USING DATA BASE MANAGEMENT SYSTEMS IN STATISTICAL DATA PROCESSING

Joan C. Veim

CENTER FOR RESEARCH ON INFORMATION SYSTEMS

Computer Applications and Information Systems Area
Graduate School of Business Administration
New York University

Working Paper Series

CRIS #14

GBA #81-02(CR)

Ms. Veim's current address:
New York University, Graduate School of Business

Permanent address:
Institute for Computer and Information Sciences
University of Bergen, 5014 Bergen-U, Norway

ABSTRACT

National and international statistical bureaus produce ca. 25,000 tables for publication each year, based on hundreds of inter-related object-types with thousands of attributes. It would appear that this environment should be well suited to the application of data base management techniques for the administration of the data.

This paper presents a data oriented model of the statistical production process which is used as a basis for a review of the state of experience within statistical offices with commercially available data base management systems. We conclude with a presentation of some important data management facilities which must be enhanced or developed in order to support statistical production processing.

1. BACKGROUND

During the last 20-25 years, national statistical officies have relied heavily upon computerized data processing techniques [Nordbotten, 1961, Kazimour, 1977]. Increasingly, attention has been directed towards improving data administrative (storage and retrieval) functions [Nordbotten, 1966, 1967a 1967b]. More recently the usefullness of data base management techniques has been studied [Claringbold and Smith, 1973, Veim and Sundgren, 1979]. In 1978, the Conference of European Statisticians reviewed a report on the use and future need for data base management in national statistical services [Group of Rapporteurs on DBM, 1978]. According to a follow up survey, [Davies, 1979], data base management systems are being used or tested in the national statistical offices of 18 European and American countries.

Viewing this interest and activity, it seems safe to observe that data base management systems will become a part of the computerized data processing services offered to statistical offices. The objective of this paper is to review this exploratory activity from the view point of the data management requirements for statistical data processing in order to estimate the value of these systems to the statistical production environment. Further, those data management facilities required but not fully supported by currently available data base management systems are outlined with the intent of directing attention to the development of these facilities.

2. DATA MANAGEMENT IN STATISTICAL DATA PROCESSING

Traditionally statistical production systems have been based on the life cycle of the statistical survey. Figure 1 (adapted from [Veim and Sundgren, 1979]) provides a model of survey processing viewed as a sequence of activities:

1. Survey formulation
   stems from the definition of a 'problem' or a request for information about some real world entities. The survey formulation process defines a real world model, or entity model, describing the entities of concern then defines the data required to describe the model, the desired respondents, any alternate data sources and finally, the desired aggregations/statistics to be realized.

2. Data gathering
   selects a respondent group for the questionairs, processes the responces and gathers the required data from the specified alternate sources. The data is merged and placed into storage structures representing the original entity model.

3. Coding and Editing
   involves the translation of verbal data to predetermined code values and the test and correction of all data. Data editing normally is preformed in a series of passes through the data collection, each pass checking and correcting one or a related set of the attributes within the data.

4. Data aggregation
   involves the statistical processing of the micro data into the desired aggregate information elements. The aggregation routines each process some subset of the total micro data collection. New aggregation routines may be defined throughout the life of the data.

5. Presentation
   covers the retrieval, analysis and publication of the aggregated statistics. Information requests concerning aspects of the original entity model should first be addressed to the presentation processes, which search the macro data for relevant information before resorting to an aggregation process.

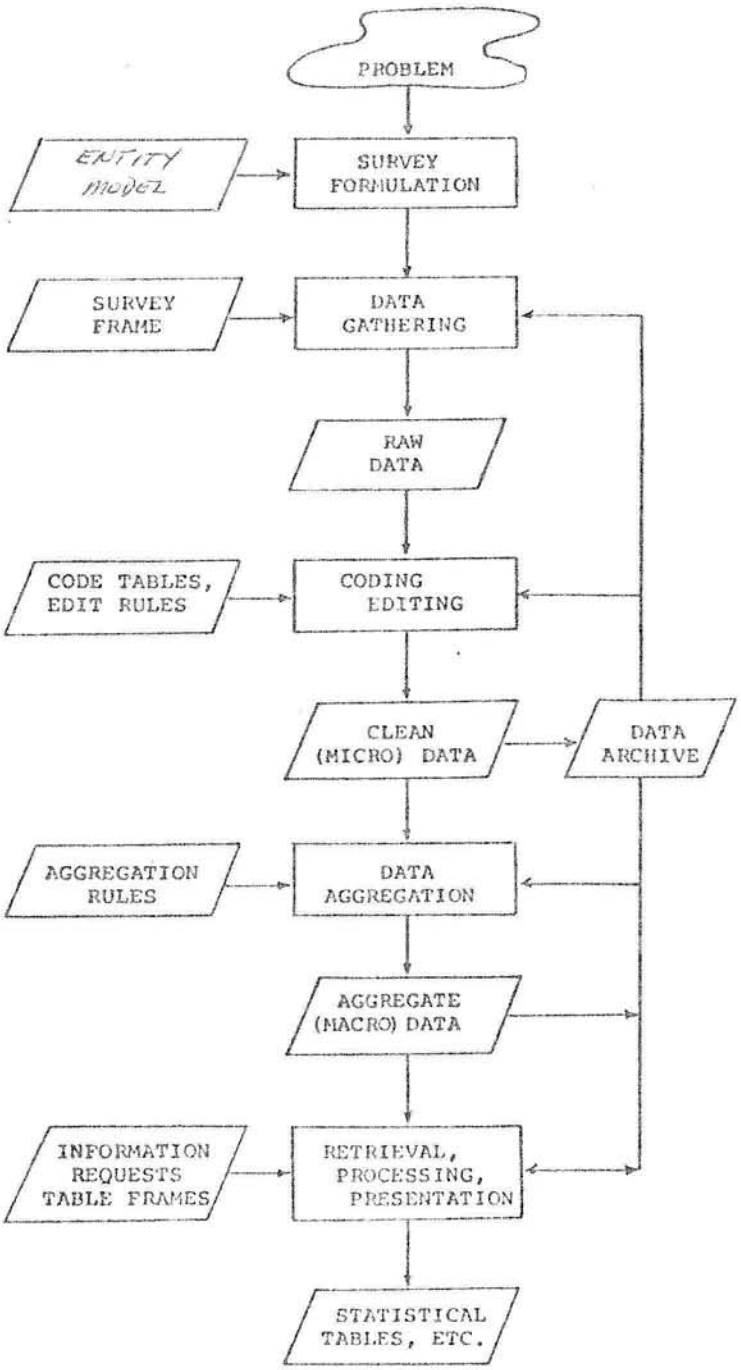

FIGURE 1: THE STATISTICAL PRODUCTION PROCESS

2.1 Statistical Data

During survey processing, the survey data is continuously refined, producing as 'by-products':

1. Raw data
   which includes the initial data collected from the respondents, with any data from alternate sources such as administrative data collected by other, non-statistical agencies.

2. Micro data
   which is the result of the coding and editing processes on the raw data, merged with any micro data retrieved from previous surveys. Micro data is archived for use in the current and future aggregation processes. From the archive it may also be used to supplement new survey data.

3. Macro data
   which is the primary statistical data, aggregated from the micro data of the survey. The macro data is archived acording to a storage structure representing the entity model specified for the survey. The analysis and presentation processes will retrieve their data from the macro data (the data archive).

The data archive of figure 1 represents the total data collection for a survey, both its primary micro data and the aggregate macro data. This archive is frequently organized as a set of more or less loosely related files. Our interest is in the degree it is applicable to store this data in one or more related data bases administered by a data base management system, DBMS.

Figure 1 also indicates the directive data, or meta data, which is used to describe and direct the processing at each of the production stages.

1. The Entity Model
   describes the entities and their attributes which are of interest in the framework of the survey. This model outlines the data types which need to be gathered for the survey.

2. The Survey frame
   describes the respondent group to which the survey questionairs are sent.

3. The Code tables and Edit rules
   constitute the translations and allowable relationships for the collected data,

4. The Aggregation rules
   describe the aggregation routines acting upon the micro data.

5. The Table frames and Information Requests
   describe the layout of the resulting tables and the activating queries recognized by the system.

These meta data describe and permit interpretation of both the micro and macro data resulting from the survey processes. As such it is necessary to record and retain access to these data when processing the micro and macro data.

2.2 A Data Oriented Model of the Statistical Production Process

From the previous discussion we note that the data archive, as illustrated in figure 1, actually consists of two 'levels' of statistical data, the micro and macro data. We know that micro data, once accepted for accuracy, will seldom be subject to recoding or further editing. This data is archived for the current as well as for future aggregation processes. Inorder to be able to interpret the micro data properly, the code tables and edit rules must be available. Also the knowledge imbedded in the original entity model as well as the survey frame must be known. Normally it is also necessary to have a detailed knowledge of the storage structures used for the micro data, particularly if the following processes occurr some time after the generation of the micro data archive.

The aggregation proceses actually produce the macro data from previously collected micro data. Occasionally aggragation processes can be run up to several years following the generation of the micro data which serve as input to these routines. It is imperative for these routines that the descriptive data for the micro data be readily available. Stored macro data will also need to be described if it is to be properly understood and used in other processes following its generation.

Based on the time separate processing procedures for the micro and macro data with the need for recording meta data describing each as well as the relationships between the two, a triple point data base model has been proposed [Sundgren, 1978, Veim and Sundgren, 1979]. This model is presented in figure 2.

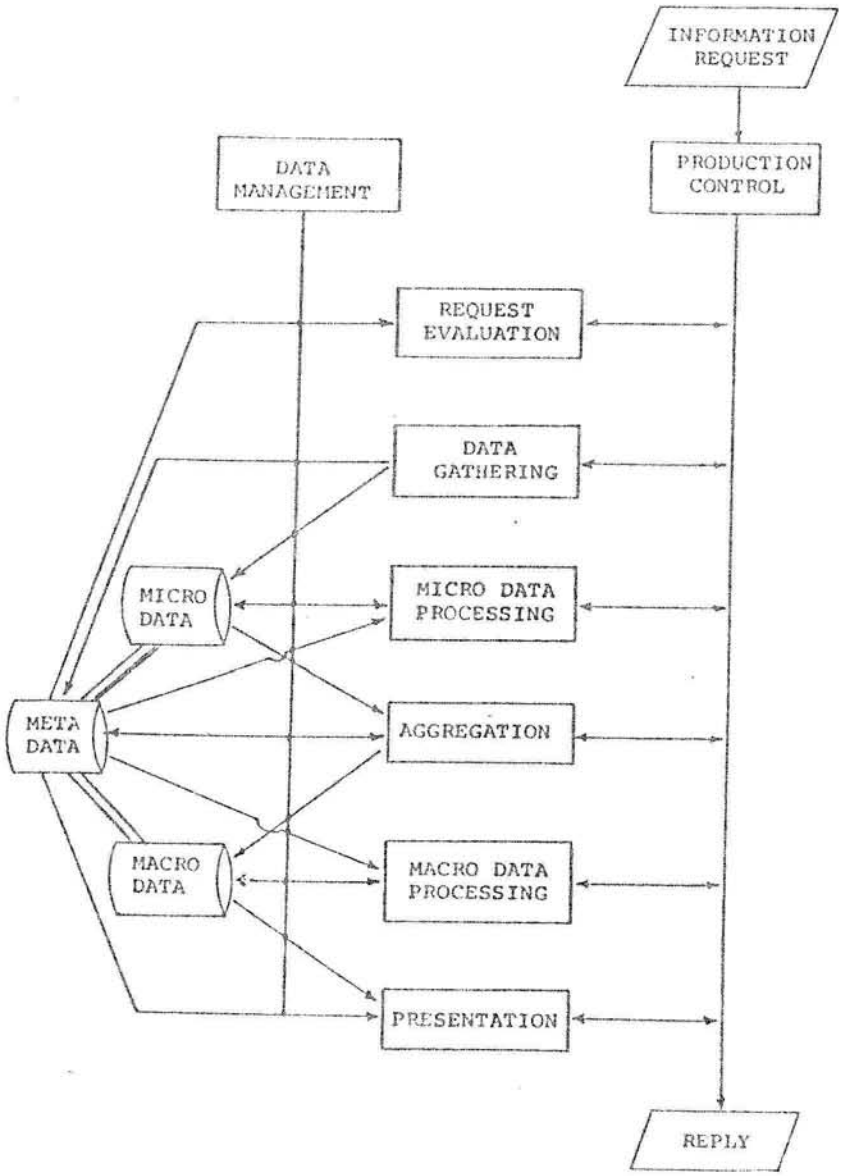

FIGURE 2: A DATA ORIENTED MODEL OF THE STATISTICAL PRODUCTION SYSTEM

According to this model of the survey production process:

1. The basic survey production processes remain as before. Control of their functions is given to a production control module which is capable of initiating 'out of sequence' processing, for example the initiation of a (new) aggregation routine, combining several sets of micro data. The request evaluation process would use the meta data base to determine the existance of relevant data.

2. All data administration, including the definition of both the micro and macro data, generation of storage structures, retrieval of data and execution of access control is provided by a common data management system. *) The meta data base connects and describes access paths and relationships between the micro and macro data bases, thus providing the descriptive and control information necessary to service the multiple processes.

Our interest now centers on the data management facilities required within this system.

2.3 Data Management Requirements

Data management refers to the functions of data description, storage, retrieval and structure maintenance as well as access and integrity control. The data storage structures selected for an application process, here the statistical survey, should minimize the retrieval and maintenance processing times, while allowing sufficient flexibility for various data access requirements. The data management facility should also provide support for the data definition requirements of the application system, here, for recording and providing access to the meta data. Studies of data management requirements in statistical systems [Group of Rapporteurs on DBM, 1978, Veim and Sundgren, 1979] identify

---

*) Possibly using some commercially available data base management system, DBMS.

the following requirements:

1. provide definition facilities for the meta, macro and micro data,

2. manage multiple, inter-related data archives, representing the data for multiple surveys,

3. provide basic data management functions: storage, retrieval and replacement of data in any one of the archives,

4. allow separate as well as combined processing of the micro and macro data bases,

5. allow multiple, complex and ad-hoc queries to the meta as well as the macro and micro data,

6. allow batch and on-line processing,

7. provide data security, (data privacy and integrity).

8. allow data base growth in number of recorded instances, and in addition of new data types

From reviews of the data base litterature [Veim and Sundgren, 1979], we know that this requirement list corresponds well with the goal statements of data base management systems. The requirements list, and its comparable DBMS goal set, has not yet been completely acheived, however powerful tools are available for many of the facilities required.

3. USING DATA BASE MANAGEMENT SYSTEMS

The previous section indicates a system environment which is suitable for data base management technology. We will here review the data base management experiences of some statistical offices with respect to the system environment modeled in figure 2. Our model is of a future, data oriented system. For the most part the state-of-the-art in both data base management systems and their application in statistical production systems is at the more simplistic level of figure 1, i.e. at the single application (survey). Therefore we will look at DBMS usage for data administration of each of the three levels of statistical data: micro, macro and meta, for survey production processing.

3.1 Data Base Management System Structure

Generally, a data base management system, DBMS, can be viewed as a system of three parts [Veim, 1977 and 1980] which interact to provide an integrated data management tool, see figure 3.

1. The data base management routines, DBMR
   contain the user communication routines such as the data definition and manipulation languages, DDL and DML, and the data management routines for data storage, retrieval, storage maintenance and access control. This component provides the capabilities and facilities of the particular DBMS. It would be in this component that the facilities for fulfilling the goal list of section 2.3 would be found.

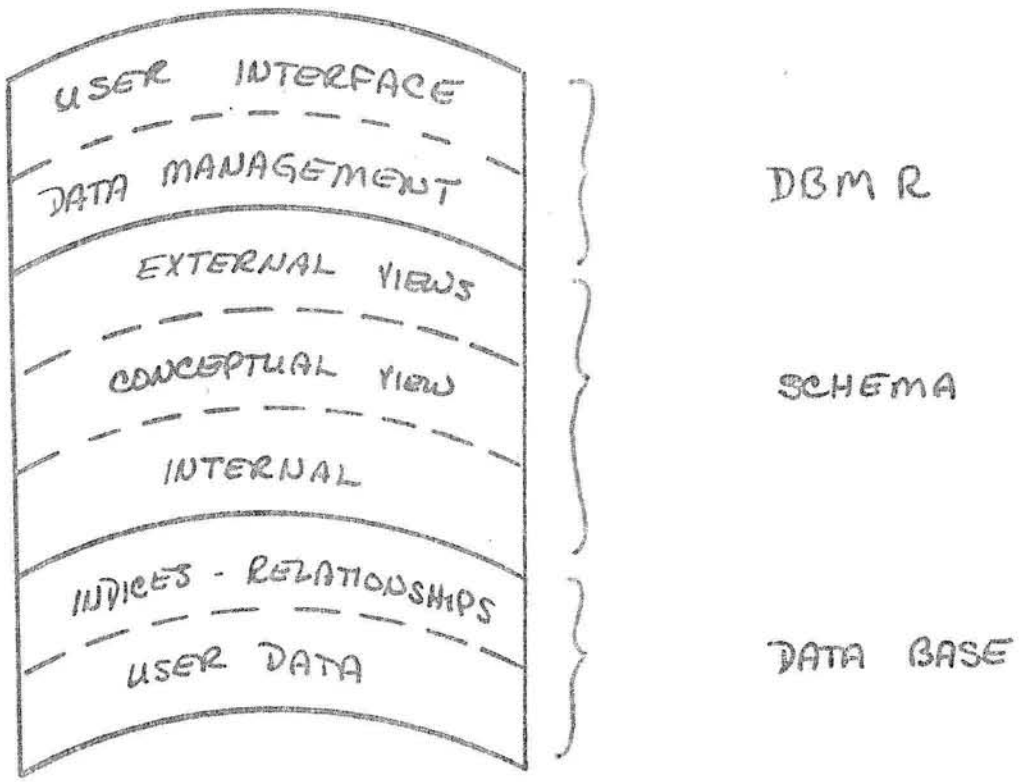

FIG. 3: DBMS STRUCTURE

2. the SCHEMA
contains the data definitions specified by the user and required by the host language processors and the DBMR for access to the data stored within the data base. Depending on the specific DBMS, these definitions pretain to three levels of description:

   1. the external schemas, which describe the data requirements for each user,

   2. the conceptual schema, which provides the union of the external schemas, supplemented with any applicable control information, and

   3. the internal schema, which describes the physical layout of the data within the data base.

   The DBMS/SCHEMA provides a facility for defining the data oriented meta data, i.e. the data describing the data names, formats and data element sets required for the processing routines (users).

3. the data base
contains the user data with any access data (indices and relationship representations) required. The DBMS assumes that the data base will contain data collected for two or more inter-related object-types.

Depending on the DBMS, the data base is structured as a set of files or as an integrated set of record-types.*) In either case, the data base structures will be fully described within the schema and the DBMR will contain the appropriate service routines.

---

*) For a review of the file stuctures used by data base management systems see [Martin, 1977, Cotton et.al.,1979 or Veim, 1980].

3.2 Data Modeling and the Meta Data Base

A Statistical survey, or set of surveys, builds upon a model of the object system being studied [Nordbotten, 1980], shown as the entity model in figure 1. The survey collects and processes its data according to this object system model. The collected data is further defined by the meta data (illustrated in figure 1) which describes the the survey frame, the code tables and editing rules to be applied to the micro data, the rules for generation of the macro data and the product table formats. Implicit in the meta data is a description of the data model and the data formats to be used. We can divide these meta data into three parts:

1. a model description,
   giving the entity and attribute names, definitions and inter-relationships which are to be represented by the survey data,

2. a process description,
   describing the code tables, edit rules and generation algorithms to be used when processing the data, and

3. a data description,
   giving data names and types and inter-data element relationships be used when accessing the survey data.

Current, commercially available DBMSs, (IBM's IMS, UNIVAC's DMS-1100, Cincom's TOTAL, etc.), support data modeling only at the stored data element level, corresponding to point 3 above. Primitive model description facilities, corresponding to point 1 above, are provided as the translation from a graphic/table description to the schema layout. The model types assumed and thereafter described include one of hierarchic, network or relational structures [Martin, 1976, Date, 1977].

Commercially, work on defining the entity model, its data value sets with their generation and presentation rules as well as their legal domains, is considered a function of the data dictionary system, i.e. as a strict (and separate) library facility for data terms. Before a true meta data base can exist, the existing schema data must be extended to include the data dictionary data.

Within the framework of statistical offices, work is being done to develop data management systems which offer an extended meta data base facility. Of particular interest is the Decision Information Display System, DIDS, which currently supports queries of the basic schema data and is being extended to include higher order meta data [Duncan, 1980] and the RAM system development in Sweden [Sundgren, 1977 and 1978].

3.3 The Micro Data Base

The storage structure for the micro data base is of primary importance as virtually all processing of the survey data will have to work within the restrictions set by this structure. The data capture processes, i.e. those translating questionair data to machine readable form, function serially. Initial coding and editing are also frequently performed serially, or serially within assigned groups, for example, within geographic areas. Editing for consistancy between attribute values requires the ability to select a subset of the attributes, via a PROJECT *) facility and then serial processing through the data.

---

*) Describing the data required for a particular process can be stated as one or a combination of the relational operators SELECT, a subset containing a given value set, PROJECT, over a subset of attributes or JOIN two or more relations containing a common attribute, see [DATE, 1977].

Traditional serial/sequential data storage, provides an unordered/ordered list of all record occurrences. This is basically satisfactory for the initial micro data processes, however no support is provided for the selection of subsets of the data by specified attributes, using a PROJECT facility.

Survey data, considered as one record per respondent, frequently may have as many as several hundred attributes. of which only a small sub-set, <10, will be required for the cross attribute editing or the later aggregation processes. Both the editing and aggregation processes use a large number of the data records, normally >10%. These attributes have lead to the definition of a statistical query as distint from a report, using many attributes of 'all' data or an information request for 'all' attributes of one record. [Cotton, 1979].

A DBMS can improve the micro data processing, principally by the maintenance of the schema data allowing the support of PROJECT facilities. Several commercial systems are in use for micro data processing, including TOTAL, ADABAS and the DBMS available from the central computer vendor (IBM's IMS, UNIVAC's DMS-1100, Honey-Well's IDS, etc.) [Davies, 1979]. Experience with these systems indicates that their emphasis is on providing support for information queries, as described above, or queries relying on the SELECT facility. Therefore, their performance for the statistical query is not as good as desired. From these observations, Statistics Canada is developing a DBMS called RAPID [Cotton, 1979] which will principally support the statistical query via the PROJECT facility on data stored as transposed files.

3.4 The Macro Data Base

The macro data base most frequently resides in the publications printed by the various statistical offices. New macro data are generated by reprocessing the stored micro data, usually after programming a new aggregation procedure. This procedure can take many weeks. If some time has passed, this process may be made difficult to impossible due to the loss of meta data.

Several statistical offices are investigating the feasibility of providing an on-line macro data base for interactive query and analysis. Interesting systems are being developed in Sweden, the UK, [Davies, 1979] and in the United States [Mendelssohn, 1979, Duncan, 1980]. In order for these systems to be truly user oriented, the data management system must provide support for an extended meta data base.

Current experience indicates that the relational model of the entity model provides the best user interface. The underlying DBMS must then support this view. Within statistical offices (e.g. Canadian, Swedish, US Bureau of Labor Statistics), the Canadian system RAPID or its predecessors is being tested for the data management of the macro data base. In other offices, the commercial DBMS, available from the vendor of their computer equipment, is being used. Positive experiences are reported for UNIVAC's DMS-1100, and Cincom's TOTAL [Davies, 1979].

4. CONCLUSIONS

It seems feasible to view the production of statistics as a set of processes aimed at developing and maintaining a comprehensive statistical data bank, encompassing the micro and macro data, described by the meta data, as illustrated in fig.2. The stated objectives of current data base management system software provide a potentially powerful tool for the administration of this data. However further development must be made on:

1. incorporating meta data into the total data base system, and

2. providing support for statistical queries.

Experience indicates that such an ambitious plan as suggested in figure 2, can be attempted by developing one data set, of micro and macro data, at a time, corresponding to the separate surveys as they are preformed. This leads to a final major requirement for data management services, that of:

3. providing for data base growth or extension.

5. ACKNOWLEDGEMENT

A special appreciation is extended to Svein Nordbotten without whose patient encouragement and support, this paper would not have been written.

BIBLIOGRAPHY and REFERENCES

Claringbold,P.J., and Smith,J.L., 1973
Data Base Aspects of Statistical Computing,
39th Session of ISI, Vienna

Codd,E.F., 1970
A Relational Model of Data for Large Shared Data Banks,
CACM vol.13, pg.377-387

Cotton,P., Turner,M.J., and Hammond,R., 1979
RAPID: A Database Management System for Statistical
Applications, 42nd Session of ISI, Manila

Date,C.J., 1977
An Introduction to Database Systems,
Addison-Wesley Publ. Co, Massachusetts

Davies,B.N., 1979
Data Base Management Systems in National Statistical
Services, 42 Session of ISI, Manila

Duncan,J.W., 1980
The Establishment of National Information Systems,
Office of Federal Statistical Policy and Standards,
Washington, D.C., USA

Fellegi,I., 1977
Functional Analysis of an "Ideal" Statistical System,
"Statistical Services in Ten Years Time",
Pergamon Press, Oxford

Graves,R.B., 1979
Data Base Modeling as an Aid to Data Editing,
Proc. 1979 ISIS Seminar (CES/SEM.9/10)

Group of Papporteurs on Data Base Management, 1978
Report on the Use and Future Need for Data Base
Management in National Statistical Services,
Statistical Commision and Economic Commission for
Europe, Conference of European Statisticians

Kazimour,J., 1977
Computer Hardware and Software: Its Use in a Central Bureau
of Statistics, "Statistical Services in Ten Years Time",
Pergamon Press, Oxford

Martin,J., 1977
Computer Data-Base Organization, 2nd edition,
Prentice-Hall, New Jersey

Mendelssohn,R.C., 1979
LABSTAT: A Database and Information System for
National Statistics, 42 Session of ISI, Manila

Nordbotten,S., 1961
Statistical Data Processing in the Central Bureau
of Statistics of Norway, 33 Session of ISI, Paris

Nordbotten,S., 1966
A Statistical File System,
Statistisk Tidskrift, 1966:2

Nordbotten,S., 1967a
Automatic Files in Statistical Systems,
Statistical Standards and Studies, Handbook no.9,
United Nations, N.Y.

Nordbotten,S., 1967b
Purposes, Problems and Ideas Related to Statistical
File Systems, 36th Session of the ISI, Sydney

Nordbotten,S., 1980
Systems Analysis and Design of Computerized
Information Systems, working paper
UN Statistical Office, New York

Sundgren,B., 1977
Meta-Information in Statistical Agencies,
Proc. 1977 ISIS Seminar, (CES/SEM.9/5)

Sundgren,B., 1978
RAM - A Framework for a Statistical Production System,
Proc. 1978 ISIS Seminar (CES/SEM.10/3)

Veim,J.C., 1977
On Data Base Theory, working paper
Institute for Information and Computer Science
University of Bergen, Norway

Veim,J.C., and Sundgren,B., 1979
Data Base Techniques in Statistical Data Processing,
42nd Session of ISI, Manila

Veim,J.C., 1980
An Introduction to Data Base Management,
working paper, GBA/CAIS
New York University, New York