A MULTIPURPOSE SHARED DATA BANK - CANSIM

P.N. Triandafillou
Director, CANSIM Division
Marketing Services Field
Statistics Canada
Ottawa, Ontario
K1A 0Z8

ABSTRACT

The advent of computerization and the emerging sophistication of users have led to demands for more data which are more disaggregated, but integrated in machine-readable form as the basic input into computers for analysis. We believe that the greatest benefit to CANSIM will center around the data it contains and the manner in which data are "packaged", rather than the sophistication of either its hardware or software. Future emphasis will be placed not only on new relevant series but on the organization of data for retrieval and comprehensive analysis.

A MULTIPURPOSE SHARED DATA BANK - CANSIM

A quick skimming of the Statistics Canada catalogue will demonstrate the enormity of the problem facing a researcher who wants a sample of information on one area of economics, or one industry. It is quite possible that such a person would have to sift the relevant information from about 20 different publications, and perhaps numerous issues of any given publication. Worse still, he may have to make several phone calls to different offices within Statistics Canada merely to establish what information is available. So far, the user has only seen the tip of the iceberg.

The influx of new graduates with "computer" training in user institutions is accelerating rapidly, and the volume of data that can be brought to bear in analytical and decision-making processes is increasing tremendously. These two facts alone have naturally led to the more extensive use of computers. In addition, a striking feature of early computer technology was arithmetic speed. A number of statistical and mathematical techniques, well-developed in theory but simply too time-consuming to use by manual methods, are now possible or feasible. In summary then, the advent of the "computer" and the emerging sophistication of users have led to demands for more data which are more disaggregated but integrated in machine-readable form as the basic input into computers for analysis. While the "ideal" computer-based statistical information system is not a reality, it may be useful to conceptualize it, in order to proceed in the right direction. The "ideal" system would allow a researcher sitting at his terminal in a remote location to search for, locate, access and manipulate data from a single central source which is multi-dimensional in subject matter content.

In other words, the challenge we are faced with is to make available to the right user, the right information, at the right time, in the right format and at the right price, and with problem solving algorithms. By information is meant not only the data but all textual inputs making the "numbers" immediately comprehensible, and promoting meaningful understanding and their apt use.

It is this demand for statistics in machine-readable form that has resulted in the proliferation of special purpose computer-based storage and retrieval systems, each making available specialized information. These are developing independently with no apparent efforts to coordinate them. Each base is organized and structured differently, has a different objective, a different range of access elements, different indexing system and different protocols for access.

The Special Purpose Data Banks or SPDs solve the information requirements in terms of timeliness and accessibility for the researcher who has discrete requirements. However, the SPD does not meet the needs of a variety of users who require detailed and integrated information cutting across a number of subject-matter areas, i.e. information regarding Prices, Labour, External Trade, National Accounts etc.

This probably constitutes the "raison d'etre" of general purpose data banks, or GPDs. Indeed we define a GPD as a detailed and conceptually compatible set of data, covering a large number of interrelated fields, and supported by a system through which data can be easily entered, stored, maintained and manipulated in a uniform general format consistent with the input requirements of some given retrieval and analytical package.

The "ideal" GPD would be a single giant bank storing data at the micro-level (i.e. the answers to each question of each questionnaire or administrative record would be stored). In this way, a-priori determination of aggregation levels for the time series module or table format and size for the cross-sectional module would be eliminated, since a user would have the capability of creating any time series or any table using the micro-data as the building block. It is obvious that the confidentiality problems with such a system would be immense and, even if these were solved, it still would not be politically, economically, or technically feasible to establish such a bank.

One variation of a single giant data bank at the micro-level would be to establish an efficient, formal network of linked special micro data banks. However, this concept would be extremely difficult to implement, since it implies compatability in terms of software and hardware between the various elements.

The problem therefore remains of how to satisfy the requirement of the data user who wants information on a variety of subjects from one source. The CANSIM Division of Statistics Canada is working on a series of projects whose goal is to disseminate the most sought after statistics to users seeking information from non-specialized data banks.

CANSIM, Statistics Canada's computerized data bank and its supporting software, is the central depository for highly aggregative statistical information. Put another way, our concern is to provide adequate communication of what could be called "horizontal information". Many of our users are interested in the kind of information that cuts across organizational or subject-matter lines within the bureau. This, of course, implies that data needed by a specific group of users may be partially included in a specialized data bank as well as partially in another data bank such as CANSIM. In this case, partial duplication of data may be desirable. CANSIM now covers and represents a relatively large range of economic data of the publishable statistical series of Statistics Canada, as well as some material from other important suppliers of data. This adds up to an impressive mass of data, in fact, over 300,000 series or some 10 million statistical observations. The data can be called upon in a selective fashion as required and in quantities needed for particular calculations. Sophisticated users are unlikely to require more than one thousand series in the course of very elaborate analytical studies, but the rest are available should they be required.

Before describing future plans for CANSIM a brief history of its evolution and its current status may be appropriate.

In late 1966, Statistics Canada accepted the responsibility for storing and updating time series in a computerized base. Information was stored and maintained by a sequential file system called DATABANK. At that time, there was also a companion program called MASSAGER, which carried out statistical manipulations of the data. Beginning in 1967, one can distinguish three phases. The first phase provided for storage and maintenance of the entire data base in direct access memory. The system was then resident in-house with the data stored off-line. At this time only Statistics Canada personnel could access the data base. The second phase occurred between May 1972 and November 1973 when CANSIM was available at the federal Computer Services Bureau (CSB). When CANSIM was located at CSB, federal government users could access the data base directly, but other users could only access the base indirectly through the CANSIM group. This separate location outside Statistics Canada was and is intended to demonstrate clearly that the data bank, which was available to federal government departments (and which is now available to the public), had no direct links with the main computer where a great deal of highly confidential material is stored. In November 1973, CSB discontinued its service and it was decided to move CANSIM to a commercial service bureau. The third phase of CANSIM is one of continuous enhancement. The system is presently maintained at SDL and is available to all users. Maintenance and retrieval runs are performed in batch mode, and the complete base is kept on-line. The programs have been implemented on IBM equipment operating under MVS and MVT. Customers can use CANSIM software to produce computer printouts of information, or intermediate files with data structures suited to a number of manipulative programs. There is also a prototype interactive retrieval operation which initiates a batch job, generating a file which can be accessed from remote locations, for on-line APL processing or for terminal output of tables. Finally, the system can be accessed interactively. Potential customers have three choices in accessing the base. They may become customers of the Host Service Bureau, SDL, the only service bureau at the moment that has the complete base, they may become customers of any of 11 secondary distributors, or they may continue to obtain service through the CANSIM Division.

The concepts of secondary distributors, the CANSIM Mini Base and the Mini Base Supplement Service were introduced in June 1976, and evolved in response to the need to bring standardization of our product in terms of content, timeliness, accuracy and reference documentation. Prior to June 1976, any organization or individual could access the main base and then make the data available to their clients. The data sets, therefore, varied in terms of content. The timeliness varied since the individual updates were made at different times and with varying frequency. Each organization had its own documentation. Some organizations updated only the newly released data points and therefore missed the revisions, thus rendering suspect the accuracy of the data.

It must be evident that a chaotic situation existed and that the end user was understandably confused about the service called CANSIM. Indeed, we have encountered clients who thought they were accessing the whole base (i.e. CANSIM) and thus were unaware of the availability of many more series.

The mini base service consists of a daily file that updates the 15,000 most frequently accessed series, plus a periodic historical replacement file. All secondary distributors must purchase this service, and must make the mini base available to their clients. The secondary distributors have the right to access additional series from the main base and in effect the supplementary service facilitates this transaction. Upon request from the secondary distributor the supplementary series will be made available with the mini base 24 hours later. The supplementary service enables a secondary distributor to advertise the fact that his clients can have access to the entire main base with a lag of 24 hours.

A number of guidelines, rules and regulations govern the way secondary distributors are to disseminate CANSIM data. However, I shall only mention one that is very important. Each secondary distributor provides us with a list of their clients and the record of the number of retrievals per series. The number of retrievals per series is used as an important feedback to the divisions that load data on the system and secondly, they form the basis for determining the content of the mini base. In addition to this very important feedback function, secondary distributors also provide the end user with a large choice of software, telecommunications networks, and technical support staff that can be utilized in accessing CANSIM.

The system described so far is a "time series" data base, in that the system is oriented to the management of time series data. In recognition of both the increasing importance of and the demand for social information and of the fact that the most significant portion of social data is in the form of cross-classified tables, we have developed the CANSIM Cross-Classified System. This system is now operational and enables statistical tables of up to nine levels of cross-classification to be entered, stored, retrieved and manipulated. In addition, descriptive information provides titles, footnotes, explanations and definitions in order to foster proper use of the tables. Retrievals are possible in either interactive or batch mode. The retrieved data can either be printed in a specified table format or can be loaded as a matrix (table or part of table) into an APL workspace where it can be manipulated alone or with other data.

The most difficult aspect of the cross-classified module is the a-priori determination of the table format and size. In contrast to the time series module, extensive resource commitments are required from the subject matter divisions that load data on the system.

We are now negotiating with a number of divisions and plan to load datasets to the lowest and most common level of geographic disaggregation and containing comparable variables for three years of information.

The future development of CANSIM will concentrate on the SIS (Statistical Information Service) portion. At the beginning of the paper we defined the ideal system as one that would enable a researcher to search, locate, access and manipulate information. CANSIM is the data base portion of the "ideal" system, although it is not "ideal" in terms of its design or operational characteristics. The remaining and concluding portion of this paper will suggest that, although the fulfilment of user demands for machine-readable information is an extremely important undertaking, it is the search function and the subsequent creation of packages of information plus an integration function that may offer the greatest benefit to both the user and purveyor of statistics.

Databanks must be seen in their true perspective as "information disseminators". Triggered by basic needs or statistical analysis facilitated by the computer, an increasing proportion of research and analysis operates on highly disaggregated data. It is this relatively easy accessibility and comparability of highly disaggregated series that will contribute to the resolution of problems of statistical integration, standardization of concepts and coding practices and the identification of gaps in the statistical system, as well as the identification of potential areas of duplication. It is ironic but, as CANSIM has grown, it has become more and more difficult to confront related series with each other. It was initially observed that the user would have to search for the data in various publications. Now he has to search through more than 300,000 series titles to conduct an analysis. In order to facilitate the search system the CANSEL (CANSIM Data Package Selection System) module will be introduced in 1979. CANSEL will answer the question "out of all data available on CANSIM what package of data is available with the following attributes in terms of geographic location, activity classification, commodity detail, etc". In order to facilitate the confrontation of series, SIS will offer "packages" of information. For example, a user wishing to study the Canadian labour market will be able to search and retrieve a package with such data as labour force, earnings and hours, job vacancies, unemployment rates, unemployment insurance commission activities etc. The available data relating to the Canadian labour market have been integrated into a consistent set (in terms of hardware and software configurations) and statistical analysts can now use computerized statistical packages to analyse them.

The massive demands for additional information have indicated that analytical text must be added to the data released from SIS. Such text will be related current intelligence. For example, a user retrieving the latest Consumer Price Index will also be able to retrieve accompanying text pointing out significant changes in the total and component indexes, reasons for these movements, and related information on the Canadian economy.

In other words, CANSIM/SIS should provide a context for retrieved statistics, making them immediately comprehensible, promoting meaningful understanding and apt use of the information. This service will be provided for data (time series and cross-sectional) in most frequent use and essential for current policy making.

In addition, we should be developing "concepts and methods" text files for the SIS series. These files will contain summary information on various series or blocks of series, outlining definitions of components, methods of collection, and discontinuities, if any.

CONCLUSION

CANSIM is an operational system but, to date, this large system has not been as widely used as might have been expected, although our users are retrieving about 4 million series per year. It is one of those cases of government initiative, well ahead of time, in anticipation of massive new needs for machine-readable information. Although the technology of data banks will change rapidly in the future, it is not the intention to change CANSIM unless technological innovations can be added to the system as distinct modules. We believe that the greatest benefit to CANSIM will center around the data it contains and the manner in which it is "packaged" rather than the sophistication of either the hardware or the software. Emphasis will have to be placed not only on new relevant series but on the organization of data in packages for each retrieval and comprehensive analysis.

Ultimately, an extension of this integrated data information system will be its ability (through the intervention of the data provider) to warn of data of dubious quality by the use of editing checks, equation banks, econometric techniques, etc. When fully developed, CANSIM/SIS will provide the framework for a relatively more sophisticated analysis of current socio-economic conditions than is now available, create a rapid and valuable link between Statistics Canada and the policy making community, improve knowledge about the interrelationships among social and economic series and events, and, therefore, contribute significantly to improvement in data collection and integration.

The theme of this conference is "Sharing Resources - Sharing Costs." The history of CANSIM illustrates this theme very aptly. First of all, it is an excellent illustration of how Statistics Canada has been able to pull its act together and open up its total data resources through a single vehicle rather than appearing to the world as a collection of separate and seemingly unrelated subject-matter sources. Secondly, it shows how the government and the private sector (as represented by the participating secondary distributors) can pool their resources and expertise in a joint venture. Finally, to the extent we can further persuade the general user public of the relevance and usefulness of the system, we will be helping them to get a better return on the investment they have made in their capacity as taxpayers.