UNITED NATIONS E

ECONOMIC AND SOCIAL COUNCIL

Distr.
LIMITED
E/ESCWA/STAT/1997/WG.3/4
9 October 1997
ORIGINAL: ENGLISH

Economic and Social Commission for Western Asia
Workshop on the Strategies for the
2000 Round of Population and Housing
Censuses in the ESCWA Member Countries
6-10 December 1997
Cairo

DEVELOPMENT OF DATABASES AND DISSEMINATION OF CENSUS RESULTS

by

Annegrete Wulff
Statistics Denmark

UN ECONOMIC AND SOCIAL COMMISSION FOR WESTERN ASIA

31 -03- 2000

LIBRARY & DOCUMENT SECTION

- The views expressed in this paper are those of the author and do not represent those of the Economic and Social Commission for Western Asia.

- Issued as submitted.

Development of Databases and

Dissemination of Census Results

Annegrete Wulff
Statistics Denmark

Abstract

Population and housing censuses are crucial for all countries. The results are sought after with impatience. A census is typically based on questionnaires and conducted every 10 years. In some countries like Denmark, there is a possibility to utilize administrative registers and to combine them in order to create registers ready for statistical production on person level. Huge amounts of statistics can be produced this way. It can also take place much more often i.e. annually.

This paper describes experiences in Denmark. It focuses on the organisation of data in general and on dissemination of census results and census related data in particular. Dissemination is related to the users. Different users have different needs. This means that a variety of products has to be available. We have to avoid discrepancies and at the same time fulfil requirements on the level of detail, timeliness, media as well as user skill and experience.

The statistical production is continuously developed. Future challenges involve an extension of international co-operation and data exchange. These aspects are also penetrated.

1. INTRODUCTION

All statistical surveys or censuses involve data collection. The amount and type of work for data collection depends on the source. Data collected on questionnaires require other activities compared to input from registers. Irrespective of input, the continued work will consider:

- the organisation and storage of data
- the dissemination of data

It is important to make sure that these activities are interconnected. The organisation of data provides the frame for the possibilities for efficient presentation and dissemination.

Coherence is often mentioned among the substantial objectives in the dissemination of statistics. This concerns the idea to have data organised as one source from where all dissemination takes place. The different output products should be similar although presented at different aggregation levels, with different selections and on different media.

The dissemination methods have to consider the different users and their requirements as well as possible consequences on the organisation and the production process.

2. ORGANISATION OF DATA

2.1 The variety of demands

From limited dissemination...

Historically the coherence concept has not been presented with the same weight as we see nowadays. Until lately the statistics were mostly presented on paper. Discrepancy among the presented figures could be found in this situation. The figures that reached the users were however often summarized and therefore caused no serious problems.

...to a range of media...

Today, we operate with a wider range of output media. For instance:
- publications on paper
- on-line databases
- CD-ROM
- diskettes
- Internet, WWW

...and detailed statistics

The electronic media enable us to distribute large amounts of data. This means that we allow the user to pick and select the required data from different databases, diskettes or through other services. We have not restricted and connected one medium to some specific statistical information. The user will therefore be able to retrieve the same information from different sources.

A consequence is that any incoherence between statistics presented within or between media will be more obvious for the user - and it leaves the producer of statistics at least with some problems of explanation.

2.2 Typical production

Figure 1 below describes in general terms a situation which is often found in the production and dissemination process in many statistical offices. To a large extent, our organisation works in this way.

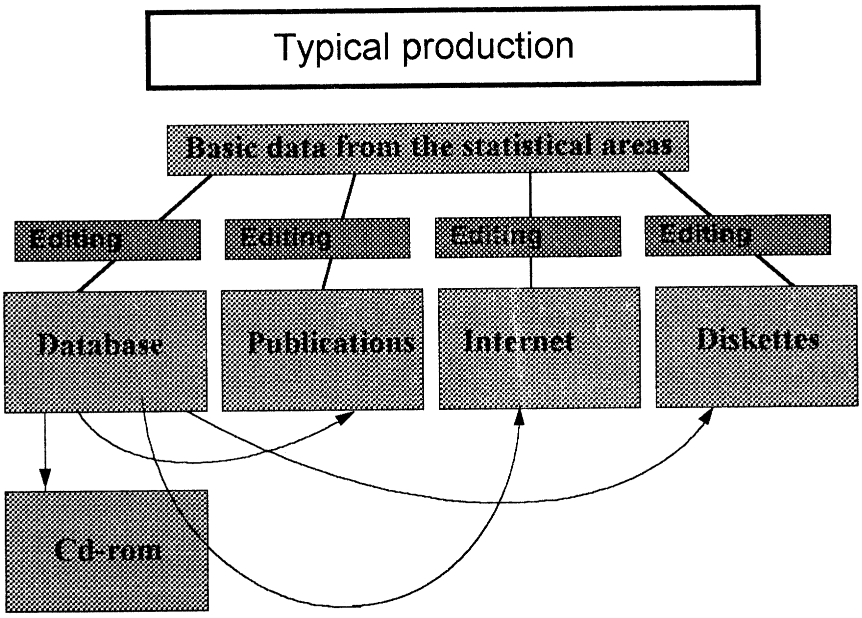

Figure 1.

Output specific compilations

One specific statistical area delivers statistics to different media. The data can be manipulated in different ways for publications, diskettes and data banks. This can cause discrepancies even between printed publications.

Subsequent dissemination based on diskettes, CD-ROM's or databases may later reveal many inconsistencies.

2.3 Coherent production

One source...

We have as producer and supplier of official statistics in the country, a responsibility not only to deliver figures we collect and calculate, but also to ensure that these figures can be used comprehensively by our users in their new technological environments. Coherence should therefore be considered a quality label on our production. To co-ordinate this, an overall system of statistical production to handle this will be indispensable. Such a system is often referred to as a ‘meta data system’.

... many media

This is the only way to ensure that the same information can be found regardless of whether it is retrieved from the databases, a diskette, a book or from the Internet.

3. CONSEQUENCES IN THE ORGANISATION

3.1 Centralisation vs. decentralisation

An increasing movement of decentralisation within our as well as other organisations has been flourishing during the last 5-10 years. The acquisition of new types of hardware and software has supported this move.

Central guidance

The assurance of coherence in our statistics production will nevertheless require some degree of centralisation or centralised guidance concerning meta information, file formats, data storage and dissemination.

3.2 Standards

We define standards. We must also make it easy to follow these standards and difficult to ignore them.

Standards are developed because they are supposed to make life easier. Unless introduced and marketed properly, the effect can be the opposite. People feel that their range of activities is being restricted, and that tailor-made ad hoc solutions tend to come closer to individual wishes and tastes. This may be true in an isolated case but at the same time we have to keep in mind that the whole statistical production must be co-ordinated.

We interact with a lot of people both inside and outside our own institution. The number of excellent tailor-made solutions are at the same time growing. It can be a drawback for effective communication and general comprehension if these services are not systematically related.

3.3 Coherence in contents

It is considered important with the same definitions and similar aggregations and classifications across subject areas and in different media. Nomenclatures are often agreed upon in UN, OECD, Eurostat and so forth. For various reasons this is not always followed. When the statistical information is presented in a book, the writer can compensate less known terminology by commenting the figures. The author has not the same influence on detailed electronic databases since he does no longer present the figures himself, instead they are selected by the individual user. It is because the electronic media can provide the user with facilities to combine and aggregate figures in a way which is difficult for the author to foresee and therefore to comment on.

It is confusing for the user when the breakdown by education used in population statistics varies from that used in labour market statistics. It may also really cause trouble when two variables with different names actually cover the same concept.

With the extensive use of electronic media like Internet such divergence's can become very visible and therefore also cause serious misunderstandings. The user has the opportunity to extract figures from a variety of sources. In this case it can be difficult to realise to which source you are connected. At least when the systems are less known and perhaps feel tricky.

The headlines and the most aggregated figures of the statistics must not deviate from what is retrieved from the more detailed publishing. This can only be obtained when the data sources are technically co-ordinated.

3.4 Coherence in format

Well-structured information on a technical level becomes a prerequisite when we want to benefit from the new ways of disseminating statistics electronically.

A coherent output format will make it easier for the user to work across different subject areas. Statistics may be distributed on-line via Internet, through different off-line products like CD-ROM:s. Supplements to printed publications or ad hoc deliveries can be made on diskettes. It is always a great advantage when the output data fits into the same software and is combined with the correct meta data. Our standards for this purpose are already defined and used to a large extent.

3.5 The output database

Our database systems are as emerged from earlier in a restructuring phase. The future *output database*[1] will consist of the *macro database*[2] and *the meta database*[3]. They are both being developed and the first versions are expected to be in production in 1999.

The *macro database* in the form of a relational database in Oracle/SQL will be accessible through different interfaces. PC-AXIS[4] will be the most important interface but other software like SAS could be used to download and manipulate data.

The *meta database* contains structured information to interpret data (both for the user and for the system). A very important part of the meta database will be a database of quality declarations. Every single statistical subject area will be described here with keywords, administrative information, collection

[1] The data base with non-confidential information for external use

[2] The data base with aggregated statistical data

[3] The data base with 'information about statistical information'

[4] PC-AXIS: Windows based software for manipulation and conversion of statistical information. Developed by Statistics Sweden in conjunction with other statistical offices.

method, relation to other areas, description of the registers and their historical or legal background, reliability, availability etc. It is essential that these descriptions are integrated in the dissemination system. Otherwise it will be difficult to assure their proper use.

3.6 Implementation

It is a difficult task to implement the necessary tools for efficient dissemination. The following are essential requirements:

- The meta data system must always be *sufficiently updated*. If we want to achieve the full use of the standards we have to eliminate obstacles. Local meta data in "private" systems which are not fed from the central system ought to be minimized.
- Everyone who need to enter or use information from the meta database must have *easy access*. The tools must support the defined standards. Menu driven assistance is i.e. normally better than allowing the entry of free text.
- *Acceptance in the organisation*. Standards must be part of the culture and the staff must understand the necessity of accepting inconveniences in some cases in order to achieve an advantageous overall result.
- *Central guidance* to complete the meta database is needed. The frames will be set here, while the meta data entry is decentralised.

4. USER REQUIREMENTS

We must know who our users are in order to develop and maintain effective statistical dissemination.

Primary users..

A statistical office might be obliged to serve some user groups rather than others. In many countries such obligations regard the Government, ministries and international organisations. These users may have preferential treatement and some services may be delivered free of charge.

...and neglected ones

Others users like schools, libraries, the press and “ordinary people" may be intentionally neglected due to restricted resources.

In the past Statistics Denmark has found itself in this position however in recent years we have tried to improve the services for all users.

One user category...

We have traditionally classified our users according to their "institutional background", for instance:

- Ministries and other governmental bodies
- Local authorities
- The private sector (banks, large industries etc.)
- Researchers
- Education sector (from secondary schools to universities)
- The press

...with different needs

This is a useful classification for some purposes but not for all. When we focus on requirements in different situations, the same user can act in various ways. The same organisation may therefore have a lot of different users and user needs. Even the same person can act as 'a multiple user'.

Timeliness, overview

Many professional users want immediate answers to their questions on different key figures. The priority is related to timeliness, accessibility and overview.

Recognizable systems

Another user requests detailed information on a variety of topics. This is a situation where the user appreciates a well-known retrieval system. It is an obvious advantage to be able to retrieve and transfer statistics from different areas (for instance the national accounts, the labour market or social conditions), using the same system. The information is often processed further in the users own local system. Therefore we have to ensure it is not only the technical system that allows for these combinations, but also the statistical contents themselves.

Regular updates

The type of user who wants 'the same time series every month' does perhaps not care so much about all fancy details we can surround the figures with. He wants a quick, reliable delivery in the expected format and in accordance with the agreed definitions. Information about breaks in time series etc. is indispensable and has to be attached to the data in a structured way. This also indicates that we need a dissemination system which is closely integrated with meta data information.

Attractive form

'The net surfer' is an occasional user who might care more about an ornate and entertaining system than about its contents. Nevertheless we shall not underestimate this user, as he might return in one of the other categories - if he found something interesting when surfing around.

5. HOW TO FULFIL THE NEEDS

5.1 Dissemination strategy

Statistics Denmark is the central producer and supplier of statistics in Denmark. The fulfilment of the user's needs is therefore one of our major duties. A dissemination strategy was adopted in 1992 and revised in 1996. A part of this dissemination strategy is to:

- Give access through the required data nets and media.
- Use paper publications for overview and text.
- Use a retrieval system that allows for import and export in standard file formats.
- Distribute statistics on different aggregation levels, but always using the same base as a source.
- Establish a user interface with easy navigation possibilities to find and select information.
- Establish easy ways for the user to repeat actions and store selections.

– Integrate the meta data system with the retrieval system for statistics.

5.2 Experience so far

Coherent databases since 1986

We have worked on this strategy for over 10 years. The basis has been the on-line databases in the AXIS system[5]. About 80 percent of our production is stored in these databases today. The policy since 1987 is that all statistics of general interest shall be stored and made available in the databases. The meta data system incorporated within AXIS is maintained and operated by the central database administration.

Reluctant accept

This centralisation was reluctantly accepted by the organisation, especially in the beginning when the data banks were established. But we considered the strategy necessary in order to keep coherence between the different statistical fields. This was confirmed by the management. Today most producers of statistics can see that consistency and coherence is advantageous.

In-house availability

The first step in gaining acceptance in the organisation is, of course, to let everyone be aware of the possibilities and requirements. Our external database systems are therefore available for our staff through our local area network. Assistance is offered by the data bank administration, and information is given regularly on e-mail and at special meetings.

From data bank...

Our present on-line databases have 250 external subscribers who retrieve over 20,000 cross tables a year and download most of them into their own PC. The export format can be Excel, Lotus, ASCII or PC-AXIS. Documentation (methods, source, definitions, contact person etc.) is transferred simultaneously.

...to CD-ROM...

The databases are used as the source for an annual CD-ROM. This is produced together with all statistical offices in the Nordic countries[6]. All files on the CD-ROM are in the PC-AXIS file format. The CD-ROM was sold in 400 copies in 1996.

...diskettes and books...

Some publications are supplied by a diskette. This diskette contains more figures than can be found in the book. The PC-AXIS format is used for the data files. A viewer is included on the diskette as well. It is a lean but compatible version of PC-AXIS. The viewer is named PX-MINI. The first success with this method was the '50 Year Review', a publication which was sold in 6,000 copies. The method has later also been used in our most popular publication, the annual '10-Year Review'. Some 20,000 copies are sold each year.

...on the Internet

The homepage on the Internet was extended with access to databases in 1997. Statistics on short-term indicators are updated every morning and released in the PC-AXIS file format. It is easy for a user to download the information in the required format.

[5] AXIS: Mainframe software developed and supplied by Statistics Sweden.

[6] Denmark, Faroe Isles, Finland, Greenland, Iceland, Norway and Sweden

Any table can be stored in HTML format through PC-AXIS. This diminishes the task of transforming files on to the World Wide Web.

5.3 Solution

Our solution in order to fulfil the dissemination strategy is shown in figure 2 below. The future output database will consist of the SumDatabase (our macro database) and TIMES2000, which will be our meta database. Please note the central position for PC-AXIS as dissemination and conversion software. GESMES is described in 6.2.

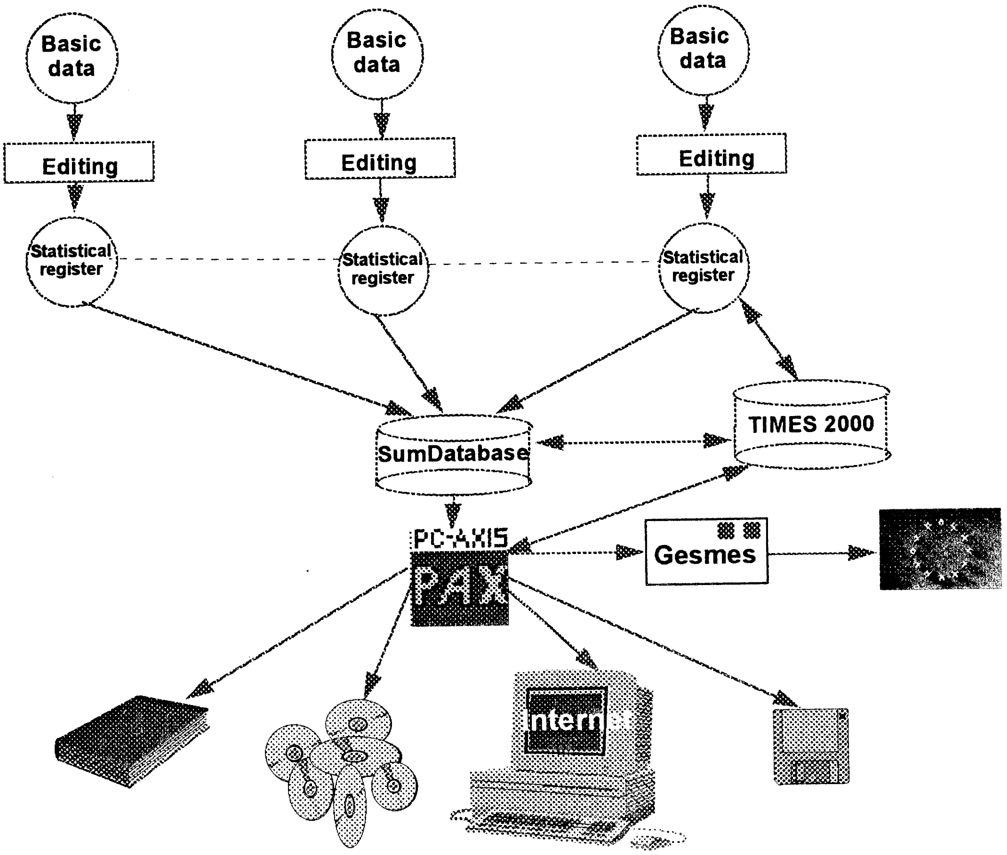

Figure 2

6. CO-OPERATION

6.1 Collaboration in Denmark

Several organisations - the same concept

Many users are very satisfied with this PC-AXIS concept. Some of our major users have shown interest in using the same format to distribute their own collected and complied statistics. The concept of including a diskette in PC-AXIS format together with the viewer PX-MINI is very attractive. This means that there is technical compatibility with the official statistics from Statistics Denmark. We look very much forward to extending this type of co-operation.

6.2 International data exchange

Nordic database

We expect further consistence in the dissemination among the Nordic countries. In 1997 it has been agreed to deliver data to a Nordic statistical database in the PC-AXIS format. This database will be the source for a printed edition of The Nordic Statistical Year Book. It will also be one of the sources for common Nordic data for the annual Nordic CD-ROM, 'Statistics Across Borders'. This saves time for the people who deliver input data to the Nordic database, especially when it can be retrieved from the national data banks in the PC-AXIS format. Moreover it promotes coherence and consistency in the content for the end user.

World wide GESMES

We hope that in the future we can include data from other countries and use these principles as well. We intend to exchange aggregated data in the GESMES[7] format since PC-AXIS can both import and export data in this format without losing any meta data information during the process.

Further countries

Several other countries and statistical offices within and outside Europe use PC-AXIS to various degrees in their statistical production. This is very satisfying. We will also be happy to share our experience with further countries and institutions in this field.

7. NEW CHALLENGES

7.1 Requirements change

It is essential to reflect very closely on new user requirements. This regards access channels as well as the work with statistics. The recently adjusted dissemination strategy stresses this, and one of the consequences is that our long-term plans have to consider to an even more detailed degree these continued changes in the user behaviour.

Communication channels must be more focused on the *Internet*. This is a demand we meet more often. However any other channel can not be abolished in the short run since many users have integrated our present services into their production systems. Some functions are also difficult to implement in the Internet for the moment.

We have to take into account changes in the demands for the delivery and exchange of information with *international organisations* such as OECD, UN, EU and other statistical bodies.

The use of statistics in electronic form is growing and *new users* appear. This makes it even more important to provide fully documented data instantaneously in various forms.

[7] GESMES: Edifact standard for the exchange of aggregated statistics. Supported by Eurostat.

7.2 Next step

Coherence regarding contents, definitions and media is central to our dissemination strategy. We owe a debt of gratitude to our old AXIS database system since it has given us the habit of working in this direction. The next step before "the big change" will be to transform even more data to AXIS matrices. Almost all social and demographic data, along with a greater portion of other data, are stored here but still some remain to be transformed.

We are in the process of converting further times series material to AXIS in addition to the hundreds of thousands we have at present. The reason for this is to facilitate the future transfer from the mainframe system to the *macro database on an SQL server*. Only one unified tool will then be required.

7.3 Conclusions

The assurance of the dissemination policy for our statistics production will require some degree of centralisation or centralised guidance concerning the concepts, and definitions, meta information, file formats, data storage and dissemination. It is critical for the success of the strategy to ensure *acceptance throughout the organisation*. This includes the full and visible *support of top management*.

We are at the starting point of a new era, where the goals are set, but where many solutions are still open. We see many opportunities to create better products and use less resources. *International co-operation* is very inspiring. We already have among others excellent experiences from co-operation with the Nordic countries.