The Impact of Microcomputer Technologies on Dissemination of Integrated Social, Demographic and Related Statistics

by Robert Johnston[1] and
Graham Templeman

Introduction - The Relation Between Integration and Dissemination

The United Nations has been concerned with general issues in measuring development and levels of living and related social, economic and environment conditions since the beginning of the organization, pursuant to the promotion of "higher standards of living, full employment, and conditions of economic and social progress and development" as set forth in the Charter of the United Nations (Article 55). Over the past five years this work has received new stimulus in the United Nations Statistical Office from the great interest at national and international levels in statistics and indicators on women and special population groups such as youth, elderly and disabled persons, and from new interest at national and international levels in the compilation and use of social statistics and indicators to design and implement social policies, to monitor the achievement of social objectives and to monitor the social impact of economic adjustment policies.

In response to these emerging interests and priorities, the Statistical Office has undertaken a substantial reorganization of its methods of compilation, presentation and organization of social and related demographic and economic statistics and indicators. This work follows up the development in the 1970s of the United Nations framework for integration of social, demographic and related statistics (FSDS), and the preliminary guidelines on social indicators published by the United Nations in 1978.[2] It has been particularly oriented to compilation and dissemination of integrated indicators drawing on a wide range of sources and aimed at non-specialists, rather than compilation of primary data, which are detailed, technical, and specialized by field. This work has been greatly facilitated by the rapid development and now nearly universal availability of highly standardized microcomputer hardware and software technologies for statistical work.

As described in the United Nations Handbook on Social Indicators[3] , the development of integrated social statistics and indicators is a wide-ranging and multifaceted process which aims to bring together basic statistics from many different fields and data collection programmes and recompile them for many different purposes. The Handbook provides a basic core of structure, concepts and methods for use in this process and thereby promote the compilation and dissemination of social statistics and indicators to better meet a wide variety of user needs through more effective integration and use of the basic data.

Unfortunately, the cost and methodological difficulties of bringing together social statistics and indicators from many disparate and often intractable primary and secondary sources have held back work on social indicators in many countries and internationally. Even where statistical services have built up a considerable volume of basic data, this has by no means ensured the ready availability to users of indicators relevant to their specific purposes and concerns, including policy issues. For such purposes, close collaboration between users and producers of indicators, and detailed attention to data requirements for indicators at the stage of designing data collection programmes, and co-ordination within a framework such as FSDS are needed. Thus much of the work on social indicators in the 1970s and early 1980s was concerned with precise identification of user interests and requirements and their translation into well-structured statistical methods and concepts.

Co-Ordination and Integration of Social and Related Classifications for Dissemination of Integrated Statistics and Indicators

One of the technical features which works on FSDS and has been stressed from its inception has been the development and harmonization of social and related classifications for integration and for indicators. However, the development and harmonization of social and related classifications is a complex process for many of the same reasons that social statistics and indicators present statistical offices with so many difficult problems of organization and methodology. In general, the subject matter is extremely heterogeneous and the relevant statistics come from a very wide range of sources, each with established traditions, procedures and objectives and often administered more or less independently of the central statistical service. These circumstances are similar at national and international levels.

The initial development of FSDS in Towards a System of Demographic and Social Statistics and in the preliminary guidelines on social indicators established a basic subject-matter and classifications framework for the further development of classifications and indicators. These early reports served to clarify what classifications were relevant to integrated social statistics and in what ways. With the current interest in improved multidisciplinary compilation and dissemination of social statistics and indicators and given the new technical possibilities of microcomputers for bringing together data in microcomputer data bases, the importance of harmonized classifications emerges all the more clearly.

A basic principle of work in the Statistical Office in this area continues to be the importance of the close linkage between so-called basic statistics and statistics for integration, and for indicators that are, basically for general disseminations. Thus it has never been suggested that new classifications should be developed for integration or for indicators, which would in any case be a technically and organizationally impossible task, given the degree or decentralization of responsibility for statistical classifications at national and international levels and the large number of competing interests and technical problems that must always be delicately balanced in preparing any kind of recommendation on classifications. What the Statistical Office undertook in the preliminary guidelines on social indicators and has now been made much more explicit in the Handbook of Social Indicators, is to recommend abstracting from existing classifications shorter forms which are needed for integration and for indicators. As the draft Handbook states, once the fields and topics for indicators have been outlined in an indicators programme at the national or international level, basic statistical classifications for use in indicators should be developed. These must, of necessity, be based on the classifications used in the basic data but for purposes of indicator compilation these source classifications often require careful adaptation.

The process of adaptation should be undertaken with three objectives in mind:

(a) Meeting specific indicator requirements;

(b) Abbreviating classifications as much as possible to simplify compilation and presentation of indicators;

(c) Devising classifications into which data from a variety of sources often using differing classifications or variants of classifications, can be fitted as consistently as possible;

(d) Identifying population groups of special policy concerns.

All of the classifications referred to in the illustrative series and basic data tables for indicators in the Handbook are listed in the table below which also shows the fields in which they are used. Sixteen of these are considered basic classifications in the Handbook. Five of these concern demographic and social characteristics (sex and age group, national or ethnic group, household size and composition, household headship and level of education); three are geographical (urban and rural areas, cities and urban agglomerations, and geographical regions); four concern activity characteristics (occupation, status in employment, socioeconomic group and time-use); and four are classifications from economic statistics (percentage distributions of household income and consumption, kind of economic activity (industry), functions of government and institutional sector).

These basic classifications can be used to provide a firm foundation for the development of indicators in all of the fields covered by the Handbook. They were selected for discussion as basic classifications on the basis of (a) their substantive importance for indicators, usually in more than one field and drawing on multiple data sources, (b) the extent of their importance and use for indicators in national and international experience, and (c) the relative detail and complexity required in their use for compiling statistics for indicators. All but one of the basic classifications are shown and discussed in the illustrative formats for basic data tables of the Handbook, drawing on the relevant international recommendations. The exception is classification by national or ethnic groups. In this case, national and experience and circumstances are so diverse that no international recommendations are feasible and even an illustrative classification could not serve any useful purpose.

Principles of Integration Applied to Dissemination of Statistics and Indicators on Women and Special Population Groups

Interest in the development of statistics and indicators on women and other population groups that are considered to be of special relevance for policy planning has given considerable impetus to a range of activities concerned with statistics and indicators on these groups. The principal groups on which work has been concentrated in the United Nations Statistical Office are women (beginning with the World Conference of the International Women's Year in 1975), disabled persons (beginning with the International Year of Disabled Persons in 1982), youth (in connection with International Youth Year in 1985) and children. There has been interest in the development of statistics and indicators on the elderly (in connection with the World Assembly on Aging in 1982 and the International Plan of Action).

In international compilation and dissemination of indicators on women, for example, a substantial quantity of data is being routinely collected in international statistical services and supplemented, in many cases, with standardized international estimates and projections. The rapid spread of microcomputers and the ease of use of spreadsheet techniques have now made it feasible to compile these data in one source, using the FSDS framework, disseminate them to users cheaply and quickly on diskettes, and prepare user-oriented software and documentation for reference, analysis, table-generation and similar uses. A special project with these objectives was established in the Statistical Office in 1984, and this work was basically completed in 1987.

The United Nations Women's Indicators and Statistics Data Base (WISTAT) consists of 72 microcomputer spreadsheet files (currently using Lotus 1-2-3) ranging in size from approximately 20kb to 150kb and totalling about 12mb. WISTAT is available from the Statistical Office on 22 microcomputer diskettes complete for 178 countries and areas or for specific regions, using the forms provided in the printed user's guide (currently available, in part, as a Statistical Office working paper). WISTAT will be fully documented in the user's guide, to be issued in final form as a sales publication of the United Nations. A listing of statistical series and topics in this data base is given in the annex below.

Using quite different underlying technical methods of organization and compilation but identical microcomputer hardware and software an international disability statistics data base was also completed by the Statistical Office in 1987, comprising detailed statistics on disabled persons from censuses and surveys in 55 countries and areas between 1975 and 1985. Like the women's data base, the Disability Statistics Data Base (DISTAT) is disseminated on diskettes. It consists of 34 microcomputer spreadsheet files ranging in size from about 7kb to 314kb and totalling about 3.3 mb. The files are described in detail in United Nations Disability Statistics Data Base, 1975-1986: Technical Manual.[4] <footnote text> The complete data base is available from the Statistical Office on 12 microcomputer diskettes using the forms provided with the Technical Manual. Version 1 of the data base (as of 31 December 1987) contains (a) information on sources and availability of statistics on disability for 95 countries or areas for various years between 1960 and 1986, and (b) detailed statistics on disabled persons from national censuses, surveys and other data sources from 55 of those countries or areas for the period 1975-1986.

Finally, the basic strategy and framework for organizing social statistics for social indicators, as set out in the Handbook on Social Indicators, were adopted by the Statistical Office for preparation of the Compendium of Statistics and Indicators on the Situation of Women - 1986 and the Compendium of Social Statistics and Indicators - 1986.[5] That is, highly simplified basic data were compiled from primary international sources into microcomputer spreadsheets. Once in the spreadsheets, new series and indicators could be calculated and data transferred within and among spreadsheets with great flexibility and minimal time and effort and, once final table formats were agreed, they could be tested and then generated in final form very quickly. On this basis, series and classifications such as those given in the Handbook have been prepared for these two compendiums, with the possibility of recalculating percentages, rates, ratios, distributions, reaggregations and the like and of juxtaposing series from different sources that may be of interest almost at will. A short, preliminary version of the social compendium was prepared using these techniques for the United Nations Interregional Consultation on Social Welfare Policies and Programmes held in September 1987[6] and generated considerable interest among delegates with no special statistical background.

Overall, it appears that microcomputer hardware and software for spreadsheets, data bases and analysis are at the leading edge of basic changes in the development of social statistics and indicators at national and international levels. The effects are now beginning to be seen on a wide scale and at the same time the technologies are advancing and spreading so rapidly throughout the world that the direction and full implications of these changes are still not completely understood or fully appreciated.

The Potential Role in Dissemination of Spreadsheets and Data Bases on Microcomputers.

Spreadsheets and their users

Spreadsheets are used extensively by people interested in statistics, and are a useful tool in most cases.

Spreadsheets are characterized by a row/column cellular approach to data in which each cell may be (typically) a number, a character string, or a numeric or logical function of the values in other cells. Cells are named by a column/row coordinate system. For example:

	A	B	C	D ...	AA	AB	AC...
1							
2							
3							
.							
.							
.							
11							
12							
.							
.							
.							

Spreadsheets are capable of displaying or printing the cells in row/column format, and manipulating the cells individually or by rows, columns, or groups of these. They are, effectively, the "cell processing" equivalent of a word processor. A word processor imposes minimal structure on its atomic units (words) except to organize them within given boundaries such as paragraphs and margins. A spreadsheet, on the other hand, maintains a positional relationship between its atomic units (cells) which is capable of being displayed as a two-dimensional row/column table.

The facilities for manipulation offered by spreadsheets and word processors have many parallels, such as "cut and paste", insertion and deletion, and search and replace. Word processors have some facilities peculiar to themselves, such as reformatting between redefined margins, text flow from line to line, and so on. Spreadsheets also have some peculiar features such as row and column transposition, default cursor movement by row or column first, and so on.

It is not surprising, therefore, that people dealing with statistical tables are drawn to spreadsheets in much the same way that typists and writers are drawn to word processing programs, by the facility for direct interactive visually-based control.

Beyond these "cell-processing" capabilities, some spreadsheets offer what they call "database capability". (See, for example, Lotus 1-2-3 Tutorial Manual, v.2.01, p. 5.1)

Spreadsheets "database capability" comes from an analogy with some aspects of relational database theory. This type of data organization can be simulated using a spreadsheet, by using the relation name as the spreadsheet name, and by treating rows as records or instances of the relation and columns as fields or attributes, as long as the user sets up the data in the format of a single relation.

```
1975/1980
persons prosecuted/convicted
male/female
total/adult/juvenile


                                          1
1975/1980                     1                  1
prosec/conv         1          1       1        1
male/female    1      1   1    1     1      1    1   1
tot/adult/juv 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
```

When the user sets up the data in "flat file" format, i.e. with one column representing the key field, all columns with unique names, and so on, the row and column manipulations available to the spreadsheet user parallel some of the simpler facilities available to a relational database user. For example, sorting rows by the key field, searching a column for a particular value or for values falling within a range.

In some spreadsheets such as Lotus-1-2-3 it is possible to use foreign key values located by a search such as a range search to extract rows from another spreadsheet. In this way the "cut and paste" spreadsheet operations simulate the linking of different relations. When such operations are expressed as a macro, i.e. a named sequence of operations which can be invoked by its name, the result can be quite efficient for some purposes. It is important to realize, however, that with spreadsheets there is only a limited connection between the column name and the data. The true column name is its coordinate. If a column is removed or added to a spreadsheet, thus pushing other columns to new coordinates, any operation which refers to coordinates will have to be redefined. This is not a problem for perfectly stable data. Spreadsheet statistical tables in presentation format. Tables with hierarchies of field names simply do not lend themselves to database-style manipulations other than those which can be simulated by "cut and paste" cell processing.

A second consequence of the affinity of spreadsheets for human readability is that spreadsheet designers tend to follow the make use of the horizontal left to right direction of reading. This results in the expansion of more variables horizontally than vertically, creating hierarchies of column headings. In the WISTAT data collection, for instance, almost every spreadsheet uses one row per country or area, with column hierarchies up to four levels deep. For example see table below:

Some simple arithmetic shows that this hierarchy generates 24 columns and that the "adult" heading will appear 8 times. An interrogation asking for listing of those countries for which, in 1975, the number of adults prosecuted exceeded a certain number, would require two of the twenty four columns to be identified, then summed, then compared to the given reference number. The first step, identifying the columns, can only be done by a person or by software capable of dealing with hierarchical field structures. For similar reasons, the task of sorting this particular data by number of persons prosecuted within age-group or sex would be daunting.

In the WISTAT data collection the tendency to horizontal expansion of variables has an even more direct problem. Many indirect users want to extract data for some or all subject areas for one country or region. Country or region is typically the only vertically expanded variable, i.e. it identifies the rows. When the relevant row from each spreadsheet is identified and extracted it is not possible to combine these rows into a new spreadsheet or an integrated listing because each row has a different column structure.

The result of extracting all data for one country is approximately 70 one-line spreadsheets. If titles and column headings are extracted as well, then we have approximately 70 spreadsheets each with a dozen or so rows. Thus the most appropriate application for spreadsheets is the manipulation of cells and blocks of cells after the table has been formed in the desired way by tabulation or data management software.

Database theory and statistical data

When people talk about a "database" they are usually referring to the type of organization of data which allows the retrieval and display of items of subgroups of the data by name or description rather than by reference to where or how they are stored. Strictly speaking the software which carries out access and retrieval is an essential component of the database.

A typical database query would be "list the names and addresses of all respondents who are hostile to interviewers". Or for aggregated data: "list the countries for which female constitute more than 50 per cent of the population". Queries for an aggregate statistical database would also be expected to extract and format tables, either for printing or for export to spreadsheet files. The particular way of expressing the question depends on the query language defined by the database software in use. One type of database organization is known as "relational". This approach is very popular, having received excellent coverage in computer magazines, and many data management packages claim to be relational. The most well-known aspect of relational data management is that its fundamental data structure is a "table". This makes it attractive to people who like to think of their data in tabular format.

The relational "table" is, however, very different from a statistical table, and in many ways the relational approach is not suited to statistical data.

In fact the fundamental data structure is a relation. A relation is like a pattern. Any type of entity for which data is to be held is given a relation or pattern, which is a list of named place-marker. For example, if we are to hold data about staff we would define a staff relation specifying the items of data to be held:

staff (id-number, name, department, position, date-hired, type of contract, etc.)

One of these items, or a group of them, must function as a unique identifier, or "key". It is shown underlined here.

The actual data consists of "instances" of the relation, e.g.:

(658889,Templeman,DIESA,,6 JUN 88,consultant,etc.)

When a set of instances is listed it looks like a table. Such a set may be stored in traditional computer terms as a file with a key field and a simple set of fields, sometimes known as a flat file.

When data is to be stored about entities which have a specific relationship to each other, relational theory specifies the mechanism for relating them. For example, there may also be a Department relation, e.g.:

Dept (Dept-id,dept-name,location,name-of-head,phone-of-head)

There is a specific relationship between staff and departments. This is expressed by including the key of the department relation as an ordinary field of the staff relation. It is known as a "foreign key":

staff (id-number,name,dept-id,location,...)

For data to function relationally it has to be set up specifically to do so. Data about one entity should not be embedded within the relation for another entity, multiple field values are not permitted, hierarchical field structures are not permitted, and so on. When data comes naturally with such impurities of structure it must first be converted into a logically equivalent set of relations of the acceptable type. This conversion process is known as "normalization".❑

[1] Presented at the IFDO/IASSIST 89 Conference held in Jerusalem, Israel, May 15-18, 1989.

[1b] The authors are, respectively, Chief of the Social and Housing Statistics Section of the United Nations Statistical Office and consultant to the Statistical Office on the United Nations Women's Indicators and Statistics Data Base (WISTAT). The views expressed are those of the authors' and not necessarily of the United Nations.

[2] See Towards a System of Social and Demographic Statistics, Studies in the Integration of Social Statistics: A Technical Report, Improving Social Statistics in Developing Countries: Technical Report, Studies in Methods, Series F, Nos. 18, 24 and 25 (United Nations publications, Sales Nos. E.74.XVII.8, E.79.XVII.4, E.79.XVII.12), and Social Indicators: Preliminary Guidelines and Illustrative Series, Statistical Papers, Series M, No. 63 (United Nations publications, Sales No. E.78.XVII.8).

[3] Series F, No. 49 (United Nations publications, in press).

[4] Series Y, No. 3 (United Nations publications, Sales No. E.88.XVII.12).

[5] Series K, No. 5 (United Nations publications, in press), and Series K, No. 6 (United Nations publications, in press).

[6] "Compilation of selected statistics and indicators on social policy and development issues" (E/CONF.80/CRP.1).