THE SURVEY OF INCOME AND
PROGRAM PARTICIPATION

DATABASE DESIGN FOR LARGE-SCALE, COMPLEX DATA

No. 100

M. H. David and A. Robbin
University of Wisconsin

U. S. Department of Commerce BUREAU OF THE CENSUS

Survey of Income and Program Participation

Database Design for Large-Scale, Complex Data

8923 – 100

Martin H. David and Alice Robbin
University of Wisconsin

November 1989

ACKNOWLEDGEMENT

Presented at the 21st Symposium on the Interface of Computer Science and Statistics, April, 1989 in Orlando, Florida. This paper will be published in the Proceedings of the 21st Symposium on the Interface of Computer Science and Statistics. The views expressed are the authors' and do not necessarily reflect those of the Census Bureau.

TABLE OF CONTENTS

DATABASE DESIGN FOR LARGE-SCALE, COMPLEX DATA

Martin H. David and Alice Robbin, University of Wisconsin

1. INTRODUCTION

High-dimensional data structures are the focus of this session.* We use the adjective *complex* in place of high-dimensionality because the problems that we describe arise both from the measurement of thousands of attributes and from the intricate logical conditioning of the measurement process. Our paper provides answers to three questions associated with these data structures.

A significant structure always underlies data collected for scientific analysis. The question is, *How do we reveal that structure to support statistical analysis?* Time is an implicit dimension of a data structure. The design of a data collection is not always identical over time. Some of our discussion is devoted to how *time* is represented when measurements are asymmetric to different time points.

Complexity connotes both technical and cognitive problems for retrieving data. Technical problems can be addressed in part by applying relational theory to simplify and clarify data structures. Faulty memory, the limited capacity to process information, and uncertainty about outcomes can be partially overcome by applying principles derived from cognitive theory to organize data for retrieval and to represent meaning. These observations are the source of our second question, *How do we make data accessible?*

Models applied to the data entail units of analysis and concepts that were not envisioned by the original data collectors. Our third question derives from a recognized tension between data collection and subsequent use. *How do we maintain the integrity of the measurements while organizing data for a variety of analysis tasks, including extension of the original data by repeated measures, auxiliary variables, and replication?*

We sought answers to these and other questions in a project that developed a prototype of an integrated information system to improve access at low-cost to large-scale, complex data by university, government, and private sector policy analysts throughout the United States (David 1985a; David, Robbin, & Flory 1988a; Robbin & David). [1] Our paper discusses aspects of the conceptual framework and strategies we applied to design a database that integrates statistical data and metadata (information about the data), including the database design and contents, survey design, collection and processing procedures.

Section 2 presents a conceptual framework for linking survey design to appropriate data structures. The remainder of the paper demonstrates how principles described in Section 2 are applied to the 1984 panel of the *Survey of Income and Program Participation* (SIPP). The generality of the principles and their success in handling the difficult

SIPP design lead us to conclude that this framework can be generally applied to social science databases.

Section 3 describes the complexity of the SIPP design and measurement. Subsequent sections elaborate on the principles applied to the SIPP database design. Section 4 discusses the value of semantic principles for organizing data. Section 5 explains how time is represented in the SIPP and why its representation in the database may conflict with the use of the third normal form as an organizing principle. Section 6 emphasizes the role of metadata in clarifying underlying data structures and improving access to complex data.

2. DESIGNING A DATABASE FOR COMPLEX DATA SETS

This section discusses four principles that guided our approach to designing a database schema for large-scale, complex statistical data.

2.1. A Conceptual Framework

Figure 1 represents two facets of data. Data are generated by a scientific design; inference from data requires interpretation consistent with the design and the procedures used to execute the design. The *design (sample-experiment-instrument)* determines what inferences can be made about populations and treatments. The *channel of measurement* executes the design. It is determined by instruments (questionnaires, etc.) that are used for measurement and by procedures that govern their administration.

The end product of design and procedures is an *image* that contains all the data values. The information in this image is a function of design and channel in precisely the same sense that information in a satellite map is a function of complex signals sent from several instruments and interpreted through computer algorithms. The image will vary according to the procedures that capture responses on machine-readable media.

We caricature the flow of information through the design and channel by the column subheadings in Figure 1: sample, stimulus, response, and data image. They convey that the purpose of measurement is to elicit information about designated entities and to represent an image of the responses in a data structure amenable to statistical analysis.

A coherent semantic principle underlies the production of the data image. Information is elicited from a well-defined sample according to known interrogative procedures (questions, observations, or real-time application of auxiliary instruments). The responses are captured and transformed by computer algorithms with single-valued outcomes. In the process messages received from the respondents may be coded, censored, or combined with other information. The messages may be attributed to other entities (proxy reports about others) or aggregated to give measures related to groups (income for families). Many null values are inserted in the data image as defaults for situations

where data were not collected on all sample elements; these are not-in-universe codes (NIU).

Interpretation, i.e., analysis, of data requires inverting the data image in two distinct senses. Samples are generalized to populations and superpopulations. The meanings of values in the data image must be translated to natural language in order to communicate information to the scientific (or policy-making) community who use the information. Aspects of this interpretive process are shown in the lower part of Figure 1.

2.2. Implications for Database Design

The concepts underlying Figure 1 have guided our development of a database to improve access to SIPP. They can be expressed as four principles that generalize to other complex data structures. They have been tested and proven in four years of developing the SIPP ACCESS facility.

Design the database schema to conform to the channel of measurement. This implies that database should incorporate the questionnaire image and all responses. It also implies that the primary consideration in database design is to preserve the semantic principle that generated the data. That is, the data are generated by a question directed at a particular population. Responses to the question reflect the object to which the question refers (referent), the bounding reference period, and the attribute that is being elicited.

Provide dynamic independence. The database must be capable of receiving additional attributes and additional observations, to permit pooling of data sources, analysis in relation to contextual variables, and reorganization of the data to a wide variety of units of analysis (Codd).

Preserve information and maintain the capability of transformation and its inverse. Any manipulation of the questionnaire image to facilitate access and retrieval must be reversible. That is, it must be possible to recover the original image. This rule assures that erroneous processing can be undone and that detail is not lost by transformation. For example, converting birth dates to century notation is information-preserving; scaling birth dates to age classes in a particular year is not.

Maintain a journal or audit trail. This document completely describes transformation of the response to the questionnaire and to the image in the database. The journal records the rules that governed decisionmaking, and thus serves as a tool for evaluating data quality as well as the assumptions that govern design and development. Journaling also creates a "template" for conducting similar activities in the future and therefore has the potential for introducing efficiencies into the transformation process. Finally, maintaining an audit trail is consistent with the scientific norms of replication and data sharing (Boruch; Boruch & Cecil; Boruch & Cordray; Hedrick).

2.3. Strategies for Database Design

We emphasize three aspects of database design: exploratory learning and prototyping, analysis-oriented data enhancements, and extended metadata.

Exploratory Learning and Prototyping. Research problems are rarely well-formulated in advance of executing query and search procedures. Nearly all research requires a period of exploratory analysis during which the scientist learns about the data collection. Orders of magnitude for the size of populations of interest, rough indications of the distribution of outcome measures, and tests of the consistency of measurements against simple logical principles reveal the feasibility of a research plan and suggest modifications. In this exploratory phase frequent interaction with the data is typical, and discoveries of undocumented data processing are common. Researchers discover errors in their understanding of variable labels and execution of simple transformations of the data.

Low cost, rapid interaction with the data for experimental analysis is achieved by storing a representative sample of the data in a schema identical to that of the entire database (David, Robbin, & Flory 1988b). Rigid adherence to identical data structures for the database and the sample assures that any "well-formed" query which produced information from the sample can be executed on the entire database without programming modifications.

What is true for the researcher is equally true for the database designer. Application prototyping assures greater implementation success through gradual learning and incremental change (Boar). We created a sample database of the complete panel data. This sample database was used to develop a compact, "restructured" database from the public use data tapes issued by the U.S. Bureau of the Census. Only when database design and testing had been completed on the sample was the complete panel reorganized to achieve efficiency through compaction.

Analysis-oriented Data Enhancements. The database designer can achieve great economies for the future analyst with "tools" that facilitate analysis. In the SIPP database these tools are summary tables that collate data from up to nine interviews for all members of the panel. These "longitudinal tables" obviate a need to retrieve data from nine different interviews, and they anticipate the structure of analysts' queries by reporting, for example, on "spells" of a state variable, dynamic relationships (family composition, and sample relevance of persons or labor force participation (Flory, Robbin, & David 1988a; Flory, Martini, & Robbin; Martini). We describe these tools in more detail in Section 5.

Extended Metadata. A third and, in our view, indispensible, strategy for improving an understanding of complex data is the creation of extensive metadata within the database (Robbin & David). Successful end-use interaction—better decisions about selecting data, correcting and locating errors, and constructing alternative access paths into the database—is the measure of successful data base design. Database design re-

quires that careful attention be given to the essential role that natural language plays in describing statistical (or any other) data and to the categories devised to organize and communicate the contents of a database (cf. Dolby, Clark, & Rogers; Fischhoff, MacGregor, & Blackshaw). A more extensive discussion of the role of metadata is found in Section 6.

3. COMPLEXITIES IN SIPP

In this section we describe principal features of the SIPP design and the resulting complexity of the data.

3.1. The Design

The 1984 SIPP panel conducted by the U.S. Bureau of the Census collects data on income of individuals and household demography through an extraordinarily complex design and channel. Some of the attributes of this complexity are described below.

- The initial area probability sample of 20,000 households is extended in time by following the adults in the household.
- The population varies after the first interview as new people are born into the sample through birth, marriage, or remarriage, and as original or new sample persons exit through death or divorce or cannot be located after they move.
- Data cluster in several natural aggregations. Information obtained from the interview may pertain to individuals (with differences between children and adults), married couples, families, households, health insurance units, employee-employer pairs, and client-program pairs (where programs are 39 different entitlements for benefits from local, state, or federal income maintenance, social insurance, medical, and welfare programs).
- Data are collected during eight or nine interviews. Respondents provide information about the four-month period preceding the interview, and a limited amount of demographic information on all members of the household is obtained by the enumerator at the time of interview. A "core" set of questions is repeated in each of nine interviews; and "topical module" questions on high-interest public policy issues supplement the core after the second interview.
- Complex conditions determine eligibility for measurement. Conditioning may depend on responses to the prior interview. For example, the initial interview screens for more than 50 different types of income. Subsequent reinterviews are conditioned on earlier reports.

These complexities in the questionnaire design are compounded by complexities in the administration of the survey. Four independent subsamples were interviewed in consecutive months ("staggered interviewing"). As a consequence, members of the

area probability sample are visited at intervals of exactly four months (for up to 36 months). However, the nine different questionnaire instruments in the 1984 SIPP were not administered to the entire sample, with consequent asymmetries in the data for persons selected into each subsample. In addition, eighteen percent of the sample was eliminated after the fourth contact, and half the sample was interviewed only eight times to reduce cost.

Following collection of data in the field, an image of the questionnaire is created by optical scanning. The image is transformed during data processing by algorithms that recode variables, impute missing items, statistically match missing interviews, and aggregate person-level information to create family and household characteristics.

The survey design facilitates a variety of estimands. Cross-sectional point estimates (e.g., for individual earnings) can be constructed for any one of the nine reference periods corresponding to the nine interviews. Furthermore, these estimates can be constructed from data collected with varying recall periods so that response error can be estimated. Data from three interviews can be aggregated to estimate annual income data; data from six interviews can be used to estimate year-to-year change. And data from the entire panel can be studied for individuals to estimate the probability of entering or leaving a status, such as unemployment. Each of these different estimators requires a different selection of sample entities. Each requires an understanding of the calendar that corresponds to the reference period in each of four independently drawn subsamples.

3.2. The Public Use Files

The public use data files are released by the U.S. Bureau of the Census for every interview. The 1984 SIPP panel includes nearly 7 million observations and about 20,000 data elements. These data are distributed on more than thirty tapes, totaling about 2.2 gigabytes of data in nine separate files representing the nine interviews. Each file is physically organized as eight record types.[2] The data must therefore be restructured and linked for studying dynamic changes in the population. Subsequent panels increase in size by about 300 megabytes of data and 2,500 data elements every four months. The 1985, 1986, 1987, and 1988 panels (each 32 months in duration with a reduced sample size) will introduce more than 8 gigabytes of data and more than 70,000 data elements into the system.

A machine-readable codebook and print questionnaires accompany the public use files. Complexity in these materials inevitably leads to errors—omissions, variations in labelling, and peculiarities in coding—that must be resolved before creating a database.

Labelling conventions used by the Bureau facilitated database development. Information taken directly from the response boxes on the questionnaire is labelled with the number of that box. Replicated questions carry the same label over the nine questionnaires. Measures related in some way (by aggregation or by time sequencing) are

often labelled with a root and prefixes or suffixes that parse to identify the relationship. For example, person earnings in reference month one are labelled *ppern1*; household aggregate earnings in reference month four are labelled *hhern4*. The role of labelling is elaborated in Section 6.2.

3.3. The SIPP ACCESS Database

The 1984 SIPP panel data reside in a relational database management system (RDBMS). The data base schema, devised by careful use of third normal form relations, reduced the scope of the original data set by 75%, while improving its logical clarity and accessibility. (For a discussion of the utility of the RDBMS for panel data, see David 1989). The logic of the RDBMS reinforces a work strategy that is iterative and exploratory, and provides an excellent tool for answering research problems that are rarely well-formulated in advance of executing query and search procedures. Direct access provided by the RDBMS improves a user's knowledge of statistical data. The logic behind the structure of the RDBMS reduces errors made by analysts and researchers.

4. THE SEMANTIC PRINCIPLE: REDUCING COMPLEXITY

A key to managing the complexity of the SIPP data is the application of a semantic principle to organize the representation or image of the data. The channel design ensures that a clear logical principle is used to select entities for observation. The design also makes certain that a group of questions pertain to that universe. Representing data in the computer storage in a conformable way increases the clarity of the information for analysis.

Attention to semantic principles requires that data be stored in relations (tables) that adhere to the third normal form (Kent). Every attribute in the relation is the property of the entity that is uniquely identified by the "key" or identifier of that case or record. Over eighty relations are required to establish the principal samples of entities measured by the design. Some are obvious and were embedded in the public use data that we received. For example, it is clear that information on jobs (employer-employee pairs) existed for a subset of adults, and that some individuals could have two jobs. In other cases, however, expediency in preparing the public use data led to illogical and confusing treatments of the principal samples.

The semantic principle ensures that users understand the relations and their content. We provide several examples below. First, we draw attention to the ways in which NIU bytes appear in an array.

4.1. Eliminating Not-In-Universe (NIU) from Sparse Matrices

Non-informative bytes arise when default values are inserted in the array to act as place-holders for cases that are not measured on a particular attribute. Typically, a subset of measurements will be made for a subset of cases. Default values are eliminated by organizing the storage according to the design which generates the measurements.

Logically, default values may appear in all attributes in a given row, in all cases of a given attribute, or in the intersection of a subset of cases for a subset of attributes. We were able to achieve economies of access by removing all three of these types of NIU. Below we give three examples of adherence to semantic principles and third normal form for organizing a SIPP database.

Deleting Rows: Children. Children under fifteen years of age are not interviewed in the SIPP. Their presence in the household and family is used to generate appropriate weights for the analysis of individuals, families, and households at a particular point of time. Nonetheless, default values commingle observations for interviewed persons with those of the children. Approximately one-quarter of all interview data entered on records for persons are therefore NIU.

Our first step in reorganizing data was to partition the person array so that weights and related demographic information on all persons were separated from substantive information obtained for persons over 14 years of age. This partition allowed us to delete one-quarter of the matrix of information supplied in the public use files for persons. This strategy reinforces the principle that responses to questionnaire items are restricted to a universe that excluded persons under 15 years of age.

Deleting Columns: Retrospective Demography. Household composition is established at the first interview, and for every succeeding month until the last interview. Data processing by the Bureau of the Census transfers household composition at the time of the first interview into attributes describing the four months prior to the first interview.[3] These values carry no information because all the measurements are generated at the time of interview (and not before). Our data reorganization removed attributes from the original data set which contained no information and were misleading.

Retaining Intersections: Reorganizing Information on Program Income. Data processing by the Bureau of the Census forces all information about program income receipt and amounts into a single array, although some questions are specific to particular types of programs. That is, different questions are appropriate for AFDC, Social Security, and Food Stamp receipt and income. We reorganized the data into separate arrays for each program type. Each array contained only the attributes appropriate to it, thereby deleting default values for persons who reported no program incomes. This is an example of deleting NIU defaults for selected cases on selected attributes.

4.2. Eliminating Redundancies

Two types of redundant information were deleted from the database.

Process Control: Interviewer Check Items and the Income Roster. The interview process requires enumerators to enter attributes measured at prior interviews (transcription items) and to recall attributes measured earlier in the current interview (check items). However, these attributes contain no new information because they repeat attributes measured earlier. Although they are important to understanding the data collection process, they are not necessary for secondary analysis. Check items and transcription items were therefore not incorporated into the reorganized data schema. This decision eliminated many redundancies in the data schema.[4]

Unchanging Attributes. A variation on redundacy occurs when unchanging attributes, such as birth dates and gender, are repeated in arrays for successive interviews. We organized all personal constants into a single array for two reasons. This eliminated repetitive entry of the same information at several points in the data set. A scan of all the values for these personal constants enabled us to eliminate response errors (i.e., we selected the modal, or edited, value when more than one value appeared for the constant).

4.3. Transformations: New Logical Universes

Another application of the semantic principle illustrates the conceptual value of retaining data in a form that corresponds to meanings in the data collection.

Joint Asset Incomes. Data on savings accounts and several other types of asset-related income are obtained in separate sequences for assets jointly held by couples and for the respondent as the sole owner. Jointly-held accounts are reported by the first interviewed member of the couple. In the public use data file an NIU is inserted in the second member's record if the first member has already been interviewed. The record of the reporting member of the couple is not necessarily the first in the sequence of data. This is confusing, because two records must be retrieved in order to discover the jointly-held amount. We reduced the storage of information on assets held by couples by fifty percent by defining the couple as the logical unit.

The semantic principle applied here assures that the design of the questionnaire is clearly represented in the data image. The task of deciding how individuals benefit from jointly-held assets can then be addressed by the analyst without being obscured by the form in which data are retrieved.

Insurance Units. A second example of how the questionnaire design can be more clearly presented in the data schema pertains to membership in a health insurance group. When some, but not all, of the members of a household are covered by medical insurance, SIPP created a membership roster. This roster applys only to those households if insurance coverage is incomplete. However, the public use files allocate space

in every person's data record, even when coverage of all persons in the household is complete. We therefore created a relation that adhered more closely to the questionnaire design. The health insurance unit membership was stored as an ordered pair in the database: identity of policyholder, identity of insured dependent.

WIC Program Data. One series of questions is used to establish whether women with children participate in the Women, Infants, and Children (WIC) program. For analytical purposes it is important to establish who participates in the program; not all members of the household will be included. This required a relation to describe membership in a WIC unit. The relation is undefined for most sample persons because few participate in the program.

5. THE SEMANTIC PRINCIPLE: TIME

Three aspects of time affect the interpretation of data. They are *reference*, *collection*, and *version* time.[5] Failure to tag each value with these times can lead to confusion and to erroneous use of data.

The *version* time is the date when the values were last altered. It locates a shared data set relative to a process of error correction that may continue after the data set was issued. Later versions may be deemed less biased or more accurate than earlier versions. The user will need to identify the version of the data to determine whether later revisions affect results.

The *collection* time may be critical, as in election polls or economic surveys where an event can modify attitudes, income, and wealth. But collection time is not necessarily the appropriate time reference for the attribute being measured. Respondents may be asked to recall events, attitudes, and behavior from an earlier time or they may be asked to give expectations about the future. Analysts need *reference* time to design estimates for their models of the data.

In SIPP reference time and the periodicity of measurement produce three modes of time series. The reference time for some data in the "core" of replicated measures is the *time of interview.* Those measures are repeated eight or nine times. Those data are dated by the enumerator's record of the date of interview. Questions that are repeated in successive interviews will be roughly four months apart, as is dictated by the survey design.[6]

Some questions in the "core" are *retrospective to each of four months prior to the month of interview.* Those measures produce a time series of 32 or 36 months for the 1984 SIPP panel. Third, measures from the topical modules are recorded *only once* (or *twice at yearly intervals*). Reference time for those measures varies widely.

5.1. Retrospective Data within the Reference Period

Three aspects of reference time within the SIPP are extraordinarily confusing. (i) A 12-month aggregate over time does not correspond to a calendar year for three of the four subsamples in the panel. (ii) Each subsample refers to months that overlap but are not identical. Data on a particular calendar month may be collected on different instruments for different subsamples. (iii) Some months appear as reference months and as interview months on a prior interview.

Interviewing for the fourth subsample began in January 1984 and continued at four-month intervals until May 1986.[7] A calendar year can be assembled from the second, third, and fourth interview contacts. Compare this situation to the first subsample, where calendar 1984 must be assembled from the second through the fifth interviews. The reference months of the first subsample follow those of the fourth subsample by one month after the first interview ("staggered interviewing").

The SIPP panel collects 38 "survey months" of data. The first four months are reported retrospectively; the last month contains only measures taken at the time of interview. Retrospective data for the interview month are collected from a subsequent interview.

An immediate confusion is that the interview month for a prior interview reappears as reference month one in the subsequent interview. A second problem is that the second and eighth questionnaire forms were not administered to all of the subsamples. As a result, different instruments were used to measure behavior at the nth contact, where $n > 1$.

Two relations were created to understand survey ("relative") and calendar time and their relationships to the four subsamples and the questionnaire forms. Information from these relations could then be used by the analyst to recode the index of the survey month to calendar dates and to draw samples of data for calendar months.

5.2. The Semantic Principle for Time Series

The time series generated by SIPP measurements can be represented in three ways:

(a)

$$\|i, t;\ x_{it}\|$$

where $i = 1, \ldots, I$ indexes entities and $t = 1, \ldots, 36$ (or $t = t', \ldots, t''$) indexes survey time. Thus t' is the earliest time at which i is sample relevant and t'' is the last time at which i is sample relevant.

(b)

$$\|i, T; x_{1,1+T}, x_{i,2+T}, x_{i,3+T}, x_{i,4+T}\|$$

where $i = 1, \ldots, I$; $T = 0, \ldots, 8$.

(c)

$$\|i, x_{i1,xi2}, \ldots, x_{i36}\|$$

where $i = 1, \ldots, I$.

Mode (a) corresponds to the third normal form: attributes refer to a particular point in the cartesian product space on entities and time. Mode (b) is the form in which data are collected on each questionnaire. Mode (c) displays a vector of attributes over the entire period for which an entity is at risk in the SIPP panel.

Mode (a) requires no placeholders for missing observations. Mode (c) requires the most placeholders. In both modes (b) and (c) missing components of the vector due to noninterviews and non-sample, or out-of-scope entities, must be distinguished because their conceptual relevance to analysis differs. Mode (a) requires more computational resources to identify persons than either (b) or (c).

The semantic principle for organizing time sequences of measurement must deal with a tension between principles one and two—designing the data base schema to conform to the channel of measurement and providing dynamic independence—and the semantic principles employed in analysis. Mode (b) most closely conforms to the questionnaire image and was used in the SIPP database. To accommodate analysis needs, however, we also created longitudinal summaries (c), which are described below.[8]

5.3. Longitudinal Tables

Our solution to this conflict was to extract certain information pertaining to changes of status from the history of an individual and to create summaries that reported the duration of particular status by individual. These "event histories" serve a double function: finding underlying data related to the event and providing information necessary for hazard rate modelling of change in status.

For example, reports of income receipt from 39 income support programs were consolidated into a matrix that identifies the type of income and the beginning and ending dates for each episode of receipt. This matrix becomes an easy focus for studies such as those that deal with AFDC episodes (Flory, Martini, & Robbin).

Missing observations in the panel create special problems for longitudinal analysis. For example, the missing observation in a recipiency spell creates ambiguity about the spell length. The number of left and right censored spells increases. Analysts have successfully used the longitudinal tables to impute upper and lower bounds to the distribution of spell duration (Fitzgerald).

5.4. Idiosyncratic Problems Related to Time: Conditioning

A complex battery of screening questions records the income received in the initial interview and in the interview following a noninterview. Subsequent interviews condition question-asking about income recorded in the previous interview and about new sources of income. As a consequence, the screening questions are asked only once for persons giving all nine interviews; and asked more than once for persons with "gaps" created by noninterview.

We recognized this interviewing procedure by concatenating the screening questions in a single array in the database. The array then clarifies the manner in which data were elicited, and highlights the fact that response error will differ between the first and subsequent interviews.

6. SUPPORT TO ENHANCE UNDERSTANDING OF COMPLEX DATA

Access has several aspects that reduce the complexity of a data set. We have discussed the design of a database schema to assure that retrieval is direct, inexpensive, and transparent. This is the most important aspect of access and relates closely to use of the semantic principle.

A second aspect of access is a testing capability. We discussed the utility of a a sample database in Section 2.3. Users need to test their understanding of the data by quickly retrieving the results of *ad hoc* queries. Database designers use a sample database to develop a working model of their innovations during the database development stage.

A third aspect of access is assistance in locating germane information about the data and the survey design, the metadata of the database. Logically, this aspect is ancillary to retrieving data and testing queries, but it is essential for successful end-user interaction with a database. Taxonomies of the attributes provide an explicit description of data relationships and data structures, which facilitates the user's orientation to the system. We devote the remainder of this section to discussing the role of language for understanding design concepts embedded in the data. (See David 1985b; Robbin & David; and David, Robbin, & Flory 1989a for a more extensive discussion of the language of data and the role of metadata.)

6.1 Communicating Meaning: Information about Data

Labelling attributes may appear to be a trivial topic for reducing complexity; it is not. Parsing characters in a label is, however, an extremely efficient way to organize information and clarify meaning. In the database parsing is used to distinguish similar

information content at several levels of aggregation, to recognize the reference period of the data, and to identify attributes with questionnaire items. Parsing also identifies replicated measurements.

Channel (Instrument) Design. Labelling attributes in the SIPP database was dictated by the questionnaire design. Labels for each question are printed on the questionnaire, and should be used for the computer image. Similarity between response labels and attribute names facilitates discovery of measures in the database.[9]

Collating the nine instruments of the SIPP panel revealed that the same label was sometimes applied to different questions in the topical modules. This caused confusion and led us to prefix topical module attribute labels with the questionnaire number on which they appear. Sometimes, different labels were applied to questions with the same meaning. To assist in identifying these repeated measures a special relation was devised to provide a concordance between attribute labels.

Classifying Entities. Another application of the classification principles reduced the complexity of families of variables that related to different entities. Earnings, for example, could appear as information for jobs, individuals, families, and households, as well as reported for four different months in the reference period. The domain for the sixteen related entity-attribute pairs is identified by a single root, and labels for each pair are generated from the root by appropriate prefixes and suffixes. This principle will be extended when 1985 SIPP data are added to the database, so that identical questions generate attributes with identical labels across the two panels.

6.2 Communicating Meaning: Information about Design and Meanings

The database should provide assists to identify the logical conditioning of the measurement instrument and to understand the database schema.

Skip Patterns. Differences in the questionnaire can be identified when identical response categories for identical questions carry the same label. Identical labeling provides no assurance, however, that the same sample of entities is included. For example, questions on training in the third interview are specific to particular groups of workers, while the same questions are administered to a broader population in the eighth interview. To clarify this problem and to describe the logical universe for each measure, we constructed a relation that displays conditioning or "skip patterns" that affect each attribute in the repeated core. [10]

The concept is simple. For any attribute on a particular interview, the relation displays the conditions that **exclude** an individual from measurement. Real measures exist for the complement to excluded individuals. When the database schema conforms to the complement, the NIU disappears from the data image; otherwise, users of the data will need to select (or restrict) cases to the relevant complement.

The Database Schema. A second assist to the user is a relation that describes the schema for the database. The principle applied to create each relation in the database

is described in this relation. Details are provided in a longer explanation in documentation that resides inside the database. This relation also identifies the logical and physical structure of each of the database tables.

7. SUMMARY

This paper addressed three questions. How do we reveal the structure of complex data; how do we increase accessibility to data; and how do we maintain the integrity of data structures? Four principles and a strategy were described. Applied to the *Survey of Income and Program Participation* (SIPP), the principles have been extraordinarily successful in reducing cognitive complexity and the costs of data management and analysis.

The structure of the survey design has been revealed by incorporating the semantic principle of the survey design into the database schema. Time is a special problem in applying the semantic principle because the analytical schema of statisticians differs from the survey designer's. This tension cannot be avoided, and we chose to preserve an image of the questionnaire and to maintain flexibility for analysts.

Accessibility has been significantly increased by using metadata to locate data required for answering a question. Accessibility has been improved by creating relations that summarize significant aspects of the dynamics of the data. Longitudinal tables provide analysts with the capability for both spell analysis and causal modeling. Accessibility was also enhanced by providing a two percent subsample of the database for exploratory data analysis on a microcomputer.

Integrity of the database was assured by maintaining an audit trail of operations on the database in the query language of the RDBMS. Integrity was increased by using consistent labeling and relations that identify repeated measures. Lastly, integrity was increased by utilizing the capability of a relational database management system (RDBMS) to identify errors in the data and to uncover logical and syntactical errors in the queries.

* SIPP ACCESS has received financial support from the National Science Foundation under grants #SES-8411785, #SES-8716448, and #SES-8701911 and from the Sloan Foundation under grants #B1986-25 and #B1987-46. Support by the University of Wisconsin Institute for Research on Poverty and the Center for Demography and Ecology is gratefully acknowledged. We want to acknowledge the major contributions made by Thomas S. Flory and Alberto Martini to developing the database and the tools that have greatly enhanced access to the SIPP.

NOTES

[1] The "integrated information system" concept was used to create an infrastructure that linked the technologies of laser disk (WORM), mainframe and microcomputers, electronic networks, and a relational database management system (RDBMS). The database is accessed by telephone or through remote login from another computer installation. Extracts can be downloaded through BITNET to a researcher's home institution.

[2] In some cases data were transformed into illogical formats which attributed data to an inappropriate entity.

[3] In fact, a failure to recognize this feature of the data led Citro, Hernandez, and Moorman (1986) to overstate the stability of longitudinal households.

[4] The redundant attributes can be accessed from an earlier version of the database for methodological work on the errors associated with the data collection process.

[5] Snodgrass (1987) proposes a slightly different terminology to use with temporal databases. The collection time is "valid time". Updates leading to new versions require "transaction time". Retrospection and time frames in the data collection protocol are "user-defined time".

[6] Different questionnaire forms may be used to elicit these questions on different subsamples. That complexity need not concern us at the moment because the four-month interval is maintained for repeated core questions across questionnaires.

[7] As we explained in Section 4.1, retrospective data for the first interview do not yield correct household aggregates for the four months of the reference period. Thus unbiased household information can only be retrieved from the reference months of the second interview.

[8] Mode (c) was used by the Bureau of the Census to prepare their *Longitudinal Research File* for the 1984 SIPP panel. It occupies five reels of high density magnetic tape. The size of this format greatly increases data retrieval costs. The semantic principle of attributing all measures to individuals misidentifies relationships and adds to the complexity of understanding the scientific design.

[9] Many users come to the data set without prior study of the design and no knowledge of the questionnaire. For them, browsing a classification of measures through a controlled vocabulary is extremely helpful. This facility was created for the repeated "core" attributes in the SIPP database.

[10] Two relations could be defined to represent correctly the different samples, and thus to conform to the third normal form rule. However, the propagation of tables according to the rule increases the burden of end-user information retrieval. A balance must be sought between strict adherence to the third normal form and the information processing capability of the analyst.

REFERENCES

Boar, B.H. 1984. *Application Prototyping.* New York: John Wiley & Sons.

Boruch, R.F. 1985. Definitions, products, distinctions in data sharing. In S.E. Fienberg, M.E. Martin, & M.L. Straf, Eds., *Sharing Research Data.* Washington, D.C.: National Academy Press.

Boruch, R.F., & Cecil, J.S. 1979. *Assuring the Confidentiality of Social Research Data.* Philadelphia: University of Pennsylvania Press.

Boruch, R.F. & Cordray, D.S. 1985. Professional codes and guidelines in data sharing. In S.E. Fienberg, M.E. Martin, & M.L. Straf, Eds., *Sharing Research Data.* Washington, D.C.: National Academy Press.

Citro, C.F., Hernandez, D.J., & Moorman, J. 1986. Longitudinal household concepts in SIPP. *Proceedings of the Social Statistics Section, American Statistical Association.* Washington, D.C.: American Statistical Association, 611-616.

Codd, E.F. 1985. Is your DBMS really relational? (Part 1) Does your DBMS run by the rules (Part 2). *Computerworld,* October 14; October 21.

David, M.H. 1985a. Designing a data center for SIPP: An observatory for the social sciences. *Proceedings of the Survey Research Section, American Statistical Association.* Washington, D.C.: American Statistical Association.

David, M.H. 1985b. *The language of panel data and lacunae in communication about panel data.* (CDE Working Paper 85-20). Madison: University of Wisconsin Center for Demography and Ecology.

David, M.H. 1989. Managing panel data for scientific analysis: The role of relational database management systems. In Greg Duncan and Daniel Kasprzyk, Eds., *The American Statistical Association International Symposium on Panel Surveys.* New York: John Wiley & Sons (forthcoming).

David, M.H., Robbin, A., & Flory, T.S. 1988a. Access to data: Handling the 1984 SIPP. *Proceedings of the Statistical Computing Section, American Statistical Association.* Washington, D.C.: American Statistical Association.

David, M. H., Robbin, A., & Flory, T.S. 1988b. *Analyzing Complex Data: A DBMS for the 1984 SIPP.* Madison, WI: Institute for Research on Poverty.

Dolby, J.L., Clark, N., & Rogers, W.H. 1987. The language of data: A general theory of data. *Proceedings of the 18th Symposium on the Interface of the American Statistical Association.* Washington, D.C.: American Statistical Association, 96-103.

Fischhoff, B., MacGregor, D., & Blackshaw, L. 1987. Creating categories for databases. *International Journal of Man-Machine Studies* **27**: 33-63.

Fitzgerald, J. *The effects of the marriage market and AFDC program parameters on recipient duration on AFDC.* Paper presented at the Social Science Research Council Conference on Individuals and Families in Transition: Understanding Change Through Longitudinal Data, Annapolis, March 16-18, 1988.

Flory, T.S., Martini, A., & Robbin, A. 1989. Attrition and spell censoring in estimating dynamic models of welfare recipiency. *Proceedings of the Social Statistics Section, American Statistical Association.* Washington, D.C.: American Statistical Association.

Hedrick, T.E. 1985. Justifications for and obstacles to data sharing. In S.E. Fienberg, M.E. Martin, & M.L. Straf, Eds., *Sharing Research Data.* Washington, D.C.: National Academy Press.

Kent, W. 1983. A simple guide to five normal forms in relational database theory. *Communications of the ACM* **26**(2):120-125.

Mark, L., & Roussopoulos, N. December 1986. Metadata management. *Computer*: 26-36.

Martini, A. 1988. *A Theoretical and Empirical Model of Labor Supply that Accounts for the Unemployed and the Discouraged Workers.* Ph.D. dissertation. Madison: University of Wisconsin.

Robbin, A., & David, M.H. 1988. Information tools to improve access to national longitudinal panel surveys. *Research Quarterly* **27**:499-515.

Snodgrass, R. 1987. The temporal query language TQuel. *ACM Transactions on Database Systems* **12**:247-298.

Figure 1. Design and channel -- Impact on data image and interpretation

	DESIGN		CHANNEL (EXECUTION)	
	Sample	Stimulus	Response	Data image
ASPECTS:				
a.	Selection Probabilities	Reference period, Question	Categories	Truncation
b.	Respondent (Proxies)	Referent	Nonresponse	NA flags, Imputation
c.	Conditioning (Skip Sequence)	Order		
INTERPRETATION				
a.	Representativeness		Response error	Inverse of transformations
b.			Proxy bias	

To appear:
21st Symposium on the Interface
Alexandria, VA: American Statistical Association

DATABASE DESIGN FOR LARGE-SCALE, COMPLEX DATA

Martin H. David and Alice Robbin, University of Wisconsin
Martin H. David, Dept. of Economics, 1180 Observatory Dr., Madison, WI 53706

KEY WORDS: database schema, data structures, Survey of Income and Program Participation (SIPP)

1. INTRODUCTION

High-dimensional data structures are the focus of this session.* We use the adjective *complex* in place of high-dimensionality because the problems that we describe arise both from the measurement of thousands of attributes and from the intricate logical conditioning of the measurement process. Our paper provides answers to three questions associated with these data structures.

A significant structure always underlies data collected for scientific analysis. The question is, *How do we reveal that structure to support statistical analysis?* Time is an implicit dimension of a data structure. The design of a data collection is not always identical over time. Some of our discussion is devoted to how *time* is represented when measurements are asymmetric to different time points.

Complexity connotes both technical and cognitive problems for retrieving data. Technical problems can be addressed in part by applying relational theory to simplify and clarify data structures. Faulty memory, the limited capacity to process information, and uncertainty about outcomes can be partially overcome by applying principles derived from cognitive theory to organize data for retrieval and to represent meaning. These observations are the source of our second question, *How do we make data accessible?*

Models applied to the data entail units of analysis and concepts that were not envisioned by the original data collectors. Our third question derives from a recognized tension between data collection and subsequent use. *How do we maintain the integrity of the measurements while organizing data for a variety of analysis tasks, including extension of the original data by repeated measures, auxiliary variables, and replication?*

We sought answers to these and other questions in a project that developed a prototype of an integrated information system to improve access at low-cost to large-scale, complex data by university, government, and private sector policy analysts throughout the United States (David 1985a; David, Robbin, & Flory 1988a; Robbin & David) [1]. Our paper discusses aspects of the conceptual framework and strategies we applied to design a database that integrates statistical data and metadata (information about the data), including the database design and contents, survey design, collection and processing procedures.

Section 2 presents a conceptual framework for linking survey design to appropriate data structures. The remainder of the paper demonstrates how principles described in Section 2 are applied to the 1984 panel of the *Survey of Income and Program Participation* (SIPP). The generality of the principles and their success in handling the difficult SIPP design lead us to conclude that this framework can be generally applied to social science databases.

Section 3 describes the complexity of the SIPP design and measurement. Subsequent sections elaborate on the principles applied to the SIPP database design. Section 4 discusses the value of semantic principles for organizing data. Section 5 explains how time is represented in the SIPP and why its representation in the database may conflict with the use of the third normal form as an organizing principle. Section 6 emphasizes the role of metadata in clarifying underlying data structures and improving access to complex data.

2. DESIGNING A DATABASE FOR COMPLEX DATA SETS

This section discusses four principles that guided our approach to designing a database schema for large-scale, complex statistical data.

2.1. A Conceptual Framework

Figure 1 represents two facets of data. Data are generated by a scientific design; inference from data requires interpretation consistent with the design and the procedures used to execute the design. The *design (sample-experiment-instrument)* determines what inferences can be made about populations and treatments. The *channel of measurement* executes the design. It is determined by instruments (questionnaires, etc.) that are used for measurement and by procedures that govern their administration.

The end product of design and procedures is an *image* that contains all the data values. The information in this image is a function of design and channel in precisely the same sense that informa-

tion in a satellite map is a function of complex signals sent from several instruments and interpreted through computer algorithms. The image will vary according to the procedures that capture responses on machine-readable media.

We caricature the flow of information through the design and channel by the column subheadings in Figure 1: sample, stimulus, response, and data image. They convey that the purpose of measurement is to elicit information about designated entities and to represent an image of the responses in a data structure amenable to statistical analysis.

A coherent semantic principle underlies the production of the data image. Information is elicited from a well-defined sample according to known interrogative procedures (questions, observations, or real-time application of auxiliary instruments). The responses are captured and transformed by computer algorithms with single-valued outcomes. In the process messages received from the respondents may be coded, censored, or combined with other information. The messages may be attributed to other entities (proxy reports about others) or aggregated to give measures related to groups (income for families). Many null values are inserted in the data image as defaults for situations where data were not collected on all sample elements; these are not-in-universe codes (NIU).

Interpretation, i.e., analysis, of data requires inverting the data image in two distinct senses. Samples are generalized to populations and superpopulations. The meanings of values in the data image must be translated to natural language in order to communicate information to the scientific (or policy-making) community who use the information. Aspects of this interpretive process are shown in the lower part of Figure 1.

2.2. Implications for Database Design

The concepts underlying Figure 1 have guided our development of a database to improve access to SIPP. They can be expressed as four principles that generalize to other complex data structures. They have been tested and proven in four years of developing the SIPP ACCESS facility.

Design the database schema to conform to the channel of measurement. This implies that database should incorporate the questionnaire image and all responses. It also implies that the primary consideration in database design is to preserve the semantic principle that generated the data. That is, the data are generated by a question directed at a particular population. Responses to the question reflect the object to which the question refers (referent), the bounding reference period, and the attribute that is being elicited.

Provide dynamic independence. The database must be capable of receiving additional attributes and additional observations, to permit pooling of data sources, analysis in relation to contextual variables, and reorganization of the data to a wide variety of units of analysis (Codd).

Preserve information and maintain the capability of transformation and its inverse. Any manipulation of the questionnaire image to facilitate access and retrieval must be reversible. That is, it must be possible to recover the original image. This rule assures that erroneous processing can be undone and that detail is not lost by transformation. For example, converting birth dates to century notation is information-preserving; scaling birth dates to age classes in a particular year is not.

Maintain a journal or audit trail. This document completely describes transformation of the response to the questionnaire and to the image in the database. The journal records the rules that governed decisionmaking, and thus serves as a tool for evaluating data quality as well as the assumptions that govern design and development. Journaling also creates a "template" for conducting similar activities in the future and therefore has the potential for introducing efficiencies into the transformation process. Finally, maintaining an audit trail is consistent with the scientific norms of replication and data sharing (Boruch; Boruch & Cecil; Boruch & Cordray; Hedrick).

2.3. Strategies for Database Design

We emphasize three aspects of database design: exploratory learning and prototyping, analysis - oriented data enhancements, and extended metadata.

Exploratory Learning and Prototyping. Research problems are rarely well-formulated in advance of executing query and search procedures. Nearly all research requires a period of exploratory analysis during which the scientist learns about the data collection. Orders of magnitude for the size of populations of interest, rough indications of the distribution of outcome measures, and tests of the consistency of measurements against simple logical principles reveal the feasibility of a research plan and suggest modifications. In this exploratory phase frequent interaction with the data is typical, and discoveries of undocumented data processing are common. Researchers discover errors in their understanding of variable labels and execution of simple transformations of the data.

Low cost, rapid interaction with the data for experimental analysis is achieved by storing a rep-

resentative sample of the data in a schema identical to that of the entire database (David, Robbin, & Flory 1988b). Rigid adherence to identical data structures for the database and the sample assures that any "well-formed" query which produced information from the sample can be executed on the entire database without programming modifications.

What is true for the researcher is equally true for the database designer. Application prototyping assures greater implementation success through gradual learning and incremental change (Boar). We created a sample database of the complete panel data. This sample database was used to develop a compact, "restructured" database from the public use data tapes issued by the U.S. Bureau of the Census. Only when database design and testing had been completed on the sample was the complete panel reorganized to achieve efficiency through compaction.

Analysis-oriented Data Enhancements. The database designer can achieve great economies for the future analyst with "tools" that facilitate analysis. In the SIPP database these tools are summary tables that collate data from up to nine interviews for all members of the panel. These "longitudinal tables" obviate a need to retrieve data from nine different interviews, and they anticipate the structure of analysts' queries by reporting, for example, on "spells" of a state variable, dynamic relationships (family composition, and sample relevance of persons or labor force participation (Flory, Robbin, & David 1988a; Flory, Martini, & Robbin; Martini). We describe these tools in more detail in Section 5.

Extended Metadata. A third and, in our view, indispensible, strategy for improving an understanding of complex data is the creation of extensive metadata within the database (Robbin & David). Successful end-use interaction—better decisions about selecting data, correcting and locating errors, and constructing alternative access paths into the database—is the measure of successful data base design. Database design requires that careful attention be given to the essential role that natural language plays in describing statistical (or any other) data and to the categories devised to organize and communicate the contents of a database (cf. Dolby, Clark, & Rogers; Fischhoff, MacGregor, & Blackshaw). A more extensive discussion of the role of metadata is found in Section 6.

3. COMPLEXITIES IN SIPP

In this section we describe principal features of the SIPP design and the resulting complexity of the data.

3.1. The Design

The 1984 SIPP panel conducted by the U.S. Bureau of the Census collects data on income of individuals and household demography through an extraordinarily complex design and channel. Some of the attributes of this complexity are described below.

- The initial area probability sample of 20,000 households is extended in time by following the adults in the household.
- The population varies after the first interview as new people are born into the sample through birth, marriage, or remarriage, and as original or new sample persons exit through death or divorce or cannot be located after they move.
- Data cluster in several natural aggregations. Information obtained from the interview may pertain to individuals (with differences between children and adults), married couples, families, households, health insurance units, employee-employer pairs, and client-program pairs (where programs are 39 different entitlements for benefits from local, state, or federal income maintenance, social insurance, medical, and welfare programs).
- Data are collected during eight or nine interviews. Respondents provide information about the four-month period preceding the interview, and a limited amount of demographic information on all members of the household is obtained by the enumerator at the time of interview. A "core" set of questions is repeated in each of nine interviews; and "topical module" questions on high-interest public policy issues supplement the core after the second interview.
- Complex conditions determine eligibility for measurement. Conditioning may depend on responses to the prior interview. For example, the initial interview screens for more than 50 different types of income. Subsequent reinterviews are conditioned on earlier reports.

These complexities in the questionnaire design are compounded by complexities in the administration of the survey. Four independent subsamples were interviewed in consecutive months ("staggered interviewing"). As a consequence, members of the area probability sample are visited at intervals of exactly four months (for up to 36 months). However, the nine different questionnaire instruments in the 1984 SIPP were not administered to the entire sample, with consequent asymmetries in the data for persons selected into each subsample.

In addition, eighteen percent of the sample was eliminated after the fourth contact, and half the sample was interviewed only eight times to reduce cost.

Following collection of data in the field, an image of the questionnaire is created by optical scanning. The image is transformed during data processing by algorithms that recode variables, impute missing items, statistically match missing interviews, and aggregate person-level information to create family and household characteristics.

The survey design facilitates a variety of estimands. Cross-sectional point estimates (e.g., for individual earnings) can be constructed for any one of the nine reference periods corresponding to the nine interviews. Furthermore, these estimates can be constructed from data collected with varying recall periods so that response error can be estimated. Data from three interviews can be aggregated to estimate annual income data; data from six interviews can be used to estimate year-to-year change. And data from the entire panel can be studied for individuals to estimate the probability of entering or leaving a status, such as unemployment. Each of these different estimators requires a different selection of sample entities. Each requires an understanding of the calendar that corresponds to the reference period in each of four independently drawn subsamples.

3.2. The Public Use Files

The public use data files are released by the U.S. Bureau of the Census for every interview. The 1984 SIPP panel includes nearly 7 million observations and about 20,000 data elements. These data are distributed on more than thirty tapes, totaling about 2.2 gigabytes of data in nine separate files representing the nine interviews. Each file is physically organized as eight record types.[2] The data must therefore be restructured and linked for studying dynamic changes in the population. Subsequent panels increase in size by about 300 megabytes of data and 2,500 data elements every four months. The 1985, 1986, 1987, and 1988 panels (each 32 months in duration with a reduced sample size) will introduce more than 8 gigabytes of data and more than 70,000 data elements into the system.

A machine-readable codebook and print questionnaires accompany the public use files. Complexity in these materials inevitably leads to errors—omissions, variations in labelling, and peculiarities in coding—that must be resolved before creating a database.

Labelling conventions used by the Bureau facilitated database development. Information taken directly from the response boxes on the questionnaire is labelled with the number of that box. Replicated questions carry the same label over the nine questionnaires. Measures related in some way (by aggregation or by time sequencing) are often labelled with a root and prefixes or suffixes that parse to identify the relationship. For example, person earnings in reference month one are labelled *ppern1*; household aggregate earnings in reference month four are labelled *hhern4*. The role of labelling is elaborated in Section 6.2.

3.3. The SIPP ACCESS Database

The 1984 SIPP panel data reside in a relational database management system (RDBMS). The data base schema, devised by careful use of third normal form relations, reduced the scope of the original data set by 75%, while improving its logical clarity and accessibility. (For a discussion of the utility of the RDBMS for panel data, see David 1989). The logic of the RDBMS reinforces a work strategy that is iterative and exploratory, and provides an excellent tool for answering research problems that are rarely well-formulated in advance of executing query and search procedures. Direct access provided by the RDBMS improves a user's knowledge of statistical data. The logic behind the structure of the RDBMS reduces errors made by analysts and researchers.

4. THE SEMANTIC PRINCIPLE: REDUCING COMPLEXITY

A key to managing the complexity of the SIPP data is the application of a semantic principle to organize the representation or image of the data. The channel design ensures that a clear logical principle is used to select entities for observation. The design also makes certain that a group of questions pertain to that universe. Representing data in the computer storage in a conformable way increases the clarity of the information for analysis.

Attention to semantic principles requires that data be stored in relations (tables) that adhere to the third normal form (Kent). Every attribute in the relation is the property of the entity that is uniquely identified by the "key" or identifier of that case or record. Over eighty relations are required to establish the principal samples of entities measured by the design. Some are obvious and were embedded in the public use data that we received. For example, it is clear that information on jobs (employer-employee pairs) existed for a subset of adults, and that some individuals could have two jobs. In other cases, however, expediency in

preparing the public use data led to illogical and confusing treatments of the principal samples.

The semantic principle ensures that users understand the relations and their content. We provide several examples below. First, we draw attention to the ways in which NIU bytes appear in an array.

4.1. Eliminating Not-In-Universe (NIU) from Sparse Matrices

Non-informative bytes arise when default values are inserted in the array to act as place-holders for cases that are not measured on a particular attribute. Typically, a subset of measurements will be made for a subset of cases. Default values are eliminated by organizing the storage according to the design which generates the measurements.

Logically, default values may appear in all attributes in a given row, in all cases of a given attribute, or in the intersection of a subset of cases for a subset of attributes. We were able to achieve economies of access by removing all three of these types of NIU. Below we give three examples of adherence to semantic principles and third normal form for organizing a SIPP database.

Deleting Rows: Children. Children under fifteen years of age are not interviewed in the SIPP. Their presence in the household and family is used to generate appropriate weights for the analysis of individuals, families, and households at a particular point of time. Nonetheless, default values commingle observations for interviewed persons with those of the children. Approximately one-quarter of all interview data entered on records for persons are therefore NIU.

Our first step in reorganizing data was to partition the person array so that weights and related demographic information on all persons were separated from substantive information obtained for persons over 14 years of age. This partition allowed us to delete one-quarter of the matrix of information supplied in the public use files for persons. This strategy reinforces the principle that responses to questionnaire items are restricted to a universe that excluded persons under 15 years of age.

Deleting Columns: Retrospective Demography. Household composition is established at the first interview, and for every succeeding month until the last interview. Data processing by the Bureau of the Census transfers household composition at the time of the first interview into attributes describing the four months prior to the first interview.[3] These values carry no information because all the measurements are generated at the time of interview (and not before). Our data reorganization removed attributes from the original data set which contained no information and were misleading.

Retaining Intersections: Reorganizing Information on Program Income. Data processing by the Bureau of the Census forces all information about program income receipt and amounts into a single array, although some questions are specific to particular types of programs. That is, different questions are appropriate for AFDC, Social Security, and Food Stamp receipt and income. We reorganized the data into separate arrays for each program type. Each array contained only the attributes appropriate to it, thereby deleting default values for persons who reported no program incomes. This is an example of deleting NIU defaults for selected cases on selected attributes.

4.2. Eliminating Redundancies

Two types of redundant information were deleted from the database.

Process Control: Interviewer Check Items and the Income Roster. The interview process requires enumerators to enter attributes measured at prior interviews (transcription items) and to recall attributes measured earlier in the current interview (check items). However, these attributes contain no new information because they repeat attributes measured earlier. Although they are important to understanding the data collection process, they are not necessary for secondary analysis. Check items and transcription items were therefore not incorporated into the reorganized data schema. This decision eliminated many redundancies in the data schema.[4]

Unchanging Attributes. A variation on redundacy occurs when unchanging attributes, such as birth dates and gender, are repeated in arrays for successive interviews. We organized all personal constants into a single array for two reasons. This eliminated repetitive entry of the same information at several points in the data set. A scan of all the values for these personal constants enabled us to eliminate response errors (i.e., we selected the modal, or edited, value when more than one value appeared for the constant).

4.3. Transformations: New Logical Universes

Another application of the semantic principle illustrates the conceptual value of retaining data in a form that corresponds to meanings in the data collection.

Joint Asset Incomes. Data on savings accounts and several other types of asset-related in-

come are obtained in separate sequences for assets jointly held by couples and for the respondent as the sole owner. Jointly-held accounts are reported by the first interviewed member of the couple. In the public use data file an NIU is inserted in the second member's record if the first member has already been interviewed. The record of the reporting member of the couple is not necessarily the first in the sequence of data. This is confusing, because two records must be retrieved in order to discover the jointly-held amount. We reduced the storage of information on assets held by couples by fifty percent by defining the couple as the logical unit.

The semantic principle applied here assures that the design of the questionnaire is clearly represented in the data image. The task of deciding how individuals benefit from jointly-held assets can then be addressed by the analyst without being obscured by the form in which data are retrieved.

Insurance Units. A second example of how the questionnaire design can be more clearly presented in the data schema pertains to membership in a health insurance group. When some, but not all, of the members of a household are covered by medical insurance, SIPP created a membership roster. This roster applys only to those households if insurance coverage is incomplete. However, the public use files allocate space in every person's data record, even when coverage of all persons in the household is complete. We therefore created a relation that adhered more closely to the questionnaire design. The health insurance unit membership was stored as an ordered pair in the database: identity of policyholder, identity of insured dependent.

WIC Program Data. One series of questions is used to establish whether women with children participate in the Women, Infants, and Children (WIC) program. For analytical purposes it is important to establish who participates in the program; not all members of the household will be included. This required a relation to describe membership in a WIC unit. The relation is undefined for most sample persons because few participate in the program.

5. THE SEMANTIC PRINCIPLE: TIME

Three aspects of time affect the interpretation of data. They are *reference*, *collection*, and *version* time.[5] Failure to tag each value with these times can lead to confusion and to erroneous use of data.

The *version* time is the date when the values were last altered. It locates a shared data set relative to a process of error correction that may continue after the data set was issued. Later versions may be deemed less biased or more accurate than earlier versions. The user will need to identify the version of the data to determine whether later revisions affect results.

The *collection* time may be critical, as in election polls or economic surveys where an event can modify attitudes, income, and wealth. But collection time is not necessarily the appropriate time reference for the attribute being measured. Respondents may be asked to recall events, attitudes, and behavior from an earlier time or they may be asked to give expectations about the future. Analysts need *reference* time to design estimates for their models of the data.

In SIPP reference time and the periodicity of measurement produce three modes of time series. The reference time for some data in the "core" of replicated measures is the *time of interview.* Those measures are repeated eight or nine times. Those data are dated by the enumerator's record of the date of interview. Questions that are repeated in successive interviews will be roughly four months apart, as is dictated by the survey design.[6]

Some questions in the "core" are *retrospective to each of four months prior to the month of interview.* Those measures produce a time series of 32 or 36 months for the 1984 SIPP panel. Third, measures from the topical modules are recorded *only once* (or *twice at yearly intervals*). Reference time for those measures varies widely.

5.1. Retrospective Data within the Reference Period

Three aspects of reference time within the SIPP are extraordinarily confusing. (i) A 12-month aggregate over time does not correspond to a calendar year for three of the four subsamples in the panel. (ii) Each subsample refers to months that overlap but are not identical. Data on a particular calendar month may be collected on different instruments for different subsamples. (iii) Some months appear as reference months and as interview months on a prior interview.

Interviewing for the fourth subsample began in January 1984 and continued at four-month intervals until May 1986.[7] A calendar year can be assembled from the second, third, and fourth interview contacts. Compare this situation to the first subsample, where calendar 1984 must be assembled from the second through the fifth interviews. The reference months of the first subsample follow those of the fourth subsample by one month after the first interview ("staggered interviewing").

The SIPP panel collects 38 "survey months" of data. The first four months are reported retrospec-

tively; the last month contains only measures taken at the time of interview. Retrospective data for the interview month are collected from a subsequent interview.

An immediate confusion is that the interview month for a prior interview reappears as reference month one in the subsequent interview. A second problem is that the second and eighth questionnaire forms were not administered to all of the subsamples. As a result, different instruments were used to measure behavior at the nth contact, where $n > 1$.

Two relations were created to understand survey ("relative") and calendar time and their relationships to the four subsamples and the questionnaire forms. Information from these relations could then be used by the analyst to recode the index of the survey month to calendar dates and to draw samples of data for calendar months.

5.2. The Semantic Principle for Time Series

The time series generated by SIPP measurements can be represented in three ways:

(a)

$$\|i, t;\ x_{it}\|$$

where $i = 1, \ldots, I$ indexes entities and $t = 1, \ldots, 36$ (or $t = t', \ldots, t''$) indexes survey time. Thus t' is the earliest time at which i is sample relevant and t'' is the last time at which i is sample relevant.

(b)

$$\|i, T; x_{1,1+T}, x_{i,2+T}, x_{i,3+T}, x_{i,4+T}\|$$

where $i = 1, \ldots, I; T = 0, \ldots, 8$.

(c)

$$\|i, x_{i1,xi2}, \ldots, x_{i36}\|$$

where $i = 1, \ldots, I$.

Mode (a) corresponds to the third normal form: attributes refer to a particular point in the cartesian product space on entities and time. Mode (b) is the form in which data are collected on each questionnaire. Mode (c) displays a vector of attributes over the entire period for which an entity is at risk in the SIPP panel.

Mode (a) requires no placeholders for missing observations. Mode (c) requires the most placeholders. In both modes (b) and (c) missing components of the vector due to noninterviews and nonsample, or out-of-scope entities, must be distinguished because their conceptual relevance to analysis differs. Mode (a) requires more computational resources to identify persons than either (b) or (c).

The semantic principle for organizing time sequences of measurement must deal with a tension between principles one and two—designing the data base schema to conform to the channel of measurement and providing dynamic independence–and the semantic principles employed in analysis. Mode (b) most closely conforms to the questionnaire image and was used in the SIPP database. To accommodate analysis needs, however, we also created longitudinal summaries (c), which are described below.[8]

5.3. Longitudinal Tables

Our solution to this conflict was to extract certain information pertaining to changes of status from the history of an individual and to create summaries that reported the duration of particular status by individual. These "event histories" serve a double function: finding underlying data related to the event and providing information necessary for hazard rate modelling of change in status.

For example, reports of income receipt from 39 income support programs were consolidated into a matrix that identifies the type of income and the beginning and ending dates for each episode of receipt. This matrix becomes an easy focus for studies such as those that deal with AFDC episodes (Flory, Martini, & Robbin).

Missing observations in the panel create special problems for longitudinal analysis. For example, the missing observation in a recipiency spell creates ambiguity about the spell length. The number of left and right censored spells increases. Analysts have successfully used the longitudinal tables to impute upper and lower bounds to the distribution of spell duration (Fitzgerald).

5.4. Idiosyncratic Problems Related to Time: Conditioning

A complex battery of screening questions records the income received in the initial interview and in the interview following a noninterview. Subsequent interviews condition question-asking about income recorded in the previous interview and about new sources of income. As a consequence, the screening questions are asked only once for persons giving all nine interviews; and asked more than once for persons with "gaps" created by noninterview.

We recognized this interviewing procedure by concatenating the screening questions in a single array in the database. The array then clarifies the manner in which data were elicited, and highlights the fact that response error will differ between the first and subsequent interviews.

6. SUPPORT TO ENHANCE UNDERSTANDING OF COMPLEX DATA

Access has several aspects that reduce the complexity of a data set. We have discussed the design of a database schema to assure that retrieval is direct, inexpensive, and transparent. This is the most important aspect of access and relates closely to use of the semantic principle.

A second aspect of access is a testing capability. We discussed the utility of a a sample database in Section 2.3. Users need to test their understanding of the data by quickly retrieving the results of *ad hoc* queries. Database designers use a sample database to develop a working model of their innovations during the database development stage.

A third aspect of access is assistance in locating germane information about the data and the survey design, the metadata of the database. Logically, this aspect is ancillary to retrieving data and testing queries, but it is essential for successful end-user interaction with a database. Taxonomies of the attributes provide an explicit description of data relationships and data structures, which facilitates the user's orientation to the system. We devote the remainder of this section to discussing the role of language for understanding design concepts embedded in the data. (See David 1985b; Robbin & David; and David, Robbin, & Flory 1989a for a more extensive discussion of the language of data and the role of metadata.)

6.1 Communicating Meaning: Information about Data

Labelling attributes may appear to be a trivial topic for reducing complexity; it is not.

Parsing characters in a label is an extremely efficient way to organize information and clarify meaning. In the database parsing is used to distinguish similar information content at several levels of aggregation, to recognize the reference period of the data, and to identify attributes with questionnaire items. Parsing also identifies replicated measurements.

Channel (Instrument) Design. Labelling attributes in the SIPP database was dictated by the questionnaire design. Labels for each question are printed on the questionnaire, and should be used for the computer image. Similarity between response labels and attribute names facilitates discovery of measures in the database.[9]

Collating the nine instruments of the SIPP panel revealed that the same label was sometimes applied to different questions in the topical modules. This caused confusion and led us to prefix topical module attribute labels with the questionnaire number on which they appear. Sometimes, different labels were applied to questions with the same meaning. To assist in identifying these repeated measures a special relation was devised to provide a concordance between attribute labels.

Classifying Entities. Another application of the classification principles reduced the complexity of families of variables that related to different entities. Earnings, for example, could appear as information for jobs, individuals, families, and households, as well as reported for four different months in the reference period. The domain for the sixteen related entity-attribute pairs is identified by a single root, and labels for each pair are generated from the root by appropriate prefixes and suffixes. This principle will be extended when 1985 SIPP data are added to the database, so that identical questions generate attributes with identical labels across the two panels.

6.2 Communicating Meaning: Information about Design and Meanings

The database should provide assists to identify the logical conditioning of the measurement instrument and to understand the database schema.

Skip Patterns. Differences in the questionnaire can be identified when identical response categories for identical questions carry the same label. Identical labeling provides no assurance, however, that the same sample of entities is included. For example, questions on training in the third interview are specific to particular groups of workers, while the same questions are administered to a broader population in the eighth interview. To clarify this problem and to describe the logical universe for each measure, we constructed a relation that displays conditioning or "skip patterns" that affect each attribute in the repeated core. [10]

The concept is simple. For any attribute on a particular interview, the relation displays the conditions that **exclude** an individual from measurement. Real measures exist for the complement to excluded individuals. When the database schema conforms to the complement, the NIU disappears from the data image; otherwise, users of the data will need to select (or restrict) cases to the relevant complement.

The Database Schema. A second assist to the user is a relation that describes the schema for the database. The principle applied to create each relation in the database is described in this relation. Details are provided in a longer explanation in documentation that resides inside the database. This relation also identifies the logical and physical structure of each of the database tables.

7. SUMMARY

This paper addressed three questions. How do we reveal the structure of complex data; how do we increase accessibility to data; and how do we maintain the integrity of data structures? Four principles and a strategy were described. Applied to the *Survey of Income and Program Participation* (SIPP), the principles have been extraordinarily successful in reducing cognitive complexity and the costs of data management and analysis.

The structure of the survey design has been revealed by incorporating the semantic principle of the survey design into the database schema. Time is a special problem in applying the semantic principle because the analytical schema of statisticians differs from the survey designer's. This tension cannot be avoided, and we chose to preserve an image of the questionnaire and to maintain flexibility for analysts.

Accessibility has been significantly increased by using metadata to locate data required for answering a question. Accessibility has been improved by creating relations that summarize significant aspects of the dynamics of the data. Longitudinal tables provide analysts with the capability for both spell analysis and causal modeling. Accessibility was also enhanced by providing a two percent subsample of the database for exploratory data analysis on a microcomputer.

Integrity of the database was assured by maintaining an audit trail of operations on the database in the query language of the RDBMS. Integrity was increased by using consistent labeling and relations that identify repeated measures. Lastly, integrity was increased by utilizing the capability of a relational database management system (RDBMS) to identify errors in the data and to uncover logical and syntactical errors in the queries.

* SIPP ACCESS has received financial support from the National Science Foundation under grants #SES-8411785, #SES-8716448, and #SES-8701911 and from the Sloan Foundation under grants #B1986-25 and #B1987-46. Support by the University of Wisconsin Institute for Research on Poverty and the Center for Demography and Ecology is gratefully acknowledged. We want to acknowledge the major contribution made by Thomas S. Flory, our Database Administrator, to developing the database and the tools that have greatly enhanced access to the SIPP.

NOTES

[1] The "integrated information system" concept was used to create an infrastructure that linked the technologies of laser disk (WORM), mainframe and microcomputers, electronic networks, and a relational database management system (RDBMS). The database is accessed by telephone or through remote login from another computer installation. Extracts can be downloaded through BITNET to a researcher's home institution.

[2] In some cases data were transformed into illogical formats which attributed data to an inappropriate entity.

[3] In fact, a failure to recognize this feature of the data led Citro, Hernandez, and Moorman (1986) to overstate the stability of longitudinal households.

[4] The redundant attributes can be accessed from an earlier version of the database for methodological work on the errors associated with the data collection process.

[5] Snodgrass (1987) proposes a slightly different terminology to use with temporal databases. The collection time is "valid time". Updates leading to new versions require "transaction time". Retrospection and time frames in the data collection protocol are "user-defined time".

[6] Different questionnaire forms may be used to elicit these questions on different subsamples. That complexity need not concern us at the moment because the four-month interval is maintained for repeated core questions across questionnaires.

[7] As we explained in Section 4.1, retrospective data for the first interview do not yield correct household aggregates for the four months of the reference period. Thus unbiased household information can only be retrieved from the reference months of the second interview.

[8] Mode (c) was used by the Bureau of the Census to prepare their *Longitudinal Research File* for the 1984 SIPP panel. It occupies five reels of high density magnetic tape. The size of this format greatly increases data retrieval costs. The semantic principle of attributing all measures to individuals misidentifies relationships and adds to the complexity of understanding the scientific design.

[9] Many users come to the data set without prior study of the design and no knowledge of the questionnaire. For them, browsing a classification of measures through a controlled vocabulary is extremely helpful. This facility was created for the repeated "core" attributes in the SIPP database.

[10] Two relations could be defined to represent correctly the different samples, and thus to conform to the third normal form rule. However, the propagation of tables according to the rule increases the burden of end-user information retrieval. A balance must be sought between strict adherence to the third normal form and the information processing capability of the analyst.

REFERENCES

Boar, B.H. 1984. *Application Prototyping.* New York: John Wiley & Sons.

Boruch, R.F. 1985. Definitions, products, distinctions in data sharing. In S.E. Fienberg, M.E. Martin, & M.L. Straf, Eds., *Sharing Research Data.* Washington, D.C.: National Academy Press.

Boruch, R.F., & Cecil, J.S. 1979. *Assuring the Confidentiality of Social Research Data.* Philadelphia: University of Pennsylvania Press.

Boruch, R.F. & Cordray, D.S. 1985. Professional codes and guidelines in data sharing. In S.E. Fienberg, M.E. Martin, & M.L. Straf, Eds., *Sharing Research Data.* Washington, D.C.: National Academy Press.

Citro, C.F., Hernandez, D.J., & Moorman, J. 1986. Longitudinal household concepts in SIPP. *Proceedings of the Social Statistics Section, American Statistical Association.* Washington, D.C.: American Statistical Association, 611-616.

Codd, E.F. 1985. Is your DBMS really relational? (Part 1) Does your DBMS run by the rules (Part 2). *Computerworld,* October 14; October 21.

David, M.H. 1985a. Designing a data center for SIPP: An observatory for the social sciences. *Proceedings of the Survey Research Section, American Statistical Association.* Washington, D.C.: American Statistical Association.

David, M.H. 1985b. *The language of panel data and lacunae in communication about panel data.* (CDE Working Paper 85-20). Madison: University of Wisconsin Center for Demography and Ecology.

David, M.H. 1989. Managing panel data for scientific analysis: The role of relational database management systems. In Greg Duncan and Daniel Kasprzyk, Eds., *The American Statistical Association International Symposium on Panel Surveys.* New York: John Wiley & Sons (forthcoming).

David, M.H., Robbin, A., & Flory, T.S. 1988a. Access to data: Handling the 1984 SIPP. *Proceedings of the Statistical Computing Section, American Statistical Association.* Washington, D.C.: American Statistical Association.

David, M. H., Robbin, A., & Flory, T.S. 1988b. *Analyzing Complex Data: A DBMS for the 1984 SIPP.* Madison, WI: Institute for Research on Poverty.

Dolby, J.L., Clark, N., & Rogers, W.H. 1987. The language of data: A general theory of data. *Proceedings of the 18th Symposium on the Interface of the American Statistical Association.* Washington, D.C.: American Statistical Association, 96-103.

Fischhoff, B., MacGregor, D., & Blackshaw, L. 1987. Creating categories for databases. *International Journal of Man-Machine Studies* **27**: 33-63.

Fitzgerald, J. *The effects of the marriage market and AFDC program parameters on recipient duration on AFDC.* Paper presented at the Social Science Research Council Conference on Individuals and Families in Transition: Understanding Change Through Longitudinal Data, Annapolis, March 16-18, 1988.

Flory, T.S., Martini, A., & Robbin, A. 1989. Attrition and spell censoring in estimating dynamic models of welfare recipiency. *Proceedings of the Social Statistics Section, American Statistical Association.* Washington, D.C.: American Statistical Association.

Hedrick, T.E. 1985. Justifications for and obstacles to data sharing. In S.E. Fienberg, M.E. Martin, & M.L. Straf, Eds., *Sharing Research Data.* Washington, D.C.: National Academy Press.

Kent, W. 1983. A simple guide to five normal forms in relational database theory. *Communications of the ACM* **26**(2):120-125.

Mark, L., & Roussopoulos, N. December 1986. Metadata management. *Computer*: 26-36.

Martini, A. 1988. *A Theoretical and Empirical Model of Labor Supply that Accounts for the Unemployed and the Discouraged Workers.* Ph.D. dissertation. Madison: University of Wisconsin.

Robbin, A., & David, M.H. 1988. Information tools to improve access to national longitudinal panel surveys. *Research Quarterly* **27**:499-515.

Snodgrass, R. 1987. The temporal query language TQuel. *ACM Transactions on Database Systems* **12**:247-298.

Figure 1.

Design and Channel:
The Effect on Data Image and Interpretation

	DESIGN		CHANNEL (EXECUTION)	
	Sample	Stimulus	Response	Data image
ASPECTS:				
1.	Selection Probabilities	Reference period, Question	Categories	Truncation
2.	Respondent (Proxies)	Referent	Nonresponse	NA flags, Imputation
3.	Conditioning (Skip Sequence)	Order		
INTERPRETATION				
1.	Representativeness		Response error	Inverse of transformations
2.			Proxy bias	