Robert F. Teitel, TEITEL DATA SYSTEMS

I. INTRODUCTION

In essence, both papers reflect the experience of two seasoned analyzers of large and complex social science data collections, and enumerate the potential problems to be faced by a neophyte in attempting to gain useful information from surveys as large and complex as the Survey of Income and Program Participation (SIPP).

In any discussion of database systems, there must be a clear distinction between the logical model (data structures and manipulation language) and the physical model (storage structure and access methods). The logical model is the users conception of the structure of the data and the language for manipulating the data; the physical model consists of the actual, machine-oriented, storage structures and access methods used to implement the logical model. See figure 1.

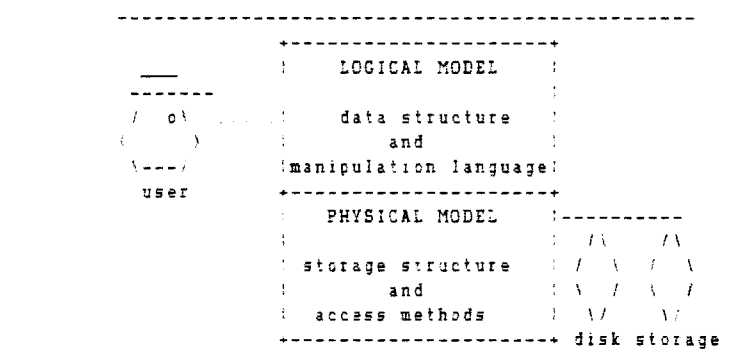

Figure 1: Logical and Physical Models of Data.

II. USER VIEWS OF STATISTICAL DATA

All database systems and even most statistical packages are now RELATIONAL systems -- at least according to their promotional material!

What is the relational model? Is it useful for statistical systems? If not, why not, and what is better? These are some of the questions we will address in this section.

This paper will address some of these issues in the context of the SIPP data collection. There exist a plethora of overlapping terms to describe the simple concepts of a logical model. The ones used here are in general common use within the computing science database research community, albeit here placed in a statistical data environment. A logical model consists of three basic components: a conceptual schema or conceptual description of the data of interest to an analyst, often with a set of consistency rules, a mechanism to create a subschema or user view of that data, and a language with which to perform the necessary data manipulations to answer user queries against the data. The Relational Model is an example of a logical model. Other common models are the Hierarchical Model and the Network Model. See any modern database management text, such as (Martin 1975; Weiderhold 1983; or Date 1981).

The Relational Model, principally defined in (Codd 1970) and extended in (Codd 1979), consists of (1) a collection of (time-varying) tabular relations, with appropriate domains, keys, and normalizations, (2) insert, update, and delete integrity rules, and (3) the relational algebra language.

In a series of papers (Teitel 1975, 1977a, 1977b, 1977c, 1982) this author has discussed the relational model, its utility, and its flaws with respect to statistical database issues. In more recent, unpublished, working papers relating to the Survey of Income and Program Participation (SIPP) an Entity- Relationship or E-R Model has been used. The Entity-Relationship Model was proposed in (Chen 1976) as somewhat of a super-model, in that the other, more conventional, models are subsumed by its properties, and is extensively covered in (Howe 1983).

Before turning our attention to the Entity-Relationship Model, and its potential role in statistical database management, it may be worth noting the (almost) present state of affairs in statistical database management. In 1981, at two separate conferences, 11 developers of statistical, tabulatory, and database systems presented solutions, based on actual computer runs, to a common set of data manipulation problems typical of those found in analyzing a large complex survey. The benchmark problem definitions are found in (Teitel 1981a) and repeated in (Teitel 1981b), and the respective solutions are in (Robinson 1981, Bragg 1981, Buhler 1981; Nagara and Nolte 1981; Schmitz 1981; Jacobs 1981; Merrit 1981; Weiss and Stevens 1981; Ilacera 1981; Maness and Dintelman 1981a; Fry 1981). The end-product of each of the data manipulation problems was a simple cross-tabulation. For one of the tabulations there were five different results, albeit several of the erroneous tables were due to "programmer" error, raising serious questions on the efficacy of contemporary statistical and database systems for processing complex data collections (Teitel, 1982b).

A study of the solution procedures for the benchmark problems reveals that, in general, the logical models employed by the varous systems have difficulty in at least one of three areas: lack of a conceptual schema for the level of complexity exhibited by the destributed data collections, lack of focus on an unit-of-analysis, or lack of handling of missing structure data.

The difficulties faced by the vendors of the packages were relatively minor compared to those expected to be faced by users of the SIPP data, for the latter is so much more complex in Within a cohesive data collection such as SIPP,

III. THE ENTITY-RELATIONSHIP MODEL

The Entity-Relationship or E-R model consists of four major components, entities, relationships, attributes, and domains.

An entity, as used here, is a collection of simple observations of a common object to be represented in the database, distinct from other objects in the database. Within SIPP, for example, SAMPLE-HOUSEHOLD, PERSON, and PERSON-MONTH-WAGE-SALARY are entities: objects about which simple observations are stored in the database. Observations are simple when the values for each attribute (also called field or variable) consist of a single numeric value or alphabetic string -- structures such as repeating-groups are not permissable within entities (The specific details of some other rules to be followed in the definition of the entities within a data collection, called normalization, will be ignored in this presentation.) Occurrences of entities are usually displayed in simple tabular form -- the rectangular data structure common to data analysis.

In statistical terms, a set of properly normalized entities represent all lowest level units-of-analysis within the data collection. Other units-of-analysis are possible, but they are aggregations (or selections) of the data contained in the normalized entities. For example, within SIPP there might be a SAMPLE-HOUSEHOLD entity consisting of observations of the original sample of households:

SAMPLE-HOUSEHOLD:

psu	hh#	structure	persons	state	...
001	01	H - W	2	NY	
123	45	H-W-CHs	4	NM	
...		...		...	

Figure 2: Example of a set of Entity Occurrences

Univariate and multivariate descriptive statistics, for example, could be presented for the SAMPLE-HOUSEHOLD entity, in each instance the "total n" would be the number of sample households. Alternatively stated, the unit-of-analysis is the SAMPLE-HOUSEHOLD segment.

One of the two reasons to determine the complete set of normalized entities within a data collection is precisely to determine all possible lowest level units-of-analysis.

The second reason to determine all normalized entities within a data collection brings us to the second component of our data model for SIPP, that is, the relationships.

the entities do not exist in isolation: there are associations or relationships between various segments. For example, the SAMPLE-HOUSEHOLD entity surely has a relationship, "consists-of", with a PERSON entity. The PERSON entity has a relationship, "earns", with a PERSON-MONTH-WAGE-SALARY entity. Furthermore, since each SAMPLE-HOUSEHOLD "consists-of" PERSONs and many a PERSON "earns" PERSON-MONTH- WAGE-SALARY, transitive relationships, such as that between SAMPLE-HOUSEHOLD and PERSON-MONTH- WAGE-SALARY, can also be defined for the SIPP data collection.

In its simplest form, a relationship is formed between two entities by associating the observations in one segment to the observations in an other based on an equal value of a like-named attribute (technically, both attributes should have the same domain). In the relational terminology, this is called an equi-join. For example, consider the following abbreviated entities, PERSON and PERSON-MONTH:

PERSON:

p#	sex	...
01	m	
13	f	...
...	.	
07	f	
...	.	

PERSON-MONTH:

p#	mnth	ms	tmy
01	01	msp	1200
01	02	msp	1200
...	...	...	...
13	11	sep	900
...	...	...	...
07	14	sep	500
13	15	div	700
...	...	...	...

Figure 3: A Relationship between two Entities

A relationship has been defined between PERSON and PERSON-MONTH based on equal values of the common attribute 'person number', or p#. The illustrative arrows between the two tables connect the PERSON observation with p# = '13' to those observations in PERSON-WAVE with p# = '13'. (This definition of relationship carries no implication as to the order of the observations within their respective tables; that is "merely" an implementation efficiency issue.)

In the Entity-Relationship model, the relationships, such as that from PERSON to PERSON-MONTH are both named and directed. Data structure diagrams can be created from the entities and their relationships to represent clearly the conceptual schema, or data model. For example, Figure 4, next page, shows that each PERSON "lives" a multiple number of PERSON-MONTHS, and that the relationship is formed via the common attribute p#.

A FAMILY segment will likely have a "contains" relationship to PERSON, and PERSON may have a "belongs" relationship to FAMILY, both based on equal values of the common attribute 'family number' or f#. Figure 5, next page, shows that each FAMILY "contains" multiple PERSONs, and each PERSON "belongs" to a single FAMILY.

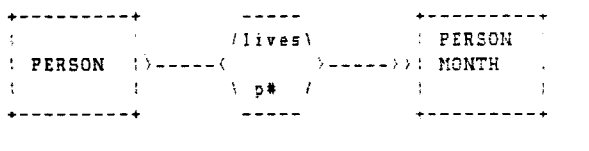

Figure 4: Sample Data Diagram Illustrating a Relationship.

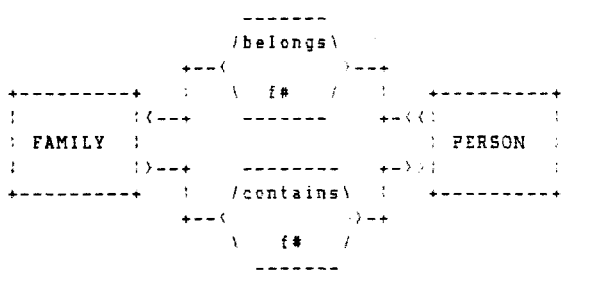

Figure 5: Sample Data Diagram Illustrating Two Relationships between two Entities.

Some relationships, such as that between PERSON and SELF-EMPLOYMENT-MONTH (a monthly self-employment earnings record which reflects family operated enterprises) are "many-to-many", that is, each SE-MONTH observation could belong to several PERSONS, and each PERSON could have several SE-MONTH observations. Such "many-to-many" relationships require explicit attributes within the relation- ship, one or more for each segment to "relate to". In effect, many-to-many relationship are transitive thrcugh the attributes in the relationship. For example, the Figure 6 diagram shows the many-to-many relationships between PERSON and SE-MONTH through a relationship called "share".

In effect, a new entity has been created, which could, in a statistical sense, become a unit-of-analysis. Though apparently not collected in the SIPP survey, a 'percent' attribute of "share" could be used to study the distribution of ownership of family held enterprises.

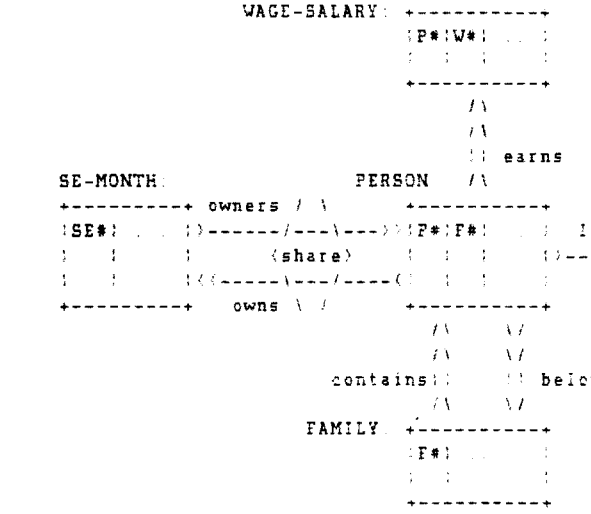

Figure 7: Some of the Principal Entities and Relationships in SIPP.

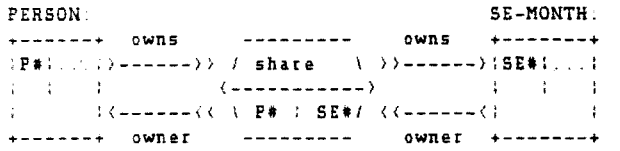

Figure 6 Sample Many-to-many Relationships between two Entities

Adding a few more entities from the SIPP data collection to those already mentioned above, the data diagram below represents some of the principal entities and relationships in that data collection.

The data diagram permits one to write down explicitly the relationships between the entities. For example, the following closely follows the syntax used the Link and Selector Language (Tsichritzis 1976):

1. LIVES: from PERSON to PERSON-MONTH on P#
2. EARNS: from PERSON to WAGE-SALARY on P#
3. BELONGS: from PERSON to FAMILY on P#
4. CONTAINS: from FAMILY to PERSON on F#
5. OWNS: from PERSON to SHARE on P#
   to SE-MONTH on SE#
6. OWNERS: from SE-MONTH to SHARE on SE#
   to PERSON on P#

The Relational Model of Data does not permit the explication of the links between entities as part of the conceptual schema; such relationships are to be activated only when necessary for the resolution of a specific query.

The Entity-Relationship Model calls for the explicit identification of the known relationships between entities; and permits the creation of new relationships when necessary in the conduct of a research task using a large and complex data collection such as SIPP.

VI. REFERENCES

Due to space limitations, the references cited in the text are not presented here. Instead the following single reference is likely to contain all the citations made in this discussion.

TEITEL, Robert F. (1985) "Statistical Databases and Statistical Database Managment", Proceedings of the First Annual Research Conference, U.S. Bureau of the Census, 108-131.