Distr.
GENERAL

WP.12
6 April 2010

ENGLISH ONLY

UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE (UNECE)
CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION
STATISTICAL OFFICE OF THE EUROPEAN UNION (EUROSTAT)

ORGANISATION FOR ECONOMIC COOPERATION AND DEVELOPMENT (OECD)
STATISTICS DIRECTORATE

Meeting on the Management of Statistical Information Systems (MSIS 2010)
(Daejeon, Republic of Korea, 26-29 April 2010)

Topic (iii): Issues for Asian statistical organizations and ways to increase international cooperation

NIS ICT Infrastructure Strategy in the Production of Official Statistics, Dissemination and Data warehouse

Prepared by Panhara OUKCHAY, National Institute of Statistics, Cambodia

I. Introduction

1. The National Institute of Statistics (NIS), which is part of the Ministry of Planning, is the focal point on statistical matters in Cambodia. The NIS compiles and consolidates statistics provided by decentralized offices and also collects primary data through households and establishment surveys and population, agricultural and economic censuses. Cambodia has a decentralized statistical structure. There are statistical bureaus and sections within Planning and Statistics Departments of the various Ministries and in planning and statistical units in the provinces and districts.

2. The NIS has recently reorganized itself to meet the increasing demand of a wide range of statistics of good quality. The NIS has now 7 departments, the new departments are: National Account Department, ICT Department, Statistical Standard and Analysis Department and Statistical Policies and Cooperation (Replacing General Statistics Department). The upgrading of staff skills through training both in-country and in international institutions, the establishment of in-house data processing capacity, the collection of primary data, and the increased capacity to serve the needs of data users have in the recent past greatly contributed to enhancing the role of NIS in the statistical system of Cambodia.

3. The ICT strategy for the NIS is very important in the production and dissemination of official statistics. There exists a need for the National Institute of Statistics (NIS) to modernize and extend its ICT to secure timely and accurate statistics serving governmental bodies and the public. The Annually Cambodia Social-Economic Survey, CPI, and some other surveys are using the Database Server (SQL Server 2005, 2008) for its data storage and the application (Developed by Visual Basic software, or others) for data entry. In order to facilitate for a smooth data capture and processing of all those surveys and other, the ICT infrastructure and

logical network domain need to be strengthened and up-to-date. To succeed in this undertaking, a number of steps have to be taken and needed to reach the goal.

4. Since a change to a new IT environment will put forward a lot of demands on budgets and manpower it must be done in steps. There are already LANs running within the agency at the headquarters and census building. These networks are peer-to-peer based and provide mainly file server solutions and Internet connection which with the future emphasis on statistical databases will not be suitable.

II. ICT Strategy in the Production, Dissemination of Official Statistics

A. Standardized Network

5. A new Physical Network Design for NIS headquarters has been elaborated. This new solution aims at making the NIS headquarter server room the information hub for the whole NIS organization and enable governmental bodies as well as provincial statistics departments and the public to communicate with NIS. Logical LAN design implementing Microsoft 2003 Domain concept:

- DNS Service
- DHCP service
- Group Policies for servers and users. Group policy rules for usage of the NIS domain

6. Domains are logical grouping of multiple Windows server-based computers that allow them to be managed and used as a single unit. In a domain users log on to a complete system instead as today at NIS where users log on to specific servers. When logged on, users can access all the resources they have rights to access including files, directories, servers, applications, and printers. Users need only to log on once to gain access to all resources they need to use.

7. All server administration is done at the domain level. Therefore in a domain network administrators can always administer all network servers. In a typical domain of the size of the NIS the following server resources could be identified.

Server type	Description
Domain Controller (DC)	Manages the domain and holds the user database in an Active Directory
DBMS	Database management system.
Fileservers	For storage of users home directories and shared files within the domain
Communication servers	Proxy and NIS Mail
Backup Servers	System and Database Backup
Application servers	Sever Application Services: Antivirus Sever, Super Star Application (Datawarehouse)

B. Data Storage

8. Data is NIS' most valuable asset and the way to safely keep this data is of high importance. It is equally important that access to this data is transparent and that data uniqueness is guaranteed i.e. if you open a dataset you should be sure that this dataset is the latest and most updated.

9. Presently at NIS data are redundantly stored on numerous locations on the network as well as on desktop PCs. Storing data this way will jeopardize data quality and give unpredictable results when tabulating and analyzing data. One way to avoid such uncertainties of data uniqueness is to introduce a Client / Server based database management system like Microsoft SQL Server.

10. Today NIS is using different storage formats i.e. Microsoft Access 95/97/2000, Microsoft SQL Server and some flat-file based systems. Some data are also stored in SPSS format but have then been further calculated and enhanced in various ways. The biggest difference between these storage methods is where and how the data is stored and the level of data security.

- Flat-file based system: Data is stored anywhere on servers and desktops and there is no data protection whatsoever. These systems are not built for concurrent connections i.e. a network environment.

- Microsoft Access: Is a file-server based data storage solution with some data protection and multi-user capabilities. Access is a great tool for smaller projects with up to a maximum of 5 concurrent users and smaller datasets for up to 50 000 rows. One major drawback is that data can reside anywhere on servers or desktops and can easily be copied or moved to another location thus creating data redundancy.

- Microsoft SQL Server 2000/2005: Is a complete client/server database system. It supports thousands of concurrent users and very large databases up to several terabytes in size. Data is stored in one place, which guarantees data uniqueness. It has built-in procedures for backup and restore of data. SQL Server is built around OLE-DB, which is an open database interface. By using OLE-DB different types of applications can easily connect to the data store. The security mechanism of SQL Server is tightly integrated with Windows 2000/2003 domain security.

Having different storage formats is not the most effective solution for an organization like NIS. When deciding on which database management system to use at NIS there are a few factors to take into consideration

- Data integrity – security
- Performance
- Backup and maintenance
- Data Warehousing

11. Microsoft SQL Server complies with all of the items above. It is currently rated as one of the best DBMS that there is. Microsoft Access is not a client/server database and is usually not compared with systems like SQL Server and Oracle. However, if one will make a comparison between the two products the conclusion will be that Access does not comply with the above items. NIS should in the first place consider choosing SQL Server in favor of Microsoft Access. However, if the survey or database is a single-user system with few records then Access is the right platform to use. SQL Server 2005 contains a complete storage solution with the following components:

- Storage Engine;
- Data Extraction, Transformation and Load (ETL) tools;
- Data Mining;
- Data Warehousing;

- Strong access security system.

C. Software

12. It is important that the ICT environment is standardized to the largest possible extent in order to secure a well functioning office information system and statistical data processing. The number of different software should be minimized. Below is a suggestion of which software that should be recommended for different tasks. It is based on the fact that this software is already in use at the NIS.

- Statistical software – Stata, SPSS (one or both)
- Tabulation and dissemination tools – SuperCROSS, MS SQL Server, SPSS, CAMInfo...
- Office software – Microsoft Office
- Software development tools – VB6, VB.Net and CSPro
- Antivirus Software: Symantec Antivirus, Trend Micro Antivirus Corporation Edition

D. Data Dissemination

13. The NIS is using Website, Data On-line (such as Census), NADA (Surveys documentation On-line), SuperCROSS, CAMInfo, to disseminate our production to all type of users:

- A dynamic website configuration for NIS has been developed last year, so all the reports and publications will be posted into the website on time and all users can access to it easily and more faster than before. The NIS website (http://www.nis.gov.kh) consisted by most results of various Censuses and Surveys, periodical publication (such as CPI, National Accounts), and some other documents which is released by NIS. However, most information that available on NIS website for downloading are in static format.

- Results On-line Dissemination: the NIS is focusing on the results of the General Population Census 2008 as On-line dissemination that is more dynamic approached. Priority tables and analysis outputs are suggested to be available on-line for wider and distant accesses. The possibilities of SQL database querying and on-line mapping will be explored. To abide by the statistics law on keeping confidentiality of respondents, however, security, authentication, recoding and aggregation shall be closely observed.

- NADA Toolkit (Web base application tool): Various surveys documentation (Meta Data) will be available on-line by using NADA toolkit to prepare this documentation in XML format and can also be provide together with Micro Data (post into the web thru this tool) to disseminate into website. This is open source platform (PHP, My SQL…)

- SuperCROSS: The National Institute of Statistics is transforming much survey and census data to a centralized storage with SQL server. To further facilitate for users to retrieve information stored in databases it has been decided to implement a Datawarehouse solution with the SuperSTAR software suite at the NIS (SuperCROSS for client component).

The SuperCROSS solution is very useful within the organization network especially all subject matter teams as well as line ministries and provincial departments. Later on in next step we are planning to establish this into the web as tabulation and dissemination tool (SuperWEB).

- CamInfo is Cambodia's socio-economic database system which provides a common platform for users to organize and present data on development indicators. CamInfo is the localized version of DevInfo, the global database system which monitors progress towards the Millennium Development Goals. The database structure consists of the following elements:

- Indicators: related to Cambodia Millennium Development Goals (CMDGs), National Strategic Development Plan (NSDP), Education, Health, Cambodia Nutrition Investment Plan and others;
- Time Periods: 1976 to 2020;
- Geographic Areas: national, provincial, district and commune level;
- Subgroups (Sex: Male and Female; Location: Urban, Rural, Remote Area, others; Age Groups: by months, quarter and years)
- Units: Number, Percent, kilometre, hours, Tone, etc, as appropriate for each indicator
- Sources: surveys, censuses and administrative data from the National Institute of Statistics, Ministry of Planning (MoP), Government Line-Ministries, UN, NGOs, etc.
- Metadata: for additional information on indicators such as definition, method of computation, limitations, notes, etc.

CamInfo is a user-friendly software application for easy presentation of indicator data in tables, graphs and maps. The system provides easy access to indicators organized by sectors, goals, themes and other monitoring frameworks. All features of the software are available in both Khmer and English language.

E. Data warehouse

14. NIS has together with donors such as WB, UNDP and Statistics Sweden/SIDA developed a work plan which is proposing a data warehouse system for dissemination and analysis of micro data for the national statistical system in Cambodia. The current Surveys and CPI data are directly stored in SQL Server; and also transformed much of its surveys and censuses data from flat files to a centralized storage with SQL Server to increase usability and security of data, and to further facilitate for users to retrieve information stored in databases it has been decided to implement a Data warehouse solution at the NIS with the Australian Software suit SuperSTAR, used by many statistical agencies around the world. the SuperStar families software suit (SuperCROSS, SuperSERVER, SuperCHANNEL, SuperWEB, SuperTABLE).

15. SuperSTAR is the tool for "normal" users to access the data. SuperCROSS is the client component of SuperSTAR. It is user friendly to be able to retrieve information. From the data security point of view SuperCROSS is outstanding. The data is stored in a SuperCROSS database and it is not possible for users to change anything in the database – deliberately or by mistake. All statistics produced from SuperCROSS and stored in SuperCROSS file format can be tracked, which means that all manipulations (contents, recodes, limitations, calculation etc) are stored with the statistical table and can be tracked at any time.

16. What is quite confusing is the use of the word "table". In a SQL database table stands for the unit where the individual data records are stored. One database might comprise several tables, for instance one table for households, one for persons, one for consumption transactions, one for sources of income etc. When producing statistics, table means a table with statistics (aggregate data).

F. Backup

17. Regular backup of server hard disks prevents data loss and damage caused by disk drive failures, power outages, virus infection, and other potential network disasters. Backup operations should be carefully planned and reliable equipment used.

- **Network Backup plan:** It is always important to make a plan for the backup system. The backup plan should answer three questions:

  - What type of backup to use?
  - When to backup?
  - Which files should be backed up?

- **Types of Backup**: There are five types of backup for shared network files: normal, copy, incremental, differential, and daily.

  - Normal backup copies all selected files and mark each as having been backed up. With normal backups, it's easy to restore files quickly because files on the last tape are most current.
  - Copy backup, copies all selected files but doesn't mark each file as having been backed up. Copying is useful if there is a need to backup files between normal and incremental backups because copying doesn't invalidate these other backup operations.
  - An incremental backup backs up only those files created or changed since the last normal or incremental backup. It marks files as having been backed up. If using a combination of normal and incremental backups, restoring requires starting with the last normal backup and the working through all the incremental tapes.
  - A differential backup copies, those files created or changed since the last normal (or incremental) backup. It doesn't mark files as having been backed up. With normal and differential backups, restoring requires only the last normal and last differential backup tape.
  - A daily backup copies all selected files that have been modified the day the daily backup is performed. The backed up files are not marked as having been backed up.

18. The most common types are normal (full), incremental, and differential. The following table lists advantages and disadvantages associated with running the most common types of backup.

Backup Type	Advantages	Disadvantages
Normal	Files are easy to find because they are always on a current backup of the system or on one tape or tape set.	Most time-consuming.
	Recovery requires only one tape or tape set.	If files do not change frequently, backups are redundant.
Incremental	Least data storage space required.	Files difficult to find because they can be on several tapes.
	Least time-consuming	
Differential	Less time-consuming than normal backups.	Recovery takes longer than if files were on single tape.

	Recovery requires only the last normal backups tapes and last differential tape.	If large amounts of data change daily, backups can be more time-consuming than incremental.

G. Staffing

19. For a statistical agency like the NIS the ultimate goal is to ensure IT competence on all levels of data processing in order to be self-reliant. Therefore decisions must be taken on what its staff should achieve in short and long-term perspective. To facilitate further discussions five areas of interest will be identified with different demands on IT skills.

Day-to-day word processing and spreadsheet handling - When trying to decide on the competence needed, it is relatively easy to state that all staff who in one way or the other are likely to do any word-processing or work with spreadsheets must receive proper training to an extent based on what could be seen as normal activities at the office.

LAN operation and maintenance - In order to ensure that the LAN will functions as intended, know-how must reside within the NIS Training in network administration (Windows 2003) and in assisting users (Windows XP / Window Vista) must therefore be given to the staff that will perform these functions.

Database design, maintenance and administration - The NIS will strongly promote the creation of relational databases in all fields of statistics and classification. Database design and implementation will require skills in using Microsoft SQL Server but also in modeling statistical information systems in general. Training in database administration will be needed as well as training in systems design for all staff that will be engaged in database activities.

Statistical tabulation and analysis - Most statistical activities will in the end involve tabulation and analysis. One of the best products available in the market today for this purpose is SPSS with its intuitive and interactive user interface. Appointed statisticians should be trained in the use of SPSS.

In-House Application development - One of the most intriguing IT task is application development e.g. designing, writing and implementing applications for data entry of large volumes of data. This is demanding and a lot of talent, skills and experience will normally be needed. One important thing to mention is that application programmers must have good abilities in reading and writing English since the development tool itself and corresponding manuals are normally in English and also that most training abroad is given in the language.

III. CONCLUSION

20. Statistical agencies should approach ICT sourcing with an understanding of the significant role it plays in fulfilling key priorities. Statistical agencies must recognize that an ICT strategy needs to support these priorities in the most cost-effective manner. Statistical Agencies must also understand the risks and challenges of ICT Strategy. Therefore, ICT Infrastructure Strategy for National Institute of Statistics is very importance role in its production of Official Statistics as well as Dissemination and Data warehouse. It is long-run strategy plan due the NIS has a limit of human resources, So capacity building for ICT staffs is needed to be done through daily job and in-house training.

All ICT staffs should have general knowledge of each area but have near specialist competence in their own field to achieve this goal substantial training within the ICT-field is necessary. Training for the ICT department staff can be provided through different means. The most important training is done *in-house on a daily basis* and sharing knowledge from staff to staff, and from consultant (Long-term and Short-term Technical Assistance). Other training could be provided through attending courses given by vendors or institutes, study tours and attachments can also be considered.

_____