ELECTRONIC DISSEMINATION OF STATISTICAL DATA

Alex van Buitenen, Anco Hundepool, Wil de Jong and Aad van de Wetering
Netherlands Central Bureau of Statistics, The Netherlands

1. Introduction

Paper is the traditional dissemination medium for statistical information. Large amounts of books are published every year to make the results of the statistical work accessible to the researchers, policy makers and other interested people. Inside the statistical offices even larger amounts of tables are produced in order to answer possible questions from the public. This way of producing statistical information is not only a waste of valuable natural resources (trees), it has become quite inefficient now that (micro)-computers have become available everywhere and can deal with this problem in a much better way.

Users of statistical information who have computers at their disposal want to use this information on their computers without having to key in the figures from the books and publications of the statistical offices. They want the information and the necessary meta-information (the meaning of the figures) in a machine readable way. Also the large amounts of tables aimed at answering possible questions from the public are nowadays being replaced by information systems capable of producing ad-hoc information when it is needed.

The Netherlands Central Bureau of Statistics has developed software to disseminate statistical information in an electronic way. STATview is now available to publish large amounts of aggregated data on diskettes and has been extended into an online information system. For more detailed (individual) data the tabulation package ABACUS can be used to produce high quality tables at high speed.

2. Individual data

As researchers have now much more facilities to do their own analysis due to the widely available micro-computers and analysis software, there is a growing demand for the release of more detailed (if not individual) data. Of course, there is a great risk of disclosure in releasing files with individual information. At the CBS the ARGUS project (De Jong, 1992) is being carried out to develop software to help identifying the possible disclosure risks. A first prototype of ARGUS is now available and is tested. The statistical offices should be aware of the risks when individual information is identified.

However if the statistical office has investigated the possible disclosure risks and is willing to release files with individual information to researchers, they are also obliged to supply the necessary meta- information. At least the researchers need to understand the meaning of the different variables, what the coding of the categories stands for, etc.

We have chosen to publish these files in an ASCII format. Not because we think that everyone wants to use this ASCII file, but it is a very general format that can be read by various programs. Besides this ASCII file with the data we can supply a system which is capable of converting the meta information into a format required by the user. This format can be a plain record description but also a setup for transforming the data into the format of the various analysis packages like SPSS, SAS and Stata or the tabulation package ABACUS.

3. Aggregated data

Aggregated data fall apart into two categories: first, information that is composed of a collection of (small) tables as in the traditional publications, and second, larger databases which contain large amounts of data like municipal data or large collections of time series. In the first case the user is at a certain moment only interested in a few of the available tables, while in the second case the user wants to select a specific part of the database/ matrix.

The program STATview has been developed to serve as a shell around both the data and the meta-data. Especially the coherence between the data and the meta-data is a very important aspect of publishing statistical information. Data without the relevant meta-data is meaningless. It is a responsibility of the publishing statistical office to take care that the user can make optimal use of the available statistical information.

The first version of STATview has been developed to serve as a shell around data-sets, which are in the form of large more-dimensional data matrices, like the municipal database. STATview allows the user to select a subset of the large matrix by presenting the information in a hierarchical structure. During the selection process the user gets all the relevant meta-information on the screen to facilitate the selection process.

After having made the selection from the information, the user can specify the format in which he wants the selected information to be presented. At this moment we support the following formats:

- table in ASCII format
- worksheet for Lotus 1-2-3 and other spreadsheet packages
- dBase file, also suitable for Paradox etc
- ASCII file
- setups to read the ASCII file into other packages, like SPSS, Stata, SAS, Abacus and Manipula.

Abacus and Manipula originate from the Netherlands Central Bureau of Statistics and are meant for tabulation and file manipulation.

This version of STATview is used successfully for many publications of the NCBS. The second version of STATview allows for more data formats. Not only large data matrices but also publications which consist of a collection of smaller tables can be disseminated electronically with STATview. The selection process is quite similar to the matrix version. The information is presented in a hierarchical structure. Via the hierarchical structure the user selects the tables he is interested in. As an alternative the user can also use a thesaurus.

The STATview thesaurus is based on trigrams. The trigram technique is also used in the new coding module of Blaise (version 2.5) (see Roessingh, 1993). To each item (basic tables or columns of the data matrix) in the publication one or more keywords can be attached. Each keyword in the thesaurus is divided into successive three-letter groups. For each keyword the set of three-letter groups is stored. If the user asks a word to the thesaurus, this word is also divided into three-letter words and the number of matches in the data base of trigrams is calculated. If a certain level of matches is exceeded the system presents the item as a possible hit. This trigram technique makes it possible to search for keywords even if some misspellings have been made.

Also new in version 2 of STATview is the presentation of the selected table. In version 1 the selected information is presented on the screen without any possibility to manipulate the table. In the new version a table manipulation program has been added. It is now possible to perform some manipulations (add, subtract, multiply, divide, percentage) on the rows and the columns of the table and to influence the layout of the table. If the table must be printed the table can be prepared for printing by adding page breaks even at the places of the user's choice. This table presentation program adds much flexibility to the STATview package.

This table manipulation program will also be used in new versions of the tabulation package Abacus.

4. Publication media

STATview is used now to publish information on diskette. Due to the size of the publications made by the NCBS diskettes are a suitable medium.

Nevertheless other computer media are becoming available. It is to be expected that the CD-ROM is going to play an important role in publishing statistical information. The CD-ROM has a very large capacity (about 500 megabytes) and is also very reliable. Once the data have been written on CD-ROM they are very safe and secure. Although CD-ROM is a slow medium compared with a hard disk, it will be (and is already) used for publishing statistical information. At this moment STATview has not been used for CD-ROM due to the lack of very large databases, but it is to be expected that the use of STATview for CD-ROM publications will not be a difficult extension.

The disseminating of statistical information electronically on diskette has become an important part of the publication effort of the NCBS. About one third of the revenues of NCBS publications are realized by electronic publications. Besides the demand for STATview diskettes there is a growing demand for a system where the users can get the most recent information as soon as it becomes available. To meet these demands we have extended STATview into an online information system.

5. STATline

STATline is the online information system that has been build as an extension to the STATview package. This has the advantage that the users will be working with the same user interface as in STATview. The main difference is that the databases are no longer installed at their own computer but are accessed by a modem on the STATline computer at the NCBS. Of course this computer will have no direct connections with the NCBS networks for security reasons.

The STATview software has been divided in a front-end that runs on the computers of the users and a back-end that runs on the central computer. To the front-end belongs the software needed to make the selections from the database. The meta-information needed during the selection process is normally retrieved via the back-end from the central computer. However if the user is accessing the same database regularly it will be possible to install the meta-information at his own computer. The advantage is that the selection process will be quicker as there is no need to contact the central computer.

As soon as the selection of the requested information has been made the actual retrieval is made by the back-end and sent to the user. With the same table manipulation program as in STATview the information is then presented to the user. This gives the user the same possibilities to process the selected information.

Building the STATline system with the STATview software has resulted in a situation where the user will be working with exactly the same user interface whether the publications are available at his own computer or are accessed by modem at the central computer.

STATline makes it possible to release even the most recent data electronically to the public. Besides these most recent figures the users can now access all the available STATview databases and look for the information without the need to buy and install all the publications on his own computer.

It is our intention that the database in the online STATview system will be playing a central role in the publication of the statistical information. Finally it should be so that the data is published first in the online database, which will than serve as a starting point to produce all kind of publications, like still some paper publications, press reports and tailor-made faxes or e-mail messages to those who have subscribed for it.

6. Conclusion

It is very important for the statistical offices to be aware of the changing demand for information from the public. The computer will play (and is already playing) an important role in the dissemination of statistical information. The use of adequate software to disseminate the statistical information facilitates the use of this information. The use of the statistical information is the main if not only reason why a statistical office exists. The first reactions from the public with STATview publications are very encouraging. Although STATview was developed for the dissemination of CBS publications, the software is also available for other parties to make their own publications.

Bibliography

Jong W.A.M. de, Willenborg L.C.R.J., 1992, "Argus: an integrated system for data protection", *Proceedings of the International Seminar on Statistical Confidentiality*, Eurostat/ISI.

Roessing M.J., Bethlehem J.G., 1993, "Trigram coding in the Family Expenditure Survey of the CBS", *Essays on Blaise, 1993,* OPCS London.