AN APPROACH TO ENHANCEMENT OF STATISTICAL SURVEY DATABASES

Jose-Marie Griffiths
Donald W. King
King Research, Inc.

This paper deals with statistical databases that are generated from statistical surveys and that reside in organizations which perform a large number of surveys--some of which are repetitive. Examples of such organizations are Federal statistical agencies such as the Energy Information Administration, Bureau of Labor Statistics, National Center for Educational Statistics, National Center for Health Statistics, etc; state governments that have bureaus or departments that collect such data; and marketing research departments of most large consumer-oriented companies. Computer processing has provided a powerful tool for storing, manipulating, and analyzing statistical survey data. However, in addition to these advantages, computing has created a major problem in that most data analysts and users have lost touch with the data and their generation. They no longer have the feel and sense for the data that once was possible. In this paper we present an approach to database design that will directly attack this problem and enhance the usefulness of such databases as well.

In order to discuss the approach to such statistical survey database design, it is necessary to describe the entire data system within which it resides. This data system includes the following principal activities: determination of data needs, questionnaire design, sample design, data collection design, data collection, precomputer data processing (e.g., data editing and input), preliminary data analysis (e.g., data output), analysis, publication, and use. The flow of these activities is depicted in Figure 1. The first activity is determination of data needs, which presumably (but not always) comes from those who analyze the data and/or use them for decisions, policy making, etc. The three elements of survey design (sample, questionnaire, data collection) depend largely on the data needs, including which specific data values are required and from where, what statistical precision is necessary, what accuracy must be achieved, what timing is required, etc. Clearly, all the survey design activities are related, and they influence the subsequent activities as well. Data collection can be performed in many ways, including personal or telephone interview, self-administered questionnaire, or a combination of these. Precomputer processing includes manual coding and editing as well as keyboarding the data into the database. Sometimes there are two levels of databases: a microdatabase of raw data that is computer-edited and a macrodatabase, which is a database of transformed data where the transformations could be aggregation, statistical descriptions, conversions (e.g., petroleum data converted from barrels to BTUs), or models. Sometimes the Federal government develops a macrodatabase by aggregation in order to make the data available to the public and yet protect individual respondents. It is also not uncommon for there to be some human intervention or analysis of the data prior to finalizing the macrodatabase to make decisions on how to handle data such as missing values, outliers, and the like. Usually, publications are based on tabulated output and analysts and/or end users will rely on the tabulated output, publications, and--more often now--online retrieval of data.

Statistical surveys are like a chain, and their related actvities are like its links: failure during any one of these activities can destroy the validity of the entire survey. A study [1,2] of surveys performed for the Federal government suggests that such failures frequently do occur. In that study, 26 recently conducted surveys were investigated in depth to determine their validity. Seventeen of these surveys were found to produce invalid results because of failures during one or more of the activities mentioned above. Even though it is well recognized that this problem exists and serious attempts have been made to recommend solutions to it [3], little action seems to have taken place within the Federal government. In fact, the situation may be deteriorating even more; and there is no reason to believe that this situation is better in other environments.

There are many reasons for the sad state of affairs in statistical surveys. In this paper we concentrate only on those that directly involve the databases. One very important reason for the problem is that there is rarely a person who participates in all (or even a few) of the activities mentioned above or who has the responsibility to ensure quality of these activities. Least involved

in the survey design, usually, are the analysts and people who use the data. They have little understanding of what goes into the activities that produce the data they use. In our approach, we complement the databases (micro and/or macro) with metadata that will overcome many of the problems encountered in these systems.

For a statistical survey system to achieve its full potential, there must be full interdependence among the components of the system (ie., the activities performed). Furthermore, an analyst or final user must be able to interrogate any part of the system in order to determine what exactly has taken place concerning such things as response rates, editing procedures, how missing data are treated, outliers handled, transformations made, and so on. Finally, the system must provide feedback to improve the design of future surveys. Unfortunately, such feedback is rarely designed into a system.

The statistical database is complemented with two forms of metadata so that it will establish interdependence of components, permit interrogation, and create a feedback mechanism in the system. The two forms of metadata are (1) structured descriptions of data elements and their associated values and (2) tracking of data elements through the system from the questionnaire to their ultimate use. Data elements are logical definitions that have data values (or items) associated with them. For example, the number of firms in the United States in 1981 that process petroleum is a data element; the number 3,258 is a data value. One can describe data values in great detail, but never completely or entirely accurately or precisely. A structured description would include the definitions of firms (i.e., company vs. establishment), process, petroleum, sample frame from which a sample (or census) was designed, geographic coverage of the survey (i.e., does it include U.S. territories?) time for which the data is valid, and so on. Statistical precision and accuracy are affected partially by interpretation of the definitions of data elements on the data collection forms, but also by response errors, missing values, processing errors, etc. A well structured description of data elements and data values will add to the completeness as well as the accuracy and precision of the statistical database. Data tracking can be subdivided into two parts. The first part is tracing of data elements from the questionnaire through each of the activities shown in Figure 1, (i.e., precomputer processing, database processing, preliminary analysis, database output, published tables, and ultimate use). The second part of tracking involves the transformations performed on the values associated with the data elements as they pass through the systems. Such transformations on data values include data editing (either manual and/or machine), treatment of missing values, handling of outliers, aggregations, conversions, statistical descriptions (i.e., means and standard errors), modeling, and so on.

The structured description of data elements may take several forms. It may include a faceted, hierarchical classification with facets such as time, geographic area, geopolitical area, socioeconomic information, and so on. The hierarchies would be such that the lower-level nodes would correspond to the logic of numeric data and adhere to properties such as additivity. By building structured descriptions throughout questionnaires, databases, publications, etc., one automatically forms an interrelated bond throughout the system. This also permits one to gain online access to the database, if constructed properly, in a much more powerful way. Otherwise one must search exactly as the data value names (data element names, in DBMS terminology) appear in the database. An unfamiliar user would have great difficulty in determining the exact terms to use in searching for data in the system, especially as there may be no standardization of names in the system. It is emphasized that even "facts" are difficult to describe and query from, without some structured terminology to rely on.

With structured descriptions and tracking mechanisms we can easily interrogate the system to determine how data are handled throughout. Such interrogation can be used to assess quality and statistical reliability of data, validate data elements, update data elements, and enhance description of data elements. These uses are described in more detail in the second section of the paper.

Feedback can take place with all activities in the system. By monitoring use of data and their strengths and weaknesses, one can determine whether data elements should be dropped or modified. One can also establish the form (database or published) in which data should reside, where frequently used data can be output and published, and where infrequently used data can be left in the database for on-demand search or tabulation. One can also use such data to modify the structure of the database or for designing future databases. Finally, by maintaining metadata records or response rates, missing data elements, statistical variances, etc., one has useful data for modifying sample, questionnaire, and data collection design. Such data would also be extremely useful for designing future statistical systems in general.

Structured Descriptions

As mentioned previously, structured descriptions relating to data values in a statistical database have the potential to provide an improved retrieval capability for the database user. By describing the data elements and their associated data values, the database can be partitioned into small groups of similar or related data elements. The more detailed the descriptions (either in terms of the number of attributes of the data elements that are described or in terms of the depth of description for each attribute), the smaller the groups will be; and, for a database of finite size, these groups could ultimately contain only one member. The power of such a multi-attribute categorization lies in its ability to define and relate data elements. The description of a data element becomes its definition, and related data elements have similar descriptions.

The structure of the descriptions is also a significant factor determining (1) how they will be used for accessing data values and, consequently, (2) how the performance of the overall data handling system has improved. The more structure

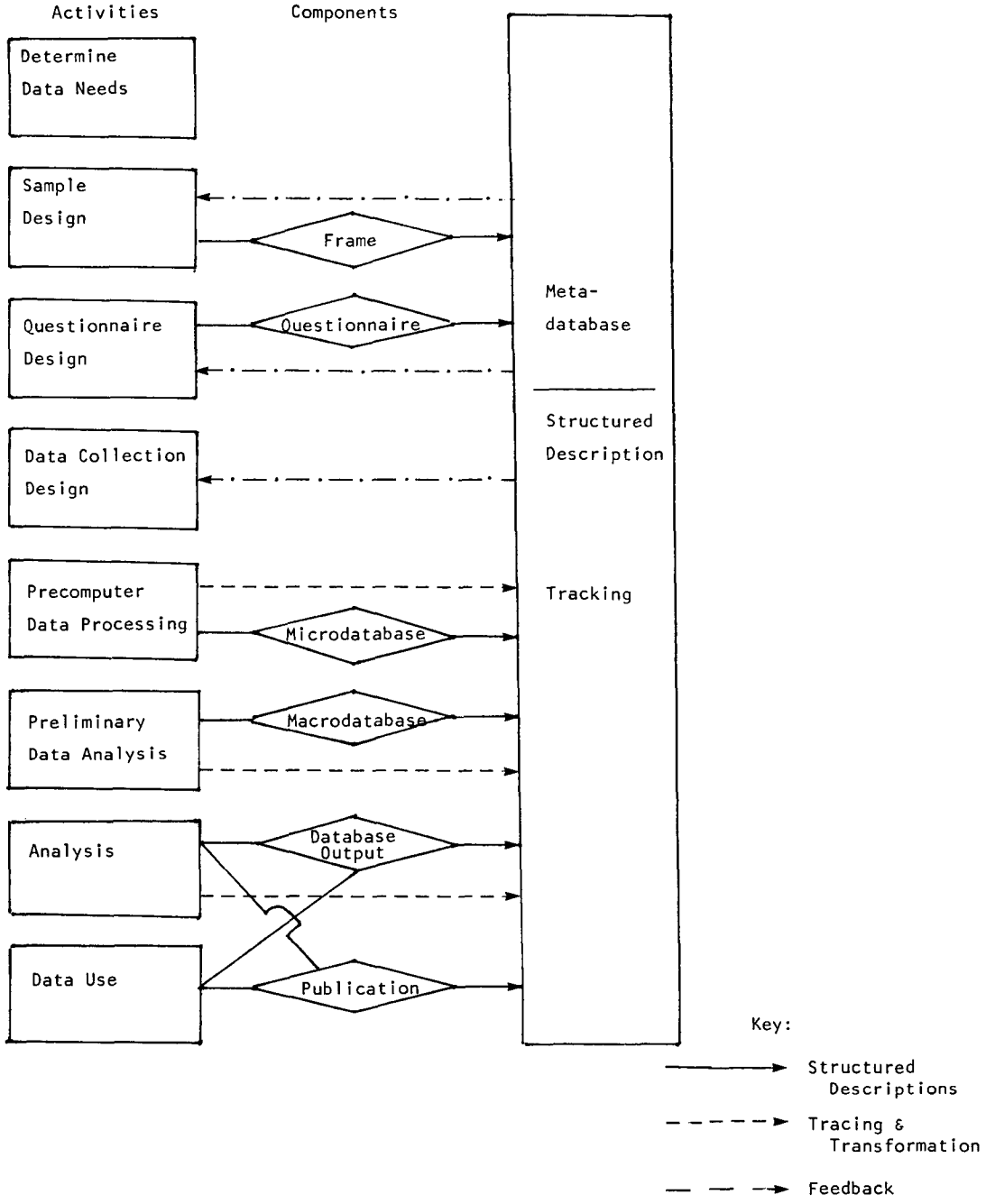

Figure 1. Flow of Activities Associated with Statistical Survey Systems

that is imposed on the descriptions, the greater their potential for retrieval purposes. However, even a totally unstructured description (like an extended abstract) would still be extremely useful to the analyst when using a data retrieval system. In fact, the contextual information provided in such a description, regardless of structure, should enable the analyst to make an assessment of quality and reliability of the data retrieved.

The advantage of structuring the description is that they can then be used for retrieving data values rather than being called up as descriptions associated with already retrieved data values. The structure that is selected will depend on the nature of the data and their ultimate use. The two aspects of structure to be considered are format (i.e., the breakdown of the description into separate items--analogous to the fields in a computer record) and vocabulary (i.e., level of vocabulary control to be imposed). A highly controlled vocabulary will improve the precision of the retrieval process; however, it will not allow for much flexibility for the user inputting requests to the system.

Such descriptions form part of a data resources directory (DRD) currently under development for the Energy Information Administration (EIA) of the U.S. Department of Energy. The data collection activities of the EIA center on over 200 repetitive surveys. EIA's primary survey mechanism is the data collection form. The DRD will provide meta-information about the data collected by surveys and, in some cases, retrieve the actual data values. The data system components and the data elements composing the data collection systems will be fully classified (using a faceted classification scheme for energy data) and described (by the four attribute categories: identification, management, statistical, and linkage).

The structured descriptions can, for convenience, be divided into two parts. The first is an in-depth classification of the data elements. For the DRD, the approach of faceted classification has been adopted [4]. The second part of the descriptions involve the remaining attributes of the element, which need to be covered in order that it may be distinguished from similarly classified data elements. For convenience, the attributes are grouped as follows: identification, statistical, management, and linkage. For example, the description of a data element on one of EIA's data collection forms is outlined in Figure 2. The type of access that users of the DRD will have to these descriptions depends partially on the vocabulary used for the individual attributes. The attributes marked with an asterisk are those that are currently envisaged as access points to the system. Once the appropriate data elements have been identified for each query, there are three possible modes of gaining access to the associated data values themselves:

1. The system can inform the user of the location of the data value.

2. The system can inform the user of the steps required to retrive the data value.

3. The system will actually retrieve the data value automatically.

The latter case is realistic only in a DBMS environment, so it may be necessary for the user to adopt one of the two other modes.

Data Tracking

The second technique that can aid the user of a data management system in determining the statistical reliability and quality of the data retrieved is that of data tracking. Additionally, the tracking can help to ease data element description and maintenance activities within the system. A tracking mechanism is being developed as part of the DRD [5]. This mechanism makes use of the linkage and statistical aspects of the descriptions to provide a powerful tool that can be employed in several different ways--for example, to:

1. Provide a means of assessing the quality and statistical reliability of data. To assess the quality of the data, the user initiates a backward tracking mechanism. For example, a user wanting additional information about a data element published in a table could use the mechanism to discover where that data element came from (perhaps from a computer printout, from a field in a computer file, from a question on a data collection form) and how the element has been transformed in moving from one data component to another (through aggregation, averaging, etc.). By allowing the user to retrieve descriptions of the data elements identified by the tracking mechanism, a much more detailed notion of the data element under consideration can be achieved by reviewing factors such as the definitions used on the data collection form, description of the respondent population, determination of response rates, data handling procedured, etc. One can also establish the statistical reliability of data by assessing the statistical design of the survey and the descriptions of the manipulations performed on the data. The tracking mechanism thus provides the essential capability to establish the data quality and statistical reliability.

2. Validate the data elements. One means of validating observed data elements is to identify comparable data from other sources in the system to see whether results are the same or similar. If they are not, then one can use the backwards tracking mechanism to establish the validity of each source and make judgements about which is best.

3. Update data elements. Data can change from time to time because respondents might alter their responses or because a field validation might indicated that data need to be corrected. This use of the tracking mechanism involved forward data element tracking which is used to locate these data elements that will be affected

Figure 2. STRUCTURED DESCRIPTION FOR DATA COLLECTION FORMS & ASSOCIATED DATA ELEMENTS

1. Identification attributes
   - Form title.
   - *OMB No.
   - *EIA No.
   - Prior form no(s).
   - New form no(s).
   - *Classification.
   - Abstract.

2. Management attributes
   - *Form status.
   - Confidentiality.
   - *EIA category.
   - Date approved by OMB.
   - *Average burden per response.
   - *Frequency of collection.
   - Reporting period.
   - Submission date.
   - *Expiration date.
   - Collection medium.
   - *Respondent type.
   - Respondent description.
   - Respondent source.
   - *Reporting requirement.
   - Geographic coverage.
   - Schedule no(s) and title(s).

3. Statistical attributes
   - Survey type.
   - Sampling methodology.
   - Sample derivation.
   - No. of respondents.
   - Estimated universe.
   - Data recording instructions.

4. Linkage attributes

ATTRIBUTES FOR DATA ELEMENTS

*Classification & Operators

Transcription

*Parent Instrument

Location on Instrument

Note

Primary Category

The term (from the classification) which is the primary focus of the data element (the entity being measured)

*Serial number

*Access point

by changes to data elements preceding them in the data handling cycle. For example, a change made to a data element on a data collection form would often require changes to appropriate data elements in a computer file, computer-produced output, and published table.

4. Enhance description of data elements. The tracking mechanism enhances the description process by providing a linkage between the description of data elements found on data collection forms and the subsequent appearance of data derived from responses to these forms. The definitions of energy-related terms found on the form, together with the transformations that occur in deriving data elements found subsequently, serve as the basis for classification.

Two alternative methods for developing the tracing and transformation information exist--manual and automatic. Manual development is an extremely time-consuming task, as it involves the identification of all data system components associated with a particular data collection form (or set of forms) followed by a detailed analysis of the derivation of data elements from those appearing on the form. Automatic development of the tracing and transformation information can only be achieved on totally automated data-handling systems and would involve the construction of a computer program to monitor data element manipulation by the editing and processing routines of the automated data-handling system.

In each case a coding scheme for recording the tracings and transformations would have to be developed. In the manually derived system the manipulations performed on the data elements could be described using natural language (e.g., data element C is a volume weighted average derived from data elements A and B). In the automatically derived system the description would be symbolic, perhaps by the development of a specialized algebra for describing data transformation (e.g., C A, B: $^{a}/_{v}$). In either case the descriptions of component data elements can be called up and displayed to provide more contextual information concerning the transformation.

The data tracking mechanism of the DRD is based on manually derived input for two reasons:

1. Not all the data collection systems are fully automated. Some of the systems are entirely manual and others include manual intervention at the editing stage or at the production of tables (for publication) from computer printouts.

2. Lack of standardization of EIA data-handling systems. Not only are the partially manual systems inconsistent in their treatment of data, but even those systems that are totally automated differ in their organization and manipulation of data (e.g., file structures differ, programs are written in different languages and use different algorithms). To develop automatic tracking, a separate monitoring program would have to be written for each system.

The problem that essentially forces us to use manually derived tracing and transformation information for the DRD relates to the fact that the DRD must describe a series of existing data sets, procedures, and manipulations. In designing a data-handling system from scratch, it would be feasible to perform totally automatic tracking of the data elements through the system by establishing a set of procedures and forcing compliance. Metadatabases could then take an active role in data management. For example, by insisting that data manipulation programs are written such that each transformation type is designed as a separate procedure known to the metadatabase along with those higher-level procedures that call the transformation procedure, then monitoring procedure calls, their parameters and the input/output segments of the main program would provide sufficient information about tracings and transformations to feed the tracking display mechanism.

A number of implementation alternatives for data tracking mechanisms can be identified:

1. Establishing the tracking patterns as a database of tracings and transformations to be searched, or tracking a data element and generating the tracings and transformation information at retrieval time. The basic tradeoff is between large volume of storage required and fast retrieval in the former case, to reduced storage requirements and longer retrieval time in the latter. Further, in the second case, it may not be necessary to generate a complete track.

2. Differences in the level of detail recorded and retrieved. It may be that just the "significant" transformations are required (e.g., may not want the details of transcriptions of data elements--moving a data element from one system component to another without changing it in any way--or conversions from one unit of measure to another).

3. There are several display alternatives, ranging from simple use of indentation of textual output to fairly sophisticated computer graphics applications. The basic tradeoff here is between ease of assimilation of information by the user, ease of operation of the user interface, and cost of design and implementation of the interface.

The selection of various design strategies from the alternatives discussed in this paper depends on the characteristics of the data being handled, the organizational environment within which the data are processed, and the types of users having access to the data. Nevertheless, structured data descriptions, coupled with data tracking mechanisms, have the potential to significantly enhance the performance, operation, and use of statistical data management systems by pro-

viding the opportunity for qualitative assessments by their users of the data they handle.

The Data Resources Directory has been partially implemented at the Department of Energy, Energy Information Administration and further work on the concept is now being done for the National Science Foundation from whom King Research has been awarded two research grants. One grant, under the Small Business Innovations Research program, is planned to develop an integrated package of system technology for numeric data-handling that will deal with all the phases of survey processing. The other related grant is to investigate the cost and performance tradeoffs of alternative methods of integrating data structures. It is anticipated that these two investigations will help advance the capability of handling survey data from different sources as well as through the many forms and processes that the data elements are subjected to.

REFERENCES

1. Bailar, B. A., and Lanphier, C. M. "A Report of the American Statistical Association Project on the Assessment of Survey Practices and Data Quality--Surveys of Human Populations." NSF Grant No. S.O.C. 74-22902. 1977.

2. Bailar, B. A. "Progress and Problems in the Assessment of Survey Practices." 1976.

3. King, D. W.; Bailar, B.; Ladd, B.; and Reisin, P. Dowd. An Assessment of Practices in Federal Surveys Conducted Under Contracts. Prepared for the Commission on Federal Paperwork. King Research, Inc., Rockville, Maryland, May 1977.

4. Batty, D., and Travis, I. "A Classification-Based Information Retrieval System for Federal Energy Data." Paper presented at the American Society for Information Science Mid-year Meeting, Durango, Colorado, May 1981.

5. Griffiths, J-M. "Data Tracking." Paper presented at the American Society for Information Science Annual Meeting, Washington, D.C., October 1981.