ECONOMIC COMMISSION FOR EUROPE
STATISTICAL DIVISION

UNECE/UNDP Task Force Meeting on Gender Statistics Website for Europe and North America
(25-26 September 2002, Geneva, Switzerland)

Working Paper No. 4
Agenda item 3

Issues Related To Data

Paper submitted by UNECE Secretariat[*]

1. Progress made on the UNECE Gender Statistics Database

The Task Force in May 2001 recommended a set of common gender indicators for 7 policy areas to be made available on the website. In order to present these indicators for the recommended timeframe (1980, 1990, 1995, 2000, then annual series), substantial data collection was necessary, and a UNECE Gender Statistics Database needed to be set up.

Starting from the common gender indicators recommended by the Task Force, the UNECE gender statistics team identified the necessary statistical series and prepared a questionnaire for data collection towards the end of 2001. The questionnaire aimed at being self-explanatory, user-friendly and exhaustive in terms of methodological information (definitions, table lay-out, etc.). Methodology and concepts used should be in line with international standards.

The pilot version was tested on four countries (Austria, Israel, Poland, Portugal) at the beginning of 2002. As a result of the pilot collection, a few amendments to the questionnaire were made and a user-friendlier format for the collection of metadata was introduced. In order to help Russian-speaking countries, the questions and definitions used in the questionnaire were translated into Russian.

The questionnaire was sent out in March 2002 to the gender focal points in country statistical offices. That is, to 45 countries[1] out of the 55 Member Countries of the UNECE. The deadline for replies was in May. To date, we have received replies from 37 countries; some of these are partial and cover only selected areas. The countries which have not replied are: Azerbaijan, Kazakhstan, Luxembourg (data promised), Sweden (data promised for September), Switzerland (data promised for August), Tajikistan, Turkmenistan and Uzbekistan.

[*] Prepared by Tiina Luige and Sabine Gagel

[1] Albania, Armenia, Azerbaijan, Austria, Belarus, Belgium, Bulgaria, Canada, Cyprus, Czech Republic, Denmark, Estonia, Finland, Georgia, Germany, Greece, Hungary, Iceland, Ireland, Israel, Italy, Kazakhstan, Kyrgyzstan, Latvia, Lithuania, Luxembourg, Moldova, the Netherlands, Norway, Poland, Portugal, Romania, Russia, Slovakia, Slovenia, Spain, Sweden, Switzerland, Tajikistan, The former Yugoslav Republic of Macedonia, Turkey, Turkmenistan, United Kingdom, United States, Uzbekistan.
There is no gender focal point from the statistical offices of Andorra, Bosnia and Herzegovina, Croatia, France, Liechtenstein, Malta, Monaco, San Marino, Ukraine and Yugoslavia – for the time being, data for these countries are not included in the database.

In parallel to receiving the replies from countries, the database structure was prepared based on a system developed by the UNECE Statistical Division. The same system is used for several databases maintained in UNECE.[2] This allows us to save time in database development, as the basic structure, functions, user interfaces, etc. are pre-programmed. At the same time, there is less flexibility as regards the use and functionalities of the database.

One person was developing the Gender Statistics Database and transferring the data - Anne Chataignier, and three other persons were checking incoming data for completeness, consistency and plausibility – Sabine Gagel (she did most of the checking), Stein Vikan and Tiina Luige. Later also Marie Sicat joined the team.

The data checking, communication with countries to get corrections, and transfer to the database has taken about 4-5 months (from May to August/September). Next to the checking of the actual data and ensuring their availability with harmonised formats in Excel that allows for an automatic transfer of the data into the SQL-database, the harmonisation and input of metadata turned out to be very time-consuming.

We hope to transfer all data (and metadata) received by September and make the Gender Statistics Database available on the Gender Statistics Website in September/October this year. For more details on the envisaged website presentation of data, see below.

2. Data collection issues

The first data collection was a big and resource consuming effort both on the side of countries and the ECE secretariat. Therefore, a very big “Thank you!” to all focal points and staff in statistical offices who filled in the questionnaires!

There are several reasons why the amount of data requested was so big:

- First of all, according to the recommendations of the Task Force and in order to be useful for monitoring trends already now and not only in distant future, we asked for benchmarks years – 1980, 1990, 1995, 2000. From there on, it is recommended to collect data annually, so we also asked for 2001 data.
- The questionnaire tables are often very detailed (e.g., requesting data for 5-year and sometimes for 1-year age groups) – we will discuss whether it is always necessary to keep this level of detail.
- It was not clear which data we could use form other international sources – even if the requested data (i.e. same breakdown, same level of detail) were exactly available from other international organisations, it was often not possible to get data for all the benchmark years and therefore we still had to turn to countries. We will discuss the related problems in more detail below.
- The idea of pre-filling the questionnaires using data that could be retrieved from other sources proved to be unfeasible. We tried it on some countries, and it took about 1 week full working time of one person to pre-fill the data for 1 country. In addition, the retrieved data was often different from the data that the countries themselves had, so that it turned out to be easier for the countries to just fill in a blank questionnaire, and not first to compare the pre-filled data with their data.

The next data collection will already be much easier – the amount of data to provide will be much less because the data will be needed only for one year (if collected annually) and in

[2] UNECE databases are based on a Microsoft SQL server 7.

some cases with less detail (see discussion of the current questionnaire below). Also, we hope to improve on using existing international sources for the questionnaire which reliably could be used for database input (at least for some parts and some countries).

Questions for discussion concerning the data collection cycle:

- When is the best timing from countries viewpoint to ask for data so that we can get as fresh data as possible?

- How much time is needed to fill in the questionnaires (length of delay for the deadline after sending out the questionnaires)?

- To collect data each year or once in two years (asking for two years data at the same time)?

3. Potential data sources and cooperation with other international organizations

Gender statistics is one of the areas where indicators are being developed in response to the major international conferences of the 1990-ies. The development of standard data sets and consultations how to cooperate in data collection for these indicators are going on now on high level. There is not yet an agreement between international organizations for tailor-made provision of data for exactly these indicators and breakdown as needed for the Gender Statistics Database. Therefore, we had to rely on data that is readily and publicly available from different sources. The possible sources for data (international organizations, especially databases available through Internet) are listed in the Annex.

We have contacted the UN Statistical Division and OECD in order to see if some of the population, demography and education data could be received directly from them ***before*** the data become available through their public databases. Part of the health data is taken from the WHO database. We are also using data from Eurostat New Cronos and ILO databases. After the first round of data collection is finished, we will take another critical look at these possibilities to see if the data request to countries could be reduced by what is possible to retrieve from other sources. Direct investigations with focal points in each country in order to identify or confirm the exact international sources for specific tables could be envisaged.

In theory it looks like using data from international sources might be a good solution for reducing our efforts on data collection and countries' burden in providing data at the same time. In practice, however, there are several problems linked with the use of data from other sources:

3.1 Data quality

The organization collecting the data often makes estimations or adjustments to the data in order to follow an established definition, to increase the comparability between countries and for other reasons. It is not evident what has been done with the data and how much it deviates from what was originally submitted. The countries themselves make adjustments, so that the data submitted later is not corresponding to what was submitted earlier, and it is not possible to update international databases after each such adjustment.

For example, the population data available from the UNSD database and from the country replies to the questionnaire differs about 2-3%. In the case of some countries, the data for earlier years is the same and the discrepancy only concerns more recent years; in the case of other countries, there is a discrepancy in all years.

If there are several international sources, these data are very often different. It is very difficult to decide which source is best. A reasonable solution seems to be to take the data that is closest to the source, i.e. from the country statistical office. But this might mean duplication of requests.

3.2 Timeliness

Getting data from international organizations always means a considerable time delay. Collecting and checking the data, making corrections and adjustments to increase comparability, transferring the data to databases takes time. It took us about 7-8 months from sending out the questionnaires until having the data available in the Gender Statistics Database. The delay can be about the same in other organizations depending on how detailed is the data collection, how many resources can be used, etc.

E.g., when looking for education data, a good source is the OECD education online database. However, the data for 2000 will become available only in October 2002

3.3 Coverage

The ECE Gender Statistics Database is planned to include data on all ECE Member Countries, i.e. all European countries, the CIS countries, Canada and the United States (55 countries). Several available databases from other international organisations include only a limited number of these[3]. The UNSD databases cover all countries but the latest available data is often for 1998-99. The country coverage differs a lot for different topics.

Another aspect is the necessary level of detail. Some series might be available at an aggregated level but not necessarily with the requested level of detail (age, marital status, ISCED, ISCO, etc.; even the gender breakdown is very often not available but this situation is improving). Getting data for earlier years (1980-1990) can also be a problem.

3.4 Metadata

One of the important aspects with developing a Gender Statistics Database, as stated in the Report of the Task Force May 2001, was to provide well-documented data. Therefore, all data in the questionnaires has to be accompanied by a relevant description of the definition, source, coverage, special remarks, etc. When retrieving data from other international organisations' databases, the metadata is often not available at all or not at the required level of detail.

3.5 Resources

Retrieving data from external databases requires considerable resources because it is difficult to automate. Each database has a different user interface and different formats for

[3] OECD provides data for Austria, Belgium, Canada, Czech Republic, Denmark, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Italy, Luxembourg, Netherlands, Norway, Poland, Portugal, Spain, Sweden, Switzerland, Turkey, United Kingdom and United States (24 countries).
Eurostat databases provide data for the 15 EU Member Countries + occasionally the 12 accession countries and Norway, Switzerland, United States (30 countries).

downloading. We have developed an automatic procedure for data input from the harmonised Excel sheets of the gender questionnaire – therefore the most efficient way for data transfer to the database is first to convert the data into the questionnaire in Excel format. Also, our data checking procedures have been developed for the harmonised Excel layout. It is possible to develop an automatic procedure for data transfer from other database formats but it is only worthwhile if it can be used regularly over a long period of time. In addition, this also requires stability of the external database from which we retrieve the data.

Not all data of international organizations are available through databases that could be accessed online. In this case, we should send our questionnaire to the other organization and ask them to fill it in for all available countries. Or, we can ask the other organization to send us the data in their format, and try to extract the necessary subset ourselves. In both cases it means considerable work for one or the other side.

Another possible source for data are the web pages of the country statistical offices. Even more resources are needed to find the data, evaluate its relevance and download it.

3.6 The cross-cutting character of data

Using data from existing sources is further complicated because of the cross-cutting character of the gender data request - the sources are numerous and each of them requires a specific approach. Most existing joint data collections focus on a specific subject area topic such as education, labour force, demography, etc., and for most areas, we only need a sub-set of the overall set of data collected. In order to organise or participate in a joint data collection, we would need to negotiate with several organisations (and their subdivisions) to get an agreement concerning the possible bits and pieces we need out of their much bigger set of data collected.

3.7 Areas not (yet) covered by official statistics

There are very little data available in the official statistical system and international organizations for some areas, e.g. dealing with public life and decision making, and crime and violence (especially concerning victims). However, these data are very important for monitoring the situation of women and men. The gender data collection might provide an incentive for statistical offices to look for the sources of these data. Often there are some agencies or researchers exploring these topics and it is much easier to get to know about these initiatives at country level.

The above arguments are not meant to say that joint data collection by international organizations and reducing the burden of countries is not a good way to operate. It is just to show that in our specific cross-cutting case it does not work as smoothly as it could be expected, and often we have to consider that asking countries directly (although the data might be available somewhere in some format) is more viable than using existing sources.

One solution is the development of standard sets of indicators jointly by international organizations so that the data requirements to countries would be the same, and the data could then either be shared by international organizations, or countries could submit the same data to several organizations. There are several initiatives to develop such standard sets as a follow-up to major global conferences (our project being one if them), which result in a number different “standard sets”. However, the standard sets developed at the global level are often not completely convenient for our region, so our regional indicators often require more detailed information than available at global level.

4. The current questionnaire and the current list of common indicators

4.1 The gender questionnaire

The questionnaire for the 2002 data collection exercise was developed in order to be able to present the core, supporting and background indicators of the recommended common gender indicators. For these three types of common indicators, the Task Force in May 2001 concluded that both methodology as well as data availability should be well developed enough to allow compilation of these indicators for most of the countries of the ECE region, or to encourage countries to consider collection and compilation of data needed.

Correspondingly, the results of the data collection exercise 2002 are very satisfactory for many countries and indicators. Expectedly, the data situation gets poorer the further we go back in time (1980, 1990), and there are country-specific variations in data availability depending on a specific survey being regularly carried out or not (e.g. Labour Force Survey, Victimisation Survey, etc.).

However, a few tables turned out not to be feasible at all, some tables need improvements in terms of methodological clarity and lay-out, and for some tables, a reduction of the requested level of detail can be suggested.

Regardless from which sources data will be provided in the future, the gender questionnaire in a *harmonised Excel format* will remain our starting point for data collection. The reasons for this are the following:

- Automatic data loading from Excel to the Gender Statistics Database;
- Data checking procedures developed for existing format;
- Consistent and harmonised country-wise overview on data;
- Common reference platform for communication between ECE and each country.

The ECE secretariat will revise the existing questionnaire based on the experiences of the 2002 data collection and prepare a version that can be used for reference years up to 2005. I.e. for each future data collection, countries will receive the questionnaire in Excel containing the same tables; in these tables, all data provided in earlier data collection or filled in by the ECE secretariat will be send to the countries for adding of new data, filling gaps in series or revisions.

As regards the provision of additional indicators, it is suggested to postpone this to the next Task Force Meeting.

The following revisions of the current questionnaires are planned (see also Annex), and the Task Force is invited to adopt these changes. Tables not listed here remain unchanged.

4.2 Chapter 1 - Population

Table 1.1 Mid-year de facto population by age, number
This table provides reference data for several indicators (Population - core 1, core 2; Families and households - core 3; Health – core 5, background 1). Main purpose is to collect data on population by 5-year age groups.

However, in order to calculate infant mortality rates and child mortality rates, we further split the age group 0-4 into 0 and 1-4. Since this caused problems and data gaps (especially if 5-year age group data is available for the mid-year-concept but single age groups only for 1st of January-concept), we suggest asking for age groups 0 and 1-4 in an extra table.

Table 1.2 Mid-year de facto population by age and by marital status, number
This table is needed to provide Population - core indicator 1 – population 18+ by marital status (18-29, 30-59, 60+). We currently ask for a very detailed age breakdown but suggest to reduce the age groups to the ones needed for core indicator one. If 18+ is not possible, the alternative can be 15+ or 20+.

Table 1.4 Mid-year de facto population by age and by urban/rural, number
This table is for Population - background indicator 1 – population by urban/rural (0-14, 16-64, 65+). We suggest reducing the age groups in the table to the recommended ones only.

Table 1.5 Refugees by age, number
This table is for Population - background indicator 2 – refugees and displaced persons (0-14, 15-64, 65+). Data collection showed big gaps in data in general, and almost no country was able to provide the breakdown by age. It is suggested to drop the breakdown by age.

4.3 Chapter 2 – Families and households

Table 2.6 Number of persons using contraception, number
Families and households – background indicator 3 – Contraceptive use in ages 15-54: We received only very few data, and the existing contraception use surveys vary a lot from country to country. Also, the purpose of the indicator is not very clear – is it to measure birth control / access to contraception, or rather to look at health issues. In the latter case, a different table would be needed. The Task Force is invited to make a recommendation for this indicator.

Table 2.7 Total number of one parent families, number, and Children living in one parent families, by age of child, number
This table was designed to collect data for Families and households – background indicator 2 – Children living in lone parent households by sex of parent and *age of child*. Almost no country provided data "by age of child". It is therefore suggested to drop this part of the table, and to reduce Families & households / background indicator 2 to "Children living in lone parent households by sex of parent".

Table 2.8 One person households by age, number
Families and households - supporting indicator 3 – One-person households over 65 years. The current age breakdown is too detailed, we suggest to reduce to the recommended 0-30, 30-65, 65+.

4.4 Chapter 3 – Work and the economy

Table 3.10 Average annual earnings (full time, full year) by level of education completed, national currency
Work and the economy - core indicator 6 and supporting 5 –Women's average earnings (total and by level of education). The data situation is not too good for this indicator, and we received some criticism on the definition of the indicator. The Task Force is therefore invited to provide suggestions for improving the indicator.

4.5 Chapter 4 – Education and communication

Table 4.7 Total expenditure on education as % of GDP, %
The recommended indicator was actually "Other indicators 2 - public education expenditure, % of government spending". Considering the total expenditure on education as a % of GDP to be the more useful definition and in order to be in line with the Trends publication, we suggest to change the indicator.

4.6 Chapter 6 - Health

Table 6.6 HIV positive persons
For this table, it was not sufficiently clear that we want data on people living with HIV and not on new cases, so additional information to clarify this table will be added.

Table 6.7 Persons overweight/underweight by age group
For this table, data received showed that it is more likely to receive the information in percent than in actual numbers, the unit in the table will therefore be changed from number to percent.

4.7 Chapter 7 – Crime and violence

Table 7.10 Convictions
This table is designed to gather information for Crime and violence - supporting 2 - Convictions for theft; Supporting 3 - Convictions for assaults; Supporting 4 - Convictions for drug crimes. In designing the questionnaires we discovered the methodological need to break assaults further down into assaults and serious assaults. As regards the table layout, it needs to be improved – the current layout does not show clearly enough that the total refers to all convictions and not only to the sum of the 4 types of crimes. This will be improved.

5. Impact on Common Gender Indicators

Most of the above suggested changes in the tables of the gender statistics questionnaire do not have any impact on the common gender indicators (core, supporting, background) as recommended by the Task Force in May 2001. Only changes in the following indicators are to be discussed and approved by the Task Force:

Population - Background 2 – refugees and displaced persons (0-14, 15-64, 65+):

- It is suggested to drop the breakdown by age.

Families and households - Background 3 – Contraceptive use in ages 15-54:

- It is suggested to drop the indicator. Furthermore, a future indicator with regards to health issues in sexual relations could be thought of.

Families and households - Background 2 – Children living in lone parent households by sex of parent and age of child:

- It is suggested to drop the breakdown "by age of child".

Work and the economy - Core 6 – Women's average earnings and Supporting 5 – Women's average earnings by level of education:

- It is suggested to find an improved definition / indicator.

Other indicators - Other 2 - public education expenditure, % of government spending:

- It is suggested to change to “Total expenditure on education as % of GDP”.

6. Data presentation on the website

Once the UNECE Gender Statistics Database is ready for publication (expected September/October 2002), the following two options of data access will be made available on the website:

1) On-line access to the Gender Statistics Database

For on-line access, a user-friendly interface will be provided. This will allow users to extract data tailor-made according to their needs. The database contains both the “pure” statistical series, i.e. the detailed data as collected through the gender questionnaire. Furthermore, derived indicators such as the common gender indicators as requested by the Task Force are available. In principle, each indicator will be available in number, sex rate and percentage distribution.

2) Ready made presentation of common gender indicators in Excel

The Task Force recommended a simple presentation of the common gender indicators for users not so familiar with statistics. For these, we will present for each of the common gender indicators a ready-made Excel file. The information shown will be limited to mainly percentages / sex rates and only same information in absolute figures (allowing for some evaluation of the magnitude). Examples of this presentation are shown in the Annex (will be available at the Task Force Meeting).