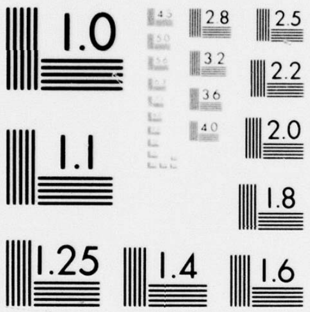
AD-A037 910 RAND CORP SANTA MONICA CALIF F/G 9/2
APPLICATION OF DATA BASE CONCEPTS IN OPERATIONALIZING THE ARCHI--ETC(U)
DEC 75 J S HURLEY
UNCLASSIFIED P-5605 NL

1 OF 1
ADA037910

END
DATE
FILMED
4-77



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

ADA037910

P
B.S.

6

APPLICATION OF DATA BASE CONCEPTS IN OPERATIONALIZING THE ARCHIVING AND RETRIEVAL OF PANEL STUDY DATA,

10

Jeanne S. Hurley

11

December 1975

12 22p.

DDC
APR 8 1977
C

AD NO. ____
DDC FILE COPY

14

P-5605

296 600

LB

The Rand Paper Series

Papers are issued by The Rand Corporation as a service to its professional staff. Their purpose is to facilitate the exchange of ideas among those who share the author's research interests; Papers are not reports prepared in fulfillment of Rand's contracts or grants. Views expressed in a Paper are the author's own, and are not necessarily shared by Rand or its research sponsors.

The Rand Corporation
Santa Monica, California 90406

ABSTRACT

The massive collection of social survey data associated with public policy research is a relatively recent phenomenon. The history of data processing associated with such data shows repeated occurrences of cost overruns, extensive time lags before data is available, and, even then, large amounts of both unusable and not yet usable data.

The Health Insurance Study, a complex, multi-year longitudinal study of the role of health insurance in the utilization of health care services, is discussed to illustrate the dimensions of the data processing task and the benefits of applying data base techniques in managing them.

ACCESSION for
NTIS White Section ☑
DDC Buff Section □
UNANNOUNCED □
JUSTIFICATION Per ltr. on file
BY
DISTRIBUTION/AVAILABILITY CODES
Dist. AVAIL. and/or SPECIAL
A

ACKNOWLEDGMENTS

The author would like to thank David Stewart and Beatrice Yormark for their review of and comments on an earlier draft of this paper.

CONTENTS

I. INTRODUCTION

The massive collection of social survey data associated with public policy research is a relatively recent phenomenon [4]. The history of data processing associated with such data shows repeated occurrences of cost overruns, extensive time lags before data is available, and, even then, large amounts of both unusable and not yet usable data [16]. Since the data processing system is a major determinant of the technical nature, sequential speed, and amount of research performed in studies of this type, implementation of a successful system is critical [13, 16].

Data base management techniques can help reduce the problems encountered to date. While the panel technique, in which data is collected by repeated interviews with the same group of people, has been used for over thirty-five years [8], the survey data processing model has remained invariant despite the growing dimensions of this type of enterprise. The Health Insurance Study (HIS) [11], a complex, seven-year project being conducted by the Department of Health, Education and Welfare (DHEW) and The Rand Corporation to study the interactions between health insurance, health care services utilization, and health status, provides an example of the dimensions of the data processing task and the benefits of applying data base techniques in managing them.

II. DIMENSIONS OF THE ENTERPRISE - THE HEALTH INSURANCE STUDY

A research project of this type is complex in a number of different dimensions:

- analysis units--the basic entities to be studied
- data content--the research interests to be served by the data
- documents (transaction types)--the vehicles for collecting data
- collection styles--sources and methods of collecting data
- data items--the individual pieces of information to be analyzed.

Analysis Units

In survey research, the most elementary unit of analysis is an individual person. For the purposes of the HIS these individuals are grouped into economic decision making units, which are for the most part the same as families. Analysis will also be performed on other groupings of the data representing subsets based on such characteristics as geographical location, episode of illness and insurance plan.

Data Content

Besides basic demographic and family configuration data, information is being collected to serve research interests in the fields of economics, health status, health services utilization, attitudes about the health care delivery system, and survey methodology.

Documents (Transaction Types)

Data will be collected on over 2,300 families or about 8,000 persons in four different locations, utilizing around 650 separate data collection document types. In a particular data collection effort, anywhere from 1 to 15 of these document types will be used, varying in number and combination from family to family. For example, initial interviews are performed in each location to obtain baseline information on the site. These baseline interviews utilize, in addition to a core document, supplements to collect information on health care utilizations, secondary jobs, and insurance coverage. The amount and type of information collected and thus the number and type of documents used will vary as widely as the health status, employment, and insurance profiles of real families.

The documents themselves range in size from 20 to 1,400 potential items of data. The occurrence of many of these items is contingent upon responses to preceding items, resulting in complex control logic that is data dependent.

Collection Styles

In addition to formal interviews, participant self-reports, laboratory tests, third party reports of administrative records such as insurance claims, and field procedures edit reports must all be linked for analysis. Both the complex structure of the documents and the different styles of data collection contribute to uneven data quality [12]. As the complexity of the structure increases, the probability that items will be incorrectly skipped or inappropriate items answered increases. A trained interviewer is more likely to reliably follow a complex structure than a respondent unfamiliar with survey techniques. A busy doctor or nurse is more likely to rush when filling out a form and not complete it exhaustively.

Data Items

Some data items will be collected only once; others repeatedly, at varying intervals throughout the study. As a further complication, the critical analysis unit, the person, may be related to different economic decision making units at different points in time. This will occur, for instance, as children set up separate residences, couples separate, or persons living alone form new family units.

III. SUMMARY OF PROBLEM TYPES

Generally speaking, then, the difficulties encountered in attempting to use a data base like that of the HIS revolve around the number, variety and logical complexity of the data collection documents, the variable number of occurrences of particular data items, the need to link pieces processed separately, and the longitudinal tracking of changing analysis units. These problems are exacerbated by the size of the files and the resulting costs associated with the use and restructuring of these files, the variety of end users, the length of time needed to gain sufficient familiarity with the data to use it successfully, the need for rapid access to new data to provide research results that represent current social situations, the uncertainty associated with drawing conclusions from incomplete and still accumulating data banks, and the varying quality of data collected by survey methods.

IV. THE TRADITIONAL SURVEY APPROACH

The traditional approach to managing survey data utilizes the matrix format that is the standard input to most packaged statistical routines. In a matrix format, the completed document for each analysis unit becomes a record. Where more than one analysis unit type is to be studied, a separate file is created for each type [5]. Each data item is referenced by a fixed location in a particular file. For example, using a card and column notation, the data item, "health status" would be located in the same position, say card 06 - column 25, on each record for one type of document. The same item might be located on card 10, column 68 for another document type. Associated with each data collection document and its resulting data file or files is a codebook describing each data item [10]. The contents of a codebook include the text of the question used to elicit the response, the meaning of coded responses, the meaning of codes for non-response such as 9 for "refused" or 0 for "not asked due to skip logic," and the location of the data in the record. This traditional approach is efficient for a small survey but rapidly becomes unwieldy as the number of documents and the size of the records increases [12].

V. THE DATA BASE APPROACH

In planning the data processing effort for the HIS, a data base systems approach was selected [14]. The concepts of an integrated data base, centralized data descriptions and an applications administration appeared to solve many of the problems inherent in the traditional survey approach. The Joint GUIDE-SHARE Data Base Requirements Group specifies seven basic objectives for a data base management system [6]:

- data independence
- data relatability
- data non-redundancy
- data integrity
- security
- performance
- compatibility.

The ANSI/SPARC Study Group on Data Base Management Systems refers to three levels at which a data base can be viewed: internal, external, and conceptual [1].

The current state of the art for public policy research is such that no one is able to fully understand the information needs of the enterprise or sufficiently define the uses of the data collected [7]. Therefore, it is impossible to define a conceptual schema for a data base of this type. As a result, a centralized external schema has been defined for the HIS data base, in order to provide a common reference for the data base system.

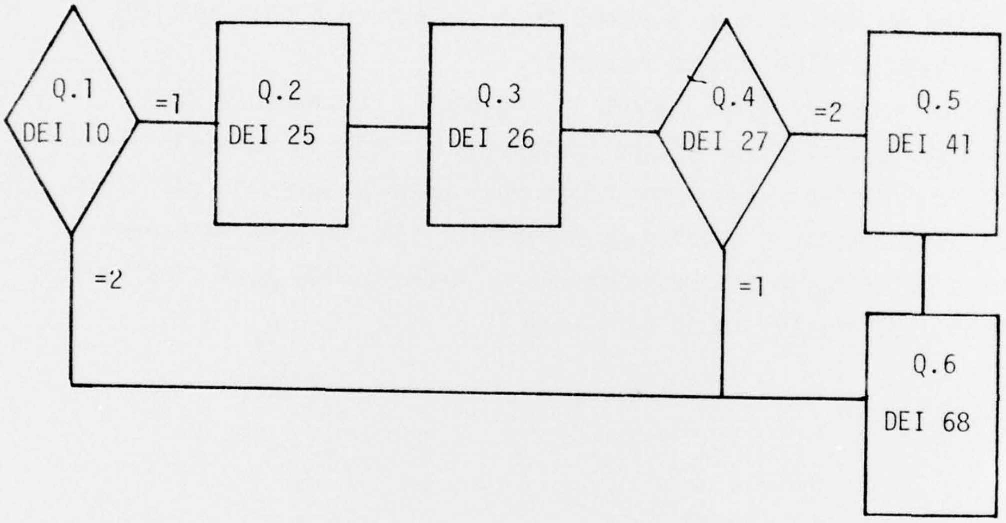
VI. SPECIFIC DATA BASE APPLICATIONS - THE HEALTH INSURANCE STUDY

DEFINING THE EXTERNAL SCHEMA

One of the steps toward achieving the data base goals was to divorce the identification of any particular data item from its physical location in a record. To accomplish this, unique numbers called data element identifiers (DEIs) are assigned to each different data element. In this way, the answer to a question asked on two or more document types is assigned the same DEI which forms an automatic link across documents.

The meaning of a survey data element is frequently dependent on the context in which it is collected, so the assignment of DEIs is not sufficient to create a full definition. Two other tools have been designed to fill this gap. A logical description is created to define the flow through each new document. Such a description can be represented schematically by a flow chart or by a set of logical statements as in figure 1.

FIGURE 1



10 mandatory

if 10=2 then 68

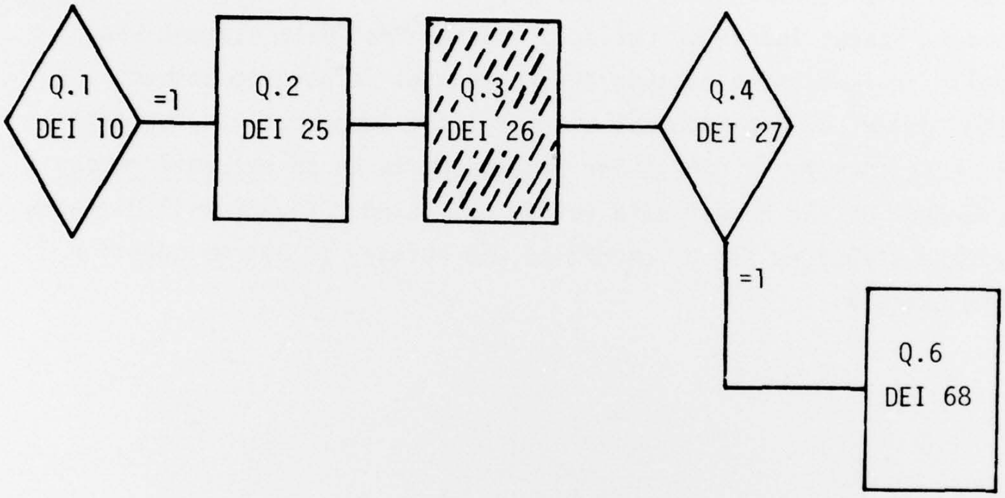
if 10=1 then 25, 26, 27

if 27=2 then 41

68 mandatory

This external document structure can then be used to explain the intra document dependencies among data elements. This works well in the abstract, but to account for the non-occurrence of an item required by the document flow, e.g. when a real respondent refuses to answer a particular question, a second logical entity needs to be defined. To do this, for each real respondent, a set of data status indicators is created associated with only those data elements that would be required for that respondent if the document flow was followed correctly. The value of a data status indicator reflects whether that data element was collected normally or is missing because the respondent refused to answer, did not know the answer, or the document was completed incorrectly. The set of data status indicators for a particular respondent forms an external record structure independent of the actual data values collected. Figure 2 illustrates the external record structure for a respondent who refused to answer question 3 in the preceding example.

FIGURE 2



DEI	Data Status Value
10	normal
25	normal
26	refused
27	normal
68	normal

We now have an external schema that links items both between and within data collection documents and functions in both an abstract and a real sense.

IMPLEMENTING THE EXTERNAL SCHEMA

The Data Dictionary

Central to implementing a centralized data description in a computable form is the concept of a data dictionary [9, 15]. The data dictionary for the HIS [2] contains two files representing the two abstract components of the external schema: a file containing information on each data element identifier such as

- full question texts
- answer texts and their code values
- a set of document-related information for each document on which the DEI occurs
  - field site
  - question number
  - allowable value ranges
  - storage mode
- a facility for specifying relationships between DEIs e.g., same basic information but different question text,

and a file containing the external document structure for each of the data collection documents. In addition to functioning as the central driving mechanism for the computable system, listing the information contained in the data dictionary fills the documentation role served by the codebook in traditional survey processing.

The Unit Tracking System

It is also necessary, within the context of the external schema, to describe the relationships between analysis units. To assure that all the data on a single individual is linked, permanent personal identifiers are assigned. Associated with each individual identifier is a family identifier representing a unique configuration of individuals. If a family configuration changes, new family identifiers are specified for the data collected under the new configuration. All analysis unit identifiers contain an imbedded check digit to control potential transcription errors. The logical relationships between family configurations are described by a tracking system which contains pointers mapping each individual identifier to every family identifier with which it has been related in the data base. In addition, the tracking system includes information on which individuals prompted a given change, the reason for each change, and the source of the information that a change has occurred.

IMPLEMENTING THE INTERNAL STORAGE SCHEMA

Another step in creating a functionally integrated data base requires a storage schema that allows record variations to be dealt with in a manner that is transparent to the user. To accomplish this for the HIS, a generalized record structure has been used. The structure designed contains three parts:

- a header section containing identifying and control information
    - document number
    - respondent identifiers
    - record length
    - data element count
- a pool of the actual data values
- a data directory associating each DEI with pointers into the value pool.

Using the generalized record structure as a mapping between the external and internal schemas, a particular data element is accessed by referencing its DEI. In this way, a data element appearing on various documents having different external structures is always retrieved in the same fashion. In contrast to the traditional survey approach which requires frequently changing software to directly manipulate data files to achieve the data orderings required by new external schemas, the use of the generalized record for the storage schema permits standardization of software for archiving and retrieving data collected in diverse ways.

IMPLEMENTING CENTRALIZED DATA ACCESS

Data Base Administration

In order to maintain an integrated data base, control of the archiving process needs to be centralized [3]. This has been done on the HIS through a data base administration function. The data base administration staff has responsibility for creating the centralized external description of the data base, maintaining the data dictionary and unit tracking system, and certifying the integrity of the data base. Since successful use of a data base as large and complex as that being collected for the HIS requires a substantial learning period to gain sufficient expertise, control of the retrieval process has also been centralized in the form of an applications [1] or data request administrator (DRA).

Data Request Administration

The data request administrator serves as a technical liaison between the end users of the data and the data base administration function. Utilizing standard software, the DRA can rapidly retrieve needed data elements from the data base and create data files suitably organized for each analysis. In this way, end users need only gain familiarity with the data elements relevant to their particular research interests. Centralization of data access, therefore, not only protects the integrity and security of the data base, but facilitates use through more rapid access and shortened learning periods.

VII. IMPACT OF THE DATA BASE APPROACH

The difficulties associated with processing panel study data can now be reviewed in the context of the data base design implemented for the HIS. The number, variety and logical complexity of the data collection documents still pose problems for the novice user faced with a multiple document analysis. However, once the user has determined which data elements to use, the problem of access has been greatly reduced. Whether or not a data element has occurred on more than one document is readily apparent through reference to the centralized data element dictionary, antiquating the search through multiple code books for reused items. Problems still exist in reconciling data processed separately, but linkage is handled automatically by the standard retrieval software. Unchanging units of analysis can be automatically separated from ones that have changed by using the unit tracking system.

The costs of using such an extensive data base are reduced since the user works with only those data relevant to a particular analysis. The variety of end users and the length of time needed to gain familiarity with the data is less of a problem when the user need only be familiar with a subset of the data and then only with the external schema. Access time is improved through the use of standard software and additional savings are realized through minimizing custom programming with its associated debugging costs.

The uncertainty associated with drawing conclusions from incomplete and still accumulating data banks and the varying quality of data collected by survey methods are not simply data processing problems. However, integration of the data base and centralization of data access allow for consistent documentation of the state of the data base at any point in time.

VIII. SUMMARY

Through implementation of centralized data definition tools, centralized data access and a generalized storage schema, the seven basic requirements of a data base system have been achieved to the extent that:

- the mapping between external and internal schemas provides for a level of data independence unachievable in the traditional survey data processing model
- data can be related across sources and through time
- data used in different external schemas by different users are stored without redundance
- integrity of the data base is maintained by the Data Base Administration, over time, for a changing user community
- the data base is protected by the Data Base Administration from unauthorized or premature use
- performance continues despite constant change and update
- use of a centralized external schema provides compatibility over a wide range of users.

By achieving these data base goals, the data processing problems encountered in large panel studies have been controlled, enabling the analytical research effort to proceed concurrently with the accumulation of the data base.

REFERENCES

(1) American National Standards Institute Committee on Computers and Information Processing, *SPARC/Data Base Study Group Status Report,* Document No. X3/74-1, February 1, 1974.

(2) Dunn, W. C., Yormark, B., *Use and Maintenance of a Data Dictionary,* The Rand Corporation, P-5324, November 1974.

(3) *Codasyl Data Base Task Group Report,* New York: Association for Computing Machinery, April 1971.

(4) Harrar, W. S., Bawden D. L., "The Use of Experimentation in Policy Formulation and Evaluation," *Urban Affairs Quarterly,* Vol. 7, June 1972, pp. 419-430.

(5) Institute for Social Research, *A Longitudinal Study of Family Economics: A Brief Description of an Ongoing Panel Study,* The University of Michigan, August 1969.

(6) Joint GUIDE-SHARE Data Base Requirements Group, *Data Base Management System Requirements,* November 11, 1970.

(7) Juster, F. T., "Microdata Requirements and Public Policy Designs," *Annals of Economic and Social Measurement,* Vol. 1, No. 1, January 1972, pp. 7-16.

(8) Lazarsfeld, P. F., "The Use of Panels in Social Research," *Proceedings of the American Philosophical Society,* Vol. 92, No. 5, November 1948, pp. 405-410.

(9) Martin, G. N., "Data Dictionary/Directory System," *Journal of Systems Management,* December 1973.

(10) Nasatir, D., *Data Archives for the Social Sciences: Purposes, Operations and Problems,* Reports and Papers in the Social Sciences No. 26, UNESCO, 1973.

(11) Newhouse, J. P., *The Health Insurance Study - A Summary,* The Rand Corporation , R-965/1-OEO, March 1974.

(12) Primus, W. E., "Data Collection and Processing Problems Associated with Social Experimentation," presented at the 41st National Meeting of ORSA, 1972.

(13) -----, "Data Organization and Structuring Problems Associated with Micro-Economic Survey Data," University of Wisconsin, unpublished paper.

(14) Stewart, D. H., *A Design for Information Processing in the Health Insurance Study,* The Rand Corporation, P-5229, September 1974.

PRECEDING PAGE BLANK-NOT FILMED

(15) Uhrowczck, P. P., "Data Dictionary/Directories," *IBM Systems Journal*, November 4, 1973.

(16) Watts, H. W., "Microdata: Lessons from the SEO and the Graduated Work Incentive Experiment," *Annals of Economic and Social Measurement*, Vol. 1, No. 2, April 1972, pp. 183-192.

P-5605 APPLICATION OF DATA BASE CONCEPTS IN OPERATIONALIZING THE ARCHIVING AND RETRIEVAL OF PANEL STUDY DATA Hurley