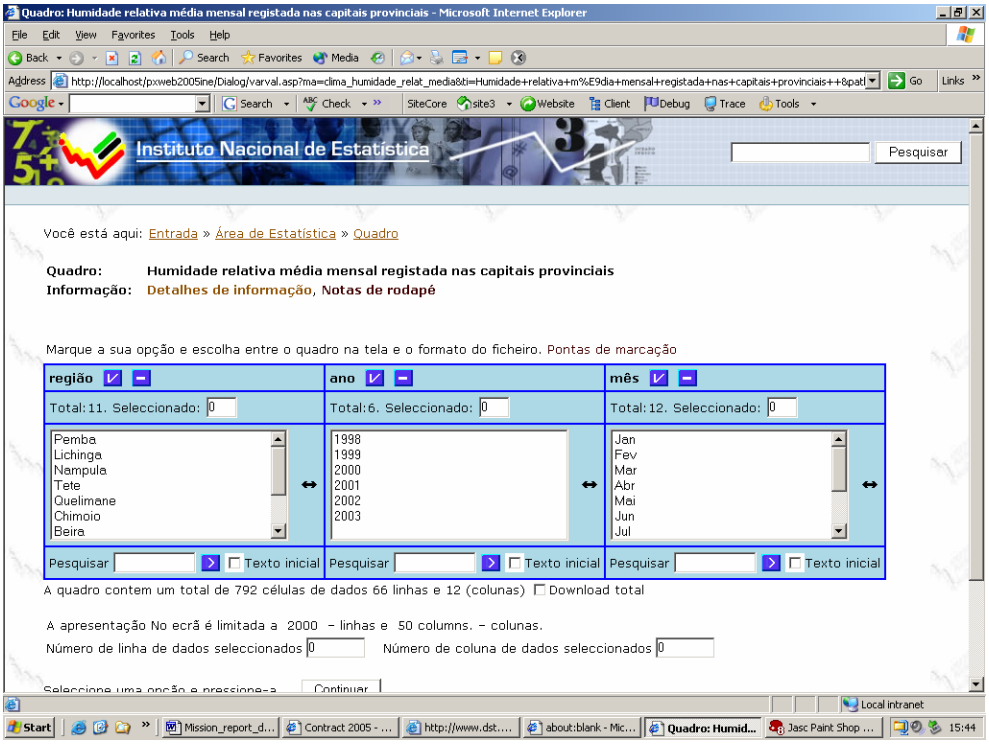


Illustration of PX-Web 2005 after localization to the INE portal design.

Direct download of files in Excel format

INE felt that a large number of their internet users are more interested in Excel files than either seeing tables on the screen or on downloading them in PC-Axis.

The presentation functionality of PX-web was therefore modified so that it is now possible to download Excel files directly from the table selection menu. The changes in presentation functionality also necessitated changes in the logging function and in the search function as it should also be possible to download Excel files directly form the search menu. Normally the use of PX-web (screen presentations and downloads) is counted after the selection process, but direct downloads of Excel files are also counted in the INE installation. Although PX-Web is designed for easy customization, this task did take a significant part of the mission time.

Illustration of Download total (Excel)

Statistics of usages

Both the old and the new PX-Web installations have a logging function where all presented and downloaded tables are tracked in an MS-Access database file. It was discussed how local (inside INE) and outside users could be identified using the Access database. It is highly recommended that a report on the use of the PX-Web database is distributed to senior management at regular intervals.

The consultant and Mr. Nhane vent over usage statistics and in general the numbers must been seen as satisfactory.

Although it can always be discussed what constitutes a satisfactory number of visits it is important to remember that the figure should always been seen in relation to circulation figures for printed matters and also to pay specially attention to the number of external users abroad which may have little or no access to printed publications.

Farmers

Also a detailed analysis of the web statistics should make it possible to identify users who extract data with a high dissemination frequency. I.e. it should be possible to identify users of the monthly consumer price index. Such users are often described as “farmers” in contrast to so-called “tourists” who use the INE portal in with an irregular frequency.

Connectivity problems

As already mentioned, the original PX-Web installation was intended mainly for a pilot study. Due to a shortage of production servers PX-Web and PLONE (The content management system of the INE Portal) is running on the same internet server. To enable this coexistence between two different systems the PX-Web has be configured to use port 82 and INE Portal is on port 80. Do to the increasing amount of unwanted internet traffic in the form of hacking, viruses and denial of service attacks a number of internet providers has chosen a policy were they either limit or totally ban traffic directed to port 82. This policy has the unwanted consequence that users connected to the internet through these internet service providers cannot connect to INE’s PX-Web installation.

Dedicated server is highly needed

To ensure that all internet users can in fact connect to INE’s internet database it is important that a dedicated server and associated IP address is obtained for the Internet Database. For standardization purposes it is important that a new server run Windows 2003 and thereby comply with INE infrastructure architecture. This also follows the recommendations from the 2004 mission.

Multiple languages

The PX format has been extended so that it now supports multiple languages. However, the aggregation tools PX-MAKE and PX-Edit are not yet ready to support the new multiple language facilities. This will happen around the end of 2006. When the tools are ready INE should consider introducing multiple languages into the output database and perhaps also into the Data Warehouse model it self. It is the impression of the consultant that INE senior management would have a strong interest in a multiple language setup.

This transformation could be supported by a STA when the tools are ready.

3.2 Structure of the Dissemination Database

Another purpose of the mission was to discuss the information structure of the PX-Web database. The present information architecture is originally sourced from the INE Yearbook and its structure. Also most of the data in the Database originally comes from the Yearbook.

Best practices

When working with output databases the generally accepted best practices is to follow as closes at possible the overall structure of all other products compiled inside the organization, as this is the most logical for the experienced user. The most important thing is that the information architecture always must reflect "user needs" and not the organizational diagram of the organization.

During the mission it was demonstrated how a different information structure can be created either instead of the present structure or as a supplement to it.

Parpa / millennium goals

As discussed, an organization of the data along the indicators known as the Millennium Development Goals could be a possibility as it would give users in Mozambique an easier access to the much sought after data on poverty reduction.

However such a reorganization is only possible in very close collaboration with the relevant subject matter departments, as the IT department can not take responsibility for defining the relevant data.

3.3 Training

During the localization of PX-Web and the different pilot exercises training and advice on different aspects of the software was given to the main counterparts. These now have the necessary understanding of the relevant tools and tasks.

Need for additional persons

However, as INE moves towards the Data Warehouse, the number of persons with the sufficient knowledge is clearly too small for an efficient production process. Also, the general amount of knowledge of the PX-family and acceptance of the Data Warehouse project inside INE is too small.

Also outside the IT department

It is therefore highly recommended that additional persons inside the IT department has given both basic and advanced training in the PX-family of programs. Also staff outside the IT department should be given basic training in PC-AXIS and PX-Make. This would both increase awareness and acceptance of the Data Warehouse project and help improve quality as the matrix structure of PX-files helps highlight the quality of Meta-data.

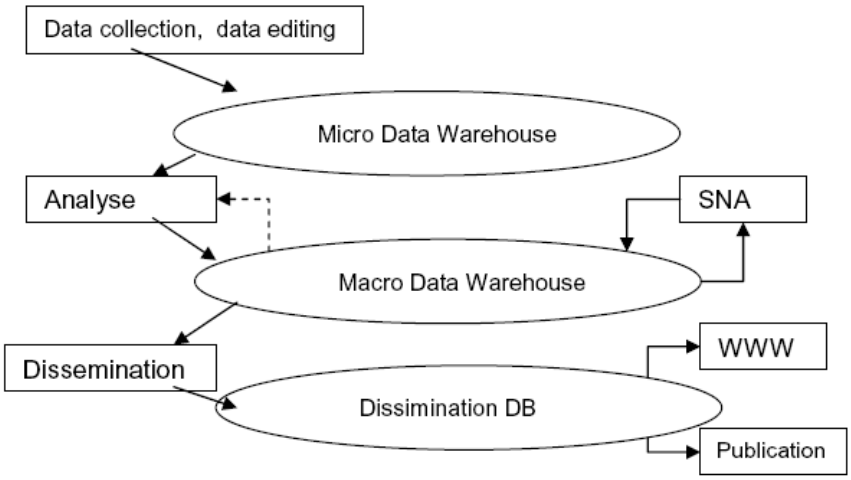
In the TOR it was hoped that additional people inside and outside the IT department should receive training during the missions. This was however not possible do to a shortage of available staff.

3.4 Data Ware House

An architecture for the development and implementation of a Data Warehouse solution has been developed by the IT – department of INE in corporation with the LTA on IT Mr. Bormann and Mr. Netterstrøm (from Statistics Denmark).

PX-Web as output interface

In the proposed architecture PX files / PX-Web is intended to function as the output interface for the aggregated side of the Data Ware House. During the mission the flow of data from survey / sample to PX-Web was discussed and simulated together with the LTA on IT and Mrs. Anastasia Honwana. Tests were done using IFTRAB data and showed that it is possible to move either directly from cleaned micro data to PX- files / PX-Web or from aggregated files to PX-files / PX-Web. In both cases using PX-Make as a tool.



3.5 Pilot projects

It is recommended that a number of pilot projects taking data from the earlier surveys / registers and converting them into data warehouse files is undertaken. As mentioned above a few pilot conversions were done during the mission. The pilots should be based around the lessons learned and the theoretical framework presented at the workshop conducted during the mission on *Data Modeling* by Søren Netterstrøm 31 January – 4 February 2005.

Tools

The "pilots" should also help to identify and develop relevant tools for the different transformation steps in relation to the Data Warehouse. The "pilots" were generally conducted using standard tools inside the SPSS and MS-Access programs. However, in the production setup around the Data Warehouse there could be a need for the development of "helper applications". These needs would become clear through the pilot conversions.

Quality of existing data

A number of pilots will not only help INE to prepare for a real implementation of a Data Warehouse but will also help to build analytical skills and improve the general awareness regarding quality issues.

Inconsistencies

During the mission a test transformation into PX-WEB using data sourced from the ESDEM CD-ROM indicated a number of inconsistencies in existing meta-data. One of the major strengths of DevInfo (ESDEM) is its high degree of flexibility regarding meta-data. This unfortunately also means that "anything goes" in terms of descriptive meta-data. In the pilot conducted during the mission we experienced a number of nearly identical texts describing variable and variable values, in turn disrupting easy access to time series of data.

The problem with such inconsistencies in the use of titles and text is that the "end user" have problems deciding if data related to the same dimension in ESDEM is actually comparable or not. These types of data inconsistencies are primarily brought to the surface when data from different surveys / sources are entered into a system of multidimensional tables or cubes, such as PC-AXIS.

The quality project

A pilot working through existing data and Metadata inside INE will therefore help to increase the quality of both existing and future data as it will bring focus on both the reliability and the validity of the indicators. The work on going through existing meta-data should been seen in the context of the work on quality improvement which is also a part of the Scandinavian assistance to INE. In general, it is the opinion of the consultant that the potential synergies between the two projects should be explored and taken advantage of.

3.6 Process and organization

As INE progress towards the proposed Data Warehouse architecture a number of organizational issues must be addressed. At the present moment the Data Warehouse project very much belongs to and originates inside the IT – department.

Shift ownership

For the project to be successful in medium to long term it is important that ownership and initiative of the project is shifted from the IT department to the subject matter departments. The process of defining, describing, storing and maintaining meta-data both regarding production and dissemination is ultimately only possible inside the relevant subject matter divisions. Only the subject matter divisions / experts have the necessary detailed understanding and knowledge of the involved concepts and variables.

Role of the subject matter departments

So in the ideal world the subject matter departments must take care of defining and describing variables, values and their associated metadata. They should also work on the coordination of the metadata / variables which cut across INE and the SEN. Gender, Age and Geographical classifications are some of the most typical classifications in need of coordination.

As mentioned above, the pilot conducted during the mission using data sourced from ESDEM indicated that age and geographical variables has room for improvement. With the present organization and division of work between IT and subject matter departments these quality and coordination issues are mainly seen by the technical staff at the IT departments. But the technical staff does not have the statistical knowledge and the organizational authority to correct and coordinate the metadata. And in most organizational modalities it will not be inside their area of responsibility.

Instead, metadata is usually coordinated by a methodological department or some other department with authority for coordination across the organization.

Role of the IT-department

The "proper" role of the IT department is to provide the subject matter divisions with the IT infrastructure in the form of software tools and the Data Warehouse and training in the use of these tools.

Necessary for the Data Warehouse

If a Data Warehouse is to actually function at both the micro and macro data level meta-data must be coherent and coordinated across surveys and subject matter areas. It is clearly evident that the organizational procedures around the Data Warehouse at INE must be addressed.

"Pragmatic approach"

However, it is recommended to follow a very pragmatic approach when defining the organizational setup around the Data Warehouse. Special attention should be paid to the fact that the different Subject matter departments at INE has very different survey / publications schedules and that it therefore can be argued that a number of work processes are better undertaken at a centralized level, at the IT department, instead of a decentralized processes at the subject matter departments.

4 RECOMMENDATIONS

The consultant would like to make the following recommendations based on the work undertaken during the mission:

Dedicated server

PX-Web should be moved to a dedicated Internet server running on Windows 2003. The use of a dedicated server will ease access to the database and solve the problems associated with running PLONE and PX-Web on the same server.

Participate in the reference group

As a user of PC-Axis / PX-Web, INE should take part in the International reference group on PC-Axis. The group usually meets once every year with the purpose of discussing and prioritizing future developments in the software family. The next meeting will be on Iceland in late august.

Compile and circulate statistics on usage

As it is possible to generate a number of statistical information on the use of PX-Web it is recommended that a short internal memo describing the use of the database is compiled and circulated at least on a quarterly basis. The analysis can be made either through a simple webpage or through direct analysis of the MS-Access database collecting the user statistics. It is recommended that this is done together with other statistics that describe the use of the INE portal. The memo should be circulated to senior management and to the subject matter departments, so they can all see that their products are in fact reaching end users through the internet.

Build an additional structure in the internet database

It is recommended that a structure representing the Millennium Goals and its associated indicators are entered into the PX-Web database. This exercise can be conducted in cooperation with the relevant Subject-matter division(s) inside INE. To the extend that data is not already available in the PC-Axis format they can be sourced from ESDEM.

A navigation structure based on internationally recognized indicators would increase the value of the Database to external users, especially outside Mozambique quite significantly, therefore the IT department should prepare a prototype and present it to the other stake holders. The prototype should be ready to deploy or as close as possible to ready, when presented. Otherwise the project will be in danger of losing the necessary momentum.

Move to a multiple language setup

It is recommended that INE moves to a setup with both English and Portuguese in the output database. However this can only happen when PX-Make / PX-Edith are ready to support the use of multiple languages. This will most likely happen at the end of 2006. PX-Web already supports multiple languages. It is important that a bilingual setup is fully integrated into the Data Warehouse architecture and its development plan.

Validation of production pipeline

Different options and scenarios for the operational implementation of the suggested / planned data warehouse architecture were discussed during the mission. It is recommended that at formal pilot study is conducted to validate the architecture.

Continuation of IT – Infrastructure consolidation

As INE moves to a more centralized and controlled storage of data and Metadata, it is even more important that the IT-infrastructure is physically and logically safe and considered as such by the INE staff. It is recommended that priority is given to the consolidation efforts already in the pipeline. This

will involve Windows 2003 migration to ensure logical safety and implementation of more secure backup procedures to ensure physical safety.

APPENDIX 1 List of persons met

INE
Dr João Dias Loureiro, Presidente do INE
Ms Destina Uinge, Program Director of the Scandinavian program
Ms Anastasia Honwana, Head of IT
Mr Thomás Bernardo, DICRE/DISI
Ms Fatima Zacharias, Director of Social Statistics
Mr. Anselmo Nhane

Scanstat Consortium, LTA:
Mr Lars Carlsson, Team Leader
Mr Karsten Bormann, LTA on IT

Scanstat Consortium, Project assistant:
Ms Isabel Novela, Project Assistant

Danish Embassy
Ms Lola Lopez
Mr. Peter Engbo Rasmussen, Counselor

APPENDIX 2 List of Literature

All mission reports from the Scandinavian programme are available online on: *www.dst.dk/mozambique*

More information on PX-Make can be found on *www.dst.dk/pxmake*
More information on PC-Axis and PX-Web can be found on the PC-Axis website maintained by Statistics Sweden see *www.pc-axis.scb.se*

MZ-2005-8 Report from a short-term Mission on *Data Modeling by Søren Netterstrøm 31 January – 4 February 2005*

MZ-2004-17: Report from a short-term Mission on Internet Database Pilot Project by Jesper Ellemose Skou Jensen

MZ-2003-24: Report from a short-term Mission on The Creation of an Output Database by Jesper Ellemose Jensen and Annegrete Wulff

APPENDIX 3 Terms of Reference

TERMS OF REFERENCE

for a short-term mission
on
PC Axis and the Macro Data Warehouse

June 6 - June 16 , 2006

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

Consultants: Jesper Ellemose Jensen
Counterparts: Anastácia Honwana, Anselmo Nhane, ???

D R A F T

Background
INE needs to guard data in a way that is both robust and which makes easy the sharing and analysis of data. For this, a strategy involving a 'Statistical Data Warehouse' has been envisaged. Due to the relatively static nature of statistical data, the technical requirements are rather low, and emphasis is placed on creating simple, standardized storage formats and a simple storage architecture for easily supporting work without creating an unnecessarily large administrative overhead.

The Statistical Data Warehouse is conceived as consisting of three different parts that each supporting a different part of the Statistics Production Pipeline.

The Micro Data Warehouse will guard cleaned micro data (e.g. from a survey). The format will be that of a comma separated value file with the first line giving the variable names and each subsequent line containing the data from one record (e.g. one 'household' or one 'person'). This 'Fact File' will have a number of associated 'Dimensional Tables', or classifications, or value sets that describe the values that dimensional variables can attain. The relationship between the two files is documented in a Logical Data Model (alike to the CS Pro Data Dictionary). Because newer and future (survey) data is expected to come from CS Pro, the process of taking CS Pro Data to the should be streamlined. The reason for adding a Micro Data Warehouse (Format) to the CS Pro Storage Format is that the latter is tied to the questionnaires, not to the objects of study and also that some kind of transformation is needed to place objects in a format that is easily loaded into statistical analysis packages or relational databases. It is also very simple to extract sub data sets from a comma separated file.

The Macro Data Warehouse is a store for analysed information (aggregated data), ideally in the form of annotated time series of statistical tables. The same is true for the Dissemination Database. The latter is currently placed on the PC Axis platform, at there is no compelling reasons to change this. To allow for easy integration of Macro DW and Dissemination DB it is thus

suggested that the Macro Data Warehouse be stored as a PC Axis database also. We are, however, not satisfied with the current structure of our PC Axis database - organised, as it is, around the organisational structure of INE. Instead, the structure of DevInfo/ESDEM is considered superior, organised around Policy Themes - Goals- Indicators, with the possibility of viewing the database according to sectors, if so desired. This, goal centered structure makes (ad hoc) searching for data relatively easy. DevInfo, however, is constrained in that it can hold only time series of indicators classified according to area and subpopulation. Time series of tables having an indicator grouped according to other variables is not possible (without detrimental proliferation of subpopulations).

Thus the idea is to restructure the Dissemination Database according to the principles used by DevInfo. For ease of integration this structure then should also be imposed on the Macro Data Warehouse.

The metadata store is expected to be shared among various Warehouse components as one should be able to trace the origin of all final or published to its roots, to see associated data gathering and analysis methodologies, quality assurance measures, and reports that have relation to the data in question. (Both reports written by INE and by external agents).

Thus, in this conception, the production pipeline looks more or less as the following chain (the first and the last link in the chain are only relevant to a subset of the data):

CS Pro --> Micro Data Warehouse (Survey Objects in .csv files)
--> Analysis (SPSS) --> Macro Data Warehouse (PC Axis)
--> Dissemination Database (PC Axis, on the Portal) --> ESDEM/DevInfo

3.6.1.1.1 Objective

The mission has two technical objectives strictly related to PC Axis (and the Portal)

- Upgrade PC Axis to the newest version (to improve search capabilities).
- Restructuring of the Dissemination Database around Policy Themes
- Training in newest PC Axis and PX-Make versions
- Aggregation csv files through PX-Make
- Enable statistical staff to place data in PC Axis without DISI assistance

The other objective is to use the Statistics Denmark experience with Data Banks to give input to the Macro Data Warehouse, envisaged to include the following:

- PX Make and the PC Axis file format - Metadata representation
- Data Organisation
- Data Management

3.6.1.1.2 Expected results

- PC Axis Upgraded
- PC Axis (new version) and PX Make training
- First steps to enable statistical staff to work with PC Axis
- Input to Macro Data Warehouse construction

Activities

- The PC Axis upgrade is to be made in cooperation with Anselmo Nhane (and, if relevant, with new personnel to be assigned to the portal).
- Prerequisite steps for allowing statistical staff to work directly with PC Axis also has Anselmo Nhane as primary counterpart
- Training on PX Make and the PC Axis file format involves personnel that is to work with metadata (for the Macro Data Warehouse and Dissemination Database).
- Macro Data Warehouse discussions will involve assigned personnel with Anastácia Honwana as main counterpart.

Tasks to be done by INE to facilitate the mission
- Elaborate ToR for the mission
- Supply the consultant with necessary documents and information.
- Supply good working conditions for the consultant
- Make sure that involved personnel is available

Consultant and Counterpart
Consultants: Jesper Ellemose Jensen
Counterparts: Anastácia Honwana, Anselmo Nhane

Timing of the mission
Two or three weeks (June XX - June XX, 2006).

Report
The consultant will prepare a short draft report to be discussed with the counterparts before leaving Maputo. The final version will be sent to INE within one week of the expert having returned to Denmark. The Counterpart then has to provide, also within one week, at least a summary in Portuguese (if the main report is in English – or else; vice versa) to be included in the final printed report. Statistics Denmark, as Lead Party, will print the final version within three weeks of the end of the mission. The structure of the report should be according to Danida format.

These Terms of Reference were prepared by

Day / /
.....

Approved by/in the name of the President of INE

Day / / .....