ED 183 178 IR 008 119

TITLE	Report of the Conference on Development of User-Oriented Software (Alexandria, Virginia, November 8-10, 1977).
INSTITUTION	American Statistical Association, Washington, D.C.
SPONS AGENCY	National Science Foundation, Washington, D.C. Div. of Social Sciences.
PUB DATE	Nov 77
GRANT	NSF-76-15271
NOTE	257p.; Figure 2, page 235, is nct legible.
EDRS PRICE	MF01/PC11 Plus Postage.
DESCRIPTORS	*Census Figures; Data Bases; Data Collection; *Data Processing; *Disclcsure; Information Processing; *Information Systems; *Statistical Data; Use Studies

ABSTRACT

One of four projects conducted by the American Statistical Association (ASA) in cooperation with the Bureau of the Census, the conference explored the most important and fruitful research and development topics within the user-oriented software domain. Its objectives were to (1) develop recommendations on mechanisms to improve access to and use of machine-readable Census Bureau data; (2) identify software systems needed to assist the user community to more easily organize, tabulate, and present census data; (3) review possible additional means for user access to census data; (4) identify and recommend specific research and development activities that would lead to improvements in the access to and utilization of such data; and (5) develop specific recommendations to ASA for proceeding with an expansicn of its program. This report summarizes each day's session, as well as discussions and recommendations of the conference groups and sub-groups. Appendices list the participants, provide background and biblicgraphic material, describe the conference agenda, contain the papers submitted, and offer a Census Bureau view of the activities discussed by the participants. (FM)

REPORT OF THE CONFERENCE ON DEVELOPMENT OF USER-ORIENTED SOFTWARE.

Old Town Holiday Inn
Alexandria, Virginia

November 8-10, 1977

AMERICAN STATISTICAL ASSOCIATION
806 - 15TH STREET, N.W.
WASHINGTON, D.C. 20005

CONTENTS

REPORT OF THE CONFERENCE ON THE DEVELOPMENT OF USER-ORIENTED SOFTWARE[1]

I

INTRODUCTION

The Conference on the Development of User-Oriented Software was held at Stouffer's National Center Hotel in Arlington, Virginia on November 8, 9 and 10, 1977.

Background

This conference is part of a 3-year program conducted by the American Statistical Association (ASA) in cooperation with the Bureau of the Census, and supported by the National Science Foundation and the Bureau of the Census. Its purpose is to explore ways of improving the national data base through a program of research at the forefront of statistical techniques applied to the social sciences, and by supplementing and sharing with researchers in a large data collection agency the experience of senior social scientists and the training of graduate students in statistics, economics, demography, computer science and related areas. The Conference on User-Oriented Software is one of four projects being conducted under this program. The other projects are in the research areas of (1) seasonal adjustment of economic time series, (2) edit research of computer output and (3) the development of new population projection methods for States and metropolitan areas.

Purpose

The conference sought the advice of experts outside the Census Bureau on the most important and fruitful research and development topics within the user-oriented software domain. Five specific objectives were posed:

1. To develop recommendations on mechanisms to improve access to and use of machine-readable Census Bureau data, especially through the development of user-accessible software.
2. To identify software systems needed to assist the user community to more easily organize, tabulate and present Census data.

[1] The conference is supported by NSF grant #76-15271. The views and recommendations herein are solely those developed by the conference and not necessarily those of the National Science Foundation.

3. To review possible additional means for user access to Census Bureau data other than the three identified software areas of data-base management systems, graphic systems and generalized tabulating systems.
4. To identify and recommend specific research and development activities that would lead to improvements and simplifications in the access to and utilization of Census Bureau data.
5. To develop specific recommendations to the ASA for proceeding with an expansion of its program.

Participants

Conference participants were selected and invited jointly by the ASA and the Bureau (Appendix A). The selection process balanced participants by professional backgrounds as well as by areas of application. The final list included statisticians, demographers, computer scientists, sociologists, geographers and others; their experiences ranged through business, government, academic and research applications. Some 35 people attended from outside the Bureau, with another 20 coming from inside the organization.

Conference Format

The format of the conference was organized around a view of generalized software for the Census user in three parts--data organization, data tabulation and data presentation (Appendix B). Data organization encompasses public-use microdata files and summary files in terms of their preparation and organization for better access by the general user. Data tabulation is, of course, a large part of the special processing of Census files. Data presentation is viewed as including microform output, graphics, mapping and all types of publication-quality presentation forms.

The first day of the conference was devoted primarily to the presentation of invited papers. The second day, the conference participants divided into three groups under the headings given above of Data Organization, Data Tabulation and Data Presentation; each group separately prepared recommendations to be made to and from the whole conference. The third day was devoted to the presentation, discussion and refinement of the recommendations.

Much of the original content planning for this conference was accomplished by William Alsbrooks and Kam Tse of the Census Bureau; further planning also included Bruce Carmichael, Lawrence Cornish, James Foley and Melroy Quasney. Michael Garland, Warren Glimpse and Paul Zeisset, of the Bureau's Data User Services Division, also made substantial contributions. Daniel Relles, of the Rand Corporation, and George Heller, of the Census Bureau, served as co-chairmen of the conference and were involved at all stages, including responsibility for this report.

Organization of Report

Following this introductory section, the first day's session is summarized in section II. It was not the intent of the conference to include a verbatim transcript of all proceedings, although the formal papers and other materials presented by the speakers the first day are reproduced or completely referenced in Appendix C. Nevertheless, it is important that the reader be given some sense of the range and spirit of the sub-group discussions the second day and during the presentation of their recommendations to the plenary session for review and perfecting the third day.

Accordingly, section IV summarizes for each sub-group highlights from its day's discussion and formation and presentation of its recommendations. Section V covers the discussion and acceptance of final recommendations on the third day.

Section III summarizes the final recommendations of all three groups and relates the conference's findings to the objectives posed at the beginning.

Appendix A lists the conference participants, providing appropriate background and bibliographic material as well. Appendix B describes the conference agenda. Appendix C contains the papers submitted by all of the speakers and some of the participants. Appendix D is a "Status Report on Selected Census Bureau Activities," to provide the reader with a Census Bureau view of many of the activities discussed by the participants.

II
OPENING OF CONFERENCE AND PRESENTATION OF PAPERS

Opening

The conference was opened and a welcome extended by the Directors of the American Statistical Association and the U.S. Bureau of the Census.

Fred Leone, Executive Director of ASA, traced the historical effort to improve the social science data base, of which this conference of prime movers in that field is but one facet. The dual purposes of the conference, he explained, are directed toward developing and perfecting software to enhance the use of Census Bureau and other data by the social sciences and to examine future needs in terms of users' requirements and necessary research.

Manuel Plotkin, Director of the U.S. Bureau of the Census noted that among the Bureau's top goals and priorities are (1) the uses and applications of data and (2) the improvement of data-processing systems to enhance timeliness, accessibility and relevancy of the data. He called attention to the Bureau's increasing workload and corresponding pressures; preparation for the 1980 census is under way, the 1977 Economic Censuses are about to be taken, there will be a census of agriculture for 1978, the Current Population Survey is being expanded, etc. There have been software meetings in the past, but this conference is the first one in which there has been a joint meeting among data users outside the Bureau and data users and computer hardware and software staffs from within the Bureau. All have different perspectives to contribute. In the area of generalized software, the Bureau hopes for some innovative developments that will increase usefulness and productivity..

Presentation of Papers

Following the conference opening, ten speakers presented, in full or in summary, papers which had been prepared and distributed to the participants and are reproduced in Appendix C. The first four papers were designed to give a general view of the state of the arts, the need for, and availability of, user software as seen by the Census and the uses and needs as seen by other governments and in the private sector.

William Alsbrooks, who is in charge of the programming staff that develops software for use within the Census Bureau, presented (in a paper written with James Foley) an overview of the three topics to be addressed by the conference, namely, data organization, tabulation and presentation.

Warren Glimpse, Assistant Chief of the Data User Services Division, Bureau of the Census, focused on the supply of, and demand for, software for improving data use. He reviewed the availability of machine-readable resources and existing software. He emphasized that, while there are some unmet needs for user software, there are many related requirements for effective use of Census Bureau machine-readable products other than software. Major problems in using these products are not only software but also the file structure, documentation, and archiving procedures followed by the Bureau, or the absence of them.

Harold King, who directs computing services for the Urban Institute, talked about the software needs of State and local governments. He observed that in 1969 meetings with the Census Bureau to request software support for users had negative results, but the Bureau's position has changed since that time. There are approximately 38,000 general-purpose governments in the United States, roughly 35,000 of which are small municipalities and townships that need data to meet Federal grant-application and other requirements. It must be assumed that many of these governments have little or no computer capabilities, although there is a rapid expansion in the use of mini-computers. Users still will need guidance on how to apply census data to local problems.

Richard Ellis, a marketing manager for the American Telephone and Telegraph Company, emphasized in a review of his full paper the variety of corporate uses of census data he had covered. These are good analogs for general business usage of research information.

The last six papers of the opening day provided opportunity to hear from a representative of the user or technical software community and a Census Bureau speaker on each of the three topics the individual sub-groups would be working on the second day. The first two speakers had prepared papers on the organization of data.

Mervin Muller, Director of the Computing Activities Department of the World Bank, posed a number of questions about data organization and outlined research areas that would lead to fruitful discussion within and beyond the time period of the conference.

Bruce Carmichael, leader of the Central Data Base Group at the Census Bureau, discussed the importance of data organization and the need for more sophisticated data schemes and accessibility, and stressed the Bureau's need for users' help in this direction.

During the remainder of the first day the next two speakers addressed statistical tabulation and the final two speakers, statistical presentation.

Hugh Brophy, Chief of the Systems Development and Programming Unit at the United Nations Statistical Office, noted the magnitude of the processing involved in a national census. The resultant information should be regarded as a valuable national resource. In practice, there tends to be a loss of information in summarizing the statistics, difficulty in linking with local data, and great expense when special tabulations are required. Flexible tabulation systems are a partial solution. He then summarized his paper, dealing with generalized tabulation systems.

Melroy Quasney, of the Census Bureau's Systems Software Division, said that the Bureau, in solving its own problems, hopes to supply users with tools that may include a problem-oriented language.

Robin Williams, of the International Business Machines Corporation (IBM), discussed a user-oriented systems approach for software and hardware developed in IBM's Research Division. He illustrated with slides IBM's Geodata Analysis and Display System (GADS): it builds and maintains files, extracts data, and then analyzes and projects them in tabular or graphic form on a color display. Topics suggested for further discussion were: graphic-terminal functions research, software for interactive graphic-terminal support, the feasibility of supplying data in the format and form requested by users, and the provision of services to a requester, e.g., on-line query facilities, plotting facilities, etc., for census data.

Lawrence Cornish, of the Census Bureau's Systems Software Division, pointed out that the classical approach to data publication is to deliver them in non-machine-readable form. Using materials from an internal Census Bureau study, he described the hardware now available for a wide variety of alternative data delivery systems including graphics.

III
SUMMARY OF THE MAJOR RECOMMENDATIONS

Before going on to the discussions and detailed recommendations of the three sub-groups of the conference, which are set out in section IV, it would seem helpful to try to pull together and highlight the most important of those recommendations. The reader, of course, is urged to consult section IV for the full effect. Particularly noticeable at the onset is the high level of overlap and concurrence in the three sets of recommendations, the more so in view of the separation of the three groups when their recommendations were being drafted and the three distinct software areas represented.

The recommendations of the Conference were far-ranging but certainly not beyond the general guidelines set as objectives for the conference. In covering the Census Bureau's user software, and the distribution of that software, it was inevitable and natural to discuss the products and objects of that software--the data. At times also it was necessary to discuss and cover the associated areas of user documentation and user training. It could not be expected that all of the participants would be fully cognizant of all of the Bureau's efforts and plans in all of these areas. While an attempt was made in some of the Bureau's first day presentations to present some of this background, all of the relevant areas could not be anticipated. So, in order to give the reader of this record a brief overview of these activities, Appendix D has been included. This appendix is not an answer to the conference recommendations or a solution to the problems raised, but is background material that could have been provided prior to the conference.

As the general discussion of the third day showed, and as reflected in the recommendations, there was a strong sense of concern by many attendees that the Bureau would not be adequately prepared to meet user demands in the 1980's. Many felt that an examination of the entire data delivery system was necessary, not just the software development component. To the extent that these concerns are actual, the conference participants will await a Bureau response; to the extent that these concerns represent a lack of knowledge of the Bureau's activities, the participants will expect a better educational effort by the Bureau.

The recommendations fall into three types: institutional, which involve largely improved communications between users and the Census Bureau; technical, which deal with the actual software development; and those particularly appropriate as further ASA/Census endeavors.

Institutional Recommendations

Strengthening the Interface

The need to strengthen and broaden the interface between users of census data and the Census Bureau suggests that:

* User needs should be monitored.
* An ongoing assessment of user needs for software should be conducted.
* User comments and evaluations of software should be compiled.
* A users' group on user software should be formed.
* User education and training must be expanded.
* Materials and training courses for user education should be developed.
* User-oriented documentation and training material on data and software capabilities should be geared to various levels of technical and processing proficiency.

Serving the User Community

To better serve the user community, it should be determined which of the following should be considered:

* A national census data center.

* A consortium of users.

* A national network.

Technical Recommendations

Possible technical solutions to a wide variety of user problems merit the examination of:

Data Dictionaries

Machine-readable data dictionaries must be distributed for each distributed data file. Then dictionaries must be accurate, up-to-date and machine-portable. The dictionary should include definitions and common recodes and provide easy mapping to data elements. The Census Bureau needs to work with existing groups such as the Association of Publication Data Users (APDU), the Federal Statistical Users' Conference (FSUC), etc., that have already addressed the subject of terminology, conventions and definitions, to ensure that the data dictionaries are meaningful to users. The Census Bureau also should provide as detailed information as possible on its data dictionary plans to the user community as soon as possible. All software developed for users ought to access data via a data dictionary to remove format dependencies from programs associated with reading census files.

Data Extraction

Efficient mechanisms and procedures should be established to extract data for users and to manage the response to such requests. The Census Bureau should support the development, with an eye to subsequent portability, of generalized extraction software that will automatically provide a modified data dictionary.

Software should be developed and made available by the Census Bureau for handling the most basic and simple types of data retrieval and presentation.

Research should be conducted to determine the special machine-readable files (extract files) and extraction programs that should be produced for special program compliance.

The extraction in machine-readable form of the full array of census data aggregated according to user-defined geographic areas should correspond to the full range of information now available for standard Census-defined geographic areas.

Geographic Base Files and Other Geographic References

User specification of tabulation areas in terms of coordinates should be allowed. This would require a uniform high standard for coordinates in geographic base files (GBF's), and GBF coordinates should be corrected topologically and cartographically.

A machine-readable data base should be developed that defines changes and equivalencies in statistical areas.

The Census Bureau should provide separate machine-readable files of spatial definitions (e.g., polygonal coordinates or raster) for all statistical areas.

Generalized Tabulation Systems

The tabulation group made recommendations for research, development and general support in the area of generalized tabulation systems. While approving the Census Bureau's efforts to elicit users' needs for this type of software, they listed a number of areas that would require research before a system could be put in place. For example, a generalized user system would have to interface with data dictionary systems, and these dictionary systems have not been defined for the Census users.

Data Base Methodology

As a vehicle for promoting research on advanced data base management technology, it was recommended that an efficient access and transmission system for user requests concerning specific places, types of persons and characteristics be investigated. A capability to flexibly combine persons into alternative social units was described as highly desirable and technologically worthy of research.

Time Series

The data organization group recommended that the costs and benefits of a time series capability be explored.

Hardware

The Census Bureau should investigate the potential role of minicomputers and microcomputers for data portability and for access and analysis of census data by users with limited resources.

Possible Areas for Future ASA/Census Cooperation

All three groups recommended National Science Foundation support for a research fellow at the Census Bureau. One proposal involved basic research projects and feasibility studies in data organization and delivery systems. Another recommendation was that NSF support research and development into efficient, effective and statistically useful techniques for the generation of statistical tables.

IV
GROUP DISCUSSIONS AND RECOMMENDATIONS

As noted in the Introduction, it is not intended to reproduce in this conference report a verbatim account of the proceedings of each of the three sub-groups. What is attempted in the following pages is to give the reader a feeling of the matters each group addressed, how they covered them during their discussions and finally, in each group's own language, the recommendations they agreed to and how they presented them to the full conference on the third day.

Data Organization Group

Discussion

The data organization group began by pointing to certain areas that it would like to cover and directions it might wish to take in developing its findings and recommendations. Included in these were:

* More flexibility in the organization of census data to accommodate the broad spectrum of user needs.

* More detailed information and links between relevant data at a person or block level (base level).

* Easier access and utilization of data; data should be made available more quickly to users who request it. Better documentation would reduce the amount of time spent interpreting census data.

* Census data should be able to accommodate and be accessible to both sophisticated and unsophisticated users, or large vs. small organizations.

* What the current state of the art is and what advances can be made based on technology available today.

data that would be available in different formats and still preserve confidentiality. The Bureau might provide detailed and specific information tailored to particular needs without violating disclosure rules.

* Requires data from demographic, housing and economic areas; uses the 1970 summary tapes and would like more software for them. At present, much tape handling is required before getting information; very often has to regroup data. It would be helpful if the data contained different codes for such things as school districts and police precincts. Needs public-use data provided faster than it is. SPSS is satisfactory for much of the analysis. Is unfamiliar with new graphics developments and applications.

* Expressed concern for the unsophisticated users in small organizations or small branches in large organizations, that require a lot of assistance in utilizing census data. There should be some way for a user to produce some quick exploratory work in only a day or two of planning. Would like much faster access to data; delivery is slow. Also would like more flexibility in data and approves of the notion of a small common denominator. Often needs different sub-populations and geographic boundaries for different purposes and has a problem with Census's divisions. Complained about having to alter the existing data too much to meet his specific needs.

* Would like to have the data available faster. Produces cost estimates of legislation proposals and needs the best available data at the current time. Has a limited amount of time and so has to focus quickly. Perhaps an on-line system for non-programmers with an easy access to a big data base might be the answer. Another problem is trying to locate the data. Suggested some sort of data library, perhaps another on-line system which points to where the data could be found would be helpful. Would like the Census Bureau to maintain its professional integrity, as well as treat its users more equally. Recognized that units of analysis are always changing and that constant updating is

* What can be accomplished by the time of the 1980 or the 1990 census; what research tools exist and can be immediately utilized.

* The Bureau should be more concerned with user needs.

* The group should think about goals far in the future.

* How to approach idealistic goals with finite resources.

* What research should be done by Census, NSF, or others; should currently fundable research projects be considered?

Following this general discussion, individual users in the group were each given a brief time to explain their own interest in, needs for, and concerns about, data organization. Their responses can be summarized as follows:

* Wants a more detailed public-use sample at a smaller geographic level, but does not need any more software. The Census Bureau should not get involved in selling software; what seems most important is getting results as quickly as possible. Also, the available data are getting farther and farther away from what a city needs.

* Involved in planning for the 1980 census. Interested in data organization suggestions, how the Bureau can satisfy its users more fully and what its job should be in research. Would like ideas for improvement of the census internally as well as for services users need externally.

* Works with population samples ranging anywhere from 2,000 to 1,500,000, in a planning capacity. Would like to see more consistency in data and better documentation of census data and how they are organized. Frustrated when determining the difference between census first and fourth counts due to confusing documents. Would also like greater distinctions in race, such as black vs. brown. Interested in more detailed information at the census tract level.

* There is a barrier when dealing with the smallest geographic common denominator; interested in county information but that is not always the case. Called for a diversification of

required to keep abreast of what is going on.

* Does not want the Census Bureau to get into the software business. Reorganization and greater consistency of data are needed.

* Concerned about what kind of software is needed to help users. Some of the Bureau's internal work might be useful to outsiders in solving some of their problems. The Bureau has a number of areas where improvement is needed, especially in its relationship to outside users.

* Interested in reorganization of data. Has problems with public use and summary tapes of the nature discussed by other participants.

* Data content is insufficient. Would like to see the data edited and documented more effectively, and users be advised promptly of data changes. Would also like more group data and tapes available on more of different structures by different characteristics and areas. Has specific and varied interests, and structures are always changing. Asks for more data and more flexibility in the data available.

* Has problems converting data from non-machine-readable to machine-readable form and would be interested in software that could make this conversion. Sees the household as an important unit of analysis; has a great need for more household data at many different levels of geography.

* Users need specialized information for specific areas produced by people qualified at manipulating gigantic data bases, and flexibility that allows the aggregation of people and geographic units. There is a need for greater detail at smaller area levels, plus the ability to strip off specific things of interest from census data, all as soon as possible. It takes too much time to wade through unwanted information to extract needed data.

* There is a need for software research to provide flexibility of data; a data base is an important step for this. Data

base technology is very important in relation to user-defined areas of analysis. The Canadian approach is data bases with links, although sequential files are lacking. There are microdata down to household and person levels, and normalized rectangular files are produced for users. Some link keys are provided that users can tie into. A lot of customized work can be done that includes data provided by users.

* Much of the technology, e.g., geographic base files, UNIMATCH, etc., already exists to solve the problems discussed. The Census Bureau should avail itself of this technology in solving many of its users' data problems. The time has come for the Census Bureau to get more involved in distributing specialized data to its broad spectrum of outside users. There is a need for data at the person-within-household level, in which the pexson is the basic unit but has a link to his household with some kind of identification of the type of household. The structure of the 1970 housing file is unsatisfactory; the person file and household file appear to have been done by two completely different groups, i.e., more cohesion is needed. The Bureau should have a data dictionary similar to one provided by the NSF.

* Users need flexibility of data, available quickly, aggregated in a variety of ways, and in small and large groups. Users want to be able to submit a request to the Census Bureau and get exactly what is ordered. Cross-tabulations are fine, but availability and accessibility are the keys. There is a need for rectangular files. Cross-case analysis would be a useful tool.

* The Census Bureau must get into a data base system so it can handle users' requests. In view of the significant time lag involved in this process, perhaps there could be a public-access data base system through which users could get directly at the data without having to go through Census bureaucracy. The Bureau cannot presume to guess the cross-tabulations that people need.

* A data base system should be subjected to cost/benefit analysis, and the state-of-the-art in data base technology

should be examined. Data might be considered along with data organization. What is the cost of multiple structures and adding identifiers to data? Should the Census Bureau transmit raw data or the results of processing? (There are ways to measure this). Should data be given on a magnetic tape or perhaps over a network, such as telephone lines? Is mylar tape wanted, and how would this be decided? What about modes of storage? In any case, there are many things to consider before making any significant changes. Better documentation is in order, perhaps in the form of software, and data definitions for processing should be included.

Presentation of Recommendations

In presenting their recommendations to the full conference for review and general approval, members of the data organization group pointed out that the objectives set for discussion of software were flexibility, accessibility, time dimension and modes of storage. The group did not specify long-term or short-term goals, nor apply these measures to any programs. Expense, difficulty and serious technical constraints are involved in this area, but there should also be an awareness of the research already reported in the literature. Further, the conceptual differences among techniques need to be understood. There was a consensus that there should be more work in the area of disclosure analysis to determine how more data can be released and still maintain acceptable levels of confidentiality.

Recommendations

Bureau of the Census data are an invaluable national resource. Our recommendations are intended to achieve modern and efficient use of this resource by the broad and varied spectrum of users dependent upon it.

There is a real concern that, failing aggressive and well planned changes in the Bureau's perceived mission and procedures, there is a significant risk that it will be unable to meet the obligations placed on it in the 1980's. The specific areas of concern include:

* Incomplete knowledge of the needs of present and future external users of census data.

* Lack of forceful developmental efforts to ensure that state-of-the-art technology is brought to bear on meeting defined user needs.

* Lack of a systematic delivery system geared to a diversity of users with a wide range of technical and professional capabilities.

In order to serve these users, we therefore make a set of recommendations including a set of technical innovations which would lead the Bureau of the Census to take advantage of modern data organization techniques. We also recommend establishment of an equally innovative institutional setting which will insure access to Bureau of the Census data by all segments of society requiring such use. The thrust of the technical recommendations detailed below is toward greater usability of and access to the full complement of Census Bureau materials. Although we fully recognize that no data organization schemes, delivery systems, or presentation techniques can be allowed to violate individual confidentiality statutes, we nevertheless believe that current access to microdata can be greatly expanded while protecting this confidentiality.

Further, we are aware that no existing security system is failsafe, including the present one. However, careful security systems can be constructed, while permitting greater access than is currently the case, to the socially critical information contained in the data files within the Census Bureau.

The thrust of the institutional recommendations was essentially:

* Monitoring user needs.
* Providing user training.
* Giving timely service.
* Pricing to support user access.

Whether organized inside or outside the Census Bureau, the institutional setting might entail:

* A national census data center and/or
* A consortium of users, and/or
* A national network.

Each of the above should be considered and justified in terms of cost and the best ways to serve the user community.

Technical Recommendations

The Census Bureau should research alternatives so as to develop and implement techniques and software to provide the following capabilities:

1. Flexible reconstitution of data about people into a variety of significant social units, such as families, households, dwelling units, etc. This will entail developing and retaining data that relate the person to the designated social units. An example of one step in this direction is the recent Bureau of Labor Statistics concept of a "person in a family."

2. Extraction in machine-readable form of the full array of census data aggregated according to user-defined geographic areas. This data-extraction capability should correspond to the full range of information now available for standard Census-defined geographic areas.

3. Efficient access to and transmission of selected user requests concerning:

   * Specific places.
   * Specific types of people.
   * Specific characteristics.

   This will require that the Census Bureau aggressively promote research on advanced data-base management technology.

4. Deployment of timely, accurate, portable, machine-readable data directories.

5. Provision of user-oriented documentation and training material on data and software capabilities geared to various levels of technical and processing proficiency.

In addition, the Census Bureau should explore the costs and benefits of developing and maintaining a time-series data capability on both a forward-looking and an historical basis.

NSF/CB/ASA Research Programs for Fellows

Individuals should be assigned to explore technical as well as cost benefits and alternatives for:

1. More advanced disclosure analysis techniques to allow larger volumes of detailed public-access data.

2. Development of time-series data base capabilities.

3. Gathering and publishing information on present and projected Census

data use in order to determine alternative data organization strategies and delivery systems.

This work should review previous Bureau data-use research, and recognize that projected data use is influenced by present data organization strategies.

Data Tabulation Group

Discussion

In opening the discussion the group recalled the general goals set for all the groups and stated them as:

Short-term: Role of the research fellow to visit the Bureau of the Census. What would you have him do?

Long-term: What ought the NSF to fund in order to promote civilized analyses of Bureau of the Census data? Are there enough research projects having common needs that general software development will pay off?

It was noted that the responses should be based on what users (rather than the Bureau) want to do, but the Census Bureau would have input to the dialogue also.

Rudolph Mendelssohn, Assistant Commissioner of the Bureau of Labor Statistics (BLS), discussed a paper he had prepared and which had been distributed, on his agency's experience with generalized tabulating systems. They use TPL (table-producing language); it may be inefficient, but is widely used in place of programmer resources. He felt that it is essential to (1) identify the end use of the data, and (2) develop the necessary software. The BLS writes the user manual first, then the language, then the routines.

Gary Hill, Director of Information Systems for CACI (Consolidated Analysis Centers, Inc.), who had also prepared a brief paper for the group's discussion, said that his firm has generalized information systems that emphasize processing efficiency and has had favorable experience with data base dictionaries and interrecord analysis. He noted the problems of statistical accuracy inherent in correlation analysis, and suggested there be research in correlating household and person variables.

In the discussion that followed, it was felt that the problem lies in the basic statistical assumptions (interpretation of values), where the unit is the same for the observer, but differs at various hierarchical levels. Some statisticians are working in this area now. Several data users reviewed their approaches to census data and how editing and extraction were carried out to arrive at end products. Among the needs and individual recommendations voiced were the following:

- * Photocomposition software.
- * Portability of data between software systems.
- * Hard copy available at the local level (for small governments).
- * Clear statements for the end users regarding the Census Bureau's allocation, imputation, and suppression practices.
- * Indications of inherent problems in the data or the software (users often lack hardware/software compatibility).
- * Cross-tabulations wider than the Bureau's printed output.
- * Cross-tabulation in such a way that further work with the data is possible.
- * Attention to the microtechniques used in tabulation algorithms; time vs. space tradeoffs.
- * Make users aware of nonsampling error.
- * Software packages through which the tables come close to tabular analysis and capture multiple-regression coefficients.
- * Tools that involve use by noncomputer specialists.
- * Focus on types of software, and determine what can be done in these five fields:
  - Tabulation from basic records (use the Census Bureau's system if it is quick and cheap).
  - Provide a general tabulation system for the public-use sample.
  - Make the basic record tapes available for tabulation.
  - Computer mapping and charting.
  - More sophisticated statistical analysis.
- * Identify the user and the software available to him.
- * Good, documented quality checks in software that has the ability to check and impute.
- * A table package for generating machine-readable files and dealing with the missing data.

It was noted that tabulation definitions vary, and it was suggested that it would be more appropriate to consider all functions in processing, such as maintaining the universe, sampling, response control, editing and screening returns, cross-tabulating, analysis and presentation. This would allow dealing with each more efficiently. Perhaps what is needed is that the Bureau's tabulation be done in such a way that other things can be done with the results.

The need for software to handle hierarchical files and the time-series processing and analysis was noted, and it was also pointed out that the National Bureau of Economic Research, the RAND Corporation and the Massachusetts Institute of Technology all have software for (2).

Users were asked to define what is "acceptable" data and how they should be presented, and do the same thing for software. Should the Bureau work on existing packages available outside and act as a clearing house for them? Responses could be that the Bureau simply should organize its data in such a way that they can be used with existing packages or that the known tabulation packages and their respective characteristics be listed. A visiting research fellow might try to identify the commonalities or uniquenesses of user needs so they could be linked with package capabilities; he could help simplify the match and decide what training (if any) would be needed so that the user would be best served. He also could identify what the Bureau would need to do in providing the data. This could involve both economic and demographic programming.

The discussion included the subject of installation and training needed for systems, the costs involved and what happens when the user complains about a system in place. Questions were raised. Does the discussion imply that the NSF should stimulate the supply or the demand? Are there too many systems and too few users for each? Should users be informed about the packages that are available and their problems with them be investigated? Discussion then turned to the Bureau's generalized tabulation system proposal, in which it was cautioned that the Bureau should allow the system to evolve locally, and to what a visiting fellow might do at the Census Bureau.

The question of the Bureau developing a data base dictionary that is readable by various systems was raised. This dictionary would require continual updating and the problem of how to make this automatic could be addressed in a research project. These things are being done, but recommendations are needed on exactly how. One suggestion was to put the dictionary in codebook format and make it available for reading through interface packages that users might have without having to use codebooks themselves. Vendors will produce codebooks, but the Census Bureau should be motivated to enhance the utility of its data. Should the Bureau take on the task of making these data more usable or let the vendors do that, since the Bureau has its own needs as well?

The group then turned to formulation of its recommendations, focusing on (1) the areas for research and development needed to better satisfy users' requirements, and (2) tools or access to tools for further use of machine-readable data, either directly or through a distribution center. It also was felt that it should be made possible for a user to designate a submodel when suppression occurs. There was a discussion of suppression, random rounding and "noise" injection, and there was sentiment in favor of research for alternatives to all of these. It was felt that a system can be devised that permits greater detail than is presently available and still preserve confidentiality.

There was general agreement that data should be as portable as possible, and that there should be a machine-readable dictionary in well documented format (e.g., compatible with SPSS) and well tied to the data elements. A subset of the dictionary could be used for translation programs and a format statement. There were differences of opinion as to whether it should be possible to run this dictionary on all kinds of computers.

There also was disagreement as to whether the Census Bureau should distribute generalized systems, because this might entail servicing them as well. It was suggested, however, that the Bureau should create a system, implement it and then consider the problem of distribution. If the Bureau develops extraction software this should be made as portable as possible, being written in ANSI COBOL or COBOL. The group thought that the Bureau should develop a generalized extract program and a modified data dictionary with an eye to their subsequent portability. It also should be able to respond efficiently to demands for extracts. There was some dialogue over the cost of a tabulation program equipped to do extract work, with estimates running from $300,000 to $600,000. While this was deemed to be expensive, the alternative might be anywhere from 500 to 5,000 Federal contracts in various parts of the country that would have to include funds for independent software for this purpose.

It was felt elsewhere that a Federal agency such as the Census Bureau has an obligation to make its software known to the public, but that it should not be in the software dissemination business. On the other hand, the agency uses tax money to build a system for its own use, so the system ought to be usable outside the agency for maximum cost benefit. There was no agreement on this topic. One possibility is that vendors should be stimulated to produce their own interfaces with an agency system. Somewhere, however, there should be an effort to bridge the gap between a Census Bureau system and local users.

This discussion led to tentative recommendations that there be an investigation of the need for software to transform data for use in a generalized tabulation system, and of the need for corresponding dictionaries. It also was suggested that the Bureau generate various recodes of the items in its delivered tapes; this would avoid repetitive recodes that might be reflected in the dictionary. The recodes and associated headings and stubs could be supplied in the dictionary, together with a hierarchical key understandable to the system. This is partially available in the START system, but not in UNIVAC. One member recommended that the Bureau proceed to make generalized tabulation software available to users, either in the form of access or programs with support. A visiting fellow might be asked to assess the demand for such software, or at least evaluate the potential. Several participants called for documentation of this software so that users could implement it without difficulty. There was some disagreement as to whether the Bureau would be obligated to document beyond its own needs for internal use.

Possibly if four or five heavy, knowledgeable users of census data jointly advised the Bureau on the development of usable extract and other programs, the NSF might be interested in underwriting some of the group costs. There were divergent opinions on this, but a consensus that someone should make this possible.

It was suggested that generalized tabulation software in the Bureau should be developed with an eye toward it becoming part of the public domain, and the group was told that this is one of the Bureau's objectives, given input from users as to the directions such software might take. There was a feeling that the Bureau should make a greater effort toward this end, and that users should be assured that there are adequate resources for providing the detail they need once the software is available, e.g., output tables for further analysis, additional computations (medians, order statistics, etc.), and the capability to handle as input records that require file manipulation.

There was a discussion of how all this could be brought about, and it was suggested that the NSF might take up the issue with an ongoing organization such as the Association of Public Data Users (APDU). This could be a vehicle for interaction with users concerning the tables required to meet their needs. It was suggested, on the other hand, that the Bureau already has channels for such dialogue. One proposal was that the NSF might make it possible for users to spend time at the Census Bureau so that they and the Bureau staff would have a better grasp of each other's operations.

Looking toward 1980, the initial investment in generalized systems would be very great unless the files are made available in more usable forms than they were for 1970; vendors would hesitate to fill gaps between the Census product and user capabilities. Might the NSF establish and support an activity that would ensure adequate planning and appropriate allocation of funds to obviate these gaps? The activity might be lodged in the APDU to ensure wider involvement. There was a discussion of whether the APDU is capable of such a function.

A possible general recommendation that would take into account "exploding" technology, and the need for technology for minicomputers was discussed briefly.

Presentation of Recommendations

In presenting his group's recommendations to the conference, the group chairman stated that they had rejected a comparative evaluation of tabulation systems because the variables--bounds, environment, objectives, equipment, etc.--are too great. It was felt that there is a residual gap between the development of needs in the market and of services in the Census Bureau; this gap merits further investigation. It would be valuable for users to visit the Bureau for short periods, and vice versa, to go through a variety of work using census data; further, there should be interchange involving such organizations as the APDU to try to solve data problems.

Recommendations

The group had discussed the Census Bureau's plans in the field of generalized statistical tabulation. There was a strong feeling among users outside the Bureau that the situation with respect to availability of generalized software and data (other than published tables) was likely to be little better than the most unsatisfactory situation which was obtained in the past. Special mention was made of the need to improve services and products of the 1980 Decennial Census compared to that of 1970.

In the short term (for the next 3-4 years), the group appeals to the Bureau to maximize its efforts to respond to user needs with respect to machine-readable data and appropriate tabulation software. Failure to do this will lead to continued problems such as those that existed ~~after~~ the 1970 census--namely, continued parallel and redundant efforts by many users (often supported by Federal funds) to overcome deficiencies, loss of information, failure to use information, etc.

Special mention was made of machine-readable data dictionaries, which this group felt to be of fundamental importance, especially for the 1980 census. The group requests that the Bureau work with existing groups such as the Association of Public Data Users (APDU), the Federal Statistical Users' Conference (FSUC), IASSIST, etc., that have already addressed the subject of terminology, conventions and definitions, in order to ensure the data dictionaries are meaningful to users. The Bureau should also provide as detailed information as possible on its own data dictionary plans to the user community as soon as possible.

For the longer term (1980 and beyond), the users among the group agreed to work through their professional organizations to bring the needs of the user community to the highest possible forum. It was felt that the U.S. Congress must improve its perception of the value of Census data.

The group recommends that the Bureau continue its efforts to close the gap between supply and demand for Census products (other than published data) in order to avoid the problems outlined above.

This sub-group recommends that the NSF support an investigation into ensuring the adequacy of planning and allocation of appropriate resources to meet identified user needs.

Further specific topics and conclusions of the sub-group are as follows:

Impact on Bureau of the Census Summary Tabulation Plans for Proposals to Meet User Needs

* Ensure that general tabulation software provides tabulations needed, i.e., no information is lost in the treatment of suppressed data (privacy versus maximizing information at detailed geographic levels); all information necessary for subsequent analysis, including (a) output tables for further analysis (provided in useful formats), (b) capability for additional computations developed while tabulating (medians, order statistics, etc.), and (c) capability to handle (as input) records that require manipulation.

Data Portability

* Produce a machine-readable data dictionary that includes recodes, definitions, etc., and provides easy mapping to data elements.

* Ensure efficient and effective management of updates to the data dictionary and of its distribution to users.

* Support the development, with an eye to subsequent portability, of generalized extraction software that will provide automatically a modified data dictionary.

* Investigate the need for software to transform data and create dictionaries to use generalized tabulation systems.

* The Bureau of the Census should generate various recodes of items in delivered tapes to avoid repetitive recoding (needs to be reflected in the dictionary).

* Efficient mechanisms and procedures should be established to extract data for users and to manage the response to such requests.

* Minicomputer applications should be considered in planning for data portability.

Modification of Generalized Tabulation Software Development Toward Eventual Dissemination To and Use In the Public Domain

* The group applauds Census Bureau plans to elicit information on the needs for features and documentation to facilitate this, but strongly confirms its recommendation that the NSF support an investigation into ensuring the adequacy of planning and the allocation of appropriate resources to meet identified needs.

The Group Requests the NSF to Support Research and Development Into Efficient and Effective Techniques for the Generation of Statistical Tables

* Such research ought to consider what statistics (e.g., cell medians, quartiles, etc.) can be easily computed along with the tables to give a more complete description of the data's patterns.

Data Presentation Group

There are two distinct areas of data presentation, the first dealing with machine-readable forms such as tapes and the second, the noncomputer-readable final product such as microfiche, film and paper-copy graphic displays. There is a need to focus on users' requirements for census data as well as on software that should be developed. It was determined by the group that software for data presentation falls into three categories: routines that produce graphics, those which organize the data, and routines that prepare data for graphics. Sophisticated software already exists to produce graphics but is needed in the remaining categories.

Education and Communication in the Area of Data Presentation

A lack of education and/or communication with respect to the area of data presentation is a major problem. In the discussion it was noted that data presentation is not a visual process alone, but that an understanding of the data needs to be included. Footnotes and explanations that accompany visual material tend to be shortcut. One hazard noted was that the printed report is an excellent means of promoting an understanding of data, but that it is ignored when it accompanies graphic material. Computerized documentation is a partial solution to the problem, but often users will ignore a more detailed printed report in favor of condensed, automated documentation. In the absence of documentation, users interpret graphic output as they see it. A well organized, readable book might be sponsored, showing a broad spectrum of Census data uses; perhaps a comic book and/or film approach would be appropriate. Interaction and involvement were cited as good vehicles for education, and perhaps the concept of the Census Bureau's DIME workshops could apply to the area of the use of census data.

Various methods of computer-assisted education and communication were discussed. Microfiche could be produced at a central facility and distributed among users. The benefits of microfiche include low cost and easy accessibility. The use of data machines with a CRT (cathode ray tube) and cassette capability as a means of disseminating information was suggested. These units are inexpensive and have the benefit of analysis as well as display functions. An interactive system that could lead the user through requests for Census data as well as provide educational facilities was suggested. Problems with the interactive system approach include greater expense, limited accessibility, and a reluctance on the part of State and local users to fund timesharing rather than a capital investment.

The Census Bureau might well take advantage of the motivation that exists at local levels to aid in the implementation of an educational process. The Bureau could supply educational support to a State that commits itself to the program and the State then would be responsible for the distribution of information to local users.

Data Selection and Requests for Data

The areas of data selection, presentation and education are inseparable, as shown by two different directions that the data presentation process takes as a result of a lack of knowledge. It was noted that the uneducated user often requests a "dump" of all available data in a rough form in order to determine which subset of the data upon which to focus. Once the subset has been determined, the user then requests more sophisticated displays. The other extreme is the user that initially requests a small subset of data to be presented, only to learn that more is available, resulting in further requests. Education as to the availability of data and the means of presentation would offer a partial solution to the problem.

Several methods were suggested to aid in the selection process of the subset of data to be presented; one was that software should be developed to select subsets of data. Problems with this include hardware limitations of some users and the expense involved in developing and implementing a software solution. Microfiche was suggested by another participant as a possible alternative in light of the expansion of microfiche capabilities. Data from summary tapes could be stored on microfiche, enabling a user to select from the available data. A participant suggested that a regional processing center could exist with the hardware and software necessary to provide data to the community.

It was observed that the selection process controls the level of presentation and also the analysis that can be performed on the data. Data possibly should be presented without analysis, leaving that for the user to do.

Another problem seen is in the timing of requests for data. Following some Federal announcements, many requests were received. Software could be developed to facilitate the handling of data requests, which would also avoid duplication of effort in the case of commonly used reports. An interactive system could supply the requested data. It was suggested than an area of research might include defining the classes of commonly used data and also the means of their presentation.

Another means of facilitating the processing of data requests might be to sponsor a legislative analyst at the Census Bureau who would be responsible for surveying all legislation and guidelines pertaining to data requests by users. He could also determine the Federal programs that the user might qualify for.

Data Editing

There were several complaints about the lack of software in the area of data editing, i.e., getting the data in a format that is useful for their purposes. A relationship needs to exist between graphic packages and a data base management system, which would facilitate the use of the existing graphic software. One area of research could be the problem of organizing large amounts of data for graphic presentations.

Different data areas by Census and the user are a major problem. One application that was mentioned was that of forecasting future equipment and manpower needs or demand for a product. This requires the ability to overlay Census and user data, and the procedure is very difficult when the two data areas overlap. Perhaps a smaller census data tabulation unit could be determined which would allow users to aggregate Census data up to their particular data area. It was noted, however, that a trade-off must be made between more data for large areas and less data for smaller areas. The smaller the area, the greater the occurrence of suppressions to avoid disclosure.

The problem of differing data areas is further compounded by the poor coordinate quality found in Census files. The user typically must convert Census DIME files to user polygonal-area files. Several participants complained that the coordinates found in the Census GBF/DIME files are very inconsistent and that a good coordinate system is one of their functional requirements. It was stated that the process with which coordinates are edited at Census is too cumbersome to be practical, and that the Bureau lacks incentive in this area because coordinates are not used in its own applications of DIME files. Research exists in this area; the Arithmicon system,presently in the research stage at the Census Bureau, provides an interactive capability for editing and maintaining DIME files.

It was suggested that research should be conducted in the area of Census data presentation form. A different form might result in easier conversion to user data areas. Raster form was discussed as a possible alternative, as that field is rapidly expanding. Valid areas of research would be to investigate the level at which Census should distribute data in raster form, as well as raster vs. polygon vs. DIME forms for distributing data. Data files could exist at different levels, perhaps at as many as five. It was mentioned that perhaps the Bureau should not get involved in the area of providing data for areas other than an agreed-upon unit of issue.

Color and Graphics

Graphics are the final end product for many data requests and are a very popular means of presenting data. Although the group agreed that sophisticated software already exists to produce graphics, it was suggested that research needs to be conducted in this area. One participant suggested research into the most frequently requested types of graphs and visual presentations. Another suggested research to determine which subsets of data should be graphically presented.

The concept of color with respect to data presentation was discussed and research was suggested in this area as well. Research might include experimenting with color and making comparisons to determine what is most effective. Everyone has a different concept of color; the same color can imply different meaning to different people. Another noted that users often state exactly which colors they want in their presentations. It was pointed out that quantitative scale mapping is not adapted to color. A participant felt that the advertising field has already performed much research in the area of color, and perhaps what is needed is research of research.

User Interface and Service Organization

Many participants expressed desires for an automated user interface to ease the process of presenting Census data. An interface is needed between local, State and Census data. The need for development and marketing research in the area of a common user interface was discussed. Such an interface would ease the problem of using Census data for the nonsophisticated user. It was suggested that the existence of an appropriate level of standardization vs. a limit in flexibility should be investigated. A user interface with a query capability would provide facility between Census data in a raw form, subsetting and aggregation routines, and graphic-analysis routines. It was recommended that the development of user software should be keyed to a data dictionary, which would enable it to be flexible in case of format changes. User software should be machine-independent.

The idea of a service organization to provide software services and a user interface was discussed. The service organization would be responsible for distributing data in various forms that would facilitate matters for users of Census data. Listings of software applicable to the use of Census data could be maintained by the organization, in order to refer users to appropriate consultations. It was questioned as to whose reponsibility such an organization would be--government or industry. Concern was expressed that perhaps government might be interfering with private industry in this area. There was some feeling that the Census Bureau's first obligation is to provide data and that software development must be at least secondary.

Presentation of Recommendations

The group asked that consideration be given to instructing local users how to cope with Federal program applications that require census data for small areas. If all software were to access data via simple dictionaries or more complex data base management systems, there would be far-reaching effects on software development. Also stressed was the fact that training and education are major requirements for effective use of census data and for the development of useful, user-oriented software.

Recommendations

Needs of users and alternative modes of presentation are both extremely diverse. Some can be directly addressed by short-term recommendations for user-oriented software, while others require longer-term efforts in which information must be gathered before software recommendations can be formulated. The Data Presentation Group considered a broad range of possibilities, and its recommendations reflect concerns shared by the other two groups. The overall theme is flexible and effective public access to census data. We have identified two major areas in which public access can be facilitated--user education and technological improvements. Under these major topics we have listed a number of specific gaps or omissions to be dealt with. We also feel strongly that the technical program should be integrated with the communication program, and that the integration of specific technical activities is essential to the objective of facilitating public access.

User Education

1. Materials (various multi-media forms) should be developed for the purpose of educating/communicating the use of Census data. Training courses should be developed involving computer-assisted instruction, movies, videotape, programmed learning texts and case studies.

2. We recommend that the research fellow be a trainer to develop a specific training program for census data use (technical and professional). See recommendation 3.

3. Investigation should be done on users' needs and desires for output media, in order to determine products (e.g., slides, paper maps) to be produced.

4. Research should be encouraged in display techniques (e.g., color) for quantitative information.

Hardware

1. Research should be conducted on the potential of new processing technology (e.g., terminal access and mini- and micro-computers) in the analysis of census data by users with limited resources, and the implications of that potential on prospective Census data-documentation techniques.

Software

1. All software developed for users should access data via a data dictionary to remove format dependencies from programs associated with reading census files.

2. Software should be developed and made available by the Census Bureau for handling the most basic and simple types of data retrieval and presentation.

3. Software should be developed to present data about change through time. (A data base should be developed which defines changes and equivalencies in statistical areas).

4. The software developed by the Census Bureau for its processing should be documented, and also made portable and available where feasible.

5. Geographic base files should be developed to facilitate time-series-analysis of small-area data and to allow direct access to census data via independent geographic coordinates.

6. Research should be conducted to determine the special machine-readable files (extract files) and extraction programs that should be produced for special program compliance.

Data Requirements (Geographic)

1. Higher standards are required for coordinates in geographic base files (GBF's) in order to allow user specification of tabulation areas in terms of coordinates. Specifically, GBF coordinates should be corrected topologically and cartographically.

2. A machine-readable data base should be developed which defines changes and equivalencies in statistical areas.

3. The Census Bureau should provide separate machine-readable files of spatial definitions (e.g., polygonal coordinates or raster) for all statistical areas.

Organization

1. Investigate the possibility of a user clearinghouse(s) for the availability and development of user software. Set up a clearinghouse for user software and investigate the possibility of developing and supporting user software.

An ongoing assessment of user needs for software should be conducted. Compile user comments and evaluations of software, and form a users' group on user software.

2. We support the concept of summary "tape" data processing centers.

V
DISCUSSION AND ACCEPTANCE OF RECOMMENDATIONS BY THE CONFERENCE

Submission of Preliminary Group Recommendations to the Conference

The third day of the conference began with the submission of the preliminary group recommendations to the plenary session. During the opening discussion concern was expressed that the Census Bureau still is using 1950's techniques and needs modernization. Some portion of the members wanted to say that the Census Bureau is "in trouble," and that the cost to catch up, in the face of political and social needs, is increasing rapidly. These needs cannot be met with current technology.

There was a discussion of the respective responsibilities of users and the Bureau with respect to filling the technological gaps foreseen. The consensus appeared to be that the average user needs to be trained to use the tools at hand, and that the Bureau, as it develops techniques and software, should constantly recognize users' needs and abilities to keep pace.

It was observed that the Bureau plans to replace its hardware completely by 1982, and this hardware will be geared to data base management systems. The Bureau would like users to spell out in detail what their data needs are so that the Bureau's specifications can match them.

It was agreed that it would be helpful to recommend the first explicit step(s) to the NSF, and the groups returned to their individual sessions for further considerations.

Acceptance of Final Group Recommendations by the Conference

Upon completing the additional deliberations by individual groups, each group's final recommendations were read and discussed by the conference as a whole. Some language was modified to reflect consensus positions, and the approved texts appear above in Section IV. The deliberations in the final individual group sessions were not reported; only the plenary discussion which follows below.

Comment was made that what users tend to do is limited by the technology available. The history of extensive analysis that led to research and development in the Census Bureau during the 1960's and still conducted by its Center for Census Use Studies was cited, but note was taken that some projects that should have been carried forward were not.

There was a discussion as to whether the recommendations should be time-oriented. It was felt that the conference may have the 1980 Decennial Census in mind, whereas there are economic censuses, surveys and other statistical programs being carried out in other years. It was agreed that "short-term" might be interpreted as 3 to 4 years, but with emphasis on 1980.

Question was raised whether this conference or the presentation group might be the beginning of a user group to address in more detail the various items suggested. Another suggestion was the establishment of a clearinghouse to follow up on the conference agenda items, noting that the Census Bureau, its oversight committee in Congress, and the Office of Management and Budget are only some of the "actors" involved. Perhaps there might be a follow-up conference in a year or two. The review process that has been set up for the ASA, the Census Bureau, and the NSF's four joint projects' results was mentioned and also that there will be general meetings with the Census Advisory Committee of the ASA. It was noted that there will be efforts to formalize user support as much as possible and the report of this conference will be given wide circulation. An offer was made to monitor progress a year from now and report through a user journal.

It was suggested that a good use of the conference resources would be to look at the purpose, process and impact of the 1980 census data products and software on data processors. Training modules may be needed for various user groups, together with data and use guides.

A question was raised as to whether the Bureau would feel the conference's attitudes are unjustified or distorted, and whether the Bureau is worried about its software products and their distribution. In reply it was stated that discussion from all standpoints is being encouraged. The Bureau will receive the recommendations and be glad to state what is being or can be done to carry them out. Another participant felt that the conference is supportive of improvements. It would be helpful, however, for the Bureau to tell how it will use the conference information and what it is doing.

The following resolution was then passed:

> "The conference expressed its desire that the Bureau of the Census be asked to advise participants through the American Statistical Association of its plans to respond to the various recommendations contained in the report of the proceedings of the conference."

APPENDIX A

NAMES, AFFILIATIONS, ADDRESSES AND BACKGROUND OF CONFERENCE PARTICIPANTS

WILLIAM T. ALSBROOKS, Assistant Division Chief, Systems Software Division, U.S. Bureau of the Census, Washington, D.C. 20233. M.S. (Computer Science), Purdue University, 1970. Formerly Programming Branch Chief of Statistical Methods Division of the Census Bureau.

MICHAEL J. BATUTIS, JR., Principal Demographer, New York State Economic Development Board, P.O. Box 7027 - AESOB, Albany, New York 12225. M.A., Duke University, 1972. Has served as demographer with New York State since Duke.

PATRICIA C. BECKER, Head of Data Coordination Division, Planning Department, City of Detroit, 801 City-County Building, Detroit, Michigan 48226. M.S. (Sociology), University of Wisconsin, 1964. Before going to Detroit in 1968 did academic survey research at the universities of Michigan, Wisconsin and California (Berkeley).

JOHN BERESFORD, President, DUALabs, 1601 N. Kent Street, Arlington, Virginia 22209. M.A., University of Michigan 1952. After military service he was with the Bureau of the Census until founding DUALabs in 1969. He is presently Chairman of the Association of Public Data Users Census Committee.

WILLIAM M. BRELSFORD, Supervisor, Statistical Computing and Methodology Group, Bell Laboratories, Holmdel, New Jersey 07733. PhD (Statistics), Johns Hopkins University, 1967.

HUGH FRANCIS BROPHY, Chief, Systems Development and Programming Unit, United Nations Statistical Office, Room 3114 United Nations Plaza, New York, New York 10017. B.Ec. (Hons), Australia National University, 1965. Held Deputy Director of Computer Services and other posts with Bureau of Statistics, Australia and was Project Manager of a computing research centre in Czechoslovakia.

LARRY CARBAUGH, Data Users Service Division, Room 3624 - FB #3, U.S. Bureau of the Census, Washington, D.C. 20233. B.S. Duke University, 1964.

BRUCE CARMICHAEL, Group Leader, Central Data Base Group, U.S. Bureau of the Census, Room 1373 - FB #3, Washington, D.C. 20233. PhD (Computer Science), University of Maryland, 1976. Consultant to General Electric Space Flight Division, systems analyst at NIMH and technical staff member at Bell Telephone Laboratories.

WILLIAM S. CLEVELAND, Member Technical Staff, Bell Telephone Laboratories, 600 Mountain Avenue, Murray Hill, New Jersey 07974. PhD (Statistics), Yale University, 1969. Assistant Professor, University of North Carolina (Chapel Hill) before joining Bell Laboratories.

LAWRENCE E. CORNISH, Chief, Graphics Software Branch, U.S. Bureau of the Census, Room 1529 - FB #3, Washington, D.C. 20233. Michigan and Michigan State Universities.

JACK DANGERMOND, Director, Environmental Systems Research Institute, 380 New York St., Redlands, California 92373. MLA, Harvard University, 1969. MA (Urban Design), University of Minnesota. Was a teaching research associate at Universities of Minnesota and Harvard and served as project manager with Scientific Systems, Inc. and as director of the Environmental Systems Research Institute.

PETER DICKINSON, Director, Data Processing, Center for Demography, University of Wisconsin, 1180 Observatory Drive, Madison, Wisconsin 53706. MA (Sociology), University of Wisconsin 1975. Was programmer analyst with the Center for Demography and photogrammetric surveyor with the U.S. Forest Service.

RICHARD B. ELLIS, Marketing Manager, Information, American Telephone & Telegraph Co., 295 N. Maple Avenue, Basking Ridge, New Jersey 07920. B.A., Hamilton College 1950. Held other marketing positions with AT&T and was supervisor, Corporate Staff, with the New York Telephone Company.

CARL E. FERGUSON, JR., Director, Center for Business and Economic Research, Box AK, University of Alabama, University, Alabama 35486. PhD, University of Missouri 1975. Before coming to Alabama as Assistant Director of the Center was Assistant Director of the Public Affairs Information Service, University of Missouri.

LAWRENCE FINNEGAN, Data Users Service Division, Room 3069 - FB #3, U.S. Bureau of the Census, Washington, D.C. 20233.

JAMES FOLEY, Associate Professor of Electrical Engineering and Computer Science, George Washington University, Washington, D.C. 20052. PhD, University of Michigan 1969. Was assistant professor at University of North Carolina, and with the Graphics Software Branch of the Census Bureau.

WILLIAM H. FREUND, Leader, Systems and Programming Group, Data Services Center, ERS, U.S. Department of Agriculture, Room 456 GHI Building, Washington, D.C. 20250. B.S. University of North Carolina, 1963. Has held a variety of positions in economic analysis and systems design with the Department of Agriculture after graduation from North Carolina.

SHIRLEY GILBERT, Consultant and data analyst, Princeton-Rutgers Census Data Project, Princeton University, 87 Prospect Avenue, Princeton, New Jersey 08540. M.A., University of Oregon, 1946. Was an instructor in mathematics at New Jersey College for Women (Rutgers) and University of Oregon.

WARREN GLIMPSE, Data Users Service Division, Room 3069 - FB #3, U.S. Bureau of the Census, Washington, D.C. 20233. B.S., University of Missouri, 1969. Was Director of Public Affairs and taught at Missouri. Consultant to industry and government on software design and evaluation.

SCOTT B. GUTHERY, Principal Software Engineer, Mathematica, P.O. Box 2392, Princeton, New Jersey 08540. PhD, Michigan State University 1969. Worked previously in applied statistics and data base management system research with Bell Laboratories.

ROBERT D. HARRIS, Deputy Assistant Director, Congressional Budget Office, 2nd and D Street, S.W., Washington, D.C. 20515. B.S., Ohio State University 1960. Prior to joining the Congressional Budget Office was Chief of Information Services with the Office of Management and Budget and held a number of posts in the Department of Agriculture.

GEORGE M. HELLER (Conference Co-Chairman), Principal Researcher, Statistical Research Division, U.S. Bureau of the Census, Washington, D.C. 20233. M.A., Columbia University 1949. Has held a variety of positions with the Bureau of the Census since coming there from Columbia.

GARY L. HILL, Director, Information Systems Department, CACI, Inc., 1815 N. Ft. Myer Drive, Arlington, Virginia 22209. MBA, Indiana University 1961. Has been an officer of Data Use & Access Laboratories, Computer Resources Corporation and project manager at IBM.

DAVID C. HOAGLIN, Senior Analyst, Abt Associates, Inc. and Research Associate in Statistics, Harvard University, 55 Wheeler Street, Cambridge, Massachusetts 02138. PhD, Princeton University 1971. Has been on the faculty at Harvard since 1971 and also served as senior research associate at NBER Computer Center for Economics and Management Service.

HAROLD B. KING, Director, Computing Services, The Urban Institute, 2100 M Street, N.W., Washington, D.C. 20037. B.A. (Mathematics), San Jose State, California 1959. Helped to establish the Association of Public Data Users and was with the interuniversity Communications Council.

FRED C. LEONE, Executive Director, American Statistical Association, 806 15th Street, N.W., Washington, D.C. 20005. PhD (mathematics and statistics), Purdue University 1949. Taught at Iowa, University of California (Berkeley) and Case Institute of Technology. Visiting professor at University of Sao Paulo, Brazil and was on Ford Foundation Education Team in Mexico.

RICHARD G. MAYNARD, Acting Manager, Policy Support and Special Studies Division, House Information Systems, 3641 HOBA #2, Washington, D.C. 20512. M.A. (Economics), University of Pennsylvania 1969. Was with EDP Technology, Inc. and the Department of Defense.

MARK D. MENCHIK, The Rand Corporation, Santa Monica, California 90406. PhD (Regional Science), University of Pennsylvania 1970. Was with New York City-Rand Institute and taught in the geography department at the University of Wisconsin.

RUDOLPH C. MENDELSSOHN, Assistant Commissioner, Bureau of Labor Statistics, Room 2047, 441 G Street, N.W., Washington, D.C. 20212. A.B., University of Chicago 1938. Prior to becoming Assistant Commissioner in 1967 was in charge of various Bureau employment, hours and earnings statistics. Edited the Bureau's journal in that field.

JULES MERSEL, Senior Operations Research Analyst, Community Development Department, City of Los Angeles, 200 N. Main Street, Room 1404, Los Angeles, California 90012. M.S. (Physics), University of California (Berkeley) 1951. Was with the National Bureau of Standards and has had a broad range of computer consulting positions in private industry.

PETER A. MORRISON, Member, Senior Research Staff, The Rand Corporation, 1700 Main Street, Santa Monica, California 90406. PhD, Brown University 1967. Formerly assistant professor at the University of Pennsylvania and a special consultant to the National Commission on Population Growth and the American Future.

MERVIN E. MULLER, Director, Computing Activities Department, The World Bank, 1818 H Street, N.W., Washington, D.C. 20433. PhD (Mathematics), University of California, Los Angeles 1954. Taught and was Director of the Computing Center at the University of Wisconsin. Managed Project WELD at IBM and has been on the faculty at Princeton, Cornell and the University of California.

DAVID M. NELSON, Acting Program Director, Computer Information Systems, 415 Coffey Hall, University of Minnesota, St. Paul, Minnesota 55414. PhD (Economics and Statistics), Kansas State University, 1968. Has been a visiting professor at Boise State University and Hamline University.

NORMAN H. NIE, President, SPSS, Inc., Suite 1236, 111 East Wacker Drive, Chicago, Illinois 60601. Currently Senior Study Director, National Opinion Research Center and Professor at University of Chicago. Was Senior Fulbright Fellow, University of Leiden, The Netherlands and Woodrow Wilson Fellow, Stanford University. Principal investigator for a number of political science projects.

MANUEL D. PLOTKIN, Director, U.S. Bureau of the Census, Washington, D.C. 20233. M.B.A. (Statistics), University of Chicago 1949. Came to his present position from the corporate headquarters of Sears, Roebuck and Company where he was Associate Director, Corporate Planning and Research. Managed the Economic and Market Research Department of Sears and also served as Chief Economist. Was earlier with the U.S. Bureau of Labor Statistics in the Chicago and Washington offices and taught in the evening division of several Chicago colleges.

JOE W. PYLE, Director of Physical Planning and Development, Houston-Galveston Area Council, 3701 W. Alabama, Suite 200, Houston, Texas 77027. PhD, University of Houston 1973. Previously held positions with Boeing Company, Philco-Ford Corporation and the University of Houston.

MELROY QUASNEY, Systems Software Division, Room 1061 - FB #3, U.S. Bureau of the Census, Washington, D.C. 20233.

LAWRENCE C. RAFSKY, Statistician, Chase Manhattan Bank, 18th Floor, 1 Chase Manhattan Plaza, New York, N.Y. 10015. PhD (Statistics), Yale University 1974. Formerly at Bell Telephone Laboratories.

DANIEL A. RELLES (Conference Co-Chairman), Statistician, Rand Corporation, 1700 Main Street, Santa Monica, California 90406. PhD (Statistics), Yale University 1968. Was a member of the technical staff of Bell Telephone Laboratories.

ALBERT H. ROSENTHAL, Rand Corporation, 1700 Main Street, Santa Monica, California 90406. With Rand since 1953. Currently Senior Analyst.

ALFRED J. TELLA, Special Adviser, Office of the Director, U.S. Bureau of the Census, Washington, D.C. 20233. M.B.A., New York University 1959. Has been Research Professor of Economics, Georgetown University and Director, Office of Labor Force Studies, The President's Commission on Income Maintenance Programs.

ANTHONY G. TURNER, Mathematical Statistician and Census Coordinator for ASA/Census Research Program, U.S. Bureau of the Census, Washington, D.C. 20233. B.S. and graduate work, University of North Carolina. Has been sampling consultant to FDA and Population Research Council and was with the Statistics Division of LEA. Served in Census previously as Chief of the Special Surveys Branch.

MEL TURNER, Assistant Director, DBMS, Systems Development Division, Statistics Canada, 12-P, R.H. Coats Building, Ottawa, Canada K1A 0T6. B.Sc. (Hons) (Physics), Queen Mary College, University of London 1966. Has been in several programming posts with both Statistics Canada and IBM (UK), Ltd.

HARVEY WEINSTEIN, SPSS, Inc., Suite 1236, 111 East Wacker Drive, Chicago, Illinois 60601

FORREST B. WILLIAMS, Manager, Marketing and Information Systems Group, CACI, Inc., 1815 N. Fort Myer Drive, Arlington, Virginia 22209. PhD (Geography), Ohio State University 1975. Has been a research analyst with the Census Processing Center, Battelle-Columbus Laboratories and Special Projects Manager for the Behavioral Sciences Laboratory at Ohio State.

ROBIN WILLIAMS, Manager, Display Systems Architecture, IBM, K 54 - 282, 5600 Cottle Road, San Jose, California 95193. PhD, New York University 1971. Worked in optical character and memory systems with Philips research laboratories in England and Briarcliff Manor, New York. Taught at New York University.

PAUL T. ZEISSET, Chief, Data Access and Use Staff, Data User Services Division, Room 3540 - FB #3, U.S. Bureau of the Census, Washington, D.C. 20233. M.A., University of Texas 1969. Has been with the Data Access and Use staff since college.

- - - - - - - - - - -

Management of the conference has been under the direction of John W. Lehman, ASA Conferences Director, with the assistance of Barbara Lindell. Additional services have been provided by the ASA office. The conference was reported by Fred Bohme of the History Staff of the U.S. Bureau of the Census, assisted by Cynthia Agard and Patricia Griffin. Anthony Turner served as Census coordinator for the program.

APPENDIX B

FINAL PROGRAM FOR CONFERENCE ON DEVELOPMENT OF USER ORIENTED SOFTWARE

Stouffer's National Center Hotel
Arlington, Virginia

November 8, 9, 10, 1977

TUESDAY, NOVEMBER 8, 1977

8:00 - 9:00	Registration	
9:00 - 9:30	Welcome and Introduction FRED C. LEONE, Executive Director, American Statistical Association MANUEL D. PLOTKIN, Director, U.S. Bureau of the Census	(Potomac Room)
9:30 - 10:15	Overview of software state-of-the-art in information delivery WILLIAM ALSBROOKS, Systems Software Division, U.S. Bureau of the Census	
10:15 - 10:30	Break	
10:30 - 11:15	Current plans and activities of Census Data Users Division WARREN GLIMPSE, Data Users Services Division, U.S. Bureau of the Census	
11:15 - 12:00	Needs of users from the viewpoint of local governments and other public agencies HAROLD KING, Urban Institute	
12:00 - 1:15	Lunch	(Charleston Port Room)
1:15 - 2:00	Needs for users from the viewpoint of economists, market researchers and others in the private sector RICHARD ELLIS, Market Research, American Telephone & Telegraph Co.	(Potomac Room)

Organization of Data

2:00 - 2:30	Summary of user paper and questions MERVIN MULLER, World Bank
2:30 - 3:00	Summary of Census Bureau paper and questions BRUCE CARMICHAEL, Systems Software Division, U.S. Bureau of the Census
3:00 - 3:15	Break

Tuesday, November 8, 1977 - Continued

Tabulation of Data

3:15 - 3:45	Summary of user paper and questions HUGH BROPHY, U.N. Secretariat	
3:45 - 4:15	Summary of Census Bureau paper and questions MELROY QUASNEY, Systems Software Division, U.S. Bureau of the Census	

Presentation of Data

4:15 - 4:45	Summary of user paper and questions ROBIN WILLIAMS, IBM Corporation	
4:45 - 5:15	Summary of Census Bureau paper and questions LAWRENCE CORNISH, Systems Software Division, U.S. Bureau of the Census	
6:00 - 7:00	Reception	(Charleston Port Room)
7:00 - 8:30	Dinner	(Charleston Port Room)

WEDNESDAY, NOVEMBER 9, 1977

Simultaneous sessions by the Organization (Room 204), Tabulation (Room 110), and presentation (Room 104), sub-groups according to the following schedule:

9:00 - 10:15	Opening statements without interruption	
10:15 - 10:30	Break	
10:30 - 12:00	Discussion of invited papers and opening statements	
12:00 - 1:30	Lunch	(Dewey I Room)
1:30 - 3:00	Proposing and discussion of recommendations	
3:00 - 3:15	Break,	
3:15 - 5:00	Completing recommendations for submission to the full Conference	

THRUSDAY, NOVEMBER 10, 1977

(Resume full Conference)

9:00 - 9:30	Submission of Organization sub-group recommendations to full Conference Discussion	(Potomac Room)
9:30 - 10:00	Submission of Tabulation sub-group recommendations to full Conference. Discussion	
10:00 - 10:30	Submission of Presentation sub-group recommendations to full Conference. Discussion	

Thrusday, November 10, 1977 - Continued

10:30 - 10:45	Break	
10:45 - 12:00	Individual sub-group meetings to review any proposed changes and prepare final recommendations *	
12:00 - 1:30	Lunch	(James Room)
1:30 - 4:00	Acceptance of final recommendations from sub-groups by full Conference	(Potomac Room)

* For this period the Tabulation (Room 110) and Presentation (Room 104) sub-groups will meet in the same rooms they used on Wednesday. The Organization sub-group will stay at the front of the Potomac Room.

APPENDIX C

The Organization, Tabulation and Presentation of Data

State of the Art: An Overview

William T. Alsbrooks
Bureau of the Census

James D. Foley
George Washington University

1. Introduction

The purpose of this paper is to survey the state of the art, from both a hardware and software technology point of view, of the technical and delivery capabilities for

- Data Organization
- Data Tabulation, and
- Data Presentation.

These areas are central to improving access to and use of machine readable Census Bureau data. In the area of data organization, we will talk about the state of the art in Data Base Management Systems (DBMS); in the area of data tabulation, we will talk about the state of the art in Generalized Table Generator Systems; and in the area of data presentation, we will talk about the state of the art in Photocomposition and Computer Graphics.

The sections that follow examine functional capabilities of each of the three individual components; the integration of the three components into a total system; and the delivery of the system capabilities to the end user.

2.0 Functional Capabilities

2.1 Data Organization

The term "database" can be viewed from many different vantage points: its access, purpose, description, content and integration. But all definitions seem to contain three essential and practical characteristics -

- An organized, integrated collection of data.
- A representation of the data which is natural and convenient for users, with few restrictions or modifications imposed to suit the computer.
- Capable of use by all relevant applications without duplication of data.

A data base management system (DBMS) is simply the software that supports such a database. The purpose of a DBMS is to allow users to deal directly with data and relations of data rather than be concerned with sometimes complex storage structures.

As summarized by (PALM 75), the facilities that a DBMS can be expected to provide are:

1) The controlled integration of data to avoid the inefficiency and inconsistency of duplicated data.
2) The separation of physical data storage from the application logic using the data to aid flexibility and ease of change in a dynamic environment.

3) A single control of all data permitting controlled concurrent use by a number of independent on-line users.

4) Provision for complex file structures and access paths such that relevant relationships between data units can be readily expressed and data can be retrieved most efficiently for a variety of applications.

5) Generalized facilities for the rapid storage, modification, reorganization, analysis and retrieval of data so that the use of a database system imposes no restrictions upon the user.

6) Security controls to prevent unauthorized access to specific units of data, types of data or combinations of data.

7) Integrity controls to prevent misuse or corruption of stored data, and facilities to provide complete reconstruction in the event of hardware or software failure.

8) Performance both in a batch mode and on-line, that is consistent, measurable, and capable of being optimized.

9) Compatibility with major programming languages, existing source programs, a

variety of hardware systems and operating systems, and data external to the database.

Figures 1,2, and 3 summarize the capabilities of various DBMS's.

The data base approach is more than merely a different computer technique involving the storage of data and the use of additional generalized software. It involves a new approach to designing and operating information systems and has far-reaching effects well beyond the data processing activities. Data base is a philosophy that regards data as a resource to be managed just as other resources of the organization are managed.

Described in terms of the CODASYL model, this is accomplished by defining to the DBMS, through the facilities of a Data Definitional Language (DDL), the structure and format of data in the data base, the names and descriptions of the data, relationships among units of data, and the methods of access to the data. This definition of the data base is called the schema. Data requirements of applications programs are also defined using the DDL and are called subschema. This can be thought of as the user's view of the data base.

Operations of retrieval, modification, storage and deletion of data are accomplished through a Data Manipulation Language (DML).

The DBMS is directly responsible for the physical placement of data on the storage devices. A Device Media Control Language is used by the system programmer to determine:

1) choice of device by data type
2) physical block size
3) record placement
4) overflow strategy.

Fundamentally there are only two ways of accessing data from the mass storage device. Either the physical address is known so that it can be retrieved directly, or if not known, the relevant part of the data base must be searched. The fundamental physical structuring alternatives are quite limited, although they can be combined in a myriad of ways. The most simple is sequential where the next record required is the next record on the file; it is defined by its position, and its address is of no consequence. Records can be chained together, with the address of the next record in the current record.

Hashing and indexing are both techniques which allow direct access to the desired record, in some cases with just a single access to the file.

The basic physical access methods available to a database system are limited and do not, of themselves, provide the necessary complex file structures. Instead these are implemented by the use of logical structures defined in the schema and interpreted by the system software in terms of the basic structures. Logical data structures can first be classified as any of the following:

1) Simple:. All units of data are independent and of logically equal significance. They can be either ordered or unordered.

2) Hierarchic: Units of data are dependent and can be logically arranged in a hierarchy of levels in which units have a single owner and/or own one or more other units. A hierarchical file is always ordered.

3) Network: Units of data are dependent, but in a more complex structure than in a hierarchy, in which units have more than one owner, as well as own one or more other units.

A variety of file organizations are supported by database management systems for both simple and hierarchical structures. These can be thought of as second-level, or logical structures, since each corresponds to combinations or extensions of the fundamental physical structures. Such organizations include indexed, inverted, multilist, ring, tree, and network structures.

These logical data structures are then used to implement the data models supported by various DBMS's.

A hierarchical data model is a collection of trees in which the nodes are the record occurrences -- in other words, a one-to-many relationship.

This data model can be used in two ways:

1) The selection criteria can be specified as a path through the tree. Some or all of the records along the path are the desired records. Example - IMS (IBM).

2) The selection criteria are specified independently of the tree structure. The tree is then searched through the facilities of an inverted index for the desired records. Example - System 2000 (MRI).

The principal disadvantages of this type of model is that it is often inadequate to accurately model the data. An example of its weakness is its inability to model a geographic lattice. Also, the tree structure makes many retrievals difficult. If, however, a hierarchy is an accurate data model and if most accesses can be expressed as straightforward tree searches, it can be very efficient.

The network data model allows for many-to-many non-hierarchical relationships. The best known of the network systems are those based on the CODASYL (CODA 71) reports.

Superimposed on a variety of physical storage structures is a logical structure called a set-ring structure which links record occurrences. An owner record can have many members. A member record can be associated with many owners in different sets. The primary advantage of the network is that a wide variety of physical and logical structures are provided and they can model most collections of data very well. There are many choices to allow for optimizing performance. There are also disadvantages. For with all the alternatives, come the complications. A network model is very complex, and a user must know a great deal about the actual storage structure to program efficiently.

Examples -	DMS 1100	(UNIVAC)
	IDS/II	(HONEYWELL)
	IDMS	(CULLINANE)

The relational data model is an approach developed largely in the IBM Research Laboratories at San Jose, California. The most significant papers have been by E. F. Codd (CODD 70). The original motivation for this approach was the need for data independence and the need to identify inconsistencies within the database. But it soon became apparent that the relational model, because of its basic simplicity, could well provide a unifying structure for the design of any database system and manipulation language. The user is presented with only one logical structure with which to design a schema and need not be concerned with the complexity of linkages, networks, repeating groups and indexes.

The relational model is a mathematical approach built around two basic concepts. The logical storage structure used is a relation in third normal form, which is a type of relation with the optimal properties for use in a database.

All data in the relational model is viewed logically as a simple table. This is easily understood by the layman and is suited for display on terminals. Mathematically these tables are known as relations. A relation of degree 'n' has the following properties:

1) contains 'n' columns (known as domains);
2) all elements in a given domain are of the same type;
3) each row represents an n-tuple of the relation and contains 'n' elements;
4) the ordering of rows is immaterial;
5) all rows are distinct (there are no duplicate tuples); and
6) columns (domains) are assigned distinct names.

In conventional terms, a relation can best be equated to a serial file containing one record type of fixed length. Thus, a tuple is equivalent to a record; a domain, to all data-items of a particular type in the file.

Tuples are identified by their keys, which are formed from a combination of one or more elements. A tuple can contain more than one combination of elements that uniquely defines it. Each combination is termed a candidate key; the one arbitrarily selected to identify the tuple is its primary key.

A relational model subschema is very concisely defined. It need name only the relations and domains and indicate the primary keys. The user is not concerned with ordering, indexing, or access paths so they need not be defined. In addition, such aspects of the physical data can be altered without impairing the applications using it.

From the user's point of view, and to a lesser extent the implementor's, the major advantage of this approach is its basic simplicity. It is not a system that has grown simply in an attempt to meet user requirements, but an approach from first principles with a rigorous mathematical basis in relational calculus.

The relational calculus is powerful in its simplicity, and its conciseness and clarity make it easy to amend. Programming effort is reduced, particularly in updating, because entire relations can be processed with one relational calculus statement. It is well suited to query handling but it is not concerned with output formatting. Because only relations and domains can be addressed, access control problems are reduced. The relational calculus is claimed to be better suited to optimization and to augmentation with improved facilities than procedural languages based on relational algebra.

By removing many decision-making responsibilities from the user, the relational model imposes additional problems upon the implementor.

The user cannot define network or hierarchical structures. This does not mean that they cannot be used by the system if it is the most efficient means of physical storage. Relations in third normal form could be stored as serial files. However, the number of extraneous fields would produce a great deal of data duplication with possibly unacceptable storage overheads. The problems of amending such duplicated data have not been eliminated. Unlike the CODASYL set structures, there is a wide choice of methods of representing relations in physical storage. For example, a relation can be stored by tuples or domains, or can exist only as pointers from other relations. The ideal implementation should be sufficiently flexible to provide the structures best suited to the particular data and its usage. If it is not, the database administrator will need control over the physical storage structures used for each type of relation.

The disadvantages of the relational model are not clear at this time since there is a lack of practical experience of commercial systems to draw upon, the notable exception being the Honeywell Multics Relational Data Store. Statistics Canada has developed a relational system in which they are quite pleased called RAPID, specifically for processing their 1976 census. INGRES is a relational DBMS that has been developed at the University of California - Berkeley.

Why are we so interested in the relational model? The answer is simple; most DBMS's available today are designed to optimize the retrieval of a large amount of information from a small number of records. In statistical data processing, most often what we need is a small amount of information from a very large number of records.

2.2 Data Tabulation

Tabulation of data is an integral and inevitable part of any statistical task. Whether the tables be created by experienced programmers for large scale censuses or by subject matter analysts for studies involving small samples, this task is complicated, tedious and repetitive. In most cases, a generalized tabulating system simplifies the effort and enhances the final product.

A generalized tabulating system is a series of parameter driven computer programs designed to select, to restructure, to cross-tabulate and to display statistical data. The system is highly user-oriented through the utilization of a nontechnical, nonprocedural, compact, English-like command language that is easy to learn and easy to use. Users need not have experience with conventional programming languages in order to produce a wide variety of tables with minimal programming effort.

The four important components in determining the success of a generalized tabulating system are its

1) tabulating power,
2) ease of use,
3) environmental adaptability, and
4) acceptance.

Tabulating power refers to the ability of a system to produce tables as requested by its user. For example, the computational and formatting ability, and the lucid and aesthetic display capability are fundamental to this criterion. On the other hand, the clarity of the documentation and the design of the user language are central issues concerning the system's ease of use. Environmental adaptability may also play an important role in the decision of choosing a tabulating system for installations which do not possess large scale computers. Transportability, memory requirements, and processing efficiency have effectively eliminated many tabulating systems from being considered for adoption. Other functional features of a tabulating system such as statistical capabilities, linkage to data base management systems and graphical display systems may also be critical.

Finally, a generalized tabulating system can have the power, be easy to use and be adaptable to the environment but then it must be accepted by its potential users. In most cases, this means a change from the practice of custom coding complete programs to the coding of simple parameters. Statistical and economic analysts like this because it means that they can produce their analytical tables independent of programmers. Programmers and programming managers seem not to like a GTS because it stifles their creativity and minimizes their independence in the statistical production process. But, in order for a GTS to be effective, it must be used; therefore, it must be accepted.

What is the state of the art in Data Tabulation systems? Figure 4 shows a selected list of tabulation systems and some of their characteristics. Figure 5 shows a selected list of statistical packages with tabulation capabilities. Much of this information comes from (FRAN 76).

Packages like SPSS, BMDP, DATA-TEXT, and SAS are well accepted and widely used. They provide limited tabulation capabilities in the sense of the number of cells that can be tabulated in one data pass and in their data display options. But, they do provide the analyst with a broad range of statistical routines.

Also of concern is the ability to tabulate large census micro and macro data files, and to format the_tabulation ready for publication.

Several national statistical offices are active in the data tabulation area. Statistics Canada is using four generalized tabulating packages. CASPER, STATAPE, STATPAK and TPL. CASPER was developed in the late 1960's and caught on slowly, but it still has limited use. CASPER has been largely replaced by STATAPE with its expanded capabilities and improved user language. STATPAK supplements STATAPE by providing interface capabilities with Statistics Canada's data base management system RAPID, mentioned in the preceeding section.

Statistics Canada estimates that 70% of all tables are currently being produced using generalized tabulating systems. This figure includes the tabulations for their 1976 Census of Population and Housing.

The U. S. Bureau of Labor Statistics released their Table Producing Language (TPL) in 1974. Today, this appears to be the most widely used generalized tabulating system in the world. It has been distributed to over 150 installations. Recently introduced at Statistics Canada, TPL is already gaining widespread usage.

As do many such systems, TPL uses a codebook or data dictionary to define data variables, their names and their descriptions. This codebook is usually coded by a programmer familiar with the data. It is then used by analysts or other programmers for their table preparation. Data is then referenced by data name, just as with DBMS's. This is a very important feature for table generators, because it allows for data independence and consistency between programs and programmers.

Usage of TPL has increased to a level where today there are over 3000 references each month at the NIH computer center. It is now normal practice at BLS to perform all new tabulations with TPL.

Sweden is using their TAB68; France their system called LEDA; and Czechoslovakia, ISIS. In May, 1977, the Census Bureau released a generalized tabulating system called GTS1.

Although these table generators may be different in their language, the machines they run on, and their internal design, they possess one common thread - they are all working, parameter driven generalized tabulating systems.

2.3 Data Presentation

Data Presentation, using the computer, comes in many forms - charts, graphs, photocomposition, microform, and publications. The objective is the display of data in graphical or pictorial form to help users of the data discern relevant patterns, trends and relationships. Very few people who have used good charts and graphs would argue with the proposition that "A picture is worth a thousand data values."

This paraphrasing of the old adage not only reveals the power of graphics, but the problem with graphics: for graphics technology to be useful, there must be data values to be displayed. This is best achieved by integrating graphics and DBMS's, a goal which is much-discussed and little-achieved. This integration theme will be further pursued in this and the following section.

Understanding the state of the art in graphics requires recognition of the dichotomy between graphics for data analysis and graphics for publication. There are substantial differences in quality, precision, and aesthetics of the data presentation. At the level of preparing graphical output, publication-quality graphics is more expensive and time-consuming than is data-analysis graphics. Yet both sorts of graphics are relevant to the use of Census Bureau data.

In data analysis, the emphasis is on quick interactive specification and production of scatter plots, empirical and theoretical probability density functions and cumulative probability functions, regression fits, and time series. The purpose of the analysis is to aid both understanding of the data's statistical phenomenon (type of distribution, correlations) and the data's significance and meaning (demographic trends, relation between various social and economic indicators, etc.).

The aesthetics of the data presentation are not overly important. What is important is the provision of easy to use, uncomplicated systems whose use can be quickly mastered by analysts with little or no computer programming experience. Ease of use includes integration of the system with a general and powerful data base system, so that any and all data of interest can be easily accessed.

A number of such systems exist. The success of the systems is much less a function of the straightforward graphics technology they use than of their integration of graphics and data.

In publication of statistical data, the aesthetics, quality and resolution of computer-generated images become very important, even critical. Crude plots which might satisfy and be useful to an analyst are unsatisfactory to many of the end users of Census Bureau data. Decision-makers and policy-makers in the public and private sectors who use the data have neither time nor inclination to work with anything but the best that can be offered.

The state of the art in data presentation is schizophrenic. There is the data analysis - data publication dichotomy. In addition, there is a broad gap between state of the art and common practice: a gap broader than in most technically evolving areas. On the one hand, there are numerous examples of magnificient computer-generated charts and graphics, many of them in full color. On the other hand, there are precious few commercially available turn-key systems. As a consequence, state of the art work is done in but a few research labs, universities, and government agencies.

There are several reasons. Doing graphics work requires the integration of numerous hardware and software components - more so than regular interactive computing. Major investments in time and equipment are usually necessary. As discussed in a later section, most graphics software is not especially portable, so program sharing is difficult. Investment in graphics is often treated as discretionary, so graphics development has lagged areas, such as DBMS, seen as more central or crucial to many organizations' goals.

For these various reasons, the state of the art in graphics is rather diffuse, quite unlike the DBMS and tabulation areas. The state of the art will be described from the viewpoints of hardware technology and software/system technology.

Graphics Hardware

Available hardware for interactive graphics (for data analysis or preparation of publications) ranges from the $4000 direct-view storage tube to the $100,000 high-performance line-drawing or color raster display system. While even better price and performance are desirable and expected, what we have is quite usable for the tasks at hand.

This is also true for graphics plotting devices. Small, inexpensive pen plotters and electrostatic matrix printer/plotters produce very usable plots for data analysis and for proofing of some types of publication material. Costs are often well under $10,000. High-quality proofing and final output can be had using precision plotters or COM devices, which cost from $150,000 to $300,000. It is possible, from the hardware viewpoint, to produce complete camera-ready copy of pages including charts, graphs, maps, text and tables. Color-separated negatives can also be produced.

Graphics Software

Graphics software technology has two major focuses: general-purpose graphics subroutine packages, and the applications which are built using the packages. The packages are of two sorts: those whose exclusive or major emphasis is plotting, such as DISSPLA (DISS) and CALCOMP (CALC), and those whose major emphasis is interactive graphics, but still with the possibility of producing hard-copy plots, such as GPGS (CARU 77) and GCS (PUK 76). The distinction between these two types of packages has already begun to blur: most of the newer packages, while still identifiable as one type or the other, also provide some (perhaps limited) capabilities of the other sort. These packages will continue to evolve, but they are already quite usable. Their basic purposes are firstly to hide all details of the display hardware from the programmer (much like a compiler hides a computer's details), and secondly (in most but not all cases) to allow production of complicated charts such as timeseries, barcharts, and pie charts with just a few subroutine calls to the package. The packages allow programs to draw simple plots to be written and tested in a few hours or less. The packages also allow simple interactive programs to be prepared in days or at most weeks.

Unfortunately, little of the hardware and software technology has been translated into turn-key systems which can be purchased, installed, and put into use solving real problems without first requiring non-trivial investments in system integration and application programming. There are few exceptions. The first is Tektronix hardware and software, which can readily be used for some types of data analysis. Costs are low, and there is a wide user community. General Electric's Genigraphics system can be used interactively to produce impressive color slides for presentations, and could be modified to produce output suitable for publications. It is a specially-programmed, minicomputer-based, stand-alone system which would be difficult to integrate into an overall publication or data analysis system, and costs in excess of $300,000. The final exception is in the drafting and design area, but such systems are not usable for analysis and publication of Census type data.

Resources for Graphics

Simple graphics can be done with little investment in people or equipment: $15,000 for a Tektronix terminal and hard-copy unit tied to a large time-shared computer, plus a programmer to work with the people who have the problems to be addressed. Quite a bit of leverage can be had in a programmer-rich environment, simply by having the "graphics programmer" train other programmers. But if programmers are scarce or nonexistent, little can be done with such equipment beyond the use of "canned" packages for plotting small quantities (tens or twenties) of data values.

Significant graphics, be it for data analysis or publication, requires significant people and equipment resources. Most major computer graphics installations have 5 to 10 staff members, and equipment valued from $250,000 to over $1,000,000. There are naturally a few exceptions to these staffing needs: a few installations manage with two or three exceedingly dedicated and self-disciplined people.

Reemphasizing what has been stated earlier, graphics is unlike DBMS and Data Tabulation because it requires more integration of system components, such as terminal hardware, data communications, plotters, systems software, and application programs. Thus it can be expensive and technologically challenging, especially when the graphics is further integrated with a data base, as described in the following section.

3.0 Integration

Data Base Management Systems, Data Tabulation Systems, and Data Presentation Systems are all useful in their own right. But to develop user-oriented software for dealing with very large statistical data bases such as the Bureau of the Census, an integration of the systems is absolutely essential. Figure 6 shows the general sort of integration which is required.

Data is entered through the DBMS into the data base where it can be edited and imputed. The Data Tabulation System accesses the data through the DBMS, and stores the resulting tabulations back into the data base. The tabulation results can of course be immediately printed as tables for examination, and can be used as input to the Data Presentation System for the preparation of charts and statistical maps. The common user interface allows users of the total system to deal with single, uniform sets of concepts, terminology, and procedures for carrying out data tabulation and data presentation.

Each component of the integration is important. The DBMS - Data Tabulation link allows all data being tabulated to be represented, stored, and accessed in a uniform way. It is not necessary to write special conversion programs for data to be tabulated. The DBMS - Data Presentation link permits serious graphical data analysis and chart and map presentation to be done. Some such link is essential, because the volume of data involved can be quite high. For instance, a county-level choropleth map contains in excess of 3000 data values. A ten-year trend chart of several monthly economic indicators contains 120 data values per trend line. These data values cannot realistically be manually entered into a data presentation system. In an environment where the emphasis is on ease of use and high volume use, it is unreasonable to require or expect the writing of one-time special-purpose programs to convert data from a DBMS or access method representation to that needed by a particular plot package. This simply consumes too much programmer resource, and juxtaposes a psychological and financial barrier between the data analyst or publication producer and the computer data base. What we need is a system with which the data analyst, publication producer, and perhaps in some contexts, the decision-maker can sit down at a terminal, specify any required tabulations, and then interactively examine the data in tabular and graphical form, with sufficient flexibility to allow experimentation with the data presentation.

The high-level model of Figure 6 can be further refined in two directions - one for data analysis, the other for data publication. Figure 7 shows an expanded data analysis system. Data can be retrieved, tabulated, analyzed, and presented in various ways. With the possible exception of the data retrieval (which might be quite slow), all these steps would be carried out interactively.

For publication work, the integration needs are actually more complex, as shown in Figure 8. This figure reinforces the centrality of the data base, and shows that a number of subsystems (only some of which directly involve graphics) are needed for total computerization of the publication (that is, Data Presentation) process.

Existing Integrated Systems

A number of partially-integrated systems have been developed, but we know of none that are completely integrated. In the data analysis domain, University of California at Berkeley's GEOQUEL/INGRES, IBM's GADS and Los Alamos Labs' oil-lease system represent varying degrees of integration. GEOQUEL (BERM 77) (Figure 9)*, a geographic information system, is built upon the relational data base system INGRES (STON 76). Maps and data about geographic areas defined by the maps are stored in the relational data base, and can easily be displayed.

If "mapofusa" is a state-level map of the USA, then

```
MAP mapofusa ON population
```

causes the map to be displayed with state population figures. A statistical map of the USA, using density of printed symbols to show population and car density, can be obtained with

```
SHADE mapofusa WITH #persons IS "x", #autos IS "*"
```

Thus rather complex presentations can be obtained quite simply. In addition, the underlying relational data base system allows arbitrary retrieval and manipulation of data.

The GADS system developed by Robin Williams (CARL 74, WILL 74) and colleagues at IBM's San Jose Research Lab integrates a

*Shaded areas on the figures represent implemented capabilities.

relational data base of small area census and local data with an interactive color raster display (Figure 10). The user gives data retrieval, processing, and display commands, and quickly sees the results. The emphasis is on geo-data, and on information display using maps. Computer-generated data can, if desired, be superimposed with a local map on the display console.

Work at Los Alamos Scientific Labs (Figure 11) by Phillips, Siebert, and others (PHIL 77) is used to maintain a data base (using the S2000 DBMS) of off-shore oil leases in the Gulf of Mexico. Choropleth maps are created to show the status of various lease plots. A high-precision film recorder is used to make color slides and prints.

Statistics Canada has two partially integrated systems. The first one which produces working tables, utilizes RAPID, the relational data base system discussed earlier, with STATPAK, the table generator system that works with the data base. The second system does their photocomposition without benefit of the data base, using the table generator system CASPER with some custom coding to interface with a Videocomp owned by a private contract firm.

The Bureau of Labor Statistics' system uses CINCOM's network data base management system TOTAL with BLS's own generalized tabulating TPL. Photocomposition is done using PCL, their print control language within TPL. The resultant output is phototypeset using the Government Printing Office's Linotron. Graphics work at BLS consists primarily of the production of trend charts using DISSPLA. BLS is working toward a completely integrated system.

The Census Bureau has two partially-integrated systems (Figure 14). The graphics systems are oriented toward data publication but used also for some data analysis. There are data presentation systems for dot, choropleth, and statistical maps (JONE 77), and for bar (FREE 76), pie (JOHN 76) and time-series or trend-line (SPAI 76) charts. The time-series chart system is integrated with a special-purpose DBMS for maintaining the time series (BUSC 76).

The second system, GTS (Generalized Tabulating System) (GENE 77) tabulates sequential files according to retrieval and processing requests. There is a flexible capability for specifying the details of how the table is to be presented with a line printer.

It is interesting to observe that none of these systems comes close to achieving a full integration of data management, data tabulation, and data presentation systems. The highest degrees of integration are for small, limited-purpose systems. There appear to be several reasons for this.

System Integration Problems

The most fundamental problem is the quandry presented by the following two statements:

1) It is difficult to integrate existing systems which were not initially designed to be integrated.
2) It is expensive to develop new systems.

The net is that existing, already-developed subsystems are generally not directly usable in building an integrated total system. Adaptations and modifications may be feasible, and are preferred, for economic reasons, to starting completely from scratch. In fact, however, the ease of use objective is usually best met by starting a system design project without a commitment to using or adapting existing software. This allows the development of a conceptual whole with an integrity of its own, unfettered by the need to compromise the design's clarity (hence ease of use) for the sake of using existing software.

System integration is also hampered by portability and standards problems. Just the right graphics system might be available on computer "A", but unless the programs can be moved to computer "B", they are relatively useless.

We believe that it is possible to build an integrated system to use a large statistical data base such as the Census

Bureau's. The subsystems are understood, some integration has already been achieved, and the hardware, software and systems technology is understood. What is needed is the commitment and resources. It can be done!

4.0 System Delivery

Once an integrated generalized information system has been developed, the next issue to be addressed is that of delivery of the system capabilities to the end user.

There are all sorts of users of Census data - large and small, private and government, business and industry, academic and commercial, with or without technical interest and with or without computer and programmer resources. Therefore, we must consider a broad spectrum of delivery possibilities when we consider making our data available to end users. What are the possibilities?

First, the data can be made available through human intermediaries. This can be done by having some data users' service organization whereby the user's particular request for data can be satisfied.

The second possibility is the distribution of software with the data for use within the requestor's data center.

The third and last possibility would be the establishment of public-access data centers, whereby, through data communication, facilities the users would have access to the data base and the software tools to solve their data requirements.

But, there are factors affecting system delivery whatever approach is chosen. Hardware compatibility is the first problem and is affected by such things as internal formats, word sizes, and peripheral and ancillary equipment.

Another problem to be encountered is that of software portability. Delivering software to operate on different computer systems is quite a challenge. But, we know that this is possible by the example of numerous successful models, including SPSS, BMD, S2000, MARK IV, IMSL, DISSPLA, and PLOT-10. These systems have been successfully distributed by overcoming two barriers: the technological one of achieving CPU, language, and operating system independence, and the managerial one of providing a disciplined system for creating, updating, and disseminating system documentation, fixes, and upgrades. Neither barrier is trivial, although technologists tend to dwell on the former, leaving the latter to chance.

The technological problems are perhaps a bit more complex than those addressed by the distributors of the above models, because we are concerned with integrated systems which require diverse computer resources: large file systems, large main memory, various graphics devices.

Many problems are resolved by using a standard programming language, such as ANSI FORTRAN IV or ANSI COBOL. But some operating system interface matters, and word size/precision problems, still remain. They are generally resolvable by programming so as to isolate the operating system or computer dependencies to a few subroutines which are recoded for each new environment.

If a program which is to be delivered requires a DBMS, there are two choices: deliver the DBMS also, or interface to a "standard" DBMS. Unhappily there are no standards for DBMS's, although several commercial systems (such as MRI's S2000 and CINCOM's TOTAL) have been implemented with different manufacturer's computer systems. The CODASYL report (CODA 71) has had a major impact on DBMS's, and many DBMS's conform at least to the spirit of the report's recommendations. Thus there are a number of similar, but not equal, DBMS's in existence. This is not enough for software distribution, just as having ten or so FORTRAN dialects is not enough.

For passive output graphics, there are two dominant de facto standards: the "CALCOMP routines" (CALC), and DISSPLA (DISS). For interactive graphics, there are a number of widely-used device-independent packages, such as GPGS (CARU 77), GCS (PUK 76) and GINO-F (GINO). None has achieved preeminence. There is also a proposed standard, developed by ACM/SIGGRAPH, which may be officially adopted by ANSI (perhaps in modified form) within the next few years (GSPC 77).

To summarize software portability, it is fair to say that standard FORTRAN or COBOL programs which do processing and simple I/O can be "ported" to new computers quite easily. Programs requiring the services of DBMS's or graphics packages are not nearly so easily moved.

The third problem is that of data portability. In sequential summary tape form, data portability is very do-able since there are no real technological problems with the existing standards for tape format, labelling and coding. Standards also exist for data communication, therefore, data transmission poses no technological problems.

Data portability using DBMS structures is also do-able with a portable DBMS. Today, we know of none that are not somewhat machine dependent.

These portability problems largely disappear if the public-access multi-user service center approach is selected as the delivery vehicle. Statistics Canada has an interesting, unique approach to delivery of their economic time series data base through a system called CANSIM. Through a joint government/private enterprise venture, CANSIM is made available through commercial time-sharing services throughout Canada, the U.S. and even Eurpope. Statistics Canada maintains the master data base at a "parent" time-sharing center in Montreal. "Subscriber" time-sharing services contract with Statistics Canada for $1500 a month for the right to market the time series that they download each day from the parent center. Subscribers are contractually obligated to update data bases within twenty-four hours of an update of the master file in the "parent" center.

Each time-sharing service makes available CANSIM software made available to them by Statistics Canada, as well as, any software that they have developed for their users.

Statistics Canada uses an AMDAHL 470 V-6 computer which is plug-to-plug compatible with the IBM 370. Only software developed for their machine is distributed. Subscribers with machines outside the IBM 370 family must assume the responsibility of converting the CANSIM software to their environment.

This approach to delivery of data appears to work very well. It is but one of many possibilities involving a public-access system.

No matter how end users access the computer, there must be good interactive access to the integrated systems capabilities It is crucial that these capabilities be easy to use. Otherwise, they may not be used at all! We know that ease of use is easy, to talk about, but hard to achieve. Both the conceptual system model, which the users must master, as well as the details of the command language syntax, error messages, and prompts must be carefully designed.

To achieve systems that are easy to use requires careful top down design and planning. We are beginning to know how to do this (FOLE 74, CHER 76) but our skills are not nearly perfected. What we do know is that making systems easy to use is expensive of both people and computer time. We know that redesigns of command languages are sometimes necessary, and that use of general-purpose tools can make the implementation and modification tasks faster and less expensive. There is certainly the possibility of an easy to use command language patterned after English, but in a constrained form. Such systems are likely to be common in the next decade.

At the moment, only the military, large vendors/users, and a few research labs seem concerned about computer system ease of use. We must be prepared to join in their concern, study their systems, and learn their craft.

5.0 Summary

Systems have been developed for DBMS, GTS, and Photocomposition and computer generated graphics. Today, completely integrated systems do not exist but partially integrated special purpose systems do exist and have demonstrated the technological feasibility of developing a completely integrated generalized information system. All it takes is resources, a management commitment and direction. It can be done!

Host Language CODASYL Data Base Management Systems

	DMS 1100	IDMS	Honeywell IDS-II
CPU	UNIVAC 1100	IBM 370	H6000
Item Description	COBOL Oriented	Host Language Like	COBOL Like
Logical	Network	Network	Network
Physical	Pointers	Pointers	Pointers
Access Methods	Direct Hashed ISAM Network	Direct Hashed Network	Direct Hashed ISAM Network
D.B. Creation	User Programs	User Programs	Utility
Query Language	Yes	No	Yes
Report Generator	Yes	No	No
Host Language	COBOL FORTRAN	COBOL FORTRAN	COBOL
Multi-thread	Yes	Yes	Yes
Security	None	Thru Subschema	Password
Data Validation	None	None	Yes
Recovery	Full-Scale	Full-Scale	Full-Scale
Surveillance	Log Tapes and Statistics Collection	None	Yes

Figure 1

Host Language Non-CODASYL Data Base Management Systems

	Burroughs DMS II	Cincom TOTAL	IBM IMS	MRI System 2000	Software AG ADABAS
CPU	Burroughs 6700, 7700	IBM 370 ** CDC 6000 UNIVAC 70	IBM 370	IBM 370 UNIVAC 1100 CDC 6000	IBM 370
Item Description	Host Language Like	Host Language Like	Host Language Like	Host Language Like	Host Language Like
Logical	Network	Multi-list	Hierarchy	Tree Structured	Almost Relational
Physical	Pointer	Pointer	Adjacency	Adjacency	Pointers
Access Methods	Direct, Hased ISAM, Bit Vector, Network	Direct Sequential Hashed		Direct Sequential Inverted Indices	Direct Inverted Indices Hashed
D.B. Creation	User Programs	User Programs	User Programs	Utility & User Programs	Utility & User Programs
Query Language	Yes	Yes	Yes	Yes	Yes
Report Generator	Yes	Yes	Yes	Yes	Yes
Host Language	ALGOL PL/I COBOL	Any lang. with sub-routine calls	COBOL PL/I Assembler	COBOL FORTRAN	COBOL FORTRAN PL/I Assembler ADASCRIPT
Multi-thread	Yes	Yes	Yes	Yes	Yes
Security	None	None	Yes	Yes	Yes
Data Validation	Some	None	None	Some	Some
Recovery	Full-scale	Some	Yes	Full-Scale	Full-scale
Surveillance	Some	None	Some	Log Tapes	Log Tapes

**also: PDP-11
Honeywell 2000
IBM System/3
NCR Century
Varian V70

Figure 2

SELF-CONTAINED
Data Base Management Systems

	Computer Corp. of America Model 204	Meade Technology Data/Central	TRW GIM II
CPU	IBM 370	IBM 370	IBM 370 UNIVAC 1100 PDP-11
Item Description	Character String	Character String	Character string Numeric
Logical	Almost Relational	Multi-list	Almost Relational
Physical	Pointers	Adjacency	Pointers
Access Methods	Sequential Inverted Indices Hashed	Inverted Indices Sequential	Inverted Indices Hashed
D.B. Creation	Utility	Utility	Utility
Query Language	Yes	Yes	Yes
Report Generator	No	No	Yes
Host Language	COBOL, FORTRAN, PL/I, Assembler	Any language with subroutine call	COBOL and Own
Multi-thread	Yes	Yes	Yes
Security	Yes	Yes	Yes
Data Validation	Yes	No	Yes
Recovery	Some	Yes	Yes
Surveillance	Yes	Yes	Some

Figure 3

Data Tabulation Systems

Package	Organization	Machine Availability	Minimum Resource Requirements	Source Language	Cost	Comments
GTS-1	SSD—CENSUS	UNIVAC 1100	85K WORDS MEMORY 2M WORDS DISK	COBOL	—	DEVELOPED FOR CENSUS INTERNAL USE LANGUAGE—VERY GOOD EFFICIENCY—EXCELLENT
CO-CENTS	ISPC—CENSUS	MANY	IBM 370—24K BYTES	COBOL	—	DEVELOPED FOR AID FOR INTERNATIONAL DISTRIBUTION LANGUAGE—POOR EFFICIENCY—EXCELLENT
TPL	BUREAU OF LABOR STATISTICS	IBM 360/370 FAMILY	200K BYTES (NEEDS DISK SPACE FOR INTERMEDIATE RESULTS)	XPL	—	PRESENTLY INSTALLED IN OVER 150 INSTALLATIONS WORLDWIDE LANGUAGE—EXCELLENT EFFICIENCY—GOOD
CROSS-TABS II	CAMBRIDGE COMPUTER ASSOCIATES	IBM 360/370 FAMILY	80K BYTES	BAL	ANNUAL LEASE $6120 MONTHLY $600	DEVELOPED FOR CROSSTABULATION OF SURVEY DATA LANGUAGE—GOOD
CENTS AID II	DATA USE AND ACCESS LABS	IBM 360/370 FAMILY	100K BYTES	COBOL	PURCHASE FROM NTIS $600 DOMESTIC $1200 FOREIGN ——— ANNUAL MAINTENANCE $500	SIMPLE SPECIFICATION OF TABULATION FROM LARGE COMPLEX FILES LANGUAGE—VERY GOOD

Figure A

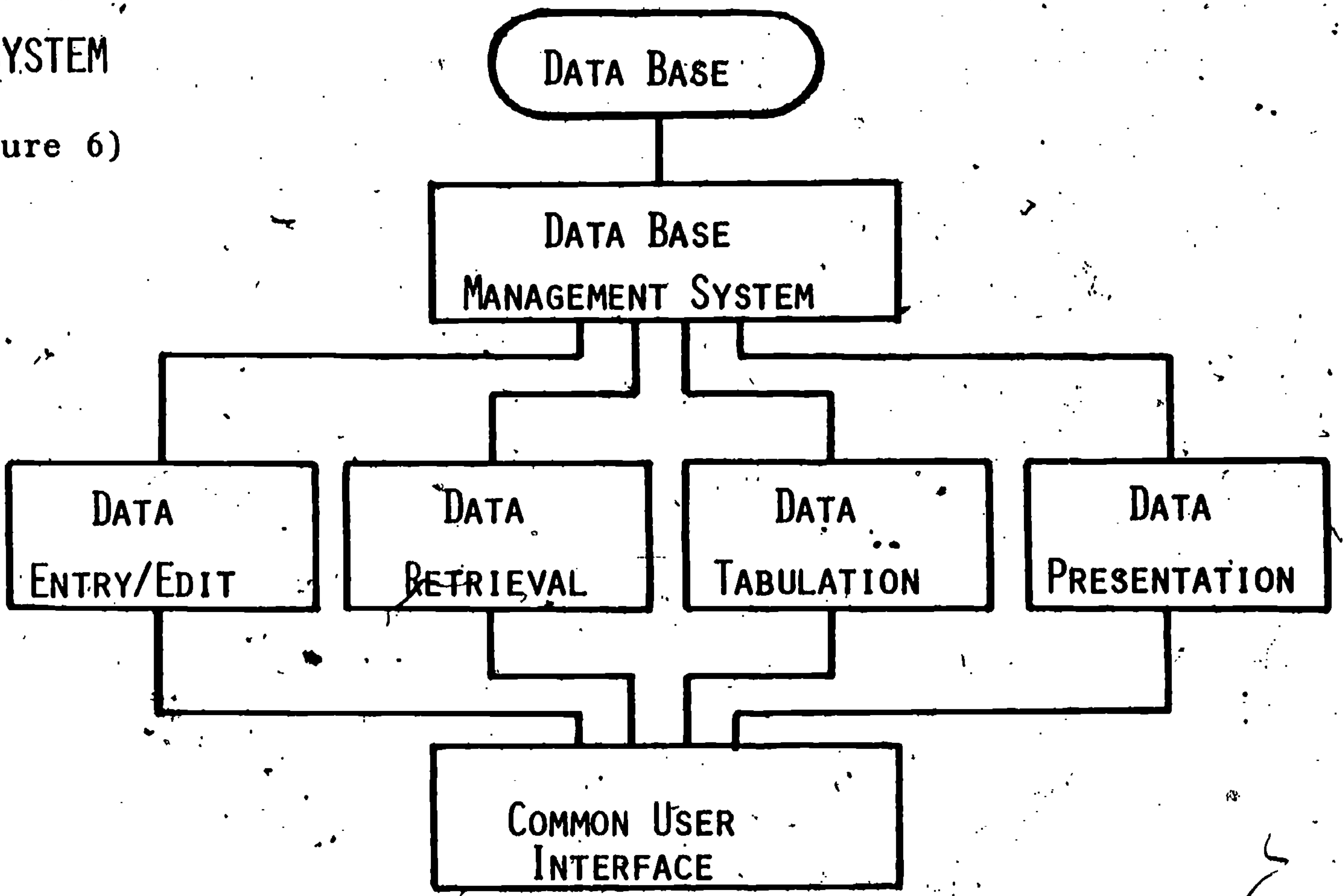
Statistical Packages with Tabulation Capabilities

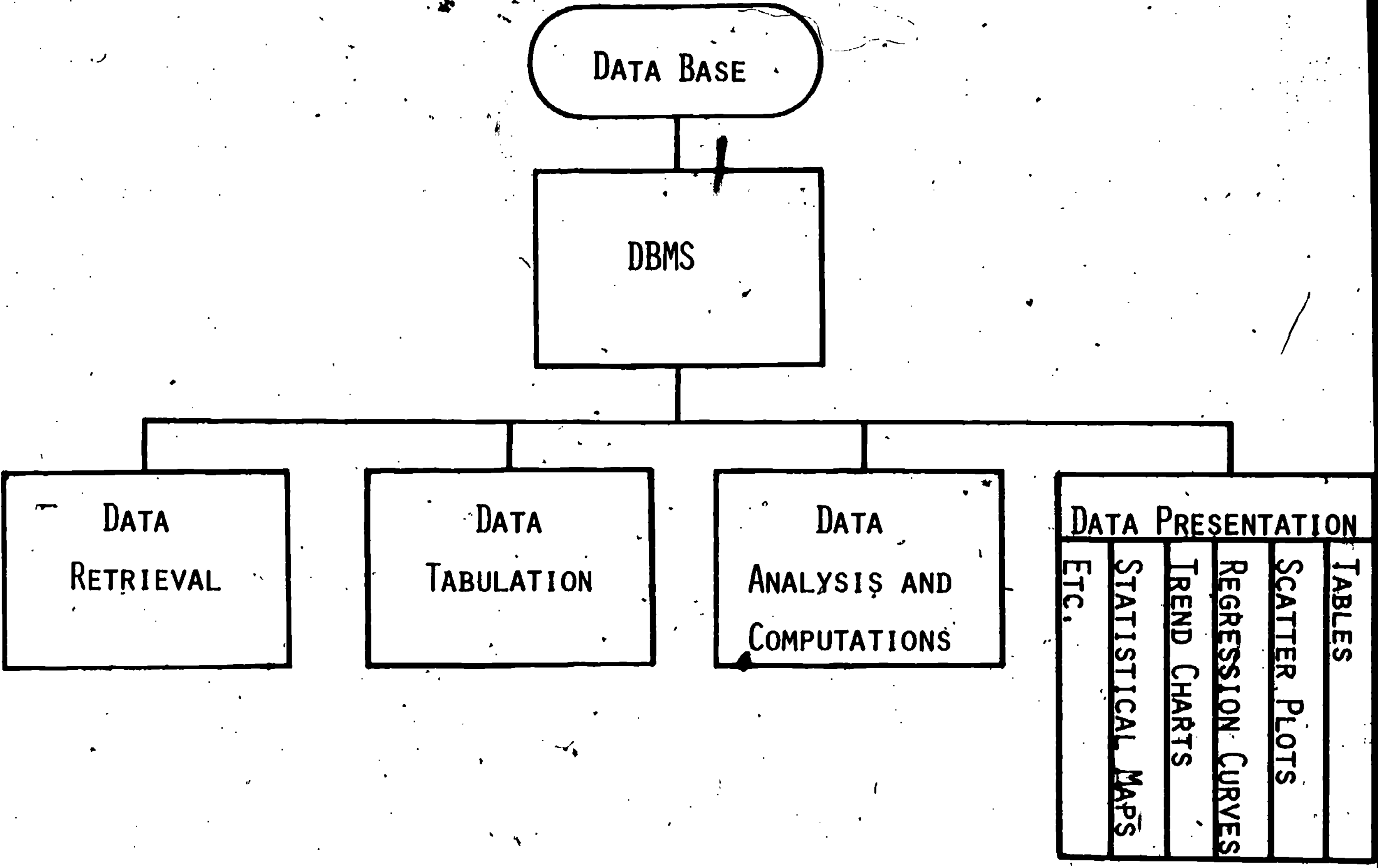
Package	Organization	Machine Availability	Minimum Resource Requirements	Source Language	Cost	Comments
SPSS	SPSS, INC.	MOST	IBM 370–150K BYTES	FORTRAN	COMMERCIAL $5000 + $2000/ YR OPTIONAL —— NON-PROFIT $1500 + $800/ YR OPTIONAL —— ACADEMIC $1000 + $600/ YR OPTIONAL	COMPLETE PACKAGE FOR ANALYSIS OF SOCIAL SCIENCE DATA LANGUAGE—GOOD
BMDP	HEALTH SCIENCES COMPUTING FACILITY SCHOOL OF MEDICINE U.C.L.A.	MANY	IBM 370–160K BYTES	FORTRAN	$100	GENERAL STATISTICAL PACKAGE FOR TOTAL DATA ANALYSIS LANGUAGE—VERY GOOD
DATA-TEXT	DATA-TEXT SYSTEMS	IBM 360/ 370 FAMILY	250K BYTES	FORTRAN	COMMERCIAL $1000 + $400/ YR RENEWAL —— NON-PROFIT $750 + $300/ YR RENEWAL —— ACADEMIC $500 + $200/ YR RENEWAL	STATISTICAL ANALYSIS AND DATA MANIPULATION OF SOCIAL SCIENCE DATA LANGUAGE—GOOD
SAS	INSTITUTE OF STATISTICS, NORTH CAROLINA STATE	IBM 360/ 370 FAMILY	150K BYTES	PL1/BAL	COMMERCIAL $3500 + $1500/ YR RENEWAL —— NON-PROFIT $2500 + $1500/ YR RENEWAL —— ACADEMIC $750 + $300/ YR RENEWAL	GENERAL STATISTICAL ANALYSIS—ESPECIALLY LIFE SCIENCES LANGUAGE—GOOD

Figure 5

INTEGRATED
SYSTEM

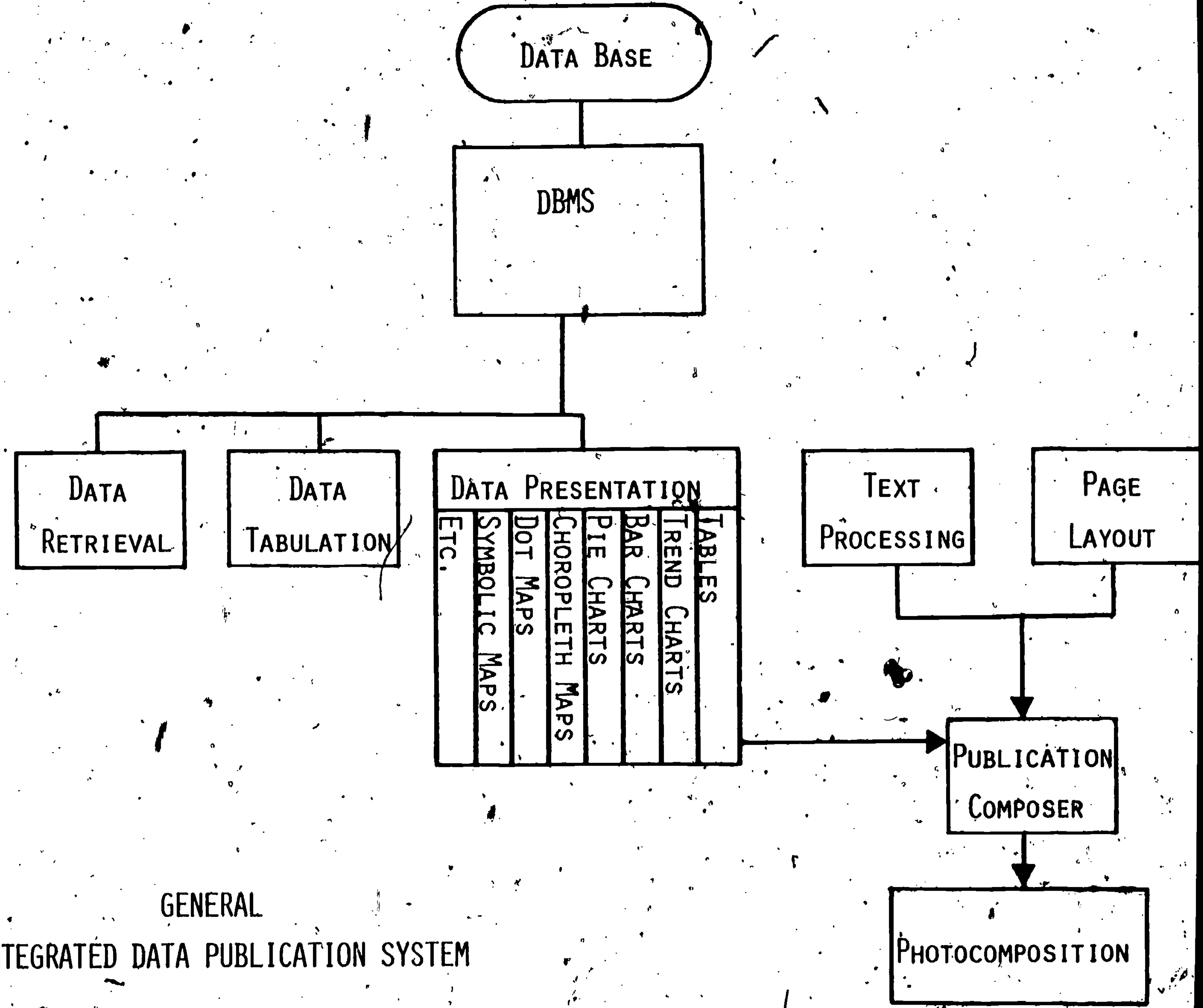
(Figure 6)





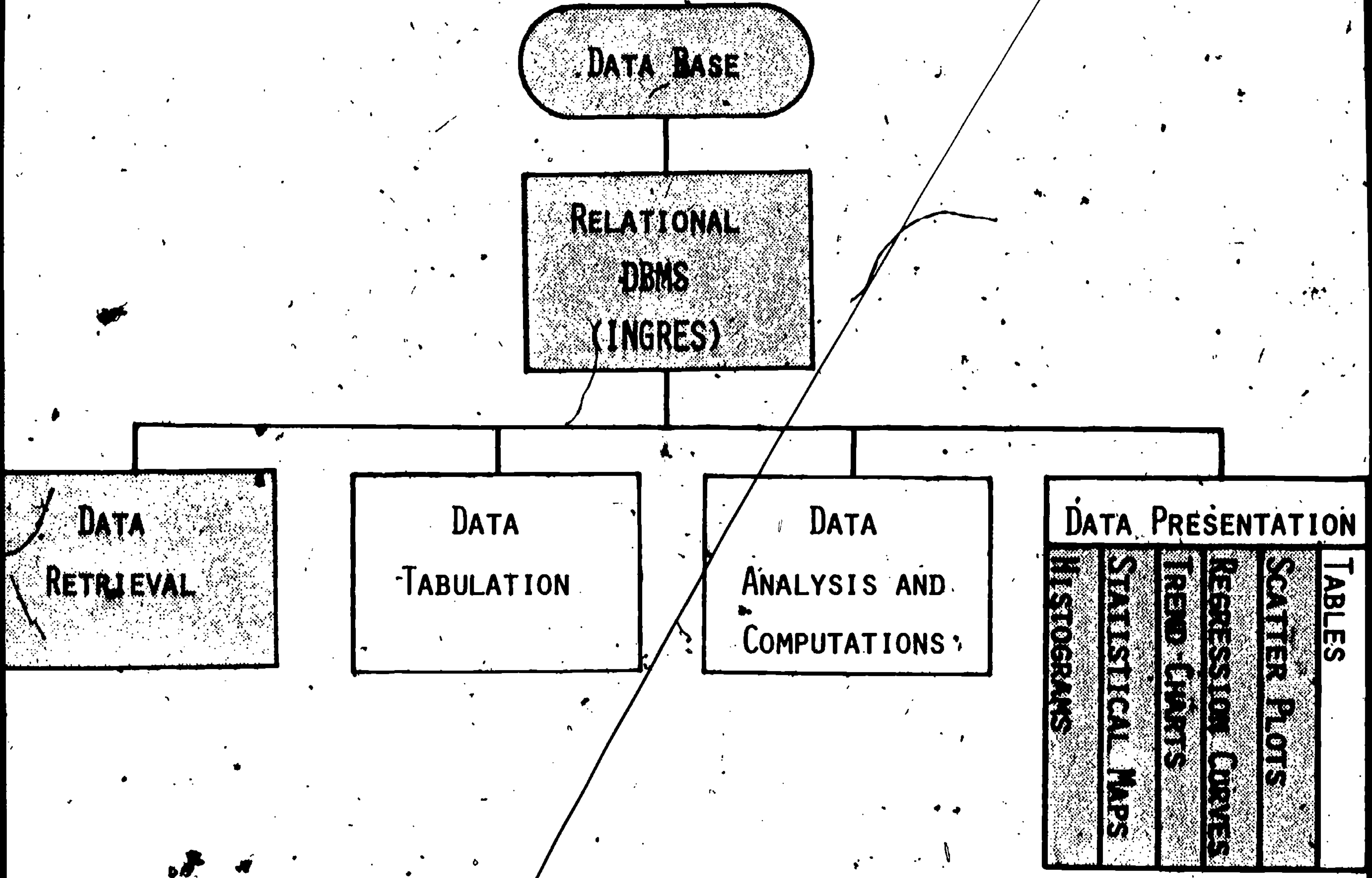
GENERAL
INTEGRATED DATA ANALYSIS SYSTEM

Figure 7



GENERAL
INTEGRATED DATA PUBLICATION SYSTEM

Figure 8

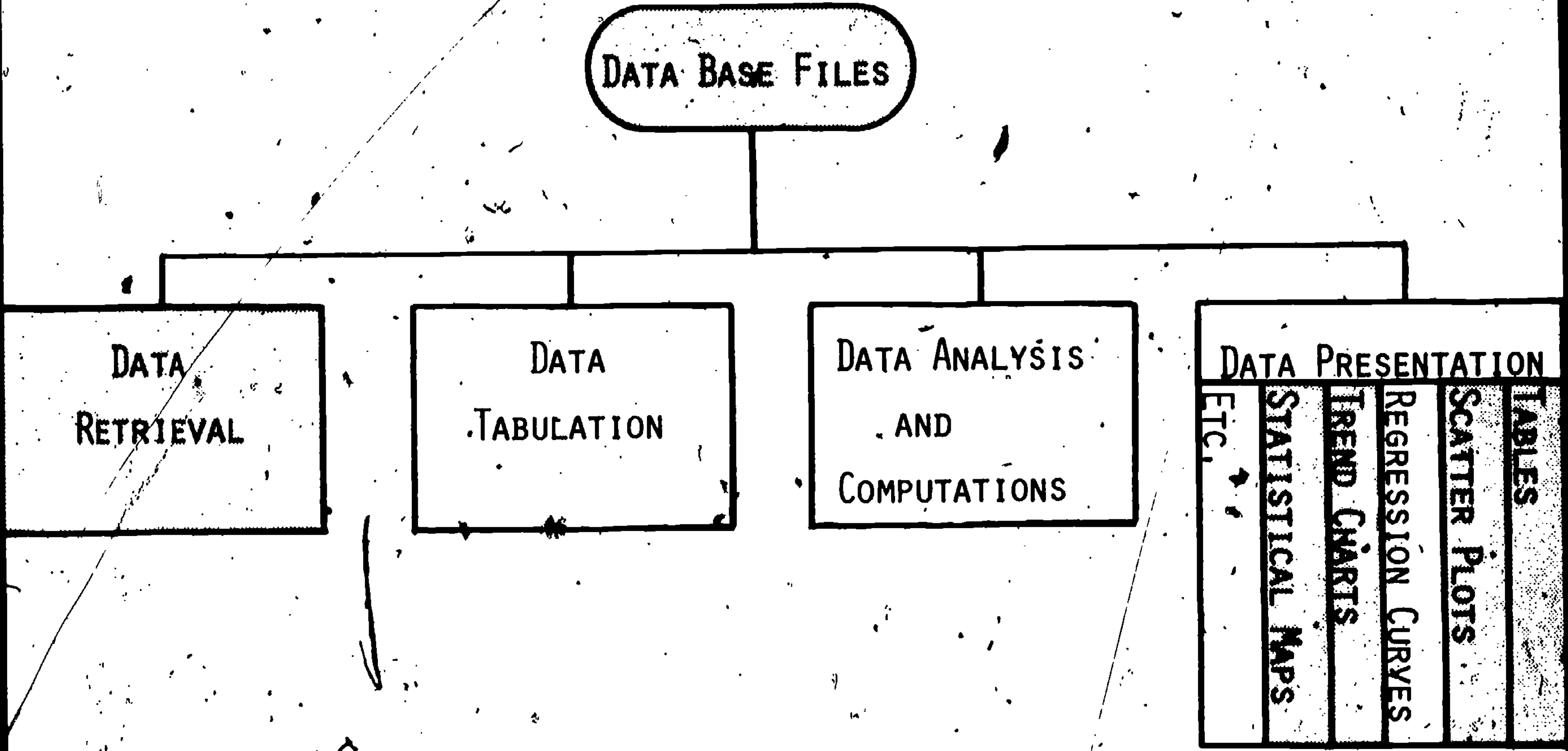


UNIVERSITY OF CALIFORNIA - BERKELEY

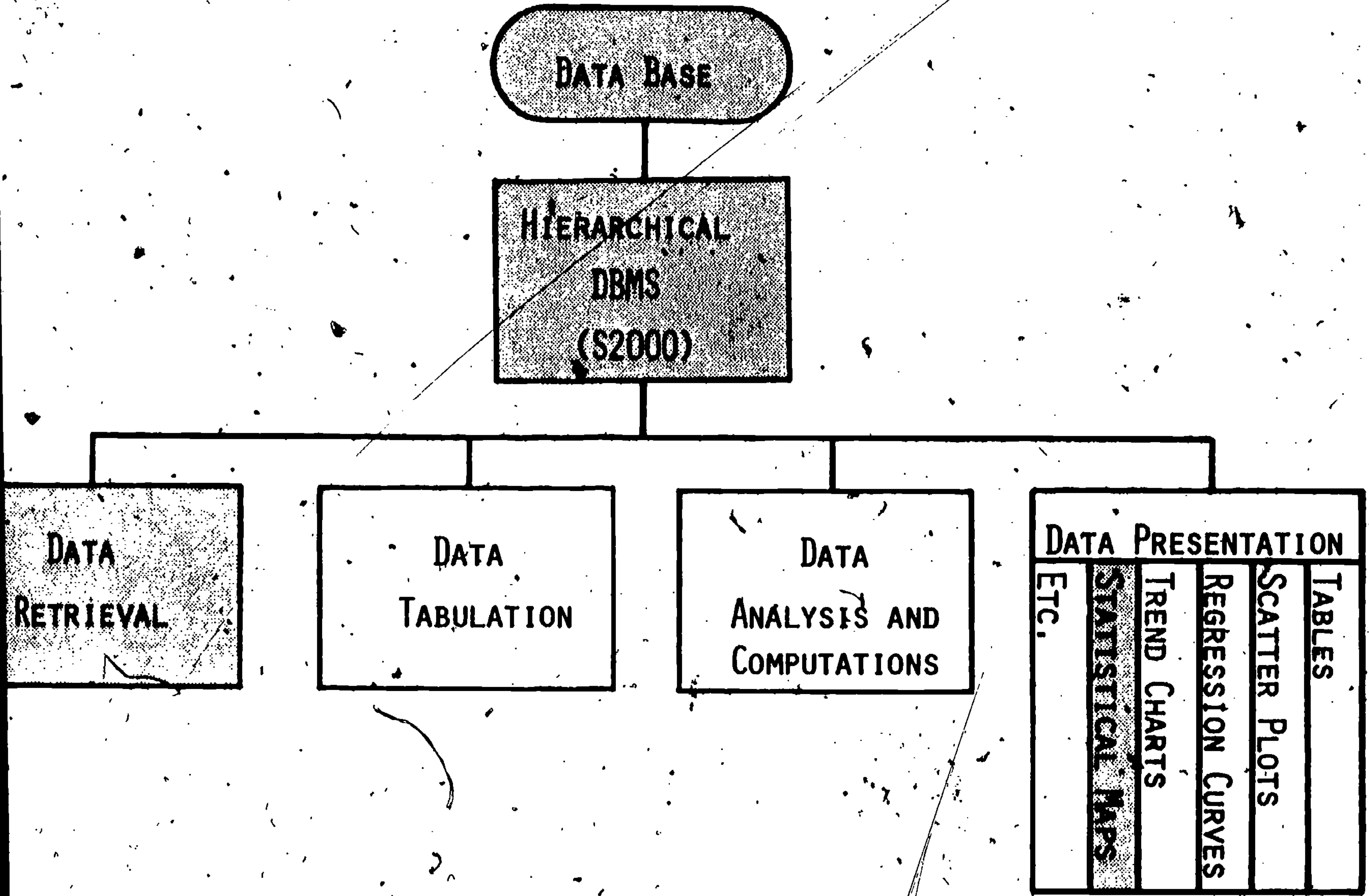
GEOQUEL/INGRES

INTEGRATED DATA ANALYSIS SYSTEM

Figure 9



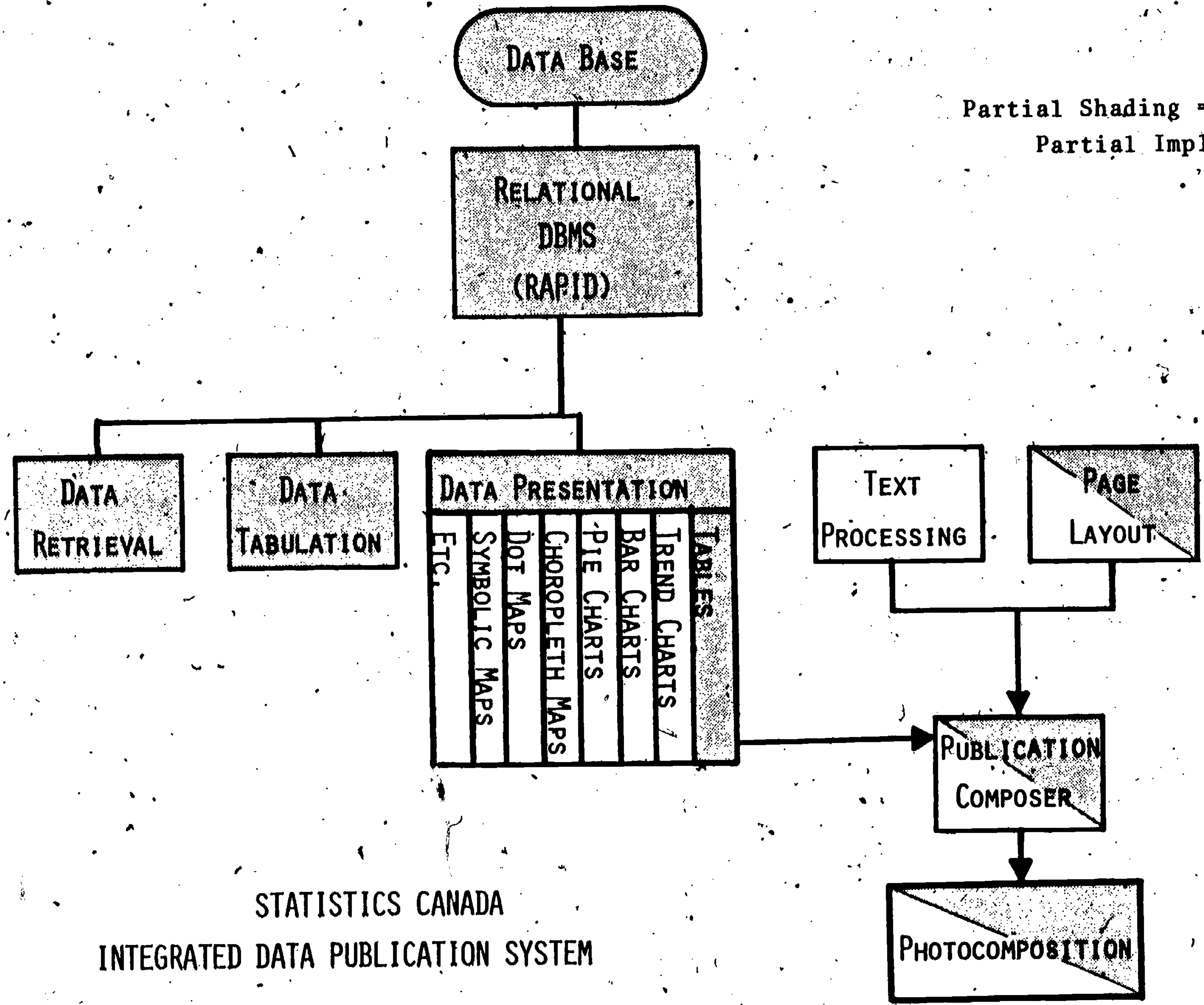
IBM-GADS
INTEGRATED DATA ANALYSIS SYSTEM
Figure 10



LASL OIL LEASE
INTEGRATED DATA ANALYSIS SYSTEM

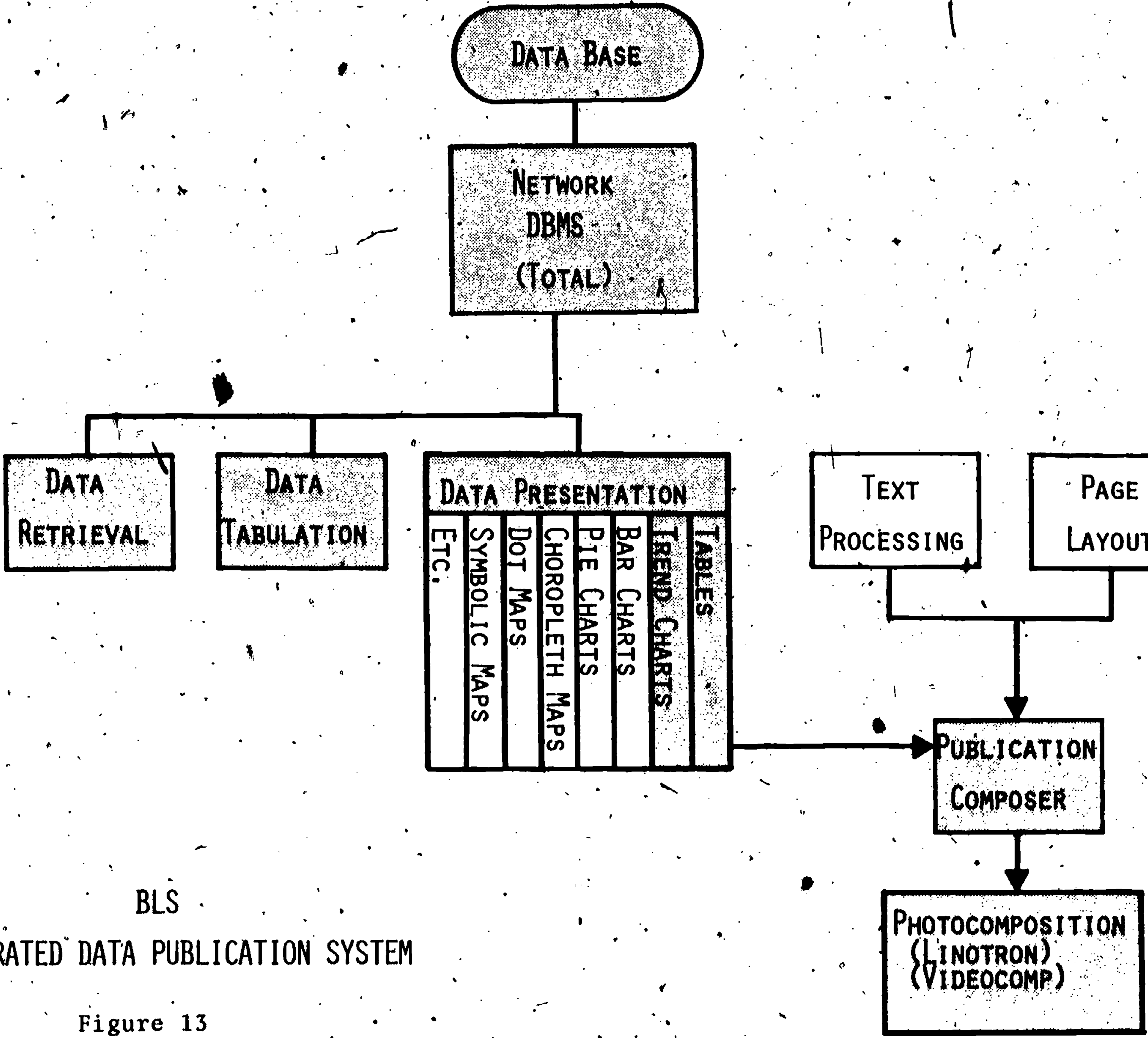
Figure 11

Partial Shading =
Partial Implementation



STATISTICS CANADA
INTEGRATED DATA PUBLICATION SYSTEM

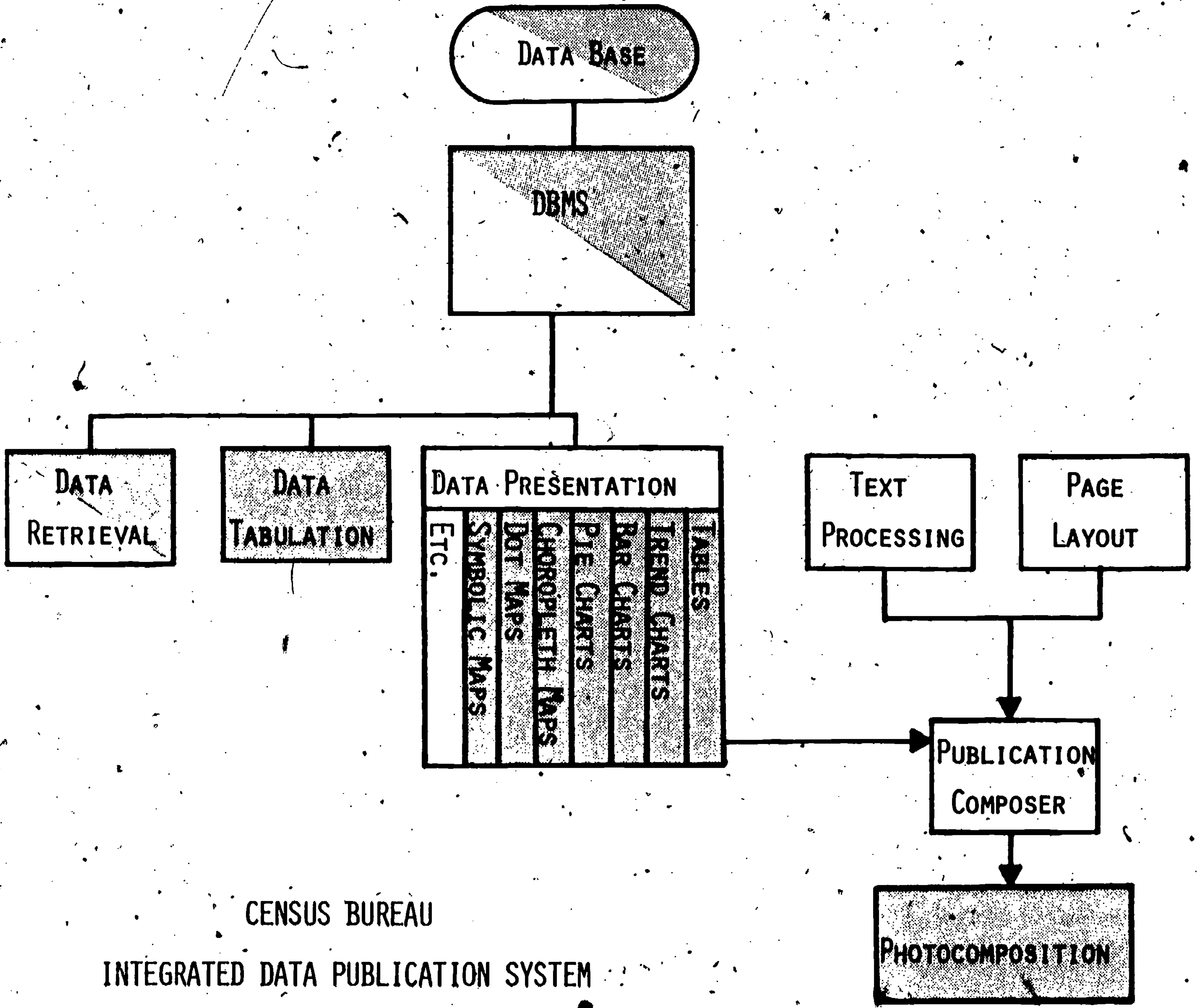
Figure 12



BLS

INTEGRATED DATA PUBLICATION SYSTEM

Figure 13



CENSUS BUREAU

INTEGRATED DATA PUBLICATION SYSTEM

Figure 14

PARTIAL SHADING = PARTIAL IMPLEMENTATION

REFERENCES

(BERM 77) Berman, R. and Stonebreaker, M. "GEO-QUEL - A System for the Manipulation and Display of Geographic Data," SIGGRAPH '77 Proceedings, published as Computer Graphics 11, 2 (Summer 1977), 186-191

(BUSC 76) Busch, R.A., TIMEBASE: Time-Series Data Base Management System, and TIMEBASE Processors User's Guide, Graphics Software Branch, Systems Software Division, U.S. Bureau of the Census, Washington, D.C., 1976

(CALC) CALCOMP Plot Package Reference Manual, CALCOMP, Inc.

(CARL 74) Carlson, E., Bennett, J., Giddings, G., Mantey, P., "The Design and Evaluation of an Interactive Geo-data Analysis and Display System," Proceedings IFIP Congress 74, 1057-1061. North Holland, 1974

(CARU 77) Caruthers, L., van der Bos, J., van Dam, A., "A Device-Independent General Purpose Graphic System for Stand-Alone and Satellite Graphics," Proceedings of SIGGRAPH '77, published in Computer Graphics 11, 2 (Summer 1977), 112-119

(CHER 76) Cheriton, D., "Man-Machine Interface Design for Timesharing Systems," Proceedings ACM 1976 Conference, 362-366

(CODA 71) CODASYL Data Base Task Group, April 1971 Report. Association for Computing Machinery, New York, N.Y. 1971.

(CODD 70) Codd, E., "A Relational Model of Data for Large Shared Data Banks," CACM 13, 4 (June 1970), 377-387.

(DISS) DISSPLA, Integrated Software Systems Company, San Diego, California.

(FOLE 74) Foley, J. and Wallace, V., "The Art of Natural Graphic Man-Machine Conversation," Proceedings IEEE 62, 4 (April 1974), 462-470.

(FRAN 76) Francis, I. et.al., "Languages and Programs for Tabulating Data From Surveys," Proceedings of the Ninth Interface Conference on Computer Science and Statistics, 129-134, April 1976.

(FREE 76) Freeman, J., BARCHART: A General Purpose Plotting Program, Graphics Software Branch, Systems Software Division, Bureau of the Census, Washington, D.C., 1976.

(GENE 77) Generalized Tabulating System - 1, Bureau of the Census, 1977,

(GINO) GINO-F Reference Manual, Computer-Aided Design Centre, Cambridge, England.

(GSPC 77) Status Report of the Graphics Standards Planning Committee of ACM/SIGGRAPH. Published as Computer Graphics 11 (3), Fall 1977.

(HEIN 77) Heindel L. and Roberto, J., LANG-PAK - An Interactive Language System, American Elsirier, New York, N.Y., 1975.

(JOHN 76) Johnson, J.A., PIECHART: A General Purpose Plotting Program, Graphics Software Branch, Systems Software Division, Bureau of the Census, Washington, D.C., 1976.

(JONE 77) Jones, P.A., MAPS, Graphics Software Branch, Systems Software Division, Bureau of the Census, Washington, D.C., 1977.

(PALM 75) Palmer, I., Data Base Systems: A Practical Reference; Q.E.D. Information Sciences, Inc., 1975.

(PHIL 77) Phillips, R., "A Query Language for a Network Data Base with Graphical Entities," Proceedings of SIGGRAPH '77, published in Computer Graphics 11, 2 (Summer 1977), 179-185.

(PUK 76) Puk, R.F., The 3D Graphics Compatibility System, U.S. Army Corps of Engineers Vicksburg, Miss., 1976.

(SPAI 76) Spaid, G.L., TIMESERIES: A General Purpose Plotting Program, Graphics Software Branch, Systems Software Division, Bureau of the Census, Washington, D.C., 1976.

(STON 76) Stonebreaker, M., Wong, E., Held, G. and Kreps, P., "The Design and Implementation of INGRES," ACM Transactions on Data Base Systems 1, 3 (September 1976), 189-222.

(TAB 75) Table Producing Language, Bureau of Labor Statistics, 1975.

(WILL 74) Williams, R., "On the Application of Relational Data Structures in Computer Graphics," Proceedings IFIP Congress 74, 723-726, North-Holland, 1974.

THE NEEDS FOR AND AVAILABILITY OF USER SOFTWARE TO PROCESS AND ANALYZE CENSUS BUREAU MACHINE-READABLE PRODUCTS[1]

Warren G. Glimpse
Data User Services Division
U.S. Bureau of the Census

I. INTRODUCTION

A steadily increasing volume of data produced by the Census Bureau is being made available to the public in machine-readable form. The user demand for these products continues to grow at an even faster pace, reflecting the high level of interest in microdata, more detailed summary data, geographic reference data, and cross-reference and descriptor type data being made available in machine-readable form. User demand is further heightened by growing sophistication of users in using computers to analyze statistical data.

Ten years ago both the supply of and demand for Census Bureau machine-readable products was quite limited with only a few reels of tape being distributed per year. Since 1971, however, more than 20,000 reels of tape have been sold representing more than $1.2 million in standard tape product sales. It is estimated that the total magnitude of these tape products in the user domain acquired through intermediaries, such as summary tape processing centers, is 8 to 10 times this volume -- as many as 200,000 feels of tape. During the same period, approximately $3.5 million in

[1] Prepared for the 1977 Joint American Statistical Association/U.S. Bureau of the Census Conference on Development of User Oriented Software, sponsored by the National Science Foundation, November 8, 1977, Washington, D.C.

special tabulation projects have been undertaken for the 1970 decennial census data alone. Most of these customized products have been delivered to the sponsor, and other interested users, in machine-readable form. These trends are expected to continue due to increasing amounts of data being made available from a larger number of statistical programs and a growing number of users making use of machine-readable products.

There is little question that user accessible software plays an important role in processing these machine-readable products for administrative, planning, and decisionmaking purposes. This paper provides an overview and perspective concerning the needs for and availability of user accessible software and related issues involving ways to improve access to and use of Census Bureau machine-readable products. In this context, users are defined to be those persons engaged in the process of acquiring and processing Census Bureau machine-readable data. While in part this group includes some Census Bureau staff, the larger universe of users are non-Census Bureau staff located in Federal agencies, State and local government agencies, colleges and universities, businesses, and professional and trade associations as well as individual researchers, and others. User accessible software includes computer programs which may be acquired by users for use on their own computer as well as software which may be accessed through terminals and time-sharing systems.

It is, however, important to stress that user software is only one of the essential ingredients necessary to achieve effective and efficient use of machine-readable products. Equally important issues, which must be addressed concurrently, include the structure of the files, technical documentation, user training, and manuals. Since the demand for user software is derived from the need to process and analyze machine-readable files, a summary of Census Bureau statistical resources in machine-readable form accessible to users is first reviewed in this paper. This summary includes an assessment of past trends and current plans for developments involving production and dissemination of Census Bureau machine-readable products.

Secondly, the need for user software, and related materials, to facilitate access to and use of machine-readable products will be considered. A review of existing Census Bureau software is presented.

An analysis of the unmet needs for user software is then considered. This includes an assessment of the problems involved with access to and use of existing Census Bureau machine-readable products with existing software including issues such as file structure, documentation, universe comparability, cost-benefit issues, etc. Along with this, plans and options for developing user software and other aids to assist users in accessing and utilizing Census Bureau machine-readable statistical resources are reviewed.

II. CENSUS BUREAU MACHINE-READABLE STATISTICAL RESOURCES

To set the stage for a discussion of what types of user software are needed, existing and planned developments for Census Bureau statistical resources in machine-readable form are first considered. The reason for this is that software are developed to process available files. To address the scope and structure of required software this framework is essential.

From the present day and historical perspective, it is important to differentiate between publicly distributable and internal, confidential machine-readable products. Publicly distributable products include those files which may be directly released to users outside the Census Bureau. Examples of these files include the 1970 Census summary tapes and public use samples, county business patterns files, intercensal estimates files, geographic base files, and machine-readable technical documentation. A more comprehensive list of these products available for sale is contained in Appendix A.

Publicly distributable files include summary statistic, microdata, and geographic and other reference files which are prepared for public dissemination. To briefly review, summary statistic files are those files containing data items which are aggregates or estimates of the number of respondents with specified characteristics, measures of activity levels, or the number of events occurring during a particular period for specific geographic areas. The common feature of these files is that of the record containing an aggregate statistic for a variable corresponding to unique geographic area.

Microdata files are those files which contain data items corresponding to characteristics of an individual respondent or respondent unit. Each record generally corresponds to an individual, household, or other type of basic survey unit. In some cases these files contain ratio scale data (such as the neighborhood characteristics 1970 Census public use sample) or family aggregates derived from person records (such as the Current Population Survey Annual Demographic File). The geographic area containing the response is identified on each record provided the area is 250,000 population or larger.

Geographic reference files contain descriptive data about selected geographic segments or areas. These files range in scope from the Geographic Base File (GBF), a computerized representation of a map with records corresponding to street and non-street segments, to the 1970 Census Master Enumeration District List, a hierarchical listing of geographic areas and names for all geography larger than blocks.

These publicly distributable files are to be contrasted from the confidential data files containing basic records with individual identifying characteristics. Basic record files cannot be released to the public in accordance with the title 13 provisions to insure confidentiality of individual information. However, basic record files are the source of many quite valuable special tabulations. As a consequence, consideration must be given to software which can be used to prepare special tabulations on a timely and low cost basis.

In review of the existing files available to the public, two of the most significant problems which are barriers to efficient and effective use are file structure and technical documentation. While these issues are discussed more extensively in a later section they should be touched upon here for an appraisal of software needs.

With regard to file structure, it is important to note that few structural and archiving standards have been set and followed from statistical program to program. Secondly, by the very nature of Census Bureau statistical files -- containing extensive data corresponding to hierarchical subject matter and geography -- logical records become quite long and are typically nested in a hierarchical fashion. As a result, user software developed specifically for machine-readable products from any given statistical program is generally not transferable to other statistical program products. In addition, due to long records and hierarchical file structures, much of the conventionally designed software is difficult, or extremely expensive, to use without modification. An additional complication to the processing of many of the decennial files is introduced due to the volume of data -- often precluding direct access methods and frequently necessitating complicated or time consuming file extractions. There are a variety of other file structural problems that frequently prevent convenient usage such as location of sample response weights, geocodes, record type codes, and others.

In the past, inadequate technical documentation and archiving has also been a problem. Sometimes users have been unable to effectively use or understand technical documentation. There have frequently been many assumptions made about the user's knowledge of the file contents. An additional problem has been the absence of a systematic approach to the archiving procedures for existing files.

Some of these problems are solvable, some are not. Looking toward the future, we envision a continued increase in the amount of machine-readable products that will be made available to accommodate more effective analysis. We are now taking steps to better identify, document and resolve those problems that something can be done about -- as will be summarized in Section IV. What is important to note here is that many of the problems associated with processing and analysis of Census Bureau machine-readable products is not user software in-and-of-itself but a number of factors which create demands for special types of software, which in a sense are artificial, as well as difficulties in understanding how to use these files.

III. REVIEW OF EXISTING AND NEEDED USER SOFTWARE

To the extent that the Census Bureau produces machine-readable products, the Bureau has an obligation to insure that users have an opportunity to make effective use of the files. Consequently, certain types of highly transferable software will be produced by the Bureau where there are potential or existing inadequacies in software otherwise available. In addition, as demands for data continue to accelerate and volumes of data continue to increase, there is the need to provide for the more timely dissemination of machine-readable data and alternative forms of access and manipulation. A partial answer to this may be a computer based, terminal oriented, public data information system which will be discussed in Section IV.

There has been a great deal of emphasis placed upon the development of table generating software both within and outside the Bureau to process summary statistic and microdata files. In their most basic form these software are oriented toward retrieval and display. Most Census Bureau summary statistic files are prepared in a tabular structure where the cells corresponding to the rows and columns are sequentially listed in the record. Table oriented, retrieval and display programs with cross-referenced data descriptor files containing English language identifiers have been in high demand. Tabular output from these programs may be generated for user defined geographic area aggregates. There are frequent requests for these types of software to process files, such as county business patterns, simply as a means of displaying specially aggregated data in a meaningful way.

The two most notable examples of this type of software, as applied to Census Bureau 1970 decennial products, are the DAUList program series prepared by the Census Bureau to process the 1970 Census summary tapes and the Data Use and Access Laboratories (DUALabs) '70 Series Automated Census Analysis System. A variety of other retrieval and display software are available that perform these functions, such as the Bureau of Labor Statistics Table Producing Language and Informatics MARK IV, although they are not specifically designed for Census Bureau products.

The Bureau has also developed COCENTS, a more generalized table generating program, capable of displaying summary statistic data and developing estimates from microdata files and displaying these data in a tabular form. More recently, the Census Bureau has been developing the General Tabulating System (GTS) in an attempt to further generalize and extend COCENTS with the possibility of public access in mind. GTS has been developed principally for internal use to provide a generalized software system for preparation of tabulations and related analyses for statistical reports and tape files. While this system is currently in partial operation within the Bureau, it is not yet generally publicly accessible.

There have been very limited efforts by the Bureau to develop more analytically-oriented software which include functions associated with modeling, parameter estimates, statistical tests, estimates and projections of variables, etc. With respect to conventional statistical methods for the analysis of relationships among variables, such as analysis of variance, correlation, regression, contingency table analysis, factor analysis, and other types of multivariate analysis, there is a large amount of software in place to meet user needs. However, there are some needs in this area. For example, there are no generalized, transferable software available to prepare population estimates using the Ratio-correlation and Component Method II techniques employed by the Bureau.

An especially important type of analytical software needed for processing Census Bureau products is that used to develop estimates and multivariate tabulations from microdata files. The reasons for the importance are many: varying types of hierarchical records from microdata file to file, variations in the type of weighting scheme employed to develop estimates or estimates of their standard error, sheer magnitude of many files causing much of the conventionally available software to be too inefficient and costly, and others. Of the commonly available statistical packages, probably the Statistical Package for the Social Sciences (SPSS) and Statistical Analysis System (SAS) are used more with these files than other packages. Certainly, there are other packages available to the user, such as OSIRIS, which also support the basic tabulating techniques. As a result of the cost to use general purpose analytical packages, more specialized microdata file processing software have been developed. Two of the most often used microdata file processing programs are CENTSAID, developed by DUALabs, and COCENTS mentioned earlier.

Important attributes of analytical software for processing summary statistic files include the capability to develop aggregates, ratio scale tabulations, trends, and compute standard errors (as applicable). The structure of many decennial files imposes particular problems due to their non-rectangular structure, suppression indicators, or the numerator or denominator for ratio scales located on different summary tape counts. For some files, such as the annual population estimates or county business patterns, a trend analysis problem is introduced as annual files are produced on a file separate basis. While indeed, SPSS, SAS, EASYTRIEVE, and other software packages support the computational algorithms, they are very difficult to use in many cases due to Census Bureau file configuration.

The Census Bureau produces a great number of geographic related machine-readable products -- extensively geocoded or geographic reference files. Many applications involving these products have required the development and use of a variety of geographic processing software. Several programs have been prepared to develop and maintain Geographic Base Files (GBF) which are computerized representations of metropolitan maps. These programs, for the most part, have been designed to permit their use outside the Bureau. This series of GBF programs does not include analytical software. There have, however, been efforts by the Bureau to develop software which could make GBF's more useful to the user community.

Several distributable programs have been prepared to permit rather specialized record-linkage, matching, and merging applications. ADMATCH was the first distributable program of this type, originally developed for use with the Address Coding Guide and the GBF to provide the capability of geocoding computer readable records containing street addresses. UNIMATCH was subsequently developed to provide a more generalized record-linkage system by employing a user specified matching algorithm. ZIPSTAN was developed as an auxiliary program to work with UNIMATCH to prepare standardized street addresses and add match keys. These types of programs are of principal importance in matching records to a specific geographic segment, or area, which can then be tabulated as a summary statistic for the area, or for an aggregated set of areas. A significant user problem with these specific programs is that they were programmed in IBM Assembler restricting their transferability.

The Geographic-Related Information Display System (GRIDS) was developed in the early 1970's by the Bureau to provide a fairly general computer mapping capability to display data by geographic area. Records processed by GRIDS contain data values to be mapped and their corresponding x, y coordinate. GRIDS has been used extensively with GBF's.

In an effort to bring these programs together into one system, the Comprehensive Manpower Planning Information System is being developed by the Bureau for the Department of Health, Education and Welfare. This system incorporates the use of ZIPSTAN, UNIMATCH, a Geographic Base File, and an address-oriented data file to develop an address-oriented data file with geocodes and x, y coordinates. This file is then aggregated into a summary statistic file using COCENTS, or may be processed by GRIDS to develop value and density maps. The system also makes use of the DIME Area Centroid System (DACS) to develop a boundary file from the GBF. The boundary file can then be used with the summary statistic file by another program (SCANMAP) to prepare value and density maps corresponding to data on the summary statistic file.

Like those described above, most applications of Census Bureau machine-readable data involve the use of other, non-Census Bureau machine-readable data. This process almost always involves the use of specially prepared programs to develop integrated files. In the most typical application, parts or all of two or more files are merged to develop a file which is then used in some type of analysis. This particular process is frequently the most time consuming and expensive phase. In more systematic approaches, data bases are developed from a variety of sources and maintained over time, such as a health planning data base.

During the past several years there has been rather considerable. interest in developing computer based information systems which make use of Census Bureau data files, geographic reference files and non-Census data to serve as a data base for continuing analysis of socio-economic behavior in a particular area. These data system have been associated with computer software, usually of a unique nature, designed to tabulate, display, and analyze trends. There are a great number of such efforts that have been undertaken in the private and public sectors. The Census Bureau provides limited technical support role in assisting Federal and State agencies to develop such systems where they can be cost-effective and useful.

Data base management systems have generally not been utilized to support Census Bureau machine-readable products outside the context of computer based information systems. The reasons for this are many; but the main ones being that they cannot be applied by the average user in a cost-effective way and that these systems are oriented toward transaction processing which does not normally apply to the typical uses of Census Bureau products. We would not anticipate a change in this situation. However, a data base management system might be very effectively applied internally to the Bureau's processing which could greatly facilitate the user's ability to access and use the data through a system such as an interactive public data information system which is discussed in the next section.

Of course many users have developed software to meet their own needs for processing and analyzing Census Bureau and related files that may be of use to others. In the past the Bureau has attempted to promote a clearinghouse for the exchange of information concerning such software, and in some cases supported the distribution of the software itself. One example of such software distributed by the Bureau is the choropleth mapping routine (C-MAP) developed at the University of Idaho. This clearinghouse function has not been used extensively by users; however, we plan to continue efforts along these lines.

In summary, there are needs for specialized user software for processing and analyzing Census Bureau machine-readable products. Most of the existing user software now available to meet these needs has been developed outside of the Census Bureau.

IV. UNMET NEEDS

A major problem in assessing unmet needs for user software is the absence of objective data on this issue. Our impression of the needs for software and problems involved in accessing and using machine-readable data is based upon extensive contact between Bureau staff and major machine-readable data users, our own staff's experience in processing and analyzing these products, both within and outside the Bureau, and limited feedback from the more general user community. Too often recommendations from this latter source do not prove useful due to failure to consider key problems such as volume, dominant types of use, frequency of access, etc.

This section provides an appraisal of unmet needs for user software and other user aids for processing machine-readable products based upon the information we do have. It also outlines current activities and plans that are underway to address these needs. The unmet need for software is of two dimensions: distributable software and access to an interactive time-sharing system. In addition, the apparently unmet needs for convenient processing of machine-readable products include improved technical documentation, file structure convention standards and practices for archiving, user orientation and training, and other user reference and technical aids.

Distributable User Software

There are at least two types of issues to be addressed for distributable user software -- transferability of the software from system to system and the type of function served by the software. Other issues that might be addressed include user convenience, costs for acquisition and use, ease of modification, etc. Clearly, application software described in this section might also be interactively accessed depending upon demand and cost-effectiveness.

Despite the conventions for developing software as set forth by the Federal Information Processing Standards (FIPS), software currently available from the Bureau does not entirely conform with standards. As a result, some of the software is not as transferable from system to system as might be possible. An example of this problem is with UNIMATCH which was programmed in IBM Assembler. Thus, one unmet need to be addressed as additional software is developed is to conform to standards for developing and documenting software that promotes maximum transferability. In cases where this may not be feasible, two versions of a particular program or system could be developed.

Turning to the second issue, several unmet functional area needs for distributable user software can be identified. As we look toward the 1980 decennial program, we are now considering the preparation of basic retrieval and display programs with increased capabilities over those available for use with the 1970 decennial products. Based upon a 1976 survey of summary tape processing centers, 78 of the 96 responding centers indicated that the Bureau should develop software for use with the 1980 decennial files. Forty centers suggested that the software should have improved capabilities over the 1970 DAUList programs. One major improvement might be the development of more generalized table generating software so that it could be used with any summary statistic file produced by the Census Bureau -- provided the appropriate machine-readable data descriptor file has been developed for the particular data file. If we proceed ahead with the development of this system it will be underway by early fiscal 1979 and may also be of use with the 1977 economic census products.

As described earlier, there is a great deal of software already available for performing conventional statistical analysis. The need in this area is for specialized analytical techniques but generalized to meet the needs of a variety of users -- such as market analyses or processes for developing estimates and projections.

County Business Patterns can be used to demonstrate this need. The County Business Patterns (CBP) data are the only source of non-proprietary, annual, county-level data containing employment and payroll characteristics of establishments at the 4-digit SIC level available on a nationwide basis.

Business firms seek to determine geographic market concentrations, either as inputs for their production processes or potential markets for their output. Typically these analyses involve aggregation of employment, payroll, or value of products for one or more 3- or 4-digit level industries and the ranking of the top few counties, SMSA's, or special market areas comprising most of the market. As substantial variations in product mixes might be analyzed by a given firm, manual analysis can become prohibitively expensive -- particularly for smaller firms. In the public sector, State and regional planning and economic development agencies analyze county-level economic activity, frequently requiring detailed industrial data. Applications involve developmental planning to strengthen or expand the existing economic base of an area as well as to provide site location information to firms which might potentially locate within the State. To meet these needs the development of an industrial analysis program is being considered. In its most basic form, the program would prepare a tabular and graphic display to analyze the top ranked areas (e.g., counties) by employment as specified by the user or areas containing user specified percent of market as measured by employment for any combination of SIC's.

Similar analytical software are needed for other machine-readable products. Another excellent application area would be the monthly construction series C-40 housing building permit/authorizations data. At present, however, the problem with these files is more basic -- the files are not properly structured, documented, nor conveniently available.

In addition, some consideration is being given to development of general purpose estimation software. Emphasis is now focused upon population characteristics although the needs exist in a variety of other areas. Several population estimation programs exist internally but are not as transferable, generalized, nor documented as the general user community requires. Some efforts are underway to make some of the population estimation software more available.

Interactive Time-Sharing Systems

Up to this point the focus of the paper has been on publicly distributable software. Due both to the state-of-the-art of computer hardware-software and to user needs which can best be served through alternative methods of data processing, dissemination, and use the paper would be incomplete without considering the possibilities of an interactive time-sharing system. To this end, the Bureau will be undertaking a study during the next year to determine the feasibility of implementing a computer based, terminal-oriented, public data information system. In its fully developed state, this system would afford access to all users to Census Bureau public use data through their own terminals. The system, undergirded by extensive documentation, training courses, and related user assistance, would provide a wide range of retrieval and display, analytical, modeling, and other capabilities to extend the usefulness of both the data base and the system. Cost-effectiveness and the improvements that can be made in extending data dissemination and use of Census Bureau products would be the key considerations in determining whether or not to implement such a system.

The system as envisioned would be in some respects similar to the CANSIM Interactive System developed by Statistics Canada. CANSIM is an interactive, on-line system which may be accessed by users of Statistics Canada data. However, there would be rather considerable differences in hardware, software, operating characteristics, and the scope and size of data base supported.

Through an on-line system users would be provided considerably improved access speed and convenience and support of an interactive dialogue for problem solving and analysis. They would also have greatly enhanced abilities to perform comparative (geographical or distributional) analyses for quicker and less costly interpretation and inference. Statistical estimates could be quickly developed from micro-data files. The need for some printed reports, or selected sections of tabulations, might even be eliminated.

Additionally, interactive facilities would provide procedural problem solving, preprogrammed self-help and tutorial aids, and other user aids oriented toward inquiry-response such as subject content indexing functions. A subject content/geographic data indexing system could be maintained to assist users in locating required data. A comprehensive bibliographic system could be maintained. A message system might be established to keep users apprised of developments and problems regarding Bureau products. Through computer assisted instruction, users could obtain instructions to assist them in accessing, interpreting, or using the data for a particular type problem. Thus, this system could provide not only improved data delivery but also a user education and technical assistance function.

Of course, there are presently interactive systems making extensive use of Census Bureau statistical resources. The most notable examples have been developed by private service bureaus and universities. These systems are both special and general purpose. However, for the more generalized ones, existing public and private efforts in this area exhibit both incompleteness and user inaccessibility. The more generalized systems have not provided sufficient revenues to adequately support and expand them in the private sector. As a consequence, only selected subsets of data are maintained and technical services are not readily available. The profit incentive forces distortion in equal accessibility of such services to all potential users even in quasi-public entities. Experience with interagency funding projects has been less than satisfactory. Private and Federal foundation support has provided for some research and development, but in general has not established a basis for a continuing operation. These considerations are some of the reasons leading the Census Bureau to consider the implementation of such a system.

Technical Documentation and Archiving

As outlined in Section I, a major need relating to the usability of machine-readable products is improved technical documentation and standards and practices for archiving these products. In the case of technical documentation, many files have been made available in the past with little more than a record layout. This, of course, leaves many unanswered questions for the user ranging from precise definitions of subject content tabulations for specific fields to methods of estimating summary statistics and their reliability from microdata files.

Steps are now being taken to improve technical documentation both in terms of comprehensiveness of the documentation -- record layout, file structure, definitions, data file dictionaries, estimation procedures, control count tallies, etc. -- as well as standardization of the formal technical documentation. We are moving toward more systematic use of machine-readable technical documentation which corresponds to standard conventions for naming and identifying fields, describing universe edit processes, identifying and defining valid codes or ranges, etc. During the past year machine-readable technical documentation has been prepared for several publicly distributable products.

A second major problem has been the absence of standards and practices for archiving machine-readable products. One result of this has been the lack of comprehensive inventory of machine-readable products available. Files prepared as special tabulations for narrowly defined uses are frequently not documented nor archived for subsequent dissemination. A more substantive problem has been the lack of a systematic approach to verifying the accuracy of data file contents and then developing a master backup copy. While this has been done in part for many of the major files, such as 1970 decennial summary tapes and public use samples, products in lesser demand have not been given the same attention. An additional problem in this area has also been a lack of standards for file structure ranging from source of geocodes used to field within record and record within file conventions.

We are now taking steps to improve archiving standards and practices. During fiscal 1978, we are developing a manual outlining conventions for developing and maintaining distributable data files. Procedures for developing machine-readable file documentation are being standardized. In addition, a tape file/computer software inventory is now being developed which will be frequently updated. An inventory identifying products prepared from special tabulations since 1970 has been developed and will soon be made available to the user community.

User Reference Aids

Even when standards have been applied for developing files so that they are processible by conventional software, and they are well described in an inventory process and technical documentation and readily accessible, many users lack the required knowledge to make effective, or even correct, usage of the files. Indeed, the lack of user reference aids which provide basic, or cookbook, approaches to the use of these files creates a barrier in some cases resulting in the user lacking a desire to acquire the files or understand how they can be utilized.

As a result, additional user aids need to be developed targeted toward specific user groups or types of uses. These products may be as basic as describing how to develop aggregates or ratio scales from summary statistic files to methods of developing multivariate frequencies from microdata files and analyzing cause and effect or other types of relations between variables.

User Training and Orientation

With the increasing number of machine-readable products becoming available, new developments in software, and increased interest in these products by a larger number of users, it is evident that there is also a need for increased promotion and marketing of available products. In addition, more user training should be provided to familiarize users not only with available files and their characteristics but also methods of making use of the files for analysis.

In the past, training opportunities for users provided by the Census Bureau have been restricted to learning what data products are available, how to acquire them, and how to locate specific data contained in them. More attention is now being given to how to use the products. Courses are now planned on assisting users specifically with the use of machine-readable products.

V. SUMMARY

The matter of primary importance that should now be further discussed and analyzed is the general issue of how to improve the accessibility and usability of Census Bureau machine-readable products. To consider only the availability of and needs for user software, while a critically important issue, focuses too narrowly on the larger issue. As stated earlier, most difficulties associated with processing and analysis of Census Bureau machine-readable products goes beyond user software to include documentation, file structure, user training, reference materials and other user aids. These factors create demands for special types of software which in a sense are artificial. In addition, their absence sometimes leads to incorrect use of the files. More extensive user dialogue on these issues is needed which can be applied to make these files easier to use as well as more useful.

However, some needs for user software can be identified both in terms of distributable user software and a more comprehensive system for accessing Census Bureau statistical resources through an interactive, terminal oriented, system.

APPENDIX A

MAJOR CENSUS BUREAU DISTRIBUTABLE MACHINE-READABLE PRODUCTS

I. SUMMARY STATISTIC FILES

1970 Census of Population and Housing
  First Count
  Second Count
  Third Count
  Fourth Count
  Fifth Count
  Sixth Count
  PC(2) Subject Reports
  Population Centroids
  Adjusted County Data
  County Migration
  Special Tabulations
1972 Economic Census
  Manufacturers
  Governments
  Retail Trade
  Wholesale Trade
  Mineral Industries
  Selected Services
  Merchandise Line Sales
1969 Census of Agriculture
1974 Census of Agriculture
Revenue Sharing Population and Income Estimates
Federal-State Cooperative Program Estimates
County and City Data Book
County Business Patterns

II. MICRODATA FILES

1970 Census of Population and Housing
  Public Use Samples
  Special Tabulations

Appendix A (cont.)

- 1970 Census Employment Survey
- 1960 Census of Population and Housing
  - Public Use Sample
- Annual Housing Survey
- Survey of Income and Education
- Current Population Survey
  - Annual Demographic File
  - Special Tabulations
- Survey of Purchasers and Ownership
- Survey of Scientists and Engineers
- Survey of Government Employment
- Survey of Government Finances
- Truck Inventory and Use Survey (1967, 1972)

III. GEOGRAPHIC AND OTHER REFERENCE FILES

- 1970 Census of Population and Housing
  - Master Enumeration District List
  - Address Coding Guide
  - Urban Atlas-Tract Boundaries
  - ZIP-Tract Cross Reference File
- Geographic Base Files
- School District Geographic Reference File
- County Group Reference File
- 1972 Economic/Geographic Reference File
- Area Measurement File
- City Reference File
- PICADAD
- DIMECO
- Spanish Surnames File

CENSUS SOFTWARE NEEDS OF STATE AND LOCAL GOVERNMENTS

HAROLD B. KING
THE URBAN INSTITUTE
WASHINGTON, D.C.

To address the software needs of state and local governments in using census data, it is necessary to initially make some very broad generalizations. The first of these is that the computing capabilities of these organizations vary from very sophisticated to non-existent. The second is that if we address the needs of these organizations by focusing on the data they will be seeking from the Census, we will be able to infer something about their software needs. The third and broadest, generalization, is that all of these organizations have similar software needs and differ only in their computing capabilities to process data and their levels of sophistication in analyzing it.

The last statement suggests that the proper approach in assessing the needs of a governmental unit might best be based on size rather than type. Studies have shown that there is a very high correlation between the size of a governmental unit, its computing capability and its analytical sophistication.

COMPUTING CAPABILITY

A report by the International City Management Association (ICMA) states, "Although there has been considerable growth in computing adoptions in cities, computer capacity is not very great except in the largest cities (500,000 population and over). The overall scale of EDP usage, which can be assessed by examining the total number of operational applications in cities, is directly related to city size."[1] Based on the 1972 County and City Data Book, only 26 cities fell into the 500,000 and over category.

A similar report on counties[2] states that large scale computer facilities normally occur only in those counties with populations over 250,000. In 1972 there were 150 counties with populations greater than 250,000.

It would seem then that our major target user group would be comprised of 50 state governments, 150 counties and 26 cities, or 226 governmental units. The 1972 Census of Governments[3] indicates that at that time there were 50 states, 3,044 counties and 35,408 municipalities and townships. Based on population size alone, our major audience would thus be comprised of only 0.59 percent of the total.

Another way of looking at the computer capabilities issue would be to examine the types of EDP tasks performed by these governmental units. James Danziger, in "Computers, Local Governments and the Litany to EDP",[4] develops a typology of EDP tasks which one might find useful in describing the types of processing performed by a local government. Of interest for this discussion are two of these types: record re-structuring and sophisticated analytics.

1. Record re-structuring. This type of task is related to the re-structuring and re-aggregating of records. It indicates a level of sophistication at a local government EDP operation, which suggests they would be capable of re-formatting census tapes and performing simple descriptive statistics on the file such as crosstabs, frequency counts and aggregations.

2. Sophisticated analytics. Danziger defines this as a type of activity which includes simulation studies, regression models and geo-coded data bases. In general, these applications utilize sophisticated mathematical methods or special technical capabilities of the computer to examine data.

In the above-cited ICMA studies, cities of 50,000 or more responding to a survey indicated that record re-structuring comprised only 6 percent of their total operational applications, and sophisticated analytics comprised only 5 percent. The results for counties of 100,000 and over were similar. Of those counties responding, record re-structuring accounted for 7 percent of total applications and sophisticated analytics accounted for 4 percent.

These survey results would suggest that the majority of counties and other local governments have neither the computer resources nor the analytical capabilities to develop software to access complex census data files. This conclusion is further supported by the results of a survey conducted by the Public

Policy Research Organization (PPRO) and reported in *Nation's Cities*.[5] The survey showed that most chief executives felt [illegible] the greatest problem associated with data processing in their local governments was that data they needed for the analysis of specific questions was not available to them. They felt that the data they needed was being collected and stored, but that their computer systems were not integrated enough to present summary data to top management. Most local government computer applications continue to serve only clerical and information retrieval needs of individual departments and agencies.

The picture at the state level seems to be much brighter. *The 1976-1977 Report on Information Systems Technology in State Government*[6] identified 603 computers in use at the state level in 1976. (Florida did not report in 1976 but had 20 machines listed in 1975.) These machines ranged from some of the largest machines commercially available to mini-computers, with almost all of the major manufacturers represented. Of the 49 states reporting in 1976, twenty-three reported having ten or more computers. These varied from a high of 40 in New York state to a low of one in Wyoming.

Computer applications were also varied. Uses ranged from Driver Licensing to Resource Management. But, again, the majority of the computer applications tended to serve clerical and information retrieval needs of state departments and agencies.

What all this says is that there can be no one software solution to meet the census data processing needs of states and local governments.

Some units will have highly sophisticated processing and analytical capabilities, and will be capable of developing their own software.

Others will be more capable of using sophisticated software, and would be more than happy to receive a fully tested and well documented software package from the Census Bureau.

At the other end of the spectrum will be local governments which will have no computing capabilities or will require extremely simple software to generate descriptive statistics from small area data available from the Census.

To meet this latter demand, The Urban Institute has developed a simple multiple crosstab program.[7] This program was created to help local governments analyze survey data which they had collected in order to evaluate governmental operations. Although the package was well documented and the instructions for using it were simple, we found it necessary to supply technical assistance to the users. The Institute's experience suggests that any Census program established to meet the demand for this type of software will have to be supported by a group which will provide technical assistance for the installation and use of such software.

Along similar lines, a conference was held at The Urban Institute in 1971 titled, "Workshop on Census and the Cities." Its purpose was to determine the type of assistance needed by local governments in processing the 1970 Census of Population and Housing. In attendance were representatives from both large and small local governments, consulting organizations, and the U.S. Census Bureau.

The hypothesis around which the conference was formed was that there was a lot of useful data in the 1970 Census, and that a coordinated effort by a few foundations and non-profit organizations could result in software products which would make this data readily available to local governments. The concept was to survey the data needs of local governments, and use the resulting information to determine how best to meet those needs.

The survey was never performed because the general consensus of the meeting attendees was that most local governments would find it difficult to specify their data needs as they related to census data. Instead, it was felt that a massive training program would have to be mounted to inform potential local government users about what was available and how it could be used to answer questions and solve problems related to their own governments.

The attendees agreed that such a program would be extremely costly and would probably need a large government subsidy. To the best of my knowledge, nothing further was done along these lines in assisting local governments directly.

As a result, small local governments which attempted to use machine readable census data found the going rough. Most of the available foundation money was used to support software development to meet the needs of universities, research organizations, and large governmental units. Little was done to develop capability at the local level outside of regional presentations by the Census Bureau.

Since the Institute meeting, some work has been accomplished in attempting to determine the data processing capabilities of local governments. The ICMA and PPRO surveys mentioned earlier have been part of this.

A general conclusion which can be made, then, is that the majority of the local governments at this time are not able to make use of census-developed software products. Even though this is the case, all of these governments have a need for a mechanism which will insure timely and easy access to this data.

CENSUS DATA NEEDS

One of the major reasons states and local governments need timely access to census data is so they may evaluate it and determine it's accuracy. Many programs which make monies available to these governmental units are based on head counts and housing unit counts. If these numbers do not appear accurate, the states and local governments will be pressing for recounts.

Another major need for data will be for redistricting purposes. Many entrepreneurs had anticipated a heavy use of computerized redistricting software in the 1970's. This did not materialize because of the extremely political nature of this process. The costs involved in the use of this software and its related data bases also discouraged many from attempting it.

All states and many local governments will need social and economic data to support applications for grants. Formula grants in particular require the availability of accurate census data. The formulas are based on such data as population, income levels, and need, or a combination of these factors.

The fiscal 1976 formula for Title II of the Comprehensive Employment and Training Act (CETA) of 1973 is a good example. A three-part formula was used to determine the allocations: 50 percent was based on last fiscal year's allotment; 37.5 percent was based on unemployment; and 12.5 percent was based on the number of adults in low-income families in each prime sponsor area.

In addition, a local government must have a population of 100,000 or more to be eligible for a grant. For governmental units close to 100,000 population, accurate population statistics will be most important (assuming they indicate a population greater than 100,080).

There are a number of other grants besides CETA which use formulas for determining eligibility and allocations. Some of these are:

1. Community Development Block Grants
2. General Revenue Sharing
3. Special Food Service Program for Children
4. LEAA - Comprehensive Planning Grants
5. LEAA - Improving and Strengthening Law Enforcement and Criminal Justice
6. Industrial Development Grants

7. Urban Mass Transportation Capital and Operating Formula Grants

8. Highway Research, Planning and Construction

9. Low Income Housing Assistance Program

Data is also needed by state and local governments for such programs as urban renewal, housing code enforcement, community action, and resource allocation. Resource allocation includes locating schools, fire stations, police stations and community service centers.

Allocating educational resources requires information on family incomes, status, children age groupings, and adult education levels.

In reviewing the needs of states and local governments for census data, a few issues stand out clearly. One of these is that much of the data produced by the Census Bureau is aggregated to a unit (i.e., tract, block, block group, etc.) which does not relate to local boundaries.

Local planning units, on the other hand, need data available at such local data analysis levels as school districts, redevelopment areas, congressional districts and traffic areas. The availability of geo-coding and address matching schemes have aided in the use of census data, but they are still expensive methods for solving problems.

A more organized approach to determining data needs of local governments might be arrived at by identifying:

1. Departments which may be major statistical data users

2. Functions for which data are needed

3. Data analysis areas

4. Uses for data

5. Data types

The following lists are not exhaustive, but they are a good indication of the broad areas of statistical data needs which many state and local governments have.

Departments

General Administration
Personnel
Planning
Utilities
Police
Inspections
Public Works
Welfare
Manpower

Finance
Budget
Housing and Rehabilitation
Education
Fire
Zoning
Traffic and Transportation
Health

Functions

Industrial Development
Health
Public Safety
Employment
Education
Land Use
Urban Redevelopment
Migration

Public Works and City Engineering
Welfare
Recreation
Transportation
Commercial Development
Urban Planning
Neighborhood Development
Housing

Data Analysis Areas

States
Counties
Municipalities
Townships
School Districts
Fire Districts
Police Districts
Traffic Districts
Wards
Streets
Street Segments

Blocks
Block Sides
Households
Census Tracts
Regional Planning Districts
Redevelopment Areas
Standard Metropolitan Statistical Areas
Soil Conservation Districts
Flood Control Districts
Census Enumeration Districts

Uses For Data

Plan New Facilities
Plan New Programs
Estimate Size of Clientele
Estimate Needs of Clientele
Anticipate Staff Needs
Policy Evaluation
Program and Project Evaluation
Support Project Proposals
Continuing Research
Support Grant Applications

Data Types

Voting Records
Welfare Records
Police Records
Marriage Records
Birth/Death Records
Individual Case Histories
Union Records
School Census Data
Employment Statistics
Income Statistics
Hospital Records
Traffic Data
Migration Data
Street Location/Numbers
Family Characteristics
Land Use Data
Insurance Data
Population Densities
Population Projections
Housing Characteristics
Utilities Data
Tax Records
Federal Reserve Data
City Engineering Records
Land Values
Fire Records
Housing Market Data
Air Pollution Data

Whether a governmental unit uses any or all of these depends somewhat on size and authority. For example, only about 14 of the largest 43 cities in the United States operate welfare departments, as this is predominantly a county function.

The Census Bureau, in an attempt to clarify user needs, held a series of open public meetings. These meetings were sponsored and organized at the local level, and conducted with joint participation of local persons and Census Bureau staff members. Held between October 1974 and July 1975, the meetings were conducted in 73 cities covering all 50 states and the District of Columbia, with over 6,000 local participants. In a "Synthesis of Local Public Meetings,"[8] the Bureau presented an eleven-page description of data items and their tabulations which were compiled from these meetings.

The same type of meetings were held with state agencies. There were 16 regional meetings of this type, and all but two states (Arkansas and Colorado) had representatives at these meetings. In a "State Agency Meetings Synthesis,"[9] the Bureau again compiled an eleven-page description of data items and tabulations suggested by the participants of these meetings.

The listings from both of these reports are too numerous to be duplicated here, but they do support the hypothesis that the data needs of states and local governments are similar.

COMPUTER SOFTWARE NEEDS

The discussion so far has pointed out the wide range of computing capability from the largest state to the smallest local government. It has also presented an abbreviated description of data needs. With these computing capabilities and data needs in mind, we can now turn to computer software needs.

It is difficult to describe computer software needs of states and local governments under the categories assigned to this conference: Data Organization; Data Tabulation; and Data Presentation. In many cases an item could easily fall into two or three categories. There are also needs which do not fall into any of these categories. With this caveat, an attempt will be made at categorizing these overlapping needs. Items which seem not to fit any category will be listed under a category titled "General".

Organization

One of the most significant needs of local governments is to have data organized by geographical and political boundaries which are more meaningful to them. This could include files organized by school district, ward, traffic district, congressional district or any of the data analysis areas listed earlier. A special school district file was created from the 1970 census. More of this type of special tabulation should be made available.

It would also be helpful if tapes were made available by subject area such as economics, transportation, and health. A number of meetings were held during preparation for the 1970 census at which this idea was proposed. Nothing was done about it then. It is worth reiterating.

There is a need to supply more income data and have it disaggregated into various sub categories such as government income transfer programs (i.e., how much of a family's income is comprised of housing support payments, aid to families with dependent children and food stamps).

Along this line, there is a need for finer category breakdowns in other areas. Some of these categories should also be extended. A good example is age. The category "65 and over" isn't very helpful for planners working on problems of the aged.

Another useful addition to census data files would be the categorization of data on local governments by size. Many other resource materials present local government data by population size. (i.e., 500,000 and over, 250,000 - 499,999, etc.). This

One type of software which would be useful to state agencies and some local governments would be a program which would assist them in making inter-censal projections using census supplied data or locally generated data. The decennial census data is almost out of date when local governments get access to it. As stated earlier, only a few of these governments have the capability to write their own estimation software. Some, and perhaps most, aren't even aware of the techniques available to perform these calculations.

Another type of software which would be useful would be programs which would assist local governments in studying transportation patterns and migration patterns. This software would be extremely helpful if it could produce maps and symbols on printers which could be easily understood by local government personnel concerned with these problems.

The local public meetings and state agency meetings organized by the Census Bureau identified a number of special tabulations which these governmental units would like to have produced. Most of these tabulations could become available at a reasonable cost by a restructuring of the files. Alternatively, the federal government might subsidize the machine costs which would result from processing the files as they were structured in 1970.

Presentation

It is not clear that additional software to support data presentation would be useful for the majority of state and local governments if an output display device other than a printer would aid those interested in comparing data from various sources.

One item which would be most useful to small local governments would be the addition of means, medians, and standard deviations for most major items by geo-political and census areas. This would help those small local governments that do not produce these simple statistics. This data, accompanied by some descriptive documentation explaining the meaning and value of the statistics and some examples of their use, would be very helpful.

Another item to be considered is the development of tools and techniques that would allow users to compare data items over time for areas whose boundaries continue to change. In this same category is the need to have county groups not cross state lines. This makes it extremely difficult to aggregate county group data to the state level in order to increase sample size when working with public use sample tapes.

Tabulation

If the Census Bureau decides not to prepare files organized by geo-political boundaries, and/or subject areas meaningful to local governments, then it should be prepared to supply special tabulations to meet these needs. These special tabulations should be inexpensive to obtain and should be available in a timely manner. When these special tabulations are prepared, there should be a mechanism available by which potential users could learn of their existence.

were required. The surveys identified in this paper indicate that few governmental units have graphic terminals or plotters. There seems to be sufficient software currently available to produce tables on printers in a variety of formats.

As an example of the limitations on graphics capability, of the forty states reporting on peripheral equipment in the National Association of State Information Systems survey cited earlier, only three listed the availability of graphics terminals. Although thirty-one listed plotters, the majority of these were located in highway or transportation departments.

What might be useful would be the availability to local governments of printouts of meaningful data on their areas. This would not necessitate the development of new software, but might require the establishment of a user service sub group at the Census Bureau to respond to local government requests.

As mentioned earlier, the development of a simple multiple crosstab program by The Urban Institute was useful to some local governments. This type of software development, aimed at the small local governments that are not involved in sophisticated analysis, could be very useful. Most of the tabulation software currently available requires a level of expertise not available in these governments.

General

The first thing that is obvious from the inventories of state and local government computers is the need to develop software which is machine independent. If there is a desire to support the full size range of systems, it will also be necessary to develop software which can be run on a system with a small amount of core and few peripheral input and output devices.

There also seems to be a trend towards the use of minicomputers. Many local governments that found EDP costs to be a limiting factor in their acquisition of a computer are now rethinking the issue. The National Association of State Information Systems report shows an increasing trend on the part of state governments in acquiring minicomputers. This is an area that the Census Bureau should explore as a means for increasing access to their machine readable products.

Since the Census Bureau is releasing more files with a heirarchical structure (Current Population Survey, Decennial Public Use Sample, etc.), they should develop software which would facilitate the use of these files. Also, primary records on these files should indicate the number of sub-records following when the number of these sub-records is variable.

Data tapes should be cleaned and edited prior to their release. "Dirty" tapes could be released when access to the data is needed before cleaning and editing were completed. These tapes should be replaced when the clean versions become available. A program should be established to alert all users to new errors as they are detected. Data tapes should be treated as a planned product of the Census Bureau rather than as by-products of other functions.

Any software prepared for use by states and local governments should be available when the data tapes become available.

If this is not done, those governmental units wishing to analyze the data will again develop their own software if they have the capability. The other local governments will have to find other means for solving data problems.

Finally, the Bureau might consider putting their data files on-line and charging a reasonable fee for access. This has been done by other organizations (an example is the 1970 Decennial Public Use Sample file on the ACCESS system at the Massachusetts Institute of Technology). If the software supporting these files made retreival and analysis simple for the unsophisticated user, it would go far towards solving the data needs of states and local governments.

CONCLUSIONS

Although most states seem to be capable of using census developed software, it appears that the majority of local governments do not have the equipment or personnel to avail themselves of these proposed products.

For those that do, there is always the problem of transferability. Danziger states,[10] in reference to software transferability, that "a striking finding when particular local governments are examined is that successful examples of technology transfer are rare". This view is also supported by The Urban Information Systems Inter-Agency Committee's (USAC) experience. Of the millions of dollars of software developed through the USAC program, only a relatively few software packages were adopted by other municipalities. Conversely, the GBF/DIME package seems to have found wide acceptance by those local governments capable of handling that particular software package.

Although all states and most local governments have the need for more ready access to census produced data, only a relatively small number will be able or willing to use census produced software. This may be largely attributable to the insufficient knowledge at the local government level about how to use census data effectively. A well planned training program aimed at these governments might well raise the level of knowledge, and help to create an environment in which census produced software could be more effectively utilized.

FOOTNOTES

1. Kraemer, K. L., Dutton, W. H., and Matthews, J. R., "Municipal Computers: Growth, Usage, and Management," Urban Data Service Reports, Vol. 7 No. 11 (Washington, D.C.: International City Management Association, November 1975).

2. Matthews, J. R., Dutton, W. H., Kraemer, K. L. "County Computers: Growth, Usage, and Management," Urban Data Services Reports, Vol. 8 No. 2 (Washington, D.C.: International City Management Association, February 1976).

3. U.S. Bureau of the Census, "Census of Governments, 1972," Vol. 1 Governmental Organization, G.P.O., Washington, D.C., 1973.

4. Danziger, J., "Computers, Local Government and the Litany to EDP," Irvine, California: University of California, Public Policy Research Organization, 1975.

5. "Chief Executives, Local Government and Computers", a special report in Nation's Cities, Vol. 13 No. 10 (pp. 17-40), October 1975.

6. National Association for State Information Systems, "Information Systems Technology in State Government," NASIS, Lexington Kentucky, 1977.

7. Gueron, J., Ouyang, B., "UI-MCTAB, A Multiple Crosstab Program," The Urban Institute, Washington, D.C. 1974.

8. "Synthesis of Local Public Meetings," a report by the U.S. Bureau of the Census, March 1977.

9. "State Agency Meetings Synthesis," a report by the U.S. Bureau of the Census, September 1976.

10. op. cit. Danziger, J.

BUSINESS USE OF CENSUS DATA

Richard B. Ellis

Marketing Manager - Information

American Telephone & Telegraph Company

APPLICATIONS

Although the Bell System and its parent company, the American Telephone & Telegraph Company, are only a small portion of the vast and complex American business community, their use of census data is quite varied and, hopefully, will cover a majority of the applications generally used in business today. The Bell System's use of census data falls into three broad categories:

1) Provision of Products and Services. Many of Bell's basic products and services are currently furnished under regulated franchise which carries with it the obligation to have available what the customer wants when he wants it at a reasonable cost. Since relatively long lead times are required to manufacture and install some of the equipment to permit this, detailed demographic trends and forecasts are required for the thousands of areas we serve to predict with as much accuracy as possible future populations and their communications needs. This involves such elements as population size and make-up, migration trends, business development, household formations and constituencies, etc.

- Marketing and Corporate Management. For discretionary communication products and services, Bell is in direct and indirect competition with many other suppliers and consumer goods. Here the business objective is to optimize its product line, distribution channels and market position. Although individual market studies are often the source of the basic data, extrapolation of these findings into generalized forecasts, predictions and strategies is heavily dependent on demographic data. Typical applications include estimation of market potential for individual products or market areas, media selection for promotional activities, selection of areas for merchandizing effects and retail outlet site selection.

- Social and Labor Force Studies. As a major social and employment force, the Bell System has a requirement to track and predict changes in the society it serves and the work force it employs, in order to assess the impact of not only its own actions but various legislative and judicial mandates that may come into force. Typical problems faced in this area include the changing nature of the family/household unit, ethnic balance of the employee group, the entry of women into the labor market, and the availability and movement of skilled craft workers.

It can be seen then that Bell's need for census data is significant, quite varied, and subject to relatively rapid change over time.

There are three broad areas of concern which transcend, to some extent, the categories specified for this conference:

DATA ACCESS

As in the case of many other business users, Bell has relied very heavily on intermediate suppliers for the actual data used and has satisfied a minority of its needs by direct access to the Bureau and the original data. The comments and suggestions of these suppliers have been incorporated in this paper where appropriate. Although this was, to some extent, a planned condition for the 1970 Census and our experience has been good, there is an open question as to whether this is the best way to operate in the long run. As our needs and data volumes increase, in-house processing may become attractive. Should we obtain such data directly or indirectly? Could the Bureau organize to meet demands which, in all probability would be sporadic and subject to heavy peak loads? There do not appear to be any facile answers, but the problem should be addressed.

TIMELINESS

An endemic problem for us and most other users we are aware of is the speed with which the data becomes physically available for use. A year is the customary minimum from completion of a survey to availability. Granted the volumes are huge in many cases, but data processing technology today will surely permit a more timely response.

HOUSEHOLDS

In terms of product and service consumption, the household is a very complex unit. In the case of certain home related services or consumer durables (e.g., basic telephone service, furniture) the household itself may bé construed to be the consumer. In the case of more personal products (e.g., toll calls, clothing) the individual is normally thought of as the consumer. In fact, the distribution of purchase and acquisition decisions runs the gamut between these extremes, colored in many cases by different value systems and personal perceptions. The present household tabulations offered by the census do not adequately address this significant diversity.

Specifically, the following items deserve attention.

1. Below the national level 1970 Census households income distributions were usually broken down into families and unrelated individuals. A more useful division would be households with related individuals and those with only unrelated individuals. Since 1970 the proportion of households in the latter category has been increasing and indications are that that trend will continued thru 1980.

   If the tabulation for unrelated individuals is retained, it should at least be broken down into single-person households and persons in (non-institutional) group quarters. Furthermore, this information is of broad enough interest to warrant making it readily accessible in published form.

2. 1970 Census households were typed according to their "heads". This designation will be changed in 1980 to "the person (or one of the persons) in whose name the home is owned or rented". This suggests three classifications for each of the two household categories above: (1) joint owners/renters; (2) male owner/renter; and (3) female owner/renter.

3. The tabulations in the 1970 Summary Count did not include breakdowns by the number of wage-earners in a household. Particularly in the case of families, this information is an important determinant of socioeconomic needs and consumption patterns. With female participation in the labor force currently on the increase, it is important to measure the contribution made by working women to a family's (household's) income. It will probably be preferable to base the breakdown on full-time workers rather than all wage-earners; i.e., do not include part-time workers.

4. More research is also needed into the best way(s) to aggregate households and persons in terms of the relationships between the economic decisions they make and their socioeconomic characteristics. For instance, which decisions in households with multiple wage earners are generally made collectively and which are left to individuals.

Over and above these three general items, other areas of concern include;

ORGANIZATION

1. Summary Tapes

After the 1970 Census an additional Fifth Count Summary Tape for block groups and enumeration districts (known as File C) was processed at the expense of one of the suppliers. This tape has been used extensively by organizations which reallocate demographic data from census areas to user-defined areas. The 1980 Census, including sample questions, should be designed under the assumption that a similar tape will be made available as a standard product.

2. Public Use Sample Tapes

   a. 1970 PUS tapes had nonstandard labels (leading numeric characters rather than alphabetics). Unless an important reason for this exists, the ease of tape usage would be improved by putting standard labels on the 1980 PUS tapes.

   b. Certain of the 1970 tapes contained information for multiple states, presumably for reasons of storage efficiency. Users needing data on the last state of that tape had to read thru the records for all preceding states. If the multi-state tapes were organized into separate files for each state, the processing time could be greatly reduced.

   c. When cross-tabulations of particular census data items did not appear in the 1970 Summary Tapes, programs were written to compile the necessary data from the 1970 PUS tapes. Unfortunately, for reasons of confidentiality the smallest geographic units for which data on the latter tapes is specifically identified are individual counties of 250,000 or more within SMSA's. The 1980 PUS could be broken down to a lower geographic level, e.g., census tracts or rural counties, with a corresponding decrease in the number of data categories, e.g., income in $1000 rather than $100 intervals. If disclosure problems still existed, the Census Bureau could write a general-purpose program to produce the cross-tabulations and check the output for confidentiality problems. The usefulness of this program would be maximized if it were accessible interactively through the Summary Tape Processing Centers or their equivalent.

TABULATION

1. Racial Classification

It is unnecessary to belabor the point, but the problem of racial classification remains. We are aware that the Bureau is working to ameliorate this difficulty and it is hoped that they succeed. Accurate racial information is essential if work force targets and other population influenced goals are to be determined on a rational basis.

2. Public Use Sample

For many applications, the Public Use Sample is too small and, in many cases, it is necessary to combine several political and/

or economic areas to obtain usable statistics. These then must be imputed to the smaller areas within them which is a statistically questionable technique. A larger, more detailed sample together with the format suggestions listed under "ORGANIZATION" would produce a much more usable and credible product.

3. Households with Telephones

The need for a survey of households with telephones has been documented ("Should 1980 Census Data Include Information on Telephones?") Phil Welch, May 20, 1977) and acted upon with an appropriately worded question in the recent Oakland pretest questionnaire. This data will be most valuable to the user community if it is cross-tabulated by other selected characteristics. In particular, households with and without telephones should be cross-tabulated with the demographic characteristics of the owner/renter of the housing unit such as his/her age, race and sex. These households should also be cross-tabulated with total household income, presence and age of children, and the classifications mentioned "Households", above, i.e., families vs. unrelated individuals, male vs. female vs. joint owner/renters, and number of wage earners. These cross-tabulations should not only fulfill the needs of the telephone industry and related governmental agencies, but also allow the many public and private organizations which perform surveys by telephone to more precisely estimate the bias in the results they compile.

PRESENTATION

1. Auxiliary information - As census data users we are interested in examining demographic statistics for areas defined by our organizations rather than the census areas. The most practical and time-efficient way to establish the necessary correspondence between these areas is through the use of geographic or geodetic information provided by the Census Bureau for the census areas. At least two such compilations were provided after 1970:

   a. The Master Enumeration District List (MEDList) contains the geodetic coordinates of the population centroids of blockgroups and enumeration districts. The Census Bureau is not sure whether they will provide this information for the 1980 Census. Because of its importance and the urgency of its release, the Census Bureau should consider making arrangements to have this work done quickly and accurately by an outside organization.

   b. The maps of census tracts and enumeration districts are essential companions to the MEDList - they are used to verify the geographic translation of user areas into component census areas. While the 1970 census tract maps were made available on a timely basis, the maps for the nontracted areas have been very difficult to obtain. Both sets of maps should be released shortly after (if not slightly before) the Census Day in 1980.

c. The Urban Atlas contains geodetic definitions of census tracts. The preponderence of errors in this source indicates that the validation portion of its creation procedure was inadequate. Either this procedure needs to be improved or the Census Bureau could again consider contracting for this work with an outside organization.

2. Alternative medium - The very nature of magnetic tapes leads to inefficiencies in terms of serial or sequential processing rather than random access. The Census Bureau should seriously consider supplying the 1980 data on another medium, e.g., "floppy disk," that could be processed more efficiently.

SUMMARY

To summarize this statement of our wants, needs and concerns, we would like to offer a brief description of the "ideal" census information system from the business user's viewpoint:

1. Statistics on all census questionnaire responses from short and long forms available to the blockgroup/enumeration district (BG/ED) level;

2. Cross-tabulations among selected statistics which are defined by the user;

3. Sufficient geographic information, e.g., geodetic references for BG/ED, to allow reaggregation of census data to user defined areas;

4. Detailed migration information, e.g., cross-reference by county to aid estimation of intercensal migration; and

5. Information ready for users less than one year after its collection.

ORGANIZATION OF DATA: CONSIDERATIONS RELEVANT TO THE DEVELOPMENT OF USER ORIENTED SOFTWARE THAT MIGHT ENHANCE THE UTILITY OF DATA GENERATED BY THE U.S. BUREAU OF CENSUS[1]

by Mervin E. Muller[2]

World Bank, Washington, D.C.

[1] An invited paper to lead the discussion on Organization of Data at the joint American Statistical Association and U.S. Bureau of Census Conference on User Oriented Software, November 8-10, 1977, Arlington, Virginia.

[2] Comments made here do not represent official views of the World Bank.

CONTENTS

SUMMARY

Several questions are raised in order to identify the complexities and challenges that are involved in trying to understand better what is the problem of data organization. These questions should help the discussion to take place during these meetings by indicating areas of research and development. Some of the questions have been made in order to ensure that they will be addressed. These questions are not necessarily new but are ones that must be faced by those currently involved with statistical analyses using computers even though satisfactory solutions may not be forthcoming at this time.

INTRODUCTION

Under the terms of reference of this conference, this paper has been prepared to stimulate thinking prior to the conference and during the conference in order that we can focus more effectively on what types of software ought to be developed to aid in the area of data organization. This problem must be viewed in a rather general context in order to justify the attention given to it at this conference. It is much larger than one might first believe. It is tempting to assume that all we need to do is select from among the existing data base management systems and our problem will, in fact, be solved.

I hope this paper will generate light, rather than heat. Having stated this hope, I want to question whether we have an adequate understanding of what we are trying to accomplish, even though the objectives sent to us prior to this meeting were clearly presented. I expect to raise several questions that are provocative and hopefully useful, stimulating the kind of thinking the subject needs. I had considered and discarded several alternatives for this paper, such as: 1) summarizing the history of the subject, 2) advocating a particular approach or system, 3) evaluating existing systems, or 4) emphasizing existing limitations. I hope through considering questions we can develop proper respect for the problem and the importance of establishing priorities for a meaningful and effective research and development effort in this area.

2. For What Purpose?

The indicated purpose of the conference is for "the development and perfection of software which will enhance utility of data generated by the Bureau". The conference will also "examine the need for software improvements from the user's standpoint and help determine the extent to which the development of software is an appropriate topic for research support by the NSF/ASA." Although these statements are clear enough, I believe that we need to make them more specific in order to provide a focus for what should be considered. I think it is important for the conference attendees to discuss and refine the purpose of the conference. I hope the questions raised in this paper will help clarify the point, "for what purpose?" as well as help to focus attention on subsequent actions to be taken based on the conference.

3. Who are the users, what are their needs and what are their priorities?

The term "user" can mean different things to different people. Users could be those directly within the Bureau, or those within other parts of the Department of Commerce, other parts of government, or those external to government. It is important to know who the users are and what their backgrounds are expected to be: are they to be professional statisticians, experts in computing, or subject matter specialists who will have the appropriate supporting staff, equipment, and software to assist them in the use of data? It is necessary to identify what their needs are--particularly, what their data needs are. Can they be sure that they have useful data and data identification in the sense of the following: how will they cope with missing data? How will they be able to recognize questionable accuracy or quality? These questions will be dealt with again in Section 9 on Data Organization. Different people have different needs, and to develop appropriate software for data organization(s), it is necessary to identify who are the users, and what are their needs. Finally, what are the relative priorities of different user needs? It would be irresponsible to ignore the matter of priorities, since users clearly have finite resources. Even a government agency must also face the reality that it has neither the time nor the resources to meet all software or data needs of all users. Therefore, when directing planning and development, attention must be given to how one would go about identifying user needs and establishing priorities for what is to be done.

4. What Time Horizon?

To have proper perspective for the discussion to follow, it is necessary to look at least on two aspects of time: the time horizon of planning and development, and the time span of the data themselves. By focussing on these two aspects of time, I believe we can ask relevant questions and see more clearly how to meet the objectives of this conference. Consequently, both aspects of time are given attention before proceeding to some of the other considerations. For completeness, a third aspect of time is also mentioned.

4.1 For Planning and Development

Whenever we look ahead, there are at least two pitfalls: first, confining ourselves to the use of current technology and knowledge we possess about how to use such technology to solve today's problems; and second, restricting our thinking about the problems themselves due to conservatism or recognition of the limitations of current technology. When looking at the question of the development of user oriented software, it is not at all clear whether we are talking about what can be done this year; or three years hence at the time of the 1980 census; or at the time of the next decennial census in 1990; or 20 years ahead in the year 2000. The symbolic year 1984, indeed, falls in the early part of this broader planning period.

In looking forward we might also look back a similar time period to assess progress made.

Twenty years ago Fisher was still with us; computing was in its infancy. How far have we come since then? The breadth of application of statistical techniques has been greatly influenced by the availability of statistical software on digital computers. With few exceptions, notably in graphics, and some changes in emphasis notably towards iterative methods, the world is much as Fisher knew it. We are still, in the main, equipped analytically to handle numerical data in rectangular form (univariate or multivariate) variables by observations by time.

Although we are now able to store and retrieve non-numeric data, or data in non-rectangular interrelated structures, we lack analytical tools to support analysis directly using more complex data structures.

It is important to be realistic as to what time horizon we are addressing as we proceed in the subsequent discussion before we can really be sure what types of planning and development would be appropriate for consideration. For example, is it meaningful to consider that significant technological or theoretical break-throughs may occur in time to be of benefit? Are we looking ahead to the possibility of a data network where the hardware and/or data can be considered distributed, geographically and logically? Clearly, if this is a possibility, then more attention must be given to improved ease of access to the data in the presence of controls which recognize privacy, confidentiality and security, and this affects the selection of data organizations. According to the time horizon, I can easily imagine that we will develop different plans and approaches.

4.2 Span of Data

In looking at questions of data organization, there are two questions regarding the time span of the data: 1) are the data (actual or predicted) to be organized and maintained only for current time periods or current time periods plus historical periods? 2) are the data for each time period to be maintained separately? The influence of these considerations on data organization also depends upon the extent of data and the frequency of use. The possibility of data migration from one hardware device to another is also affected by whether the data must be currently available or available only for historical archival purposes. We will address this point in a later section.

4.3 Data by Variable vs. Data by Time Periods

If we think of data organized as time series, this type of organization is not the one naturally employed when collecting social or economic data, but it may be the desirable type of data organization for analysis or reporting purposes. Usually we obtain social or economic data for a given time point or period for many variables. This is the natural way to collect census data. However, for a given variable an analysis, even for data consistency, may make it necessary to use data by variable across time periods. The time aspects of data for a given variable raise many interesting challenges and questions with respect to data organization. When data are stored on a direct access device there can be an erroneous impression that it is immaterial how the data are organized and stored. That is, to assemble a time series of the values $\{X_i\ (t),\ \text{for}\ t = 1, 2...T\}$ for a given variable X_i , when the data are stored by time period and variable, some people may assume that it is convenient and efficient to retrieve the desired data values by searching for each time value of each variable. This assumption may be correct if for n data points the search effort can be done in less than Knlogn operations. However, reorganizing the data to be a collection of time series by first sorting the data and then using it sequentially may be a more efficient and effective approach.

Even with such brief considerations of this section, I think you will agree that it is important for data organization to take into account the many time aspects of data.

5. Modes and Frequency of Use

It is necessary to consider the modes of data use and the frequency of data use. Frequency of data use will have important ramifications for data organization, which are considered in more detail in Section 9. I find it useful to distinguish four categories of computer use, namely, production mode, diagnostic test mode, tutorial mode, and exploratory mode. As noted in Muller (1969), one reason for considering these four modes is to facilitate separating the problems of using computers into understandable and manageable parts, which may also help clarify issues and close the current gaps between hopes and achievements in use of computers. Another reason is to obtain better understanding of where to allocate research and development effort in programming and statistical techniques. Some of us still suffer from the expectation that a given "general program" can be all things to all people. Of the four modes of use, the one that most people think of is the production mode, i.e. the one the user employs to accomplish a specific computing job which no longer requires testing programs. It is assumed one knows what he wants done and how to do it (even though the user may also need help of the diagnostic test mode.)

The diagnostic mode is used to aid in testing whether or not a program or package can in fact be used for production purposes.

In a tutorial mode one may want help from a specialized computer program to learn, for example, 1) how to use a program, 2) how to understand and use available data, 3) how to use the available computer facilities, or 4) what programs or data are available. The tutorial mode is intended to support the learning of a particular body of knowledge. In the context of the current conference, the tutorial mode might enable users of Census Bureau data to explore various data bases and software that can be used, including descriptions of data structures that are available, and data coding conventions and the like which are relevant to using the data.

An alternative to the tutorial mode is to maintain and distribute comparable information by more conventional means. The questions to be answered here are those of costs and benefits of each approach.

The fourth mode, exploratory mode, allows the user to explore existing programs, computer languages, and operating systems so they can understand what they are doing. For example, what levels of precision of calculations are available? Is truncated or rounded arithmetic used in the programs?

6. Recognition of Inertia

In spite of spectacular technological achievements in hardware, it is important to recognize that development of computing techniques for improving the quality and usefulness of data suffer from inertia, in particular progress in the software that would be required to bring about changes commensurate with the spectacular improvements in hardware. If one now reviews the proceedings of the 1969 conference on statistical computing held in Wisconsin, it will be noted that most of the open research and development problems identified then are still with us. (See Milton and Nelder (1969)). There are few significant break-throughs in statistical techniques for data editing, data analyses for presentation, or data organization; the work of Fellegi and Holt on data editing, or the work on intervention analysis by Box and Tiao or on data organization by Merten are exceptional cases. Thus the lead times between identifying problems and finding practical solutions may be very long. One must recognize how difficult it can be to overcome inertia without a high priority emphasis and critical investment of people's time. Although we have on-line and interactive computing capabilities, we are far from the situation of being able to perform on-line, interactive statistical analysis.

This conference and the subsequent commitment of considerable resources may provide the critical mass needed to overcome the current intertia, if there is adequate follow-up. This inertia is reinforced by the present concern over privacy and fears of invasion of privacy, as well as by broader issues of confidentiality, including unintentional disclosure.

Another type of inertia is the failure to recognize how little progress has been made on standards for data identification and control. Until there is such progress, the obstacles to portability of software and data (see e.g. Muller (1975)) will inhibit, slow down, or preclude effective general use of available data.

7. A Necessary Pre-requisite: Data Identification

For those who were practicing statisticians before the wide use of computers, data code books were a familiar part of a well-designed data collection and analysis process. "Code" is used here to include any type of data identification. A few computer-based systems have computer-readable code books; some people refer to them as "data dictionaries" or, as I prefer, "data glossaries" (to indicate a capability richer than just a code book or dictionary, see Muller (1963)). I seriously question how data can be easily portable without a clear indication that codes can have different meanings at different times, or that at a given time multiple codes may have the same meaning. It is unrealistic to expect that this problem can be overcome by universal standards. Instead, I would urge that a necessary pre-requisite to improving the use of data is to create data-base directories which will enable the user to recognize and cope with different interpretations of data identification. Such data directories often must include the identification of the quality, source, and timeliness of the data. They may also include the identification of the various data structures used.

8. Current Data Base Management Systems: "There is still no free lunch".

There are many aspects to the current literature on data base management. There is the schema of total data base management where one looks for a way of describing the logical properties of the enterprise, or agency, the use of data, and the logical organization of the data to be used. There are some impressive capabilities, such as data definition languages. Unfortunately, many of the important stochastic considerations that influence how to design and effectively use such data bases either are not handled in existing data management systems or are ignored. The data base systems are usually designed as if to be used in a totally deterministic manner.

We seldom get anything free. Data base systems require an investment of resources to acquire or build the system as well as the cost of maintaining it, converting to it, and training people in how to use it. In some respects those advocating or using data base management systems are justifying them on the ground of increase in programmer productivity, with arguments similar to those employed to justify higher level programming languages as replacements to machine code or assemblers. There is clearly a need to increase programming productivity. In this sense, some data base management systems can provide programming tools to facilitate the input, output, and transfer of data across physical storage devices.

Associated with these tools is the expectation that there will be greater data and program independence as a result of having "the appropriate data base management system". Another expectation is that the system is extensible to changing user data needs. Although some of these systems have been around for a long time, I have not seen case histories documenting how such systems have contributed to improved statistical analyses or better portability of data. Unless one is clear about the time horizon and the needed research and development for organization of data, great opportunities for the distribution of data bases by data networks will be missed or delayed because data base management capabilities (techniques and software) are not adequate to take advantage of the hardware and telecommunications enhancements.

To face these emerging problems by means of newly-designed "data base management systems" which do not yet exist will take time and could be costly. As statisticians, we should be interested in the collection and analyses of data to evaluate how to design, use, or modify such systems of data base management, recognizing that pre-packaged systems are not likely to solve all of the important problems.

9. Data Organization and Avoidance of Fallacies

The literature is full of papers on how "best" to organize data, as if there were some set of criteria of optimum data organization. By itself, such a factor as frequency of use is an inadequate criterion for deciding how to organize the data. Even with additional information there may be no "optimum" data organization, see Merten and Muller (1972). For exclusively batch processing, one might want a data organization that would minimize the average access time, whereas in an interactive use of data one might need a form of data organization which would ensure stability of response time--for example, a minimum variance in the service access time to obtain the data. Unfortunately, there is no single optimum data organization.

Another fallacy is that there should be only a single data organization for a given set of data. This is one of the limitations associated with current data base systems. As a minimum, one may want one type of data organization for the effective and efficient maintenance of the data, but multiple forms of data organization for different types of use to be made of the data--for example, a data organization by time period and a data organization by variable to aid the construction of time series. The question of what should be "the" data organization is too general a formulation to be of much concrete value. In many respects, organizing data for effective use resembles designing a queueing system with the arrivals and possibly the service being stochastic processes. In addition to this point, will the time horizon for the research and development effort cover a sufficient span to consider distributed hardware and data bases? Is it necessary to maintain historical data? Does the frequency or volume of use warrant techniques to allow for the migration of data to various physical devices? As a minimum, the data may be organized in such a way as to be portable by having the identification of data and coding structures, and the data codes external to the data content. Current data base management systems sometimes inhibit portability of data, or make it necessary for a potential user of the data to make a large investment to acquire the entire data base system in order to use a given set of data. Furthermore, for some applications, control must be provided against unwarranted access. Such systems could be unacceptable because of the need to reprocess or even reorganize the data so that they can become portable to multiple users with different access privileges.

As in the case of hardware, it is reasonable to look forward to large economies of scale through having data bases maintained and distributed from central data services. If so, additional research by statisticians will be needed to determine what kind of data, where the data should be located, and how it should be organized. Here, again, we will need criteria to indicate who the users are, for what purposes they need the data, what are their modes and frequency of use. We also need to keep current on the relative costs of transmission and processing of data. I hope I have not disappointed anybody by recommending a relatively modest approach to these problems; I do not believe that the field has had enough research or is matured enough to cope satisfactorily with the complexity of the present situation.

Considering data organization from a purely deterministic point of view, much of the current literature which follows the results of the CODASYL Committee is relevant. This point of view treats data organization in terms of logical and physical descriptions to aid computer programmers, and several important issues of languages for data description and data structure are addressed. For example, data systems are described as network models, hierarchical models, or relational models, to mention a few. If one looks closely at these efforts, however, no criteria are being put forward in terms of how many levels of a hierarchy one should have or, in the relational model, how one describes the data internally to achieve efficient use of the data. Much of this effort is aimed at allowing data independence so that programs and data can be changed without affecting the end-users.

Although such formal descriptions of data bases can be of great help, they neglect the questions of effectiveness and efficiency, and I believe these issues are stochastic in nature. Also neglected is the matter of indicating or organizing data according to source, quality, or timeliness. We statisticians recognize that there are a wide class of problems where stratification can improve sampling efficiency. Similar advantages can be gained through using stratification techniques with regard to the organization and distribution of data bases. With stratification it may be advantageous to establish one or more data or access directories at various levels of a hierarchy or network. Stratification can also help to eliminate conflicts on data access within the computer, as well as to improve service performance with respect to average service time or variance of service time--to mention two performance characteristics which one might want to consider. It is not at all clear what performance criteria one should use. One can formulate the performance problem as a mathematical programming model and look at the question of optimization relative to some objective function, but to date I have not found this a very useful description other than to demonstrate the existence of a solution, see for example Merten (1970). In view of the sensitivity of "optima" to assumptions about data which themselves are subject to unknown changes, caution is required here. Perhaps the views of those in attendance can help clarify the priority to be given to optimization criteria.

Data organization includes the question of security and control, what types of user access will be allowed, and for what purpose. Furthermore, some parts of a record may be considered sensitive and therefore should have some type of encrypting or scrambling to protect the sensitive parts--another case where multiple files using different forms of organization may be appropriate.

10. Data Organization and Non-numeric Information

In the future some types of data organization should exist to handle non-numeric information, which I believe is necessary to consider, especially if the time horizon of the research and development effort exceeds a few years. Ordinarily, one tends to consider non-numeric information to be synonymous with text. Even this kind of data offers unexploited opportunities for data analysis. Although some types of data organization already include the facility to handle text such as footnotes, report titles, table headings, stubs, and user instructions, I believe that we need to consider more complicated data organizations and storage facilities, capable of handling digital representations of graphs, maps, and pictures. With satellite capabilities to "collect pictures" and "create maps", and with the emergence of satellite or fiber optics communications for digital transmission, statisticians now need to plan how they can improve analysis and presentation of such results. Additional statistical techniques may also be required to use this technology. The challenge of non-numeric information is here; are we prepared to accept it?

11. Use of Models to Analyze Data Organization

Analytical models to describe and evaluate the performance of data organizations or the associated software can have a place in a research and development effort if they provide useful reductions of the complexities of the real world. I believe that the models should have a stochastic orientation. Such models, to be useful, must reflect qualitative as well as quantitative factors of relevance to the user, such as ease of learning or ease of use. However, care must be taken not to lose sight of the end objective of achieving effective data organization and software. Unless one can collect real data to validate the reasonableness of a model, one should, in my opinion, suspect the conclusions or usefulness of modeling efforts.

12. Procedural vs. Problem Approaches

Most of the higher-level languages available today are effective if one is prepared to describe a problem using data (or the organization of data) in terms of procedures. The same could be said of most large-scale statistical packages that are now available to analyze data. One of the attractions of some data base systems is that they have commands which are more problem-oriented than procedure-oriented. The advantage of such a command structure depends on how important it is to adopt a problem approach rather than the procedural approach to the use of the data. The question is how much research and development effort is needed here.

The answer will depend on identifying the users, their needs, and the time horizon. If the users are experts in programming and have been trained in ways that exist today, then it would seem natural to use a procedural approach. However, in looking ahead it is not at all clear that this is what is desired if it is intended to stimulate the use of census data outside of the Bureau.

With problem-oriented software one could describe the problem rather than the procedures--for example, identify the file, the particular record types of fields within the file that one would want--and then the criteria for selection and analyses of data, rather than the detailed procedures. On the basis of the problem specifications, special compilers or translators would analyze the problem specification, either to generate procedural calls for use by conventional compilers or to translate the specifications to procedures interpretatively. I believe this is a fruitful area of research.

The problem approach has ambiguities, not so much in the syntax for problem specification as in the semantics of determining whether or not the specification of the problem permits a correct, unique and unambiguous computer execution. Without trying to prejudge what the future direction of the study should be, I think it should start with straightforward and practical problems followed by cases of greater complexity. Some of my colleagues and I have been looking at this challenge for some time, and we believe it has relevance to situations involving the need to accomplish multi-dimensional data array manipulations and transformations. In this area we feel we have been relatively successful, but it is an area needir additional research and development, see for example Muller (1977).

13. Distributed Systems and Distributed Users

It is difficult to accept the views held by some data base systems advocates that current data base management systems provide a solution to the data organization problem. Under the best of conditions, such systems may be solving some of today's problems, but these are not necessarily the problems that will be facing us for the next few years. A sizable investment is required to select and install current data base systems. Such investment may divert resources from needed research and development.

As data files become larger, it seems logical to expect that there will be an increasing need for data organization that allows the data to be distributed across hierarchical storage devices. It is logical to expect that, depending upon the time horizon under consideration, the data could be distributed geographically. Depending on who the users are, and their objectives, it seems reasonable to investigate distributed data bases as a logical and effective approach. The question then arises, is it reasonable to assume that the users need distributed data? I believe it is realistic to assume that the users will be distributed and want to use distributed data bases. Attention must be given to access control, security, and the need for tutorial modes of use to enable users to understand and use data if they no longer go to a central facility to acquire the data. This raises problems of maintenance both of the data and of software. Consequently, the question I see here is, what criteria should one consider as statisticians in making decisions about distribution of hardware, software, and the users, and what ramifications will this have on the usability of data?

As noted earlier, the question of distribution of users is related to the question of economies of scale. Large general-purpose machines have been popular because they offer economies of scale. Intelligent terminals with local memory undoubtedly will generate additional uses of centralized large-scale general purpose machines. I believe we can expect to realize economies of scale for data bases in data networks with smart terminals without necessarily having all the data in one file. One of the questions that needs clarification is how to achieve effectiveness and still enjoy economies of scale.

14. Challenges for Statisticians and Computer Scientists

It is a real challenge to bring together computer scientists and statisticians to identify who the users are, and what their needs are. A second challenge is to recognize that the design and evaluation of systems to cope with data gaps involves a problem of statistical analysis. A third is, the need for evaluations of the performance of different data organizations, the software using the data from such organizations, and the software for the statistical analysis using the given data organizations. The evaluation, I believe, should be based on carefully designed statistical experiments so that one can estimate the main effects and interaction effects of the various parameters one might have under control. I am using the term "interaction effects" in the sense employed by a statistician who has designed, say, a factorial experiment. I believe this is a very fruitful and necessary area to consider, one well worth receiving an allocation of resources for future research, and I would hope that attention will be given to this area.

15. Questions and Types of Software

15.1 Questions to be answered

The types of software research and development to be recommended by this conference depend in part upon which questions we decide should be pursued.

The questions can include:

- Time horizon
  - planning and development for: 1978, 1980, 1984, 1990 or ?
  - span of data: current only, historical only, future, or some combinations
  - data by variable vs. data by time period
- Data for what purposes?
- Who are the users, what are their needs, what are their priorities?
- What modes of use are to be supported: production, diagnostic, tutorial, exploratory?
- How frequently are the data to be updated, distributed, used?
- What data identification will be needed, how will it be distributed, and how will it be maintained?
- Will data standards be formulated and maintained?
- Will portable data directories be established and required?
- Types of data base systems: centralized, distributed, decentralized.
- Where should the data be located?
- Who should control access to the data or the data directories?
- Will non-numeric information be part of some of the data bases?
- Will statistical techniques be used to gather data or perform analyses to influence data organization?

- What software will be developed that are acceptable to different users for conversions of data from one type of data organization to others?
- Will there be an agency prepared to provide software to convert data?
- What levels of data security and confidentiality are required?
- What back-up facilities are required to ensure uninterrupted user services?
- Should problem-oriented software be developed to access and use the data bases?
- What extent of distributed systems and users are to be supported?
- What financial and human resources can be made available for various types of effort?

All of these questions have political as well as technical aspects, especially those involving security, privacy, confidentiality, the use of distributed data or networks, or use of the data by commercial service bureaus.

15.2 Types of Software

The types of software to be developed depend in part upon how the selected questions are answered. In addition, the types of software to be developed should reflect the kinds of statistical analyses that are expected to be needed and available. I am concerned that unless explicit attention is focussed on statistical questions, software development will be undertaken without an adequate underlying statistical basis. Take, for example, analyses allowing for missing data or techniques to classify multivariate data as being suspect or defective depending upon how the data are to be presented or used.

Regardless of the types of software to be developed, there are further questions needing attention, possibly by others not attending this conference. These questions include who will

- develop the software
- test the software
- distribute the software
- maintain the software
- administer requests to change the software.

I believe software is needed to:

- collect and maintain data on the use of data bases (such data can be used for evaluation purposes to influence data organization as well as clarify whether there is sufficient demand for use of the data)
- control access for the creation, modification, removal, or distribution of data, as well as determine when simultaneous use of the data can be permitted.
- maintain portable data directories for those who have different data organizations or equipment
- restrict data to forms that are compatible with the user's environment
- handle centralized or decentralized data bases
- monitor use of data so as to notify users when, subsequent to their access to the data, errors are detected in the data, including audit trails where needed.
- store, retrieve, and use non-numeric statistical image information such as graphs, maps, and pictures
- monitor use of data to estimate what data to have, where, and for whom

- develop software to allocate and reallocate dynamically the locations of the data, the amount of main memory to be used, and the access routines to be used. Such software may belong to the operating control system of the hardware network, but it should be designed in such a way as to be portable for users.
- make possible uninterrupted service or error recovery with a minimum loss of information for any users accessing data bases by means of a data network
- provide problem-oriented software as well as procedure-oriented software.

I am assuming that software to enable use of distributed equipment will be available as well as necessary software to create audit trails making possible data recovery due to environmental or equipment interruptions.

16. Basic Questions, Priorities, and Research Directions

I have raised several questions which I believe to be basic, in order to identify and understand the challenges that ought to be faced now and in the next few years. We must recognize that priorities are to be established and that resources are to be found and allocated. Depending on the time horizon selected, and the resources that can be expected to be available over the period, it may be necessary to assign relative priorities on the basis of likelihood of success, or at the other extreme, on the basis of likelihood that the projects are of such long duration and high risk that no other group could be expected to handle them. Therefore, the research directions could be the selection either of safe efforts with high likelihood of success or efforts that are the most risky, leaving the safer ones to others who do not have large staff or other resources. Sometime before this conference ends I hope that we will stimulate interest in seeking answers to the questions of who are the users, what should be done, and with what priorities. Some of these questions can be resolved by the use of cost/benefit analyses. These can be difficult since they should take into account social and economic costs and benefits as well as financial.

17. Reasons to be Optimistic

In spite of the large number of questions that I have proposed, I am optimistic, because I believe that many of the significant enhancements and developments that have taken place in computing were developed to meet the needs of statisticians at the Bureau of Census. Therefore, I believe that if we concentrate on needs and the required statistical tools, the development of the appropriate software and hardware will follow. Today it seems easier to consider hardware development. I believe that if we concentrate on the analytical statistical questions, the subsequent software development will take place.

I am optimistic because I believe that meaningful research can only result from having real and practical problems. Again, if one looks back at the influence of the Bureau of Census on development of both statistics and hardware, this was successful because it was related to real needs and real problems.

I am also optimistic because we see a joint effort between the Bureau and ASA. This is good, because many of the problems require people from more than one discipline, especially in the area of determining how to perform evaluations of software. In this sense, the existence of the ASA Section on Statistical Computing is another reason for optimism, as are some of the activities taking place outside the United States. We are increasingly looking beyond our shores in the area of computing, as we have in the past with respect to the theory and application of statistics.

I am optimistic also because of activities such as those planned for the International Association for Statistical Computing.

I am optimistic because I can see significant contributions being made by groups outside the United States that can easily influence the kinds of activities that ought to be taking place within the United States. Consider, for example, computer-based data editing, such as that which is going on in the World Fertility Survey through CONCOR, and in the efforts of Statistics Canada.

Finally, I feel optimistic because of the recognition of the need to hold such a conference as this one, composed of people prepared to meet in working groups and devote time and effort to identify what needs to be done.

ACKNOWLEDGMENTS

This is to express my appreciation to George W. Barclay, George M. Minich, and Leonard Steinberg for their helpful comments on the initial draft of this paper.

REFERENCES

Merten, Alan (1970), "Some Quantitative Techniques for File Organization", Ph.D. Thesis, Computer Sciences Department, University of Wisconsin.

Merten, A.G., and Muller, M.E. (1972), "Variance minimization in single machine sequencing problems", Management Sci., 18, No. 8, pp. 518-528.

Milton, R.C. and Nelder, J.A. eds., (1969), Statistical Computation, Academic Press, New York.

Müller, M.E. (1963), "A foundation for modern tools of management", 1963 Proceedings [illegible] International Conference sponsored by the American Institu[illegible] Industrial Engineers, New York, pp. 123-134.

Muller, M.E. (1969), "Statistics and computers in relation to large data bases", Statistical Computation, Milton, R.C. and Nelder, J.A., Eds., Academic Press, New York, pp. 87-176.

Muller, M.E. (1975), "Portability standards for software", Computer Science and Statistics, Proceedings of Eighth Annual Symposium of the Interface, ed. J.W. Frane: U.C.L.A., 173-176.

Muller, M.E. (1977), "An approach to multidimensional data array processing by computer", Comm. A.C.M., 20, No. 2, pp. 63-77.

Organization of Data for Census Users

By Bruce Carmichael, Warren Besore, and Kam Tse
Systems Software Division
U.S. Bureau of the Census

Herman Hollerith, the inventor of the punch-card tabulating machine that was the forerunner of modern computers, was a Census employee. His invention was motivated by the volume of data acquired in a decennial census that by 1890 had grown to the extent that current methods were hard-pressed to complete the processing of one census before the next was begun.

The problem of data volume is still with the Census Bureau and its users. In the 1970 census, information was collected from some 65 million households. Twenty per cent of these households completed a long form of the census questionnaire that provided a comprehensive view of their lifestyle. Today the Bureau is looking increasingly to sophisticated data organization schemes and access methods to manage this huge volume of data.

This conference was convened to examine the problem of distribution of Census data: specifically whether the distribution of software for accessing and processing Census data would make this data more easily accessible and closer to the needs of users. It is readily understood that if data is distributed in a manner that requires extensive processing to extract information in a useable form, its use is restricted to those who possess the facilities and the funds to afford the processing. This paper looks at some of the new techniques in data organization that the Bureau is using, and some of the facilities available commercially, to see if the Bureau's data organization technology can be extended to service the needs of users.

I CURRENT PRACTICE

Perhaps a good place to start is to look at the way in which data distributed by the Bureau is currently organized, and the software available for accessing it. While the subject of this discussion covers all types of data distributed by the Bureau, we will cover briefly data derived from the decennial census as an illustration of the format in which data is organized for distribution.

Raw data from a decennial census is stored on magnetic tape, grouped by geographic area. Once the raw data is edited and validated, a set of basic data tapes is created for use within the Bureau. These tapes comprise the basic, or micro, data from which summary files and special tabulations are derived. Because no disclosure analysis of confidential information has been applied, these basic data tapes must remain internal files.

External Data Organization and Use

From the basic data files, various summary files and public use samples are prepared and disclosure analysis and data suppression are applied on these files. These files are maintained on tape for distribution to Census data users. The summary files are grouped into two categories: summary counts and subject reports. The summary counts for the 1970 census occupy approximately 2200 tape reels, the subject reports about 400 reels, and the public use samples another 200 reels. These tapes are available in 556, 800, and 1600 bpi recording densities.

The summary count tapes are ordered by the type of tables contained, level of geography, and state. Thus if one were interested in the population aged 25-34 living in Suitland and earning over $15,000, this information would be located in "FILE C" of the "5th COUNT" Maryland summary tapes. Since four tape reels are required to hold this file, the user would probably have to read all four reels to locate the desired information.

A set of extraction programs called DUAList is available to assist the user in locating and displaying tables on the summary tapes. For certain cases, the extraction programs even had limited aggregation capabilities. More extensive or detailed extractions had to be performed by custom programs. Because of the difficulty of developing such programs, processing centers employed specialized staff for this work.

There are several improvements that could be made in the distribution of data. Newer tape drives permit higher recording densities, requiring fewer reels of tape to hold the files. Different tape formats can facilitate processing the data. A larger variety of extractions can reduce the amount of processing required of the users.

On the whole, though, sequential file organization, and sequential processing of data, has reached its limit. If we are to ever make any progress in reducing the cost of processing Census data, we must come up with a new way of organizing this vast volume of data that will make it manageable. This organization method should include a common logical model of the data and its structure, and a common method for accessing the data. Furthermore, this model should be compatible with the Bureau's internal structures. Compatibility would be beneficial in two ways. Data could be more timely if user-accessible data would be updated in the same form as internal data rather than having to be translated. Secondly it would require fewer resources to extend the Bureau's software systems than to go through the process of developing a new system from scratch, and then interfacing it with the Bureau's.

II. COMMERCIALLY AVAILABLE SOFTWARE, HARDWARE, AND SERVICES

A brief survey of commercially available computer products and services might be appropriate to identify items which could be utilized in distributing Census data.

Software Packages

There are well over a hundred commercially available packages that are advertised as Data Base Management Systems (DBMS) and the number is constantly growing. Packages conforming to the report published in 1970 by the CODASYL committee are available for the equipment produced by each major computer manufacturer. This report is the basis of an industry standard for a Data Manipulation Language (DML) and a Data Definition Language (DDL) for data bases built on a network model. More recently, with the emerging research by Codd and others, new DBMS are being developed and tested which present to the user a relational model of a data base.

Generally speaking, commercially available packages can be classified into the following categories:

Data Retrieval Systems

File Management Systems

Complex File Systems

Data Base Management Systems

Special-Purpose Systems

Due to the volume and complexity of Census Data, we will limit ourselves to surveying DBMS only. The following table presents basic information on 11 of the more popular DBMS. Host language packages provide a DML that is embedded in a conventional high-level programming language, usually COBOL, FORTRAN, or PL/1. Translation of the DML is generally implemented through an enhanced compiler,

Host Language CODASYL
Data Base Management Systems

	DMS 1100	IDMS	Honeywell IDS-II
CPU	UNIVAC 1100	IBM 370	H6000
Item Description	COBOL Oriented	Host Language Like	COBOL Like
Logical	Network	Network	Network
Physical	Pointers	Pointers	Pointers
Access Methods	Direct Hashed ISAM Network	Direct Hashed Network	Direct Hashed ISAM Network
D.B. Creation	User Programs	User Programs	Utility
Query Language	Yes	No	Yes
Report Generator	Yes	No	No
Host Language	COBOL FORTRAN	COBOL FORTRAN	COBOL
Multi-thread	Yes	Yes	Yes
Security	None	Thru Subschema	Password
Data Validation	None	None	Yes
Recovery	Full-Scale	Full-Scale	Full-Scale
Surveillance	Log Tapes and Statistics Collection	None	Yes

Figure 1

Host Language Non-CODASYL
Data Base Management Systems

	Burroughs DMS II	Cincom TOTAL	IBM IMS	MRI System 2000	Software AG ADABAS
CPU	Burroughs 6700, 7700	IBM 370 ** CDC 6000 UNIVAC 70	IBM 370	IBM 370 UNIVAC 1100 CDC 6000	IBM 370
Item Description	Host Language Like	Host Language Like	Host Language Like	Host Language Like	Host Language Like
Logical	Network	Multi-list	Hierarchy	Tree Structured	Almost Relational
Physical	Pointer	Pointer	Adjacency	Adjacency	Pointers
Access Methods	Direct, Hased ISAM, Bit Vector, Network	Direct Sequential Hashed		Direct Sequential Inverted Indices	Direct Inverted Indices Hashed
D.B. Creation	User Programs	User Programs	User Programs	Utility & User Programs	Utility & User Programs
Query Language	Yes	Yes	Yes	Yes	Yes
Report Generator	Yes	Yes	Yes	Yes	Yes
Host Language	ALGOL PL/I COBOL	Any lang. with sub-routine calls	COBOL PL/I Assembler	COBOL FORTRAN	COBOL FORTRAN PL/I Assembler ADASCRIPT
Multi-thread	Yes	Yes	Yes	Yes	Yes
Security	None	None	Yes	Yes	Yes
Data Validation	Some	None	None	Some	Some
Recovery	Full-scale	Some	Yes	Full-Scale	Full-scale
Surveillance	Some	None	Some	Log Tapes	Log Tapes

**also: PDP-11
Honeywell 2000
IBM System/3
NCR Century
Varian V70

Figure 2

SELF-CONTAINED
Data Base Management Systems

	Computer Corp. of America Model 204	Meade Technology Data/Central	TRW GIM II
CPU	IBM 370	IBM 370	IBM 370 UNIVAC 1100 PDP-11
Item Description	Character String	Character String	Character string Numeric
Logical	Almost Relational	Multi-list	Almost Relational
Physical	Pointers	Adjacency	Pointers
Access Methods	Sequential Inverted Indices Hashed	Inverted Indices Sequential	Inverted Indices Hashed
D.B. Creation	Utility	Utility	Utility
Query Language	Yes	Yes	Yes
Report Generator	No	No	Yes
Host Language	COBOL, FORTRAN, PL/I, Assembler	Any language with subroutine call	COBOL and Own
Multi-thread	Yes	Yes	Yes
Security	Yes	Yes	Yes
Data Validation	Yes	No	Yes
Recovery	Some	Yes	Yes
Surveillance	Yes	Yes	Some

Figure 3

a pre-compiler, or subroutine calls. Many DBMS also offer self-contained query languages for on-line interactive retrieval and update.

Hardware/Systems

In most production environments, DBMS is treated like any other program sharing the resources of the host computer. Even in those installations that dedicate a computer to data base applications, the hardware configuration and operating system software are not modified. It is common for big corporations or government agencies to employ a large- or medium-scale computer running a DBMS supplied by a software firm or by the computer manufacturer. In these installations, the data base resides on mass-storage. Numerous interactive terminals access the computer for on-line updates and instant information retrievals. Jobs for batch updates and periodic report generation are either run concurrently with on-line processing or during off-shifts, depending on the capacity of the hardware.

With the recent proliferation of minicomputers, many firms have come to possess one or more of them. There are two basic methods of employing minis for data base applications. One is a stand-alone system. Smaller companies may own or share only one mini which they use for all their computing requirements including data base.

A second method is a distributed network. Bigger corporations may own several minis and possibly some large- or medium-scale computers, in geographically dispersed locations. In addition, they may have a number of data bases of various sizes, some of which are useful only to a particular branch. In this instance, a distributed data base network would be more suitable. Each node of the network would possess a mini to handle its local data base work, In addition to the traditional approaches, there has been active research toward the implementation of a so-called data base machine. Some researchers are considering a hybrid machine in which special processors are added to the conventional general-purpose computers. For example, one such attempt was to add an Associated File Processor, implemented on a PDP-11, to perform associative (parallel) searching of a very large textual data base. Others have suggested that the architecture of the conventional computer should be changed to accomodate the functions provided by the DBMS, especially those connected with the relational model.

Computer Service Organizations

In the current marketplace, it is unnecessary for an organization to own or rent a computer in order to have access to diversified computing services, including data base packages. Many companies are in the business of providing a computing utility, much in the way the phone company provides a communications utility. One such service is General Electric's Mark III, which is described here as an illustration of the kinds of services available. This is not meant to imply that Mark III is either the best or most comprehensive of such services.

Mark III has thousands of customers on a world-wide network. Many of the customers have large volumes of data stored on Mark III. Each customer can access his data base interactively or in a batch mode, using either his own programs or a generalized software package furnished by G.E.

Local phone numbers are available in all major U.S. cities that allow users to connect to the Mark III network. Twenty-four hour, toll-free service numbers are staffed by consultants who will assist a user needing help or encountering problems.

Generalized software currently available on Mark III includes their own data base package, DMS II, which interfaces with FORTRAN as well as with specialized software packages such as plotting routines, report writers, and interactive query programs. Non-programmers can perform their own statistical manipulation of the data, such as row and column sums, averages, percentages, and deviations.

Custom Census Data Processing Services

If a user of Census data requires a more customized form of computer service, he can turn to one of a number of outside organizations equipped to perform specialized processing of summary and sample data. Some of these organizations provide a broad line of services, while others have concentrated on specialized types of work.

One such organization is DUAL Labs. Again, this description is intended as an illustration and does not imply endorsement of any organization. DUAL Labs is a non-profit corporation offering a variety of services. They provide consulting services and training as well as custom processing of Census data. DUAL Labs does not have its own computer installation, but instead buys computing services to support their work. A fair amount of generalized software has been developed by DUAL Labs, including extraction software for summary data that makes use of a data dictionary and provides aggregation capability; and software for making and documenting vertical and horizontal cuts of public use samples. This software has also been sold to users. Some DUAL Labs cooperating offices provide their software on a time-sharing basis to users. In fact, DUAL Labs provides the type of service that many countries offer through their national statistical offices.

Other organizations, such as National Planning Data, provide more specialized services, such as making ED data available on microfiche, digitizing tract boundaries, or providing population density or affirmative action information.

III. PLANNED DEVELOPMENT

The Census Bureau is working toward an integrated system for the collection, processing, and presentation of Census data. The focal point of this system will be a data base management system that will provide the structure for and access to the data.

The Bureau has selected Univac's DMS-1100 for its initial development work. A data base for administrative data is already operational under this system.

One area to which the Bureau is applying new data organization technology is disclosure analysis and data suppression. A system for automated disclosure analysis is being developed for use in the 1977 Economic Census. This system uses a highly-structured and easily accessible geographic lattice to provide containment and intersection information.

Another current data base project is the Master Reference File for the 1980 demographic census. This file will be linked to a geographic lattice. The data base will allow interactive reference and update for such pre-census activity as mailing counts and boundary and annexation changes, as well as controlling the activities of enumerators across the country during the census. Preliminary field counts will be compared with predicted counts in each geographic area to determine whether they appear reasonable, and counts that are suspect will be flagged for re-count.

From these current projects, the Bureau's aim is to develop a good model for geographic structure of its data, and to develop an in-place geographic lattice.

GTS-3, the third level of the Bureau's Generalized Tabulation System, will contain an interface for data base access. GTS-3 will use the data base both for source data and storage of intermediate results. Data base interfaces will also be built for graphics and statistical analysis systems.

Data base technology serves two functions at the Bureau. One function is to integrate data. It provides a structure for data and improves the efficiency of data access, since needed items may be accessed without passing the whole file. It separates physical storage from applications logic, providing flexibility in storage medium and allowing access software to be optimized. It avoids duplication of data, and provides a single control of all data allowing rapid distribution and correction of data while avoiding the problems of consistency encountered when data is kept in many files. Data base technology also serves to integrate software by providing a common form for passing data between processing subsystems.

IV. POTENTIAL FOR DEVELOPMENT

External Data Organization and Use

As the technology of new hardware and software systems makes the use of more highly-structured data a possibility for Census data users, it will become increasingly important to develop a common logical model of Census data, both at the summary and micro levels, that is compatible with the Bureau's data organization. A common model will also allow the distribution of pre-structured data on new mass storage media such as honeycomb cells or holograms. It will also make it possible to take advantage of new data organization technology, such as associative accessing, without modifying user programs.

The addition of a time dimension to Census data is another innovation that will be possible through the use of new hardware and software technology. Access to time-series data allows the projection of trends and patterns, but requires massive amounts of on-line storage and sophisticated retrieval techniques. The Census Bureau has developed a simple time-series data base system that is used on small economic data bases. Statistics Canada provides limited amounts of time-series data through its CANSIM system. In the future, we will probably see heavy new development in this area.

As the external user is provided with larger masses of data summarized in time-series form, the problems of disclosure analysis and data suppression become more difficult. Intersecting disclosure problems in a time-series data base have received almost no attention so far.

Shared Internal and External Data Use

Most of the data collected by the Census Bureau could be shared with data users once the disclosure and storage technology problems have been solved. If this is going to happen, the Census Bureau and its users must work jointly to come to an agreement on the best logical model of the data to be shared. The model should be as simple and as free from "computerese" as possible, so that a statistician or other subject matter analyst can work with it directly.

At the same time, new media and formats must be explored for the storage representation of data. A strict separation must be maintained between the logical and physical models so that new technology will be transparent to the user. Formats should be standardized so that data is easily transported from one site to another. Both the format and medium of data exchange should efficiently support the common logical view of the data.

In addition to format and medium, there should be a well-designed common logical model supported by a compatible data organization. Changes in the organization should be transparent to the user, and transportability between machines should be maintained. Although tape is currently the primary medium for transporting data--and hence sequential organization is predominant--in the future data may be transported as holograms, floppy disks, or bubble fields, making alternative data organizations practical.

Shared Internal and External Software

Once a common logical model of Census data is achieved, formats are standardized, and transportable data organizations are developed, it will be possible to share data management software that has been specialized to handle Census data. This shared software will need to be transportable over a variety of computer hardware. Transportability may be achieved either by producing and maintaining multiple versions of the software, each implemented for a particular machine but having identical user interface, or by producing and maintaining a single version written in a high-level language for which most machines have a standard compiler.

With shared use of data management software comes the possibility of distributing Census data in a pre-loaded data base format. This would eliminate the duplication of the time-consuming data structuring operations at every site. In fact, more complex data structures could then be feasible, since the work involved in producing the structures is done only once. At the same time, more complex data structures could provide the user with a faster and more versatile retrieval capability. With proper data structuring, micro data could easily replace summary level data in many instances, since the cost of producing special tabulations should become very low.

Shared Internal and External Data Center Use

An easier solution to the problem of sharing data management software and structured data is through the use of a shared data center. As mentioned in Section III, facilities for multiple users with diverse problems, residing over a large geographic area, sharing a common computer facility is currently available. It should be pointed out that any such facility could not house confidential data, and hence could not co-exist with many normal Census Bureau functions. It could, however, easily be a normal part of the activity of some time-sharing service. In fact, some 1970 Census data is now available on some commercial time-sharing services.

Under current disclosure guidelines, it would be fully possible to have the total 1980 Census summary data files and public use samples available through a time-sharing service to any and all interested users. More study would be required to determine the feasibility of placing the entire micro data base into a time-sharing environment. In order to make such a concept useful, one would need to be able to do special tabulations cheaply from the micro data and to insure the non-disclosure of confidential data.

Within the next ten years, hardware and software should be developed to a point that the entire micro data file and many summary tables could be maintained on-line. This will make the development of standard statistical data base packages important. At the same time, certain data users may prefer to continue to extract portions of the large data base and re-load those portions into other data base packages. This total operation could be performed within the context of a single time-sharing environment. The cost of all services would be paid by the user directly to the time-sharing service.

Such an environment would allow the development of a truly integrated system of generalized software interfaces to an up-to-date version of the Census data base. It solves the problem of standardization of a hardware/software configuration. It would allow for the expansion of Census data dissemination to include new areas such as current population surveys and economic data, perhaps even in time-series form.

V. DATA BASE IMPACT ON CENSUS DATA USERS - ISSUES OF CONCERN

Disclosure Analysis

By act of Congress, data about individuals collected in the various census and surveys conducted by the Bureau cannot be disclosed in such a way as to allow identification of the individuals. However, in certain cases, data concerning an individual person, farm, or company could be derived from unedited summaries. For example, if county A has one very small peanut farm and one very large one, then publication of data on peanut farming in County A would necessarily disclose much information on the big peanut farm. In order to protect against this kind of unwarranted disclosure, the Census Bureau spends a large amount of time and effort in editing the data to be published. In the past, this was done manually. Analysts and experts on disclosure examined the data table by table, editing it according to a set of prescribed rules. Moreover, it was necessary to sometimes modify tables for related geographic levels to protect against disclosure through inference.

In any shared data center, it is essential not only to insure that disclosure problems do not exist, but also to avoid any appearance that might imply disclosure of confidential information. For this reason, although it may be technically feasible to develop software to automatically perform disclosure analysis and data suppression, it is highly unlikely that outside users would be allowed to share a data base containing confidential information. The state of the art in data base security simply is not adequate to justify such a risk.

Accuracy of Data

One of the primary concerns of data users is the accuracy of their data. This is a particularly strong aspect of the shared data base environment. Because of the ease of correction in a data base, as post-tabulation activities reveal errors in the data, immediate correction to the shared data base can take place. In the past, correction was generally not performed due to the magnitude of the job. Instead, errata sheets were published warning users of various discrepancies whenever possible.

The availability of a shared data base also makes it much easier to perform inter- and intra-table consistency checks. Not only would such a capability help the Bureau locate and correct problems, but would also help data users convince themselves of the validity of the data.

Timeliness of Data

A user-accessible data base of Census information can improve the timeliness of data delivery in three major ways. Data could be loaded into the data base as it is processed, eliminating the normal distribution delay and making the data immediately available. Secondly, as the need for correction of summary data is discovered, those corrections can actually be made in the data base, making them immediately available to the users. Thirdly, as the original Census data ages, new survey information could be made available on a time-series basis to augment the original data. This could be extremely valuable to researchers interested in short-term trends and projections.

Cost of Data Delivery

The total cost to the user for delivery of his final data product should be greatly reduced in a DBMS environment. This is primarily due to the fact that only the exact quantity and content of information needed to supply the request must be processed. The data base eliminates repeated traversals of a large sequential file to extract a limited amount of information. It also eliminates much of the programmer cost associated with writing and debugging custom programs for summary tape processing. Finally, there should be a significant cost reduction simply because of the scale of the operation and the fact that the processing center focuses directly on the processing of Census data.

Ease of Use

One of the most important impacts of such a data base would be the easy availability of the vase amounts of Census data to users who are not computer-oriented. The user view and interface language of the data base system could be such that non-programmers would feel at ease in employing it. In addition, immediate help for such non-programmers could be made available through both HELP commands on the system and hot-line service from the center.

Adaptibility

It would be important to balance the data base carefully so that good service could be obtained by both the small request from a non-programmer and the large request from a custom program. In addition, the data base must be smoothly interfaced to other statistical software packages to provide aggregation, display, graphics presentation, and computation capabilities.

VI. CLOSING REMARKS

The use of data base organization techniques for Census user data is both feasible and cost effective. Several different approaches to the problem seem to be promising. At a most fundamental level, data tapes that are distributed to users could be reorganized to provide a limited amount of tape-oriented table indexing and chaining of data based on the structure of the internal data bases. A more useful approach would be distribution of pre-loaded data base tapes for a select group of the most popular data base packages. If it were possible to define a common set of data base software that was machine independent or easily transportable, the software and pre-loaded data bases could be distributed together. But the most viable and potentially useful approach seems to be the availability of a Census user data base on a national computer time-sharing network. This data base could be maintained by the Census Bureau and accessed by anyone wishing to make use of the data and able to pay the access cost.

If we are to pursue any of these possibilities, we need to make a decision now. Future cooperative efforts will affect the Bureau's development strategy, as well as the strategy of users' development. It will also be necessary to allocate resources to provide for future development.

GENERALIZED STATISTICAL TABULATION

By Hugh F. Brophy, U.N. Statistical Office,
New York, N.Y.

Introduction

The general subject of access to census data includes the regular programme of publication, the provision of summary tapes and software for using them and the production of ad hoc tabulations. In all cases, the task of statistical tabulation is directly or indirectly involved. Small wonder then that it is a topic receiving special emphasis in this Conference.

As one who became involved in the implementation of a generalised statistical table generator in the mid-sixties, and who considered proudly that the system produced then solved all the interesting problems, it is sobering to be involved in a Conference in 1977 that is discussing the feasibility of a project aimed at the very same software task. But, wiser now, I recognize that my efforts and those of many others have fallen short of anything approaching an ideal system, and this discussion is thus highly appropriate. I note that the discussion takes place in the framework of improving access to census data and I intend to treat that as an overriding consideration.

The Task

The task of statistical tabulation is, on the face of it, a rather mundane programming exercise - one which trainees solve fairly easily, at least for straightforward cases, early in their careers. What is involved is essentially a mapping, normally many-to-one, from the records in the input file to those in the output file. The output file is generally a series of n-dimension matrices with textual definitions and descriptors attached. That sounds simple enough. But, as those who have worked in official statistics know, the range of problems involved in defining the input, selecting appropriate records and items, and manipulating and formatting the output required for a national census presents a formidable task.

During the sixties, many organizations independently undertook, with varying degrees of success, to produce a generalised solution to the problem. The major difficulties to overcome were those presented by:

. core restrictions

. complexity

. the size of the input file

. the need for machine efficiency

The solutions proliferated in national statistical offices and other organizations. In the case of the Census and Statistics Bureau in Australia 1/, a generalised table generator was first used in

1/ "A Generalised Table Generator" L. Ion, Proceedings of the Fourth Australian Computer Conference, Adelaide, Australia, 1969 processing 1966 census results, but quickly was applied to many other fields of statistics. It had a dramatic impact on processing. Previously, 40% of CPU time was consumed by sorting. With the advent of the generator, this dropped to less than 10%. Similar results were experienced by other national statistical offices. When the UK Statistical Office decided to launch yet another effort in the early seventies, they began by taking an inventory of existing "generalised table generators". They stopped when the number had passed 100.

Many of these systems, as well as solving most of the problems above, met most of the desirable system objectives, in that they involved a user-oriented language, they were capable of producing many tables in one pass of a large file (which could be random-order) and they enabled the production of tables in a limited time from date of specification. The problem was solved many times over.

However, when one looks today for a generalised table generator for a non-trivial tabulation task, one would have reason to be disappointed with the systems available. With each system evaluated, one would find one or more of the following problems:

Size Restrictions: Many table generators are incapable of producing in a single pass more than, say, 100,000 cells. Some produce two-dimensional tables only, some have severe limitations imposed by page size, others limit any dimension to, say, 100 values, and so on. Whilst these limitations are acceptable in many if not most commercial applications, they are severely limiting in processing official census results.

Complex Language: The claims for systems of an "English-like" user language are often ludicrous, the language being instead a cryptic distorted algebra developed without regard to rigorous syntax or natural semantics.

Machine Inefficiencies: One of the objectives of a generalised package is that it should be at least as efficient in producing a given table as a program developed in a compiler language such as Fortran or Cobol. Unfortunately, some generalised systems fall short of this objective by an order of magnitude. (It is interesting to note, in fact, the incredible range of CPU times consumed in different systems doing the same job on the same computer system.)

Lack of Portability: Almost all table generators have been designed without regard to portability and are dependent on certain models of central computer, specific operating systems or compilers, certain device types, etc. A potential user can thus face the impossibly difficult task of redeveloping for his own machine or start looking for an alternative.

In addition to these problems, there is a variety of limitations that may hamper the attempt to use a generalised table generator in meeting the tabulation needs of a project. There are often restrictions on conditional manipulations, calculation of sub-totals, percentages, handling of floating-point, footnotes, treatment of "negligible" cells, and many other processes which are traditional in official statistical tabulations.

The result is that one is required to complement the use of one or more generalised packages with ad hoc programs for pre-processing data files, post-processing print files and sometimes even for performing the tabulation task itself for some tables.

The purpose of this paper is to describe the necessary and desirable features of a "complete" solution and to examine the feasibility of a project aimed at an "ideal" system. There will, of course, always be some special tabulation requirements lying outside the realm of possibilities of a generalised system, thus making words like "complete" inappropriate, but at least the elimination of major restrictions listed above should be a design objective.

It is not my intent to perform a comparative evaluation of existing systems. Such evaluations are fundamentally affected by the choice of criteria and weights, and are often biassed towards an author's own system. (However, a fairly objective and carefully circumscribed evaluation is given by Francis et al.[2]).

An "Ideal" System

It has been stated by some people that it is impossible to implement an ideal system that will meet all the design goals one might have for a single generalised generator of statistical tabulations. A short list of the major goals would be:

- ease of use.
- machine efficiency.
- applicability to a wide variety of tabulations - from simple to complex.
- capable of running on small configurations but taking advantage of bigger resources if they are available.
- producing "camera-ready" printouts with extensive formatting options.
- extensive data manipulation facilities.
- portability.

[2] "Languages and Programs for Tabulating Data from Surveys" Ivor Francis, Stephen P.Sherman and Richard M. Heiberger, Proceedings of Computer Science and Statistics: Ninth Annual Symposium on the Interface, 1976.

With the possible exception of portability, I am of the opinion that sufficient expertise and knowledge of the necessary techniques exist for the implementation of a single system meeting all these objectives. The design of such a system would have, inter alia, the following characteristics:

- a true compiler rather than a table-driven program, for the sake of flexibility and machine efficiency.

- three major modules - generation of raw tables, manipulation of tables and table print - but capable of use as a single system.

- separate definition of data structure, content and descriptors (as in the TPL CODEBOOK approach[3]).

- generation and processing of tagged cell data, rather than in-core tallying of sub-tables - both for generation of raw cells and their manipulation - again for the sake of machine efficiency.

- implementation in a high-level programming language.

- a simple but powerful user language with rigorous syntax. By simplicity is meant that the language should be easily learned to a basic level, easy to use and to extend one's comprehension. A special feature of the language should be its power. By power, I mean the amount of work one can define in a given unit of the language, not the sum of all work one can define with the language.

It is worth reflecting here that machine efficiency must remain a primary objective in statistical tabulation. When we are dealing with the scale of data files and size of tabulation involved in census data processing, machine inefficiency can render an otherwise useful package impractical.

Other Facilities

There are three additional facilities which would make a generalised statistical table generator even more useful, especially from the viewpoint of improving access to census results. These are:

- capability to produce photo-composable output. The output destined for the printer can be saved on a file which could be input to a generalised utility to produce a driver tape for the more commonly-available photo-composition devices. Relatively generalised software for this purpose is being developed in the UN Statistical Office.

[3] "Table Producing Language - Version 3.5 - Users Guide" July 1975, Bureau of Labor Statistics, Washington, D.C.

- capability to generate large multi-dimension tables on disk for scanning or "browsing" through an on-line terminal. Such a system was developed in the Australian Bureau of Statistics for Foreign Trade statistics. This avoids the printing of such large tables, for which the only purpose is availability for such occasional browsing.

- ability to link with other files. A common requirement of the users of census data is to link with the users' own data for research and analysis. Most existing generators accept either a single file only or at best files with identical format and content.

Summary

Over the last decade, there has been considerable investment of time, money and human ingenuity in the development of statistical table generators. They have had differing sets of design goals and varying degrees of success in meeting them. For a typical project, the user tends to receive rather different tables than he would prefer.

There have been some attempts at international cooperation in the field of the design of software for processing official statistics. A table generator has always been a subject of primary concern. In the Working Group on EDP of the Conference of European Statisticians, such discussions in the mid-sixties led to the establishment of a UNDP project in Bratislava, Czechoslovakia in 1969. This extensive (seven-year) project was very successful as a development project and for stimulating discussion and exchange of ideas on the general subject of official statistical information systems, with computer processing as a major element. The table generator developed in this project was, however, no better than some developed in national offices. Nevertheless, there was a very telling demonstration of portability. The system was written in PASCAL for the Control Data 3300. At a meeting of the above-mentioned Working Group in Geneva in 1974, the system was re-compiled on the IBM 370/158 (a machine with quite a different architecture) and tested and demonstrated within a week. To date, however, a PASCAL compiler exists for only a few machines - but to a certain extent the feasibility of portability of generalised software was established.

The most likely way to develop generally-useful software, it has seemed to me for some time, would be to fund a project with international input, but located in a national statistical office of an advanced country. The objectives of this Conference are thus of great importance. For the task of statistical tabulation in particular, I am confident that a team of people experienced in statistical data processing could in a matter of a few years meet the needs for appropriate user-oriented software. Such software would greatly enhance the value of census data, thus multiplying the returns to the considerable investment made in collecting the data.

Acknowledgements: The author wishes to thank the Director of the UN Statistical Office, Mr. S. A. Goldberg, for his support and interest in this paper and his colleagues Messrs. M. Lackner and P. Emerson, for their helpful comments and suggestions.

GENERALIZED TABULATING SYSTEMS AT THE U.S. CENSUS BUREAU

By: Melroy Quasney, Chief, Generalized Software Development Branch, Systems Software Division, U.S. Bureau of the Census

History of Computer Language Development:

The development of generalized tabulation systems at the Census Bureau has followed the normal development patterns of all problem-oriented software systems. It is necessary to reflect on the history of computer languages to set the stage to understand the technological advancements that permitted the development of problem-oriented software.

Stated in simple terms, any new idea has to overcome two major problems if the idea is to be implemented successfully. One, the technology must be developed, tested, and proven possible. Two, the end product must be accepted by the intended users of the product. These problems also apply to the development of computer languages.

We began the computer revolution with assembler language; it did not take long to realize that assembler languages were inhuman to the users of the computer. Then came Fortran, followed by COBOL and other higher level languages, all making the computer easier to use to accomplish a given task. All of these advancements encountered the two problems previously mentioned. All of these advancements made the job of computer professionals easier, even though the acceptance of this new technology took time. Other support software systems were developed to assist the computer professionals to accomplish their task; however, task complexity also increased.

We are now at the point where the demand for bringing the computer to non-computer professionals is upon us. This demand is leading to the development of problem oriented computer languages. These systems call for a computer language that addresses a given problem and permit the user to communicate his request in his language. Probably the single biggest technological advancement that has permitted the demand for and the development of problem oriented systems has been the access to the computer via telecommunications. This has permitted the computer user to access interactively with computer software systems, or submit work from remote stations and receive the results back at the remote site. Generalized Statistical Tabulation Systems were probably one of the first attempts to produce a problem-oriented software systems. Systems like SPSS, CASPER, CENTS/COCENTS, TPL, and others, all used as their main design objective to bring the computer closer to the end user. All of these systems contributed to the advancement of the state of the art for permitting non-computer professionals, as well as computer professionals to use the computer to produce statistical tabulations.

History of Computer Language Development at the Census Bureau:

The Census Bureau's use of computer languages has paralleled the development and use of computer languages; sometimes we have been up with the front of the pack, and other times we have been slow in taking advantage of the latest technology. We use very little assembler language in the processing of our production data processing requirements. Most production processing is done using Fortran; however, Algol and COBOL are beginning to be used for a large amount of the production processing. A more favorable point is that most of our generalized software being developed is using Algol and COBOL.

A prerequisite of acceptance for all of generalized software is to develop problem-oriented user languages that permit the users to state their request in a language most familiar to the users.

Two projects that began fairly close together in time brought the Bureau into the world of generalized tabulating systems. One system known as GENER70 which began in the late sixties and still has some limited use in the Bureau. The other project involved the Census Bureau producing a generalized tabulation system for the Department of State's Agency for International Development (AID) to be used by developing countries to tabulate censuses and surveys. This project produced the CENTS/COCENTS system.

The CENTS/COCENTS project produced a product that has been installed in over 43 countries, and in over 68 computer installations, and has trained people from 80 different countries. The system can operate on any IBM 360/370 machine, plus 12 other types of mainframe. It has been used to tabulate major censuses and surveys by computer programmers and subject specialists.

My reason for emphasizing the experience of the CENTS/COCENTS project is to demonstrate our experience in distributing and supporting software. We know the level of resources needed and the problems with using the approach of distributing software.

The main objective of this project was to produce a product that could do censuses and surveys on small computers and be programmed by both programmers and subject specialists. These objectives forced the creation of a system that was efficient, but also produced a product that received heavy criticism due to its user language being very primative.

A Complete Generalized Statistical System:

A complete generalized statistical system must be able to control the collection of data, perform editing and imputation of the data, build a data base, tabulate the data, perform statistical analysis of the data, and finally publish the data in various forms.

Currently at the Census Bureau the Systems Software Division is designing and beginning the implementation of a complete generalized statistical system. It is our objective to produce a system that will service computer professionals and also put the power of the computer into the hands of the subject specialists.

The planned system consist of six major components: 1) Edit/Imputation System 2) Data Base Management System, 3) Tabulation System, 4) Math/Stat System, 5) Graphics System and 6) Photo-Composition System. We are currently working on the Tabulation System, the Graphics System, and the Photo-Composition System; the Data Base System is being used for some projects and will be connected with the other modules now being worked on during 1978.

Some Problems in Implementing a Generalized Statistical System:

As previously stated, two major problems face us in completing our total system.

We still have considerable technical problems to overcome before the system is completed. The biggest problem is the details of communications between the components. We are designing the components to be independent units, but when the data base is introduced it will be used as the primary connection between the components. Additional control information will also have to be passed between the individual systems.

Other technical problems are the range of requirements the system must satisfy, the various size of the data files it must process, and the implementation of the latest hardware technology to process large files on-line. An ideal statistical system at the Bureau must be all things to all people, but simple to use.

The second problem is user acceptance; we need the user community to accept the individual components and to supply additional specifications to insure that the system can satisfy all of the demands of the user in its future releases. However, introducing new technology is not easy. Changes to the daily working environment of a staff can be a hard thing to bring about; proving to a staff that a new product will do a job better takes time.

Census Bureau's Generalized Tabulating System (GTS):

The Systems Software Division of the Bureau has completed the first version of a tabulating system known as Generalized Tabulation System (GTS). It is important that we explain "why build another tabulation system?"

Before making the decision to build a tabulating system for the Bureau, we evaluated most existing systems and tried to identify the pro and cons of each system. We then evaluated the minimum requirements for a first release for use by the Bureau.

None of the existing tabulation systems evaluated could solve the wide range of the Bureau's tabulation requirements. None, at the time, were operational on Univac equipment. But most of all, test showed that the basic tabulation strategy of the CENTS/COCENTS system was more efficient. It was then decided to build our own system using these proven efficient methods, but to also place major emphasis on producing a user language that is consistent with the terminology and method of operation used in the Bureau and is easy for the computer professionals and subject specialists to specify their tabulation requirements to the system.

The Least Common Denominator Approach (LCD):

The LCD approach permits the user to specify the smallest geographic level for which a table is to be displayed. Several tables can be tabulated at the same time, each with a different level of LCD being specified.. This approach permits the minimum amount of hardware resources to be allocated during the long computer runs that require the examination of millions of detail data records. This approach also permits hundreds of tables to be produced with one pass of the detail data.

After the largest part of the processing has been completed, GTS then uses the LCD blocks to build all higher levels required for display.

A cost comparison was done by DUAL Labs and demonstrates the efficiencies of the LCD approach. A file containing 17,958 records was used to tabulate a table containing 56 rows by 2 columns. DPS, Data-text Nurcros, SPSS, and CENTS were the packages selected for the test. CENTS produced the table in 18.70 cpu seconds at a cost of $3.92. The next closest system was SPSS using 42.94 cpu seconds at a cost of $16.86. The most expensive system was Data-text at 107.62 cpu seconds and cost $41.96. DUAL then took SPSS and CENTS for additional testing. Two file sizes were selected for the test: 180,047 and 1,799,888. When tabulating 180,047 records, SPSS used 459.46 cpu seconds and cost $89.38; CENTS used 92.36 cpu seconds and cost $19.00. Based on this test, only CENTS was chosen to tabulate the file with 1,799,888 records. It took 815.32 cpu seconds and cost $150.00 for CENTS to do the requested task.

BLS using their TPL system tabulated a file with 20,196 detail records and produced the same table that was used in the DUAL test. It took TPL 40.46 cpu seconds as compared to CENTS tabulating 17,958 detail records and using 18.70 cpu seconds.

This kind of efficiency must not be ignored when building a tabulation system that will be used to tabulate millions and millions of detail records for the Census Bureau. This method of process is also compatible with getting the tally matrices under the control of a DBMS.

The Bureau also capitalized on utilizing its available resources; it had the staff who built the CENTS/COCENTS system available to work on building an efficient system for the Bureau.

The first version of GTS has been completed and attached are some test results to show that we have again built a system that is efficient to use. users of the system to produce these tabulations were subject matter specialists. The total project was completed in one-fourth the time conventional processing methods would have taken.

Overall GTS Design Requirements:

Five major objectives were selected to act as a guiding force for the development of the GTS system.

1. Bridge the conflict between being easy to use and powerful.
2. Function in a conversational as well as a batch mode.
3. Exploit the availability of large core storage on the UNIVAC 1100.
4. Maintain consistency in recoding of the input data.
5. Maintain flexibility without lost of machine efficiency.

Evaluation of other table generator system was performed and some features of these systems were incorporated into GTS. A continuing effort to keep track of other systems will be done.

Evaluation of data dictionary concepts has been done; the first version of GTS uses a stand-alone data dictionary processor. The design of the dictionary language is allowing for the future connection into Univac's DMS-1100 data base management system.

GTS will be implemented in phases of capabilities. GTS-1 is now complete and GTS-2 is beginning the detail design phase.

The GTS System:

Attached is a system overview of the GTS system. GTS is designed to consist of three major segments. They are: 1) the User Processors; 2) the Execute Processors; and 3) the Display Processors. The User Processor is the only part of the system addressed by the users of the system. This provides us with the flexibility to design different user languages; and as long as these different languages follow the rules for passing control information to the Execute and the Display Processors, several user views of the system is possible. The Execute and Display Processors are designed with efficiency and simplicity as the main design goals. Any decisions that can be made by the Language Processors are made by them.

GTS-1 Design Objectives and Status:

The main criticism of CENTS/COCENTS was that the user language was too primative and resembled a form of assembler language. When designing a table generator to run on a computer with 25K of working core and a CPU that is slow as molasses on a cold day, major emphasis was placed on efficiency of running and on flexibility to produce publication output. The price was in the user interface. It should be obvious then that one of the main design objectives of GTS-1 was to produce a good Census Bureau compatible user language.

The second objective was to begin, and experiment with a data dictionary to desribe and control input to the system.

It was decided to use a computer language that would be as portable as possible to permit the Bureau to change hardware and software with minimal impact on GTS. A by-product of this decision permits the first two versions of GTS to be usable by other computer installations with a minimum of resources to adapt the system to a different environment. Using a higher level language also has advantages in the implementation and debugging of this and future versions of GTS. Unique hardware and software features of the Univac 1100 series systems were purposely not used in the first version of GTS. We wanted to maintain hardware/software independence so that converting GTS-1 to other computer systems would be an easy task.

The technical specifications of the system were distributed to the entire Bureau user community for comment. This was successful in that several critical design changes were incorporated during the implementation phase. Test projects using the system also resulted in design changes that were incorporated in the first version of the system.

It was of course necessary to maintain, or improve, the efficiency achieved with the CENTS/COCENTS system. The system demonstrated that our basic design strategies were proven to be efficient during the 1974 Ag Census Volume II test project.

The last objective to be discussed is the requirement that GTS must be capable of utilizing a Checkpoint/Restart facility. The attachment showing examples of the cost of some runs on the 1974 Ag Census Volume II project points out the reason for this to be mandatory to GTS. These large production runs were on the computer system 6 to 27 wall clock hours. The Bureau's computer systems are only averaging 12 hours meantime between system crashes. In this environment GTS must perform restart recovery.

All of the above objectives have been met in GTS-1. The first level of the system was completed in May 1977. Enhancements and error corrections have been made and the final GTS-1 was completed in October 1977. Final user documentation was completed in October 1977; training workshops will begin in December 1977.

GTS-2 Design Objectives and Status:

Major emphasis in GTS-2 will be placed on the data dictionary capabilities of the system. The major objectives of this effort will try to address the following problems:

A. Ability to store recode commands.

B. Ability to store headings and stubs connected to related stored recode commands.

C. Ability to store calculations.

D. Additional automatic documentation of data in dictionary.

E. Recode scale checking to validate recode commands.

F. Validation of a data file against the dictionary describing the data file.

G. Access to build and use dictionary from a conversational mode.

Other major design enhancements to GTS-2 will include:

A. Conversational capability.

B. Ability to process overlapping geographic areas in one pass of data.

C. Expanded statistical capabilities.

D. Improve method to process economic data when displaying data greater than four positions of the Standard Industry Code (SIC).

E. Random retrieval of geographic and SIC stub descriptors.

F. Dynamic allocation of core and I/O paging to accomplish current task.

G. Provide linkage to user programmers.

H. Capture information for Math/Stat package.

I. Begin connection to Graphics and Photo-Comp software.

The design phase of GTS-2 began in November, 1977 and will be completed by January 1978. Implementation of GTS-2 is targeted for May, 1978.

GTS-3 Design Goals:

GTS-3 will concentrate on the connecting into the data base management system. This will require GTS to use the DBMS's data dictionary and access data through the DBMS.

Distribution of Tabulation Software:

The CENTS/COCENTS project has given the Census Bureau considerable experience with the problems of distributing table generator software.

As previously stated, the CENTS/COCENTS system could run on any IBM 360/370 hardware and DOS, OS-MFT, MFT, and VS operating systems. It could also run on 12 other types of hardware with their associated software systems.

Experience has taught us that the only way software can be distributed successfully is to actually test the software on the target system. This involves buying computer time and supporting a staff in the field to install and checkout the software. If this is successful, the software must then be packaged to be as self-installing as possible. This process also requires testing to be done on the target system.

Experience also taught us that two types of training are required. A computer professional must be trained and made responsible for supporting the system at each installation where the system is installed. The second type of training involves training the intended users of the software system. The Bureau found that the best way of accomplishing this process was to send technicians to the installation to install the system and do the necessary training.

Another big problem with distributing support software is the multi-types of documentation. The basic documentation for using the system is the same. However, additional support documentation was always necessary for each unique environment for which the system was supported.

The last problem was testing new versions of the system in all of the environments for which it is supported. This requires repeating the process of testing the system in all of the environments supported. It also involves changing all affected documentation. The last phase of this process requires the distribution of the new software and documentation to all computer installations where the system was previously installed. In some cases this could mean that retraining must be done.

This total effort requires tremendous man-power and computer resources - this can be translated to a great deal of money. These resources must be allocated to the organization where the software is being developed and in each installation where the system is installed and being used.

If the Bureau is to consider distributing GTS along with Census Bureau data files the problem becomes more complicated when GTS-3 connects to DMS-1100. This forces the Bureau to keep a subset of GTS that only processes flat data files. It may also have the impact of reducing tabulation capabilities of the GTS system that is being distributed.

This distribution problem becomes more difficult because GTS is being designed as a major part of an integrated statistical system.

If the total statistical system is to become a flexible and integrated one, then it must use to full advantage the hardware/software facilities in the environment in which it is to function.

Need for Decision:

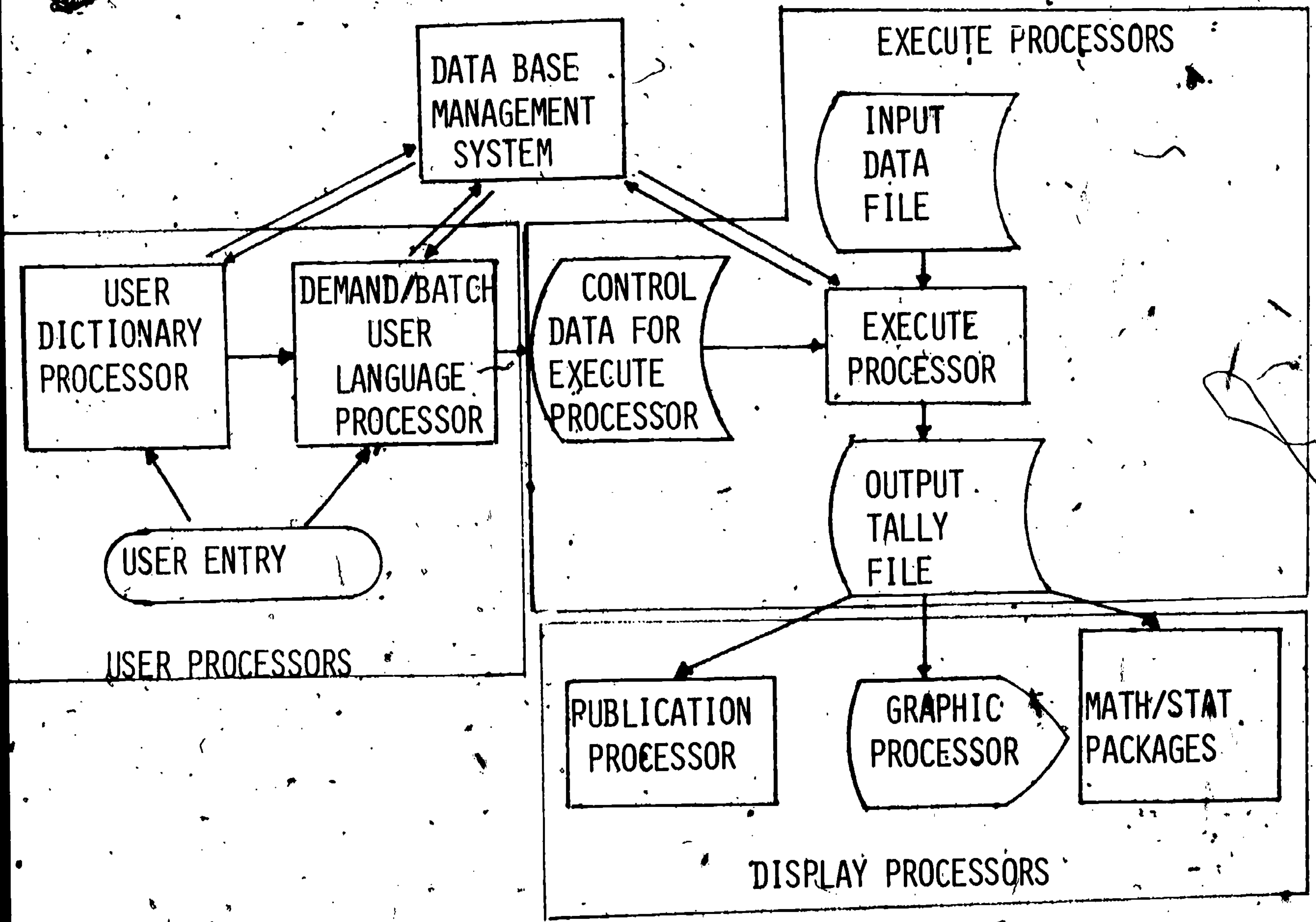
GTS-1 is technically very mobile; however, GTS-2 and beyond will become difficult. At some point distribution may not be possible.

Can the users of our data somehow use the system we are building?

Can we produce an environment that will permit access to this data and and software at a reasonable cost?

Now is the time to determine if the total statistical system, or parts of the system, should be allowed access by non-bureau personnel. If it is to be accessed by non-bureau personnel, we must define how it should be done.

TABLE GENERATOR OVERVIEW



USDA Unpub Tables

	CPU	I/O	Words	Cost
Total	22.7 HRS	721,000	300,000,000	$ 5300.00

A. Cost per table : $ 5300 ÷ 1535 = $ 3.45
B. Cost per cell : $ 5300 ÷ 1,728,000 = $ 0.031
C. Cost per Farm : $ 5300 ÷ 1,981,578 = $ 0.0026
D. Total file contained 1,981,578 farms
E. Total Wall Clock Time = 27 hours 12 mins 38 seconds

1974 Volume II Tables
Group #3

	CPU	I/O	Words	Cost
Total	6.4 HRS	473,032	292,949,768	$ 2264.00

A. Cost per table : $ 2264 ÷ 2930 = $ 0.77
B. Cost per cell : $ 2264 ÷ 244,000 = $ 0.0092
C. Cost per Farm : $ 2264 ÷ 1,981,578 = $ 0.0011
D. Total file contained 1,981,578 farms.
E. Total Wall Clock Time = 5 hours 18 mins 58 seconds

Reference Materials Used by Speakers at the Data Presentation Group

ROBIN WILLIAMS, IBM, Research Division, San Jose, California, in discussing Interacting with Data via Computer Graphics used the following three previously-published papers:

1. P. E. Mantey, J. L. Bennett and E. D. Carlson. Information for Problem Solving: The development of an Interactive Geographic Information System. Proc. IEEE International Conference on Communications. June 11-13, 1973, Vol. II, Seattle, Washington. Available from IEEE.

2. D. Weller, R. Williams. Graphic and Relational Database support for problem solving. Proc. SIGGRAPH '76. Available from ACM, SIGGRAPH, in Computer Graphics, Vol. 10, No. 2, Summer '76, pp. 183-189.

3. E. D. Carlson, G. M. Giddings and R. Williams. Multiple colors and Image Mixing in Graphics Terminals. Proc. IFIP Congress '77, Toronto, Canada. Published by North Holland Pub. Co., pp. 179-182.

LAWRENCE E. CORNISH, U.S. Bureau of the Census, for his discussion used part of an unpublished feasibility study by the "GRAPHICS AND PUBLICATIONS" Subcommittee of the "EDP REQUIREMENTS" group of the U.S. Bureau of the Census. The study was concluded in August of 1977.

MATERIALS PREPARED FOR SUB-GROUP DISCUSSIONS

Materials Prepared for the Data Presentation Group

By
Shirley Gilbert, Princeton/Rutgers
Census Data Project, Princeton University

The results of the survey of Summary Tape Processing Centers conducted by the Bureau of the Census and reported in the July 1977 Data User News clearly indicate a need for software support for processing 1980 census data tapes. How this need should be met in terms of specific program abilities to retrieve data and provide flexible report formats is important. Equally important, it seems to me, is consideration of how the production and distribution of this software will be implemented.

The Census Bureau's primary function in the area of user services should be to provide clean, well-documented data as promptly as possible. Once the data are delivered the function should be to inform data processors of problems in use of the data as soon as these problems become known. To ask the Bureau itself to write software compatible with the hardware of the great variety of computers serving Summary Tape Processing Centers is unreasonable. This conference can very usefully address the problems of how and by whom software can be produced and evaluated outside the Bureau in such a way that the Bureau can advise users of the availability of software for any particular system.

As a first step, I would like to see the members of this conference designate a committee composed of persons familiar with computer systems used by potential data processing centers. This committee could explore:

(a) How best to develop software where none now exists. (The most efficient procedure may not be the same for each of the several computer systems).

(b) How to evaluate programs so that the Bureau can make recommendations to potential users.

THE GENERATIVE APPROACH TO SOFTWARE DEVELOPMENT *

Gary L. Hill

Director, Information Systems, CACI, Inc. - Federal

ABSTRACT

The National Institute of Child Health and Human Development (NICHD/NIH) provided funding for the analysis of unique data processing problems posed by large statistical data files. One mechanism that resulted from this activity was the CENTS-AID II system, which reduces the cost of accessing large data files by as much as 80%. The generative programming techniques designed into the system are responsible for this significant cost reduction. CENTS-AID II is currently being used in over 50 computer sites around the world including the Belgian Archives, University of Heidelberg, Prudential Insurance Company, Congressional Budget Office, Social Security Administration, National Institute of Health, and the New York State Workmen's Compensation Board. The system is operational on the IBM 360/370 under OS and DOS.

1. INTRODUCTION: The Problem

Most generalized statistical access systems used by today's academic community were designed using interpretive programming techniques. That is, they were designed to scan researchers' commands and build extensive logic tables. Subsequently, as each record from the data file is processed, the contents of the logic tables are scanned and interpreted to control the execution of specific preprogrammed functions which will yield the outputs requested. As the research community developed new statistical routines, additional preprogrammed functions were integrated with minimal modifications to the basic processing methodology of the logic tables. As a result, the most popular generalized systems include a variety of analytic capabilities and require more than 200,000 bytes of core storage to execute. Even though logic tables are continuously scanned for each record on a file, and large segments of core storage must be allocated for execution, interpretive programming techniques offer an efficient mechanism for analyzing a limited set of observations. The same interpretive techniques do not however, offer an efficient mechanism for analyzing large statistical data files.

Large data producers such as the federal government provide a continuous flow of computerized statistical data. Most of these files contain tens-of-thousands, hundreds-of-thousands, or millions of records. Further, many of these sequential files are organized in a hierarchical, or tree structure format. This type of file organization provides for the definition of one or more record formats describing different units of analysis. For example, a file may contain one record format to describe the characteristics of households, another to describe persons, and a third to describe purchases.

*Material submitted for the Sub-group on Tabulation.

Additional valuable data relationships are defined by arranging the records in a predetermined order (tree structure); purchase records immediately follow the person record responsible for the purchase, and person records follow the household record in which they reside. Such a file provides researchers the opportunity to analyze the characteristics of purchases, the characteristics of people, and the characteristics of households. Further, the file enables researchers to analyze the characteristics of purchases with the characteristics of people, the characteristics of purchases with those of households, and the characteristics of purchases with those of people and those of households, et cetera, through all combinations and permutations of purchases, people, and household characteristics.

The analytic potential afforded by this type of file structure far exceeds the capacity of the punched card concept of file organization where each file has a single unit of analysis expressed in one record format. Unfortunately, most statistical access systems utilizing interpretive programming technology still require data to be organized as if they were in punched cards. In order for researchers to access the larger, more sophisticated files, data must first be reorganized to suit the unique specifications of the software system being used. This process is not only costly, but often destroys valuable data relationships defined by the original structure of the file. Whereas the utilization of interpretive programming techniques has tended to promote the general use of computers by the research community, it has also tended to limit access to large files.

The National Institute of Child Health and Human Development (NICHD/NIH) became increasingly concerned that many valuable data resources were being under-utilized by the research community. Consequently, funding was provided for the analysis of the unique data processing problems posed by large statistical files. One of the mechanisms that resulted from this activity was the high-speed CENTS-AID II System, hereinafter referred to as CENTS-AID.

2. CENTS-AID: The Generative Approach

CENTS-AID (Release 3.0) is specifically engineered to minimize the cost of accessing large data files through the use of generative programming technology. In benchmark comparisons with another widely used system designed around interpretive programming techniques, CENTS-AID's generative approach reduced computer costs by over 80%. Based upon user prepared commands, CENTS-AID generates a tailored ANS-COBOL program to process and analyze the data file. Subsequent system modules are used to format and display cross-tabulations of up to eight dimensions, produce subfile extracts complete with self-documented computer-readable Data Base Dictionary (DBD), generate and display correlation and covariance matrices, and create an SPSS (Statistical Package for the Social Sciences) Correlation Interface File upon request.

The CENTS-AID system is comprised of seven programmed modules, three standard utility sorts, and the ANS-COBOL compiler and loader. The system's generative approach can best be explained by examining the schematic diagram displayed as Figure 1 on the following page. The diagram does not depict each of the system's modules; instead it is intended to portray the system's generative nature.

2.1 Fragment Generation: Describing an application in quasi-English language commands, the user interfaces solely with the Fragment Generation module of the system. This module performs format and syntax checks on all commands, building a variety of internal tables, and organizing descriptive labels for subsequent report presentation. Once all commands are validated, the module scans the internal tables ONCE, building fragments of a COBOL program. These fragments are then combined with information from the CENTS-AID Models File to create a complete ANS-COBOL program specifically tailored to the application request.

When an application includes a request to generate a subfile extract, the Fragment Generation module will automatically create and display a computer-readable Data Base Dictionary (DBD) containing all detailed technical characteristics of the new data file, as well as descriptive labels for all variables and values of variables. The computer-readable DBD is separate from the new subfile extract itself and can be placed on any direct access storage device or alternatively, as a separate file on a magnetic tape. The Application module of CENTS-AID, to be described later, will actually generate the subfile extract according to the technical characteristics contained on the DBD. Subsequently, should the user wish to also analyze the subfile extract through CENTS-AID, all computer-oriented technical information and descriptive labels are automatically included through reference to the subfile's Data Base Dictionary. Alternatively, users can document master data files through the facilities of the Lexicographer component whose sole function is to generate computer-readable Data Base Dictionaries. This one-time documentation activity reduces the amount of technical knowledge required of statistical data users, and minimizes the amount of coding required to describe applications.

For user applications that require the generation of cross-tabulations, the Fragment Generation module is responsible for creating COBOL fragments that dimension all tabulation matrices requested. The facility of dimensioning tailor-made matrices into the generated ANS-COBOL program contributes to the overall processing efficiency of the CENTS-AID system. There is virtually no limit to the number of tabulations that can be requested in a single application. However, no single table may exceed 17 columns, or 999 rows, or 8008 matrix cells. Matrix cells can be incremented by a simple frequency count (1) or by the value of an observation variable such as income, expenditures, age, or number of live births. In order for the Fragment Generation module to dimension each table, the user must supply the minimum and maximum numeric values of each variable to be included in the table, either through CENTS-AID commands or via the DBD. Simple data transformation commands are available to manipulate variables

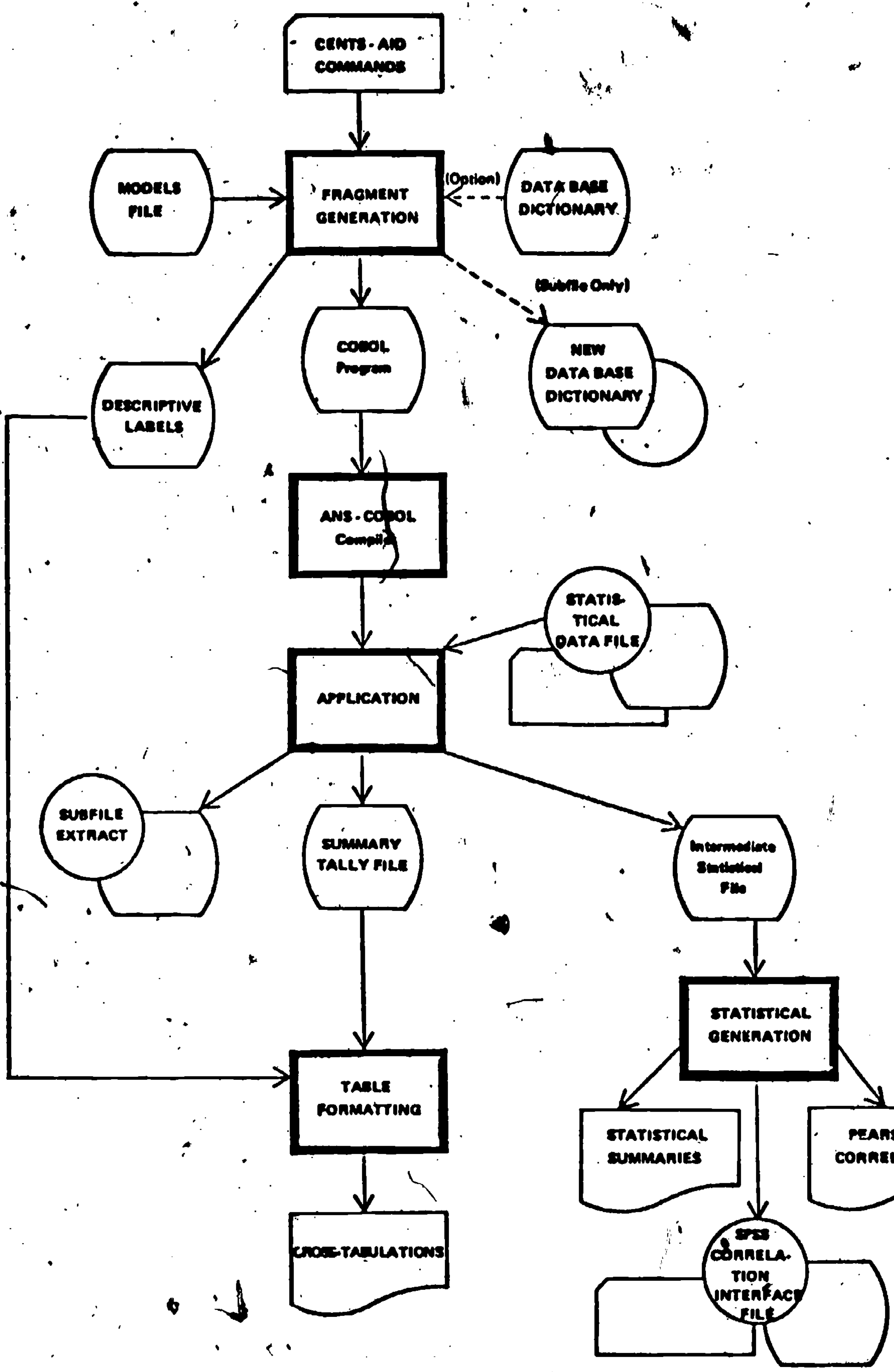

CENTS-AID COMMANDS
MODELS FILE
FRAGMENT GENERATION
(Option)
DATA BASE DICTIONARY
(Subfile Only)
DESCRIPTIVE LABELS
COBOL Program
NEW DATA BASE DICTIONARY
ANS-COBOL Compiler
STATISTICAL DATA FILE
APPLICATION
SUBFILE EXTRACT
SUMMARY TALLY FILE
Intermediate Statistical File
STATISTICAL GENERATION
TABLE FORMATTING
STATISTICAL SUMMARIES
PEARSON's CORRELATION
CROSS-TABULATIONS
SPSS CORRELATION INTERFACE FILE

Figure 1

containing alphanumeric or noncontiguous coding structures. Since each matrix shell is specifically tailored to accommodate the requirements of an application, CENTS-AID only reserves the amount of core storage actually needed to analyze the data file and perform the tabulations. In many computer billing algorithms, core storage costs are significant so that by reducing core requirements, computer processing costs can be minimized further.

CENTS-AID can also be requested to perform correlation analysis, generate variance/covariance matrices, and create a variety of other statistical measures. In those instances, the Fragment Generation module is responsible for creating COBOL fragments that define working storage areas and logic routines for the ANS-COBOL program to compute intermediate statistics for pairs of X and Y variables which will subsequently be processed by the Statistical Generation module. The working storage areas and logic routines are specifically designed to eliminate statistical error caused by accessing large data files. The intermediate statistics include the number of observations, the number of missing values, the sum of X and Y variables the sum of XY, and the sum of XY^2. All computations are performed in double precision floating point.

The COBOL fragments generated are then combined with instruction format information from CENTS-AID's Models File, reference Figure 1, to create a complete ANS-COBOL program. In a matter of seconds, CENTS-AID generates a tailor-made ANS-COBOL program designed to the specific requirements of the user.

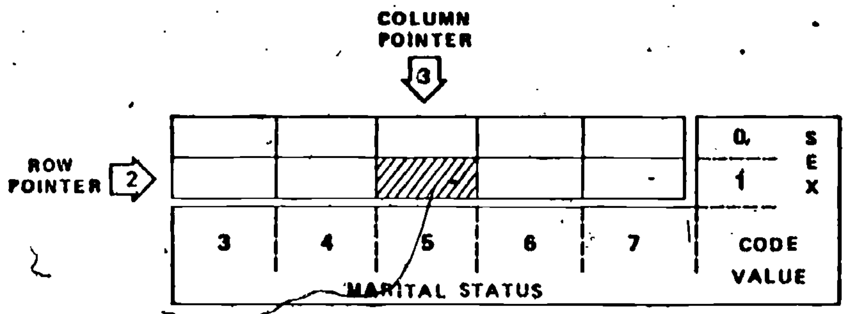
2.2 Application: Under the control of Job Control Language (JCL), the ANS-COBOL compiler and loader compiles and executes the Application program created by the Fragment Generation module. The resulting program is the only module within CENTS-AID that analyzes the statistical data file. Since the Application module is tailor-made to the specific requirements of the user, processing logic is optimized and core storage requirements are minimized. Because of the generative characteristics of CENTS-AID, most data files do not have to be reformatted in order to be analyzed. The Application module will directly process simple and complex sequential file structures whose records are fixed or variable length. Files can have up to twenty-six different record formats and a hierarchical structure of up to thirty levels, data can be recorded in binary, packed-decimal, and EBCDIC/BCD formats.

In addition to the basic generative characteristics of CENTS-AID, the processing methodology integrated into the Application module to update, or increment, matrix cells for cross-tabulations is also a major factor contributing to the efficiency of the system. Instead of continually scanning matrix dimensions to determine the proper matrix cell to increment (a technique employed by most systems), CENTS-AID uses the actual code values of the data file to compute "pointers" into each matrix. Simplified, the algorithm used to compute the "pointers" for a two-way table is as follows:

POINTER = (Code Value - Minimum Value) + 1

To illustrate the technique, suppose a user has requested the generation of a simple two-way tabulation (Sex by Marital Status); where Sex contains two code values (0 and 1), and Marital Status contains five code values (3, 4, 5, 6, and 7). A record containing a value of 1 for Sex and a value of 5 for Marital Status immediately points to the matrix intersection of (2, 3):

ROW POINTER = (1 - 0) + 1 = 2
COLUMN POINTER = (5 - 3) + 1 = 3



The processing logic of the Application module functions according to the specific requirements of the user's application. If a subfile extract is requested, records are formatted and written to an output file as the statistical data file is being processed. After the data file has been completely analyzed, the Application module then generates a Summary Tally File containing data for all cross-tabulations requested, as well as an Intermediate Statistical File. These smaller files are subsequently processed by the Table Formatting and Statistical Generation modules.

2.3 Table Formatting: The Table Formatting module is invoked solely for those applications requesting tabular output. The module combines the descriptive labels organized by the Fragment Generation module with the content of the Summary Tally File generated by the Application module. The module also computes column and row totals, as well as any optional descriptive statistics requested such as percent, mean, median, variance, and chi-square. The table formatting capabilities of CENTS-AID are extensive. Users can request simple frequency counts of selected variables, as well as more sophisticated cross-tabulations of up to eight dimensions. The TABLE command is used to identify the variables to be used in each tabulation. Variables named to the left of the keyword BY comprise row variables, whereas variables named to the right comprise column variables. The following TABLE command defined the six-way tabulation displayed as Figure 2 on the next page.

```
TABLE PLACE AND RACE AND INCGRP BY EMPST AND AGEGRP AND SEX
```

Table T007: PLACE OF RESIDENCE AND RACE AND INCOME GROUP BY EMPLOYED AND AGE GROUP AND SEX

PLACE OF RESIDENCE RACE INCOME GROUP	EMPLOYED								TOTAL
	YES				NO				
	AGE GROUP				AGE GROUP				
	18 TO 35		OVER 35		18 TO 35		OVER 35		
	SEX		SEX		SEX		SEX		
	MALE	FEMALE	MALE	FEMALE	MALE	FEMALE	MALE	FEMALE	
URBAN									
WHITE									
$0 TO $4,999	475	549	402	654	264	812	539	1,774	5,469
$5,000 TO $9,999	510	208	676	331	24	16	38	20	1,823
$10,000 AND OVER	281	14	699	53	6	2	20	2	1,077
BLACK									
$0 TO $4,999	75	86	74	121	44	138	80	167	785
$5,000 TO $9,999	63	39	82	44	6	5	8	3	250
$10,000 AND OVER	8	1	28	3	-	-	-	-	40
OTHER									
$0 TO $4,999	14	6	8	8	9	14	2	15	76
$5,000 TO $9,999	5	3	7	7	1	-	-	-	23
$10,000 AND OVER	2	-	4	-	-	-	-	-	6
SUB TOTAL URBAN	1,433	906	1,980	1,221	354	987	687	1,981	9,549
RURAL									
WHITE									
$0 TO $4,999	186	190	317	311	88	316	287	764	2,459
$5,000 TO $9,999	199	46	298	89	1	4	10	3	650
$10,000 AND OVER	66	2	157	7	3	-	6	-	241
BLACK									
$0 TO $4,999	26	16	36	26	11	25	23	59	222
$5,000 TO $9,999	6	1	9	1	1	-	1	-	19
$10,000 AND OVER	-	-	2	1	-	-	-	-	3
OTHER									
$0 TO $4,999	4	3	4	6	3	10	8	9	47
$5,000 TO $9,999	3	3	5	3	-	-	-	1	15
$10,000 AND OVER	3	-	5	-	-	-	-	-	8
SUB TOTAL RURAL	493	261	833	444	107	355	335	836	3,664
TOTAL	1,926	1,167	2,813	1,665	461	1,342	1,022	2,817	13,213

Figure 2

Descriptive labels were obtained from the computer-readable DBD. The Fragment Generation module analyzed the minimum and maximum values for all six variables referenced in the TABLE command. It then adjusted the "pointer" algorithm to automatically provide for the "nesting" of row and column variables, as well as align all row and column labels for subsequent display.

2.4 Statistical Generation: The Statistical Generation module is executed for applications requesting special statistical analysis such as Pearson's Correlation. The module processes the Intermediate Statistical File generated by the Application module and produces a variety of optional reports including correlation analysis with list-wise or pair-wise deletion, and summary reports containing such statistics as means, standard deviations, sums of squares, sums of cross-products, the number of observations, and the number of missing values. In addition, the module can optionally generate an SPSS Correlation Interface File. This file is acceptable to SPSS (version 6.0) as original input to its library of statistical functions which manipulate correlation matrices.

3. PROCESSING EFFICIENCY: A Comparison

CENTS-AID is engineered specifically to minimize computer processing costs for accessing large statistical data files. The generative techniques employed in CENTS-AID do not necessarily produce a cost effective mechanism for processing small data files. A series of benchmark tests designed to demonstrate the effect of processing increasingly larger volumes of data on CENTS-AID's generative approach and another system's interpretive approach were conducted. Although we feel that it is unrealistic to compare generalized systems that are designed for different purposes, we chose the Statistical Package for the Social Sciences (SPSS) for this comparison because it is so widely used. The benchmarks were not intended to be a comprehensive evaluation of the merits of the two systems. Whereas CENTS-AID is specifically designed to access large data files, SPSS offers a wide range of statistical analysis capabilities that far exceed the current facilities of CENTS-AID. The benchmark tests were designed by an outside consultant to meet the following specifications: 1) the test must request statistics which both systems could generate; and 2) it must use SPSS as efficiently as possible. The benchmark application used the FASTABS option of SPSS (version 6.0). The 1970 Public Use Sample Files were processed. The results of the test are presented in the following table.

	BENCHMARK TEST (IBM 360 Model 65)					
	TEST 1		TEST 2		TEST 3	
	SPSS (6.0)	CENTS-AID	SPSS (6.0)	CENTS-AID	SPSS (6.0)	CENTS-AID
Number of Input Records	27,591	27,591	277,723	277,723	2,719,249	2,719,249
Size of Universe	5442	5442	54,741	54,741	537,667	537,667
Number of Variables	9	9	9	9	9	9
CPU * Time (Seconds)	119.59	32.29	1188.17	134.08	11880.00	1113.18
Core Storage	214	94	214	94	214	94
Dollar Cost	$45.99	$10.78	$175.74	$24.48	$1543.04	$111.03

The comparative statistics generated by the three benchmark tests show that, as the volume of data increases, the computer cost of performing tabulations with software systems using interpretive programming techniques can become almost prohibitive. Subsequent to the execution of the formal benchmarks, further analysis of the processing efficiencies of the two systems was undertaken. For example, each system generated multiple tables using various combinations of user commands. Throughout these tests the variation in relative processing efficiencies remained consistent, with CENTS-AID applications costing approximately 80% less than the SPSS runs. During the testing process, an SPSS SYSTEMS FILE was created which substantially reduced SPSS tabulation costs. However, the cost of creating such a file can rapidly become expensive, and valuable data relationships may be destroyed in the process.

CONSIDERATIONS IN THE DESIGN OF USER-ORIENTED TABULATING SOFTWARE *

Rudolph C. Mendelssohn

Bureau of Labor Statistics:
U.S. Department of Labor, Washington, D.C.

The design of user-oriented software must begin with the identification of the users and the problems they wish to solve. Then, at the highest technical levels, the requirements are exclusively those of designing a language that will allow the users to communicate their problems to the computer. This is followed by the design of a generalized computer system to provide the product specified by the user.

Who are our users and what is their problem? Our mission says that the users are those who want to do tabulations. And, because the software is to be user oriented, I believe we are to assume that the user must be someone who lacks training in the computer sciences, does not care to learn either how computers work or the step-by-step procedures that get the computer to solve problems.

This may sound like a condemnation of users generally. However, I intend it as an observation of our own failure to see the computer as a tool to be given to users to operate

*Materials prepared for the Sub-group on Tabulation of Data.

in their own professional environment. The users should not be required to learn another discipline. Rather, they should be able to deal with the computer in their own technical language.

The most flexible tool that we can offer users would be a natural language. But, there are ambiguities present in natural languages. You and I can cope with these ambiguities through combinations of subtle nuances, assumptions, and prompting. Computers cannot tolerate so much freedom. A user language to talk with computers must be structured according to the demands of computers.

Knowing that computer rigidities will be a constraint, but that the language should be as close to natural as possible, we must ask ourselves what language do users employ to specify a table. Five years ago BLS undertook a study of the language used by our economists, statisticians, demographers, and other social scientists in describing and specifying tabulations.

Determining these language characteristics was not a simple matter because of the range of tables BLS users specify. These tables fall into three broad classes: Those published in the Bureau's bulletins and reports, work tables used in the production of the published data, and a third class more difficlt to observe. The BLS professional personnel is deeply involved in research and rely heavily on the Bureau's massive data files. The form of the tabulations from these files is not predictable because the analyst typically engages in an interactive process; that is, the study of one table leads to new questions which require different tables which generate new questions, and so on until the analyst is satisfied.

Our study revealed one dominant fact: There was no agreement within the Bureau on how to describe tabulation methods and table formats. Inconsistency prevailed. Among the computer systems staff, economists, statisticians, demographers, and other social scientists throughout the Bureau, commonly accepted terms and ordinary ways of expressing needs meant quite different things. Terms like variable, data element, data item, and field often were interchanged, depending on the context or the user's background. Simple words like row, line, column, table, summary, and cross tabulation had varied interpretations.

Nor did a look at other tabulation systems help. We concluded, then, that it would be best to pursue an approach that included a standardized language based on the nomenclature most commonly used in BLS. This approach would improve communication among BLS social scientists, computer science professionals, and the computer itself.

From an analysis of the study findings, it became clear to BLS that in building the standardized "language" the parts of the table had to be identified and named, and an unambiguous syntax had to be devised. This was done, and I refer you to the BLS document, The Development and Use of Table Producing Language, for a discussion of the structure of tables and the standardized language that evolved.

Upon resolution of the language problem, the BLS staff turned to the next step: the design of a generalized computer system that would respond to user written specifications for tabulations.

Briefly, the study had four goals:

1. The system should be able to produce most, if not all, of the Bureau's statistical tables.
2. It should be driven by a Table Producing Language that did not require the user to be competent in the computer science discipline.
3. It should be flexible and adaptable to changing needs for new tables and formats.
4. It should lead the way to composition of tables for publication.

The first step in system construction was to see what work others had done, particularly other national statistical agencies. A United Nations questionaire, sent to national statistical agencies in Europe, Australia, and North America in 1972, disclosed nearly 50 systems that produced tables. So much activity is certainly a demonstration that most statistical offices regard some degree of generalization desirable and possible. But two questions are raised:

1. Why so many different systems?
2. Why not use one of these in BLS rather than develop a new one?

Differences in computers and data file structures create incompatibilities that limit the use of someone else's programs, and much of the duplication of systems can be explained this way. However, this does not explain why some organizations have three or four different systems and why BLS found it useful to develop its own. The Bureau reviewed and analyzed all systems that could be found to see if they could meet its goals. Almost every system examined was capable of doing something useful. But the fact remained that no system met or even came close to meeting all the Bureau's requirements, individually or collectively.

In building our own system we relied heavily on the knowledge gained in the study of other systems. Particularly significant in this regard was the pioneering work done by the Australian Bureau of Statistics in the early and mid-1960's in the construction of their Report Generator. Another important contributor was our own BLS Information System which to some extent paralleled the work in Australia.

The work which combined the results of the twin studies of the user language and generalized tabulation program culminated in the completion of the first publicly available system in 1974. It is called Table Producing Language (TPL) and is now at work in over 155 installations worldwide.

Many users of TPL are in commercial enterprises throughout the United States and Canada. These include banks, insurance companies, computer time-sharing services, heavy industrial manufacturers, pharmaceutical houses, and research and planning organizations. But State and municipal agencies across the country, and more than a dozen Federal agencies (including both houses of Congress) are also users. Among educational institutions are over a dozen major universities.

The count of TPL installations abroad shows fifteen national statistical agencies, located mostly in Europe, but ranging geographically from North Africa and the Mideast to Australia and Thailand. United Nations installations in New York and Geneva use the system and also distribute it to member countries.

The Table Producing Language was judged to be the best in competition with eleven other leading contenders by the Committee on the Evaluation of Statistical Program Packages of the American Statistical Association. The Committee studied two principal characteristics: tabulating power and simplicity of language. When integer scoring from one to five for nine different attributes within these two categories was used, all systems evaluated scored well above the minimum figure. However, TPL scored the maximum possible, 45 while the runner-up scored 36.

The language differs from the traditional computer languages, such as COBOL, PL/1 and FORTRAN, in important ways. The latter have general application in the sense that they are used to solve a wide spectrum of problems in business and science--problems ranging from accounting, inventory, and production to weather forecasting and getting men to the moon. But in doing so, the user must give the computer step-by-step instructions on how to solve the problem being presented to it. That requires the user to know how computers work.

The Table Producing Language belongs to an emerging class of computer languages called very high level, problem oriented--very high level because they are disengaged from the computer, and problem oriented because they deal with narrow needs. TPL has limited application--it can only prepare tables, nothing else. On the other hand, this specific focus has allowed the embodiment of several advantages over the better known traditional but less specifically directed languages.

The TPL system already knows what a table is and how to generate one. It only needs to be told the particulars about the one wanted. Thus, when describing the desired table with the Table Producing Language, the user need not go through the tedious and time-consuming effort of telling the computer, step by step, how to make the calculations and lay out the table framework. Moreover, it allows Bureau social scientists who are not computer experts to use everyday common BLS language and nomenclature to describe the tables. In short, TPL has reduced a burden, speeded work, and increased the BLS capacity to respond.

I have mentioned some good things about TPL. Now, what is wrong with it. First and foremost, it will only run on medium to large-scale IBM machines, or their equivalent, such as Ahmdahl and perhaps Itel. We have had many requests for a version that would run on other machines. Unfortunately, from the viewpoint of these requestors, our mission is to serve BLS requirements. An effort to make TPL run on machines of brand names other than our IBM equipment would have been too costly.

Secondly, the system is monolithic--the user gets all or none of it. It includes special features that are closely allied with our needs. For example, there is emphasis on formatting tables for display in BLS publications through the use of electronic photo composers. This is useful to an agency that publishes most of its extensive production in table form but likely to be of little use in academic research. If the user has a small machine and small or limited needs, he can not just take the part that will help him.

Efficiency could be improved. Users can be unaware that a chosen approach is much less efficient than another that would give exactly the same result. For example, we find users breaking problems into smaller pieces than they should, resulting in extra costs at run time. We feel the system should protect them from these inefficiencies. An important goal of our project was to bring the cost of processing very large files down to palatable levels and we have reduced these costs significantly impressive amounts, compared to our alternatives. But the costs are still more than we like.

In summary, although TPL is the result of a pioneering effort and embodies important advances, a new effort should learn from its deficiencies as well. These include lack of portability across machines of different manufacturers, excessive size owing to the inclusion of special-purpose facilities, and lack of adequate protection against excessive and unnecessary running costs.

APPENDIX D

Status Report on Selected Census Bureau Activities

The conferees focused attention on a number of important ongoing and planned Census Bureau activities that were not covered in the prepared papers. Since these activities were not only discussed at length but also became the subject of several conference recommendations, a brief status report on their nature and prospects is provided in this appendix. Topics described below include market research; data delivery; training, consultation, and other user services; computer software; machine-readable data directories; computer tape files and microform.

Market Research

The identification of users' needs is always an early and high-priority activity of Bureau program planners. Many different approaches are used to determine interests in data content, tabulations, forms of data delivery, and data access and use assistance such as training and reference materials. For example, the 1980 census planners held "public hearings" in 74 cities and at several national conferences, and met with representatives of State governments to solicit recommendations. The planners also participate in the Federal Council on the 1980 Census, and maintain a mailing list of more than 7,000 interested persons to keep them informed through the 1980 Census Update, a newsletter that carries articles asking for users' suggestions on particular topics. Two planning conferences were held late in 1977 for representatives of summary tape processing centers and other tape users, resulting in more than 200 recommendations for 1980 census products and services. To obtain input to their programs, the Economic Census Staff sought suggestions concerning data content and tabulations for the 1977 Economic Censuses from hundreds of trade associations and institutes. The Bureau also maintains nine standing advisory committees.

Data Delivery

The Census Bureau is quite sensitive to the fact that effective and widespread use of its products is dependent upon an effective data delivery system which provides convenient access by novices and advanced users alike. To supplement established data access points such as the more than 1,000 Federal depository libraries and its own sales facility, the Bureau has expanded its own census depository library system, is seeking to improve the Summary Tape Processing Center Program, and has initiated a State Data Center Program. The latter program is a cooperative effort between participating States and the Bureau to improve the ability of State governments to operate data dissemination and user services facilities for the benefit of users in State and local agencies, universities, and the private sector.

Training, Consultation, and Other Services

The user services function of the Bureau is made up of such activities as product promotion, inquiry handling and user consultation, orientation and training, and provision of reference materials and other user aids. The user training schedule for 1978 includes 28 course offerings, ranging from the popular 4-day intergovernmental and librarians' seminars on accessing Federal statistics to courses on using machine-readable data files and using census data to meet Federal requirements, and on making population estimates and projections. A comprehensive inventory of guides, directories, indexes, and other user aids is available. The monthly Data User News keeps users informed about new products (also listed in the <u>Bureau of the Census Catalog</u>), training opportunities, and other relevant topics. Further, training and inquiry services have been enhanced through the placement of user services specialists in the Bureau's 12 regional offices.

<u>Computer Software</u>

The 1970 census might be remembered most for the large assortment of machine-readable products it produced. A combined total of more than 3,000 summary,microdata, and geographic reference tapes were released from that census. In recognition of the need by users for computer software to process these tape files, the Bureau developed and distributed data tabulation and display programs (DAUList 1-5 and COCENTS), geocoding software (ADMATCH and UNIMATCH), and computer mapping programs (C-MAP and GRIDS).

A study is currently underway to identify gaps in the software generally available from all sources that users need to process Census Bureau data and geographic reference files. The study results will be used to determine whether the Bureau should develop additional software for distribution to users.

A short-lived effort was made after the 1970 census to establish a software clearinghouse to provide users with a comprehensive listing of available programs for processing census files. The effort may be revised in association with the 1980 census.

Machine-Readable Data Directories

In order to be further responsive to the needs of users of computer-oriented products, machine-readable data directories have been prepared for recent products such as the 1974 Census of Agriculture tapes, Annual Housing Files, and Annual Demographic Files. Similar directories will be developed for all future public-use files.

Computer Tape Files and Microform

While the needs of users of printed reports will continue to receive a high priority, there is a definite, and deliberate, trend towards the release of more and more data on computer tape. This is in recognition of the desire for the "publication" of greater quantities of detailed data as well as the efficiencies of releasing data in this form. In addition to computer tape, microform (fiche and film) will be more extensively utilized as a data delivery medium. The combination of computer tape and microform make it possible for the Bureau to be responsive to the growing demand for additional data without contributing to the "paper explosion."

In summary, the Census Bureau recognizes that it has a responsibility beyond just collecting, tabulating, and publishing data. Its staff is aware of the large and diverse data user community and seeks in a multitude of ways, such as those outlined above, to be responsive to these users.