

Doha Workshop
2 – 5 February 2003

Role of ICT in Data Dissemination

Abdulilah Dewachi

Regional Adviser

UN-ESCWA

adewachi@escwa.org.lb

Topics


- Definition
- Importance
- IT components
- Electronic output
- Hard copy
- Data extraction and tabulation software
- Concepts, definition, quality of data, terms and conditions of supply and confidentiality
- The Internet

What is data dissemination?

- The process of extracting raw statistical information from a database or data repository to produce statistical output that can be understood and used and has relevance to the statistical query (objective) it seeks to satisfy


Importance of data dissemination

- **Knowledge base:** Statistical data form an integral component of the knowledge base required by nations, governments, regional authorities, companies and individuals to understand the constant changes that occur within a society.
- **Policy making:** Statistical data is essential to the development of informed policy formulation, implementation and evaluation. Without appropriate data, policy decisions must be made within a vacuum with little understanding of the impact or the effect on individuals or society in general.
- **Society:** The economic and social development of nations and regions within nations is very much dependent on quality and timely statistical data.



Importance of data dissemination

- **Research:** Scientific research utilizes statistical data in one form or another. Much research centres on small area issues such as deprivation, effect of government economic policy on rural areas etc., while other research projects focus on more global issues such as international migration and unemployment levels. Not only is it important to be able to compare data within a country consistently over time, but comparisons between countries within a region are also increasingly becoming the focus of research. It is clear that international comparability of data including concepts, classifications and dissemination standards need to be addressed as high priority areas.



Importance of data dissemination

- **Dissemination:** Statistical data needs to be disseminated to the widest possible audience so that the greatest benefit can be achieved from the investment made by the national statistical office. Statistical data is fundamental to the majority of public policy decisions therefore the data must be readily available in a timely and easily accessible manner in the medium of choice of the user.

IT systems transformation

- Mainframes to Personal Computers
- Networks
 - Local
 - Wide
- Intranet
- Internet
- Portals

From mainframes to personal computers

- Information technology has rapidly developed during the past decade. Previously mainframe computer systems had dominated the computer market and represented the primary computing power of most national statistical offices. Today mainframe computer systems have diminished considerably in terms of importance. While the mainframe computer still represents massive power and storage capacity it has been superseded in many areas by the greater flexibility, portability and wider range of powerful software tools available for use with personal computers.



Networks

- Personal computers can be connected together to form a network of virtually any size or capacity and may be located on a single floor of a building, multiple floors of the same building, across buildings, cities or even countries. The software used to drive the computers and manage data can be resident on the file server at the centre of the network or on the hard disk of the individual computers. Networked computers enable the storage of data either centrally or on the hard disks of the individual computers linked to the network. A file server, a more powerful computer, controls each network.



LANs and WANs

- A Local Area Network or LAN is the term used to describe a situation where two or more computers are linked together (usually by cable) within the same building. A Wide Area Network or WAN is the term used to describe a situation where two or more LANs are linked together (usually by telephone cable or wireless) across buildings, cities or localities. LANs and WANs are very powerful and connect computing systems sited both locally and at remote locations, that enable the user to send and receive data and information, manipulate, cross tabulate, graph and interrogate databases as well as present data in many meaningful ways.



Intranet

- Intranet is used to describe a situation where electronic mail, data and other information is sent via a LAN/WAN operation. The system operates very much like the Internet but is contained and accessed solely within a single organization. An Intranet operation may comprise a number of databases located on numerous file servers where access may be restricted to personnel from a specific subject matter area or to personnel within specific echelons within the organization. However, the general sending and receiving of electronic mail is usually unrestricted. A more narrow definition restricts Intranet to applications using TCP/IP (Internet) protocol for data transfer within an exclusive network organization.



Internet

- Internet is used to describe the World Wide Web (WWW) which is a global network of computing systems. The Internet is used extensively by millions of people every day for electronic mail and browsing, to obtain information about and to purchase products and services. The number of connections to the Internet is growing daily and represents an opportunity to every national statistical office to reach the widest possible number of potential customers.



Database transformation

- Database Management Systems
- Relational DBMS
- Standalone databases
- Data warehouses
- Internet-enabled DBMS



Database Management Systems

- A database is a means of storing data in the most efficient manner possible. Databases enable access to data and updating to be centrally controlled giving better security and integrity. Eliminating duplicate data and only storing essential data can reduce data redundancy. Databases make efficient use of available disk space and can be designed for the efficient extraction of data using specially developed software. A database management system (DBMS) is the software that allows you to create, maintain and report on the data.



Relational DBMS

- There are several database models including the relational, hierarchical and network models. The relational database model is commonly used when ad hoc queries need to be made and is suitable for statistical data. Data is organized into a series of relational tables divided into fields and rows. Individual tables are linked together using a common field or a combination of fields called the primary key. Referential integrity can be used to ensure that links between tables are valid for all entries in the table. Integrity is maintained by checking these links every time data is inserted, deleted or changed and only allowing changes, which will maintain these links. A database should be normalized which is the process of simplifying the database so that it achieves its optimum structure and eliminates redundant duplication.




Stand-alone databases

- Databases may operate in a stand -alone environment or integrated as a data warehouse. Stand-alone databases are easier and cheaper to establish. They store each year of each data set in a separate database. This means that to obtain time series data considerably more data manipulation needs to occur. In this operating environment, to create a new data set from several others, the user must extract the data from each database and subsequently, using another software package such as a spreadsheet or database, merge the data sets together using a common variable. However access is generally quicker and less complex than in a data warehouse situation.



Data warehouses

- The data warehouse concept does have its own set of advantages. The data warehouse is generally stored in a single area and all data sets are available at a single time. This means that different subject matter data sets may be linked together with a common variable and time series data can be easily compiled.



Mapping and graphical representation

- Uses for dissemination
 - Maps
 - Graphs
- Thematic mapping
- GIS for data analysis and dissemination



Use of maps

- Statistics may be effectively disseminated by many different methods. Tabular output is the most frequently used method of disseminating statistics but it is by no means the only method available to national statistical offices. Graphs and maps are a very effective means of displaying vast amounts of data visually. Maps are extremely effective for portraying regional differences, densities, concentrations of characteristics etc. Maps have the capacity to show in a single graphic, thousands of data that would not be practical to present in more traditional output forms such as tables.



Use of graphs

- There is a wide range of graph styles and graphing software packages (such as Excel, Lotus 1-2-3 etc.) that may be used to present statistics but there are a few simple rules that should be observed. Graphs should convey the message, bring out patterns and trends in the data and highlight key results. When designing a graph attention should be given to text, axes, colour, perception to ensure the graph is uncluttered and easy to read. Graphs should be targeted to the user's level of sophistication. It is important to ensure that the graph presents the facts in a clear and uncomplicated manner.

Thematic Mapping

- Accurate design and delineation of geographic areas is vital for both collection and dissemination of geographically referenced statistics, particularly population statistics. Detailed mapping has therefore, long been used as a prerequisite tool for the conduct of population censuses. There are a number of mapping software packages available (such as PopMap, PCMAP etc.) that can be used to generate maps that will very effectively display data. Maps are primarily used to present regional or international data. Interregional or international comparisons can be made and differences readily identified. Maps are best used to present data at a regional level. Maps may effectively complement statistical analysis in a variety of research including density studies, high and low fertility areas, changes to land use over time etc.

**ELECTRONIC
OUTPUT**

Types of electronic output

- Diskette
- Tape
- Compact disk (CD)
- Attachment to email
- Internet online

Dissemination on portable media

- **Diskettes, CD-ROM etc**
 - **Portability** There are two main modes of disseminating data in electronic form that do not involve the Internet or the electronic transmission of data. Both diskettes and CD-ROMs are portable and may be accessed by stand-alone computers or in a LAN/WAN environment. Diskettes and CD-ROMs may be posted to customers and will not suffer damage if securely packaged.

Dissemination on portable media

- **Diskette capacity** The capacity of a standard diskette is 1.4 megabytes of data unless the data is compressed. Generally the greater the amount of data provided to the user, the greater is the need to use compaction techniques such as data compression. There are compression programs available that will enable data to be stored in a high-density mode. The data must be decompressed by the user before it can be accessed. Data files compress anywhere between 0 – 99 percent depending on the file format. Already compressed file formats, such as JPEG graphics hardly compress at all, while database or word processor documents can shrink to less than a tenth of their original content.

Dissemination on portable media

- **Viruses** Extreme care must be used when using diskettes. Computer viruses are prevalent and at times may have particularly devastating effects, infecting and destroying entire computing systems. The reputation of the national statistical office must be maintained at all times therefore it is essential that all incoming and outgoing diskettes are thoroughly checked for viruses.



Dissemination on portable media

- **CD-ROM capacity** CD-ROMs have a considerable storage capacity, up to 640 megabytes and are suitable for sending large amounts of data. However, once a CD-ROM has been written it cannot be reused. It is possible to increase the storage capacity of a CDROM even more by compressing the data. In excess of 1 gigabyte of data may be stored on a CD-ROM in a compressed format.



Dissemination on portable media

- **DVD Technology** DVD's have increased storage capacity and are also able to include video footage. Their use for the dissemination of population statistics is more restricted at present because of the present state of their development and the need for standardization of formats.

Diskette pros and cons

ADVANTAGES

- Easy to create and format
- Ideal for storing many small files
- Space saving
- Portable and easy to load on many machines
- Secure
- Very common

DISADVANTAGES

- Not convenient for large files
- Prone to viruses
- Software incompatibility

Tape pros and cons

ADVANTAGES

- Suitable for large amount of data

DISADVANTAGES

- Not very common for storage and tabulation
- Not every system has a tape drive
- Easily damaged
- Sequential

Compact disk pros and cons

ADVANTAGES

- Ideal for large data
- Excellent storage media
- Files do not take up space on the hard drive

DISADVANTAGES

- Only used once
- Not suitable for small files
- Need special CD writer for writing on CD's

Scanned Publications

- Publications may be scanned onto the website and subsequently downloaded by the user. While this is a very cheap form of dissemination it is not a dynamic method in that the image is constant. This method of dissemination does not provide the user with the opportunity to manipulate the data or perform any statistical calculations without extensive reformatting of the data. However as an electronic presentation tool this method is satisfactory although not generally recommended.

Email attachment pros and cons

ADVANTAGES

- Fast delivery
- Easy transfer of data
- Ideal for small files
- Several files can be dispatched at the same time
- Very convenient

DISADVANTAGES

- Files may not go through the gateway
- Downloading problems
- Compatibility issues
- Viruses

HARD COPY

Types of hardcopy

- Hardcopy is used to prepare:
 - Customized statistical outputs
 - Reports
 - Publications – Yearbooks
 - Memos for quick distribution, e.g. faxing


Hardcopy pros and cons

ADVANTAGES

- Most useful in countries where computers and the Internet are not commonly used
- Easy to format, print and read

DISADVANTAGES

- Can get out of date quickly
- Not easy to manipulate
- Storage problems



Standards to consider when outputting on hardcopy

- Contents of reports, including tables, table numbers and page numbers must be clear
- Correct headings and labeling
- Rows and columns adequately labeled
- Footnotes to be included where applicable
- Sources of data and contact points
- Labels and page numbers

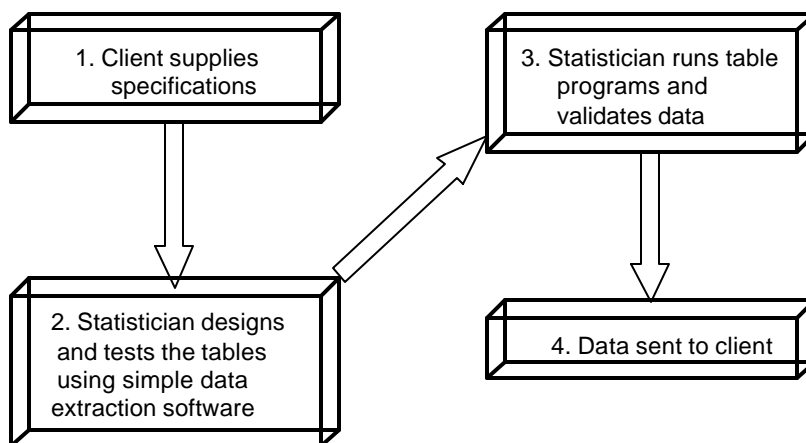


Points to consider for electronic output

- Time frame for formatting and delivering output
- Familiarity of users with output media
- Availability of correct software at the user end
- Is it a better option than hardcopy?

Data extraction and tabulation software

Electronic data extraction





Tabulation Software

□ *Wide range of software*

- There are a number of tabulation software packages available to the national statistical office. It is not necessary to develop in-house software as suitable tabulation software can be readily purchased. Examples include Supercross, TPL (Table Producing Language), Minitab, SAS (Statistical Analysis System), SPSS (Statistical Package for Social Scientists), Fastab and IMPS (Integrated Micro-processing System) from the US Bureau of the Census.



Tabulation Software

□ *Flexibility*

- All of these packages have been designed to efficiently access and tabulate data and present the results in a flexible yet print quality format. Most of the packages allow considerable manipulation of the data subsequent to table production.
- For example it is usually possible to calculate additional variables derived from the data presented within the table such as percentages and ratios.
- In addition, most packages also allow considerable flexibility in table design, such as font type and size, nesting of variables within columns, swapping of variables from columns to rows etc.



Tabulation Software

Processing speed

- One of the key considerations when deciding on an appropriate tabulating software package is the speed of data access and presentation. Speed is very important as often population databases contain large volumes of data and the user does not wish to wait hours while the software processes any table request.



Tabulation Software

Export capability

- Most cross tabulation software packages enable tables to be exported to other software packages. This is an important feature as often tables are required for presentation purposes or for placement on the Internet. Transportability between software packages is common although often it involves copy and paste techniques rather than saving tables in specific formats.

Tabulation Software

- ***Time series management*** There are generally two types of tabulating packages.
 - The first type, e.g. Supercross, access a single database at a time. To generate time series tables the user is required to run a series of programs against many databases and store the results in separate files that are later joined together using another software package such as a spreadsheet.
 - The second type of tabulation software, e.g. IRDB, while more expensive to develop and implement allows the user to select the variables, geographic area and time periods against which the program will be run. The results may be viewed either within the tabulation software itself or exported to another package.

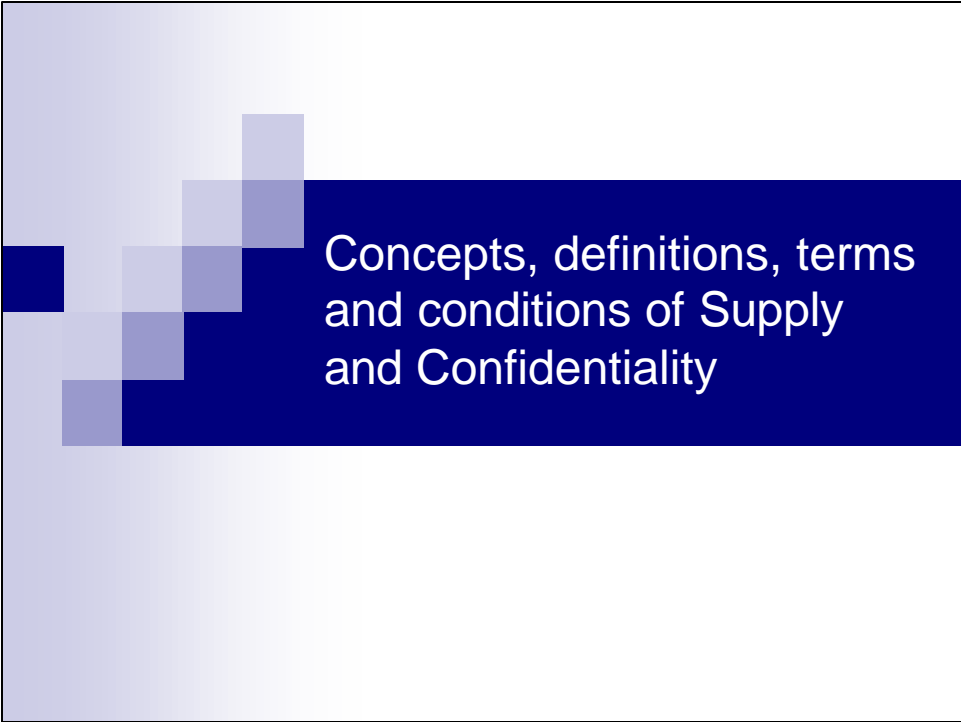
Selecting data extraction software

- Use a suitable user friendly data extraction software package
- Easy interfacing of package with the database
- Should be tested thoroughly before its introduction
- The speed and optimum size of the package



Selecting data extraction software

- Package users should know what variables are held on the output database and how the fields relate to one another
- All staff should be trained on the software
- Instruction manuals and help facilities should be available



Concepts, definitions, terms
and conditions of Supply
and Confidentiality



Concepts and definition

- Covers the more intricate details of the data and is used to assist the user understand terminology and definition of variables
- Include in worksheet or addendum to hardcopy
- Should be self explanatory and easy to read
- No jargon or acronyms unexplained



Quality of data

- Notes on data quality should be included
- Sample error tables or comment should be included




Terms and conditions of supply

- Rules for using the data or passing it to an other user
- Data liability
- Incorrect data implications
- Acknowledgement of property



Confidentiality

- Outline confidentiality provisions contained in the data
- Methods for making data confidential should be included
- Attention paid to extracting data with low value cells
- User awareness prior to use



Key points for data dissemination

- Ensure that data is extracted correctly
- Full explanation of meta data is included as footnotes or summary notes
- Confidentiality should not be compromised
- Use efficient tools to extract data
- Deliver data to the user in a form it can be used
- Deliver data on time



the Internet

Opportunities and challenges of the Internet

Establishing Internet presence

- Internet for data dissemination
- Selecting content for the web site

Establishing Internet Presence

- ***Internet and the dissemination strategy***
 - Prior to deciding to launch an Internet site and developing protocols for the release of data through this medium, the national statistical office needs to evaluate the objectives and the fundamental principles behind its dissemination strategy.
 - Issues that need to be addressed include:
 - the potential to attract new data users,
 - the adoption of appropriate charging policies,
 - the availability, efficiency and cost of the new service,
 - the suitability of the new medium for presentation purposes, security and
 - the potential effect on existing dissemination channels.

Establishing Internet Presence

■ *Internet superseding other media*

- At present the number of Internet service providers, Internet servers and users connected is relatively small in many developing countries but is expected to increase rapidly. The majority of Internet users are computer literate and may become new data users. Traditional dissemination media such as diskette, CD-ROM etc., will probably remain the forms of disseminating population data for some years to come, however, there is no doubt that the Internet will become the pre-eminent media in the future.

Establishing Internet Presence

■ *Home pages*

- The timing of the development of an Internet website by the national statistical office is an important decision. A minimum requirement of a national statistical office, even at this relatively early stage in the development of the Internet, would be to develop a home page that promotes the work and brand name of the producer. Additional pages would provide information about available data and how to obtain access to it. The Internet should be used to let the global user community know that the producer exists and what information can be obtained and download.

Principles for selecting the content for the web site

■ *Websites*

- The Internet is an efficient dissemination tool for statistical data. The Internet allows the sending of text, graphs, tables, maps etc., to data users anywhere in the world simultaneously. It also enables data providers to link their website to the website of other data providers. However, some form and structure for the website is essential if the system is to perform and satisfy user expectations.

Principles for selecting the content for the web site

■ *Web page design*

- The website content must be designed in a clear, structured and uncluttered manner.
- Material available should be confined to essential information only so as to avoid confusion in the mind of the user.
- A hierarchical structure with hyperlinks to other pages containing more detailed information appears to work reasonably well.
- Hyperlinks are links that are integrated within the text and establish connections between website pages or objects within website pages.
- The use of key word searches enables the user to move directly to the specific area of interest.
- Website needs to be practical and avoid the excessive use of memory hungry graphics that can take a very long time for the user to download to their own computer.



Principles for selecting the content for the web site

■ *Web page design*

- The fundamental aim of the national statistical office website should be to provide on a worldwide basis an up-to-date supply of statistical data designed specifically for ease of user access with provision for incremental developments over time.



Components of a website

■ *Home page*

- The Home Page is the first page that every user visits. For this reason it must be well designed and welcoming. The national statistical office logo should appear indicating whom the data provider is. A menu driven system is probably more efficient although a key word option could also be made available. In some instances it may be necessary to have a multi-lingual website. However, initial developments may concentrate on the language of first choice of each country.



Components of a website

- ***About the NSO***

- A page should be allocated to provide the user with information about the national statistical office i.e. the data provider. Issues such as the statistical collection procedures employed by the country, confidentiality procedures applied to the data, information about the nature and type of statistical activities and functions performed by the data provider will all aid the understanding of the data user.



Components of a website

- ***Products and services***

- A number of pages may be designed to present to the data user the range of products and services available from the national statistical office. This may include a list of publications including information relating to how to obtain copies and how to download, as well as sample outputs. Products should generally be available in both printed and electronic form at the choice of the user.



Components of a website

- ***Recent releases***

- There should also be a page allocated to new or recent releases such as news media releases or other public statements containing important results of official statistics. This information should be made freely available to the data user with information about how to obtain more detailed statistics.



Components of a website

- ***Online data***

- More advanced national statistical offices may consider placing important economic and social indicator data onto the website. There is also the potential to include database access software whereby the data user specifies the variable parameters and the software generates the tabular output which the data user may view on screen or download to their own computer for further analysis. This concept can logically be extended to statistical mapping, graphing and charting of the data.



Components of a website

- ***Links to other providers***

- Finally the national statistical office may consider providing links to the website of other statistical service providers. The Internet is a global communication network and the provision of links between each website assists the user to navigate their way around a series of related subject matter areas simply at the click of a mouse button.



Stages of website development

- Informational
- Enhanced
- Interactive
- Transactional
- Seamless (fully integrated)



New concepts affecting NSO presence on the Internet

- E-Business procedures
- E-Government applications
- Portals
- ASP



Representation on the Internet

- Graphics
- Maps
- GIS data



Standards for data on the internet

- Metadata
- Content



Metadata

- A critical component of the provision of statistical data via the Internet is information that describes the data i.e. metadata. This is an extremely important factor and contributes enormously to the success of the Internet data provision strategy. Metadata should provide the data user with information related to quality issues, sampling methodology, classifications used, editing procedures that may impact on data quality and time series trends, survey coverage rules and changes etc. All of this information enables the data user to become more informed regarding statistical processes and the data series themselves. Ultimately this should result in a better informed user community and consequently more sophisticated use of the available statistics.

Content

- Metadata should be presented in a structured manner. Each output requires its own metadata. A hierarchy comprising the following categories is needed:
 - Data description: purpose of the collection, general survey information
 - Data availability: start of the time series, any period gaps, update frequency
 - Questionnaires: preferably an image or else a list of the questions, response categories and any guide notes
 - Output variables: list of variables and code ranges, changes over time
 - Changes to survey: changes made and their impact
 - Data use and limitations: definitions, classifications and caveats
 - Methodology: sample errors, sample size, area exclusions etc.
 - Media availability: restrictions, list of tables and publications available

References

1. ***Guidelines on the Application of New Information Technology to Population Data Dissemination*** - Economic and Social Commission for Asia and the Pacific
2. **ESCAP Workshop -Data dissemination**