

Distr.
GENERAL

CES/SEM.40/10
4 September 1998

Original: ENGLISH

STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

STATISTICAL OFFICE OF THE
EUROPEAN COMMUNITIES (EUROSTAT)

CONFERENCE OF EUROPEAN STATISTICIANS

Joint ECE-EUROSTAT work session
on Population and Housing Censuses ¹
(Dublin, Ireland, 9-11 November 1998)

Study topic 3

DISSEMINATION ISSUES
IN THE 2001 HUNGARIAN POPULATION AND HOUSING CENSUS PROGRAMME

Invited paper submitted by the Hungarian Central Statistical Office ²

I. Overview

1. Traditionally, census information, as other statistical information as well, used to be published in printed form. This, of course, limited the users in what and how they could obtain and use for their specific purposes. Electronic data processing helped statistical services to be more flexible in what they are able to provide to users, and more recent development in EDP techniques, hardware, software and communication increased this flexibility.

2. Increasing use of statistical information also implies that users have become rather heterogeneous, with diverse needs and requirements. The statistical office has, consequently, more and more difficulty in satisfying a broad range of specific needs. Unless we are ready to face constant criticism, and fail to fully comply with our responsibility, it is in our interest therefore to find flexible and tailor made means of dissemination.

3. The majority of users in the past could hardly be considered as professionals in the use of statistics, the general consensus was therefore, that statisticians more or less defined users' needs. This situation has changed dramatically over the last 10-15 years. The demand for and the value

1 The papers which are prepared for this work session will be treated in the same manner, as papers that are prepared for seminars.

2 Prepared by Gabriella Vukovich.

of reliable data have increased in all user groups, consequently the input, on the user side, into techniques and capacities of analysis have also grown. Many users are very much professionals, they can and do define their specific data needs as well as their preferred scope and format. The production of statistics has become much more demand driven than it was previously.

4. Printed material still has a certain value especially in informing the general public on social and economic issues covered by the census but, in addition to that, users require that the statistical service provide data in machine readable electronic or optical format, partly in the form of set tables, but increasingly as data bases which allow the freedom of selection and combination of variables.

5. A number of questions arise when discussing dissemination issues.

- How to identify users' needs? When is the best time to plan dissemination? How can we attempt to serve as many types of users and needs as possible?
- How can the statistical office comply with data protection requirements if users have access to databases? Census data are special, among other things, in terms of geographic and personal detail. This implies that free access to the original census database would make indirect personal identification possible even if the files are anonymised.
- What cost recovery efforts of the statistical office are appropriate? What information should be available free of charge, and what services should be charged? How high can charges be? Should users be treated equally, or are there categories of users who should have access to information free or at low cost while others are charged higher fees?

II. Identifying users' needs

6. Identifying users' needs for census data in general is well known to all statistical offices. Various institutional solutions exist. Consultations in standing and *ad hoc* committees of internal, external and international users of census data are regular practice during the census planning process. The planning period, however, is well before the dissemination period, therefore the planners focus on what questions to put in the questionnaire, on the formulation of the questions and on coding and processing issues.

7. Little attention is paid at this stage to the actual form of dissemination. This is justified to a certain extent because the planning process is more or less completed 12-18 months before Census Day, whereas dissemination will only start 9-12 months after Census Day. 30 months is a lot of time in our information societies, certainly long enough to completely change the receiving end of the communication channel. We already find it difficult to make sure that we decide on our own IT as late as possible, without jeopardising the data processing itself, to make sure that the IT we use is not too much outdated by the time the data processing starts. Our

clients, however, have other development considerations and the information technology many of them install is well ahead our own.

8. The 1990 Census results were published in standard book form, one volume of approximately 550 pages of tables for each of the 19 counties plus the capital and thematic volumes. On the whole, the series comprised 88 volumes with a total of approximately 20 000 printed pages. The series was completed by a volume on the methodology and definitions. Census Day was 1st January 1990, the county volumes came out in June and July 1992, the thematic volumes were published after that, the last one in February 1995. In addition, a diskette with a database containing basic information for all the settlements of Hungary and a mapping programme was prepared. A CD-ROM was issued with the complete data file and a database query programme, supplemented by a mapping programme. The CD-ROM, at that time, was considered a very modern medium, so much so, that although it was very popular many interested users had to postpone purchasing it because they could not easily buy CD drives. By the time the CD drive became almost standard equipment in desktop computers, the DOS based CD publication was practically outdated, and we had to develop the Windows version. (The latter, however, still has a number of uncertainties. While it is convenient for internal use, we do not encourage clients to use it).

9. Many users do not find the combinations they need in the 1990 publication series, and since they only need a few tables, they find it easier to place special orders for tabulation with the statistical office. Towards the end of the 1990ies, we still receive an average of 20 orders per year for special combinations from the 1990 census, especially for small areas. These orders are carefully documented also for ourselves, to keep track of what users currently need from this type of data source, and what is the usual form in which they request these data. Further individual requests for statistical data other than census data are also documented and used as background information for us to plan the 2001-dissemination strategy.

10. In addition, in the planning phase of the 2001 census programme, users are asked about their preferences concerning the dissemination of census results. What we can clearly see from their replies is what was only to be expected, i.e. that standard tabulations have less and less value, and direct access to data files is the general preference of professional users. Also, users expect to see data very soon after Census Day (much sooner than we are likely to produce the data), which means that the 1990 publication schedule is out of question, and a much faster dissemination schedule has to be envisaged. This pressure for timely release also underlines the need to provide access to the data base as soon as possible, without the delay that is inevitably associated with traditional forms of publication.

11. For those users who are interested in the census results in a more general way, standard tabulations are probably the best way to provide data. A book on the shelf is still useful to many users who wish to look up or to use a few census figures. We intend to publish regional (county) and thematic volumes with standard tables of the 2001 census for these users, but the

tables will have much less detail than in the 1990 census cycle. These volumes will serve general orientation rather than sophisticated analysis.

12. There are a number of users who need data files or databases for various research or training purposes. A census microdata file seems to be the best product for them. We have not yet decided how large the sample should be. Probably a 2 % national sample of anonymised records serves the purpose and is easy to handle on PC's.

13. There is some uncertainty about the regional data. A 2 % national sample is not detailed enough for some sub-national uses, but because the extensive responsibilities of local authorities are relatively recent, and a new regional development policy was recently adopted, we still have to discuss with users about the production and the availability of regional and small area samples.

III. Protection of personal data

14. Indirect identification of respondents is an issue that really emerges when data of small geographical units are released. Various solutions have been adopted by statistical services to avoid indirect identification, like "barnardisation" (the random addition of +1 or -1), or restricting the variables that can be released for small areas, or setting a threshold, a minimum number of persons or households or dwellings that have to be in the area to which the data relate.

15. Enumeration districts have been used, in most countries, as the smallest geographical units for which data were published. In the Hungarian practice, an ED contains about 100 households (or somewhat less than 300 persons). Because of the very strict data protection regulations this threshold is usually considered too low, and we are required to provide data for more than one (usually 5-10) Ed's at a time. In the 2001 census we would like to use enumeration districts for what they were "invented" for originally, i.e. the workload of an enumerator. In the dissemination plans of the 2001 census we intend to discard the enumeration districts as geographical units. We intend to make, instead, customised delimitations from an address register (which is an address file of all dwellings, created for the census and supplemented with substantive information on the dwelling and its surroundings on the occasion of the census). This provides a lot of flexibility in fulfilling customers' needs. All delimitations will, of course, have to meet the threshold criteria.

IV. Intermediate disseminators

16. Another issue still under discussion is the transmission of the census data base or census products to intermediate disseminators. Until now, this has not been the practice of the Hungarian Central Statistical Office. The legal (data protection, copyright), practical and financial aspects of dealing with intermediate disseminators still have to be dealt with.

17. We would like to encourage intermediate disseminators who can add value to our products, thereby broadening the scope of services that are delivered to customers. Value added products that are already in demand (and not yet easily available) are GIS applications that use census and other statistical information. This certainly will be one of the main fields of interest for users and for intermediate disseminators, especially because the statistical office, for the time being, has only limited GIS capacities. We have to admit that such applications will be very useful for the statistical office itself as well, both in the planning and fieldwork of censuses and surveys, and in data analysis.

V. Cost recovery

18. Probably all statistical offices have to face budgetary constraints, some are required by law to generate revenue. This is the case in the Hungarian Central Statistical Office, too. The law on the annual budget sets a certain amount of revenue that the office has to make. The law does not encourage the public sector to make profit, only to move towards cost recovery. The regulation in the Statistical Act and in internal rules is that users have to pay the additional costs incurred by data processing, tabulation or other services that are beyond the standard range of products that the statistical office releases. However, there is no clear definition of what the standard range of products and services is. It is within the responsibility of the statistical office to plan outputs, and statisticians have a certain tendency towards traditional solutions. Many users therefore have to pay for products, which could be included in the standard range of public information in the place of others that have lost much of their usefulness.

19. The statistical office does not intend to recover the cost of data collection or the cost of data processing, up to the point of establishing a database from the census records. Beyond that point, there are certain products that are public information and others which require additional input from the statistical office. The simplest example is the case of volumes of printed tables, which are for sale but the price is set around the printing costs, without profit to the statistical office.

20. The likely dissemination cost scheme will include

- some basic information free, on the internet and in brochures,
- a set of standard tables available at low cost,
- a microdata file at low cost,
- regional and small area samples at slightly higher cost than the national sample, because of the higher analytical and practical value,
- files available to intermediate disseminators at terms to be negotiated,
- data sets tailored to specific individual needs, at the additional cost incurred by special tabulations or by producing a specific database.
- in the Hungarian Central Statistical Office practice, training institutions and research institutions are provided with data or data files at preferential rates or free of charge, the limitations as to the use of such data are stipulated in the specific agreement between

the client and HCSO. This will be the likely approach of the dissemination policy in 2001.

21. As to how users will have access to the data and the data files or databases, there is some uncertainty, because of the fast development in electronic communication. CD-ROMs may still be in demand, and the Internet seems to have conquered the heart of users, but other communication means may be developed by the time of census data dissemination. The important thing is to assess the form users prefer, and try to comply with their preferences.
