The Bureau of Labor Statistics (BLS) has gained economy and speedier results through the use of a data base manager as the central feature for most of its computer uses. Associated generalized systems are embedded in very high-level, nonprocedural languages that may be used by personnel not knowledgeable in the computer sciences.

INTRODUCTION

The Bureau of Labor Statistics started to switch to large-scale, third-generation computers six years ago. Our deliberate approach has given us time to study our past mistakes and size up new opportunities. We have taken the painful step of going from one machine generation to another, and this time we have worked out a blueprint, a master plan, to give unity to the BLS approach to third-generation computers. The work is partly done. Some will take a great deal more time and some may never be finished.

THE PLAN: DATA BASE MANAGEMENT

The Central Ring

Important to understanding, the BLS picture is the magnitude of its data base. There are about 100,000 time series and they include important measures of the Nation's economic well being, such as the Consumer Price Index, the Wholesale Price Index, and the unemployment figures. Cross-sectional summary data are also stored in the data base. These include industry employment projections and productivity studies, for example.

Central to the plan was a data base management system called TOTAL (Figure 1). This software allows network concepts for logical relationships, the way a card catalog points to the shelf location of books in a library. A search for data follows a path to the direct access hardware where the item can be found, rather than following a sequential search from the beginning of the magnetic tape file.

DATA DOCUMENTATION

Codebook and Dictionary

Computer users have common problems when processing data: They need to know what data are available, where they are located and how to get them. Whether a computer programmer or a research analyst, the variables to be used must be identified. No matter which programming language is used, somehow a path to the right storage places is mapped to locate the data that our programs must extract and process.

The BLS goal is a "Codebook" and a "Dictionary," which we and the machine can read as sources of this information. For researchers, there will be a Dictionary which names what is in the Data Base and defines the variables cited. For the programmer, there is a Codebook with technical data describing the data formats and locations. Together, the Codebook and Dictionary form a "Data Documentation" system.

SECURITY AND INTEGRITY

Responses to Bureau questionnaires are protected by a pledge of confidentiality, to be used only for statistical purposes. We take pains to insure that our pledge is kept and we must surround the BLS data base with a circle of protective software.

The protective software must recognize three types of accessibility to the core: The first is to data that are published. These are available
to anyone in BLS. In the long run, we hope to make these data accessible to anyone, whether inside or outside of the Bureau, in or out of the Government. The second need is for access by BLS employees to data that are not published. They include summary figures that contribute to published information but are not statistically reliable on their own account. Well inside the data base, in a third level of accessibility, we find the respondent reports, protected to keep our pledge of confidentiality.

Besides guarding against unauthorized use, software surrounding the data base must also protect the integrity of the figures. We must be sure that we can recover from equipment failures, such as a head crash on a magnetic disc or inadvertent erasure of a magnetic tape.

MODULAR, GENERAL-PURPOSE STATISTICAL PROCESSING PROGRAMS

In the plan to use third-generation equipment, some of the 50 or 60 "systems" were combined. For example, we have had separate statistical surveys for collecting prices of commodities in the wholesale market and in international markets. These data contribute to our Wholesale Price Index, Industry Sector Price Indexes, and International Price Comparison programs. If we had gone the old way, these separate surveys would have prompted separate, independent computer systems, even though the calculations and other computer operations were often essentially the same. In the redesign effort, the three systems were reduced to one at considerable savings in time, money, and subsequent maintenance. A similar approach was taken in combining four recurring occupational wage studies into one computer system.

The notion of combining systems with like functions brought us to the idea of generalized software to deal with these functions and replace ad hoc programming. Thus, our third ring holds general-purpose programs that are tailored to the unique needs of large-scale, statistical data processing. Although readily identified, we also found that the construction of a full line of functional modules is a tall order. More than a dozen general-purpose programs are required. Building these modules is a costly, complex task and will take time to complete.

Now, let's take a look at the functional modules that make up the third ring.

Table Producing Language (TPL)

The first outcome of the step-by-step approach was a powerful new computer language that allows Bureau staff to turn out statistical tables easily and at low cost. It is called Table Producing Language (TPL). The TPL system already knows what a table is and how to generate one. It only needs to be told about the one wanted. Thus, when one describes the table wanted with the table producing language, one avoids the tedious and time-consuming effort otherwise involved in telling the computer, step by step, how to make the calculations and how to lay out System. It will provide sample selection, mailing label generation, the table. This approach has severed the connection between the user and the computer. The user need not be familiar with how the computer works. Moreover, it allows Bureau social scientists to use the everyday common BLS nomenclature to describe the tables. In short, TPL has reduced a burden, speeded work, and increased the BLS capacity to respond.

Print Control Language (PCL)

Statistical tables form the bulk of our printed output. An extension of TPL serves table production in ways other than as a tool for tabulation. We need the results on a clear copy for photo-offset printing. For this purpose, we must surround the numbers with a clean and concise framework of explanatory alphabetic information, such as table and column headings and stubs, as well as footnotes and similar data that make sense and are readable. We call this module the Print Control Language (PCL).

Through PCL, the user can specify column and stub widths and insert his own alphabetic labels for each variable, as well as footnotes and other text. Decimal points and special symbols, such as dollar signs, are available. There are options for hyphenating and centering titles. Taken together, these features can create tables that are acceptable for direct photo-offset printing.

Reproduction of computer printout is less satisfactory than output from an electronic photocomposition device. Therefore, such devices required a new tailor-made computer program for each table to be photocomposed. A new version of the Print Control Language permits tables to be formed by a photocomposer without the need to write computer programs. The user has a wide range of print size, style, and other choices for typographic enhancement with the result that printed tables appear to have been typeset.

Statistical and Econometric Routines

There is a TPL option that allows the research user to shunt tabulated results into the SAS package. In a similar way, it is intended that other packages will be provided to accept and process data from the data base. For example, the Bureau has acquired a copy of the Harvard-KIT version of the econometric package called Time Series Processor (TSP). It is designed to carry out all the computational steps which occur routinely in statistical analysis of time series and also to provide some of the more sophisticated econometric techniques for manipulation and analysis of time series.

General-Purpose Charting System

The Bureau has acquired a generalized time-series charting system, it requires some programming knowledge. To simplify things, it has been embedded in a very high-level language in the way that cross-tabulation routines are at the core of the Bureau's Table Producing Language. The enhanced
system allows users to produce time-series charts, editorially suitable for publication. A wide choice of features controls the appearance of the chart so that it can be photographed for photo-offset printing without further editing.

Screening, Editing, and Correction System

BLS systems analysts and programmers specify and write many computer programs for testing the accuracy of data reported to us by our survey respondents and mark those which appear to be of doubtful validity or clearly erroneous. Although the subject matter may differ sharply from survey to survey, the computer steps and factors taken into account are essentially the same for all surveys. Our success in designing and implementing a generalized cross-tabulation system, some analysis of a generalized approach to screening, as well as work by Statistics Canada, suggests that the project is feasible.

File Manipulation System

The situation where the reported data, even when correct, are not ready for tabulation is a common one. Typically, at the beginning of a survey, information is organized for easy and efficient collection, entry, and screening. The resulting file organization is not usually suitable for tabulation and subsequent analysis. The capability to reorganize files will be incorporated into a proposed system, which we call the General-Purpose File Manipulation System.

Its capabilities will include selecting subsets of the original data, merging data from different files, reorganizing the data structures, and calculating new variables. This system will fill the gap between the general screening system, which is concerned with the correctness of reported data, and TER, which tabulates and displays data. Work on the file manipulation system is scheduled to begin later this year.

General Survey Collection System

For a general solution to mailing and controlling questionnaire responses, there is a module called the General Survey Collection response control, follow-up list generation, and status reports. To serve these ends, the system must support and maintain registers of names and addresses and other information about respondent attributes, such as industry code and employment size. A major factor in this system is the requirement to maintain about 4 million names, addresses, Standard Industrial Classification (SIC) codes, size, and other attributes of establishments that comprise the Bureau's universe file from which sample cases are selected.

Query and Microfilm (COM)

From time to time, there is interest in retrieving records supplied by a specific Bureau respondent or sets of respondents. Such interest may derive from discussions with the respondent about his reports or from concerns about the individual case, for example, in comparing responses with an independent quality measurement survey. A retrieval facility to query the micro data base would serve this need.

Computer output on microfilm (COM) is practiced in BLS. However, it is not clear at this point that a general solution for placing computer output on microfilm is an important Bureau requirement. At present, printed tables are the principal BLS output. In the event that a generalized COM tool is justified, it will be an extension of the Print Control Language System, which has been implemented with this possible enhancement in mind.

Application Systems: PL/1, FORTRAN, COBOL

Data processing requirements of national statistical agencies cannot be totally generalized. We know that the compilation of some statistics is based on special mathematical and statistical arguments that have applications only in a specific instance. The modified Laspeyres formula used in the Consumer Price Index is an example. Programming special instances will, of course, continue to rely on the common computer languages—PL/1, FORTRAN, and COBOL.

THE PLAN IN PRACTICE

Well, what does all this mean? Given the picture of a central data base, a data base management system and Data Documentation System, and blocks of general-purpose, functional modules, what happens? It seems clear that for future systems designers work will be reduced. The data base management software will help keep data in order. It will give assurance that corrections, deletions, and additions are processed into and out of the data base expeditiously and reliably. The general-purpose routines are available to provide for data security and integrity. If data are new, they need to be described only once for the Codebook and Dictionary. All subsequent inquiries for data descriptions may be satisfied by reference to the Data Documentation System, where the file descriptions are accurately portrayed and easily transcribed (perhaps even automatically) into the relevant sections of new or modified programs.

When an assignment calls for a statistical system from sample selection through publication of results, system designers can consider the general-purpose modules to address, edit, tabulate, and format the results for printing. They can provide for special analysis through the package of statistical routines and display results on computer-drawn charts. The research economist, statistician, demographer, and other social scientist will have these same tools available, and he need not be knowledgeable in the computer science discipline. By mixing the variables he wants tabulated and citing the kinds of tabulations to be performed, the data base management system will reach into the Codebook for the location of the value of the variable cited, retrieve and deliver these figures to the tabulating system.
for processing, and display the results in the form expressed by the user.

The foregoing possibilities might seem unreal. Well, let's see. Most BLS systems now use TPL for cross tabulation and some have used several of our other building blocks to augment tailored programs. In one graphic instance, an entire system was constructed from off-the-shelf items. Not a single line of newly written code is used. The system tests data for validity with a routine from the collection of statistical analysis programs, updates and extracts data with a prototype portion of the file manipulation system, manages its data base with TOTAL, uses TPL and its Codebook for cross tabulation, the Census X-II program for seasonal adjustments, and PCL for table display.

In summary, the exploitation of this integrated and comprehensive system depends on a willingness to seek new ways to solve old problems. The immense potential of an open-ended computational system, attached to a large data base of statistics representing a broad spectrum of the U.S. economy, is a power that should be treated with respect. Economists, statisticians, other social scientists, and systems analysts must operate as a team with communicative skill to maintain the delicate balance between innovation and uncontrolled waste of funds, computer time, and analytical effort.

REFERENCES

