Database documentation applied to the RAND Medical Outcomes Study

by Lisa Stewart [1]
William H. Rogers

Introduction

This paper describes a schema for documenting variables in a large-scale social science survey. In particular, we describe this schema as it is applied to the Rand Corporation's National Study of Medical Care Outcomes. It serves as a basic set of working instructions for documentation. The purpose of this effort is to create a database that will enable primary and secondary researchers to locate data and comprehend the scope of the database.

The system of documentation discussed here is based on a system developed by the Language of Data Project (LOD), which is supported by a grant from the System Development Foundation. Rand initially offered to be a test site for the portion of LOD that relates to documentation. However, Rand's needs turned out to differ substantially from the use for which the available materials were designed. As a result, the LOD materials were adapted to meet Rand's immediate needs.

The material with which we began represented the early LOD thinking in Dolby and Clark (1982), Dolby (1983), and developing versions of Clark (1985). Consequently the adaptation implemented at

[1] Presented at the International Association for Social Science Information Service and Technology (IASSIST) Conference held in Vancouver, British Columbia, Canada on May 19-22, 1987. Authors may be reached at the Rand Corporation, Santa Monica, CA 90406. The Medical Outcomes Study is sponsored by The Henry K. Kaiser Family Foundation, the Robert Wood Johnson Foundation, and the Pew Foundation.

Rand does not include much LOD structure developed since; this may be incorporated later. Meanwhile, Rand has developed extensions in other areas relating to the tracking of information. What is described here is the system as it has been implemented at Rand. For a more complete description of LOD structure see Dolby, Clark, and Rogers (1986) and other LOD materials.

The National Study of Medical Care Outcomes, also known as Medical Outcomes Study (MOS), is a multi-year panel study of health care process and health outcomes in various systems of medical care (e.g. fee for service or health maintenance organizations) and various specialty groups. It is intended to be a data resource for health policy research for several years.

Part 1 - Structural elements

I. Overview

General documentation needs

The MOS consists of interrelated written questionnaires given to patients and their clinicians, telephone interviews, personal interviews, and laboratory reports.

A complex study of this kind produces two kinds of documentation need. The primary user (the analyst), who is knowledgeable about the study, needs a way in which to rapidy locate those data which are important to him or her, to distinguish these data from other similar data, and to be aware of peculiarities discovered by colleagues. The secondary user, who is unfamiliar with large parts of the study, needs to be able to place particular data in the context of the whole effort.

The two roles merge when data are exchanged among researchers. Thus, when the questionnaire analysts study the lab data, they need to place the data in context.

Variables

A database is structured as a table with rows that represent the subjects of the study and columns that represent facts about the subjects. The columns are called *variables.* A *variable* is a collection of observations on the same phenomenon for a set of cases. For examples, "income" is a piece of information collected on all patients in the MOS.

Variables can be divided into two types: 1) a *raw variable* is the unprocessed datum that is collected in the field. In MOS, this is obtained either through self-administered questionnaires or through telephone interviews conducted using a system called CATI — Computer assisted telephone interviews, or 2) a *derived variable* which is computed by logical or arithmetic operations performed on one or more variables, usually calculated on a computer.

Documentation of raw variables versus derived variables is fundamentally different. Raw variables are backed up by traditional documentation such as the questionnaire itself, coding specifications and

CATI scripts. In addition, questionnaire designers created "concept keys" that describe groups of closely related variables. Although these documents are voluminous, they serve primary user needs well. Secondary users involved with details need to invest a great deal of effort to understand these materials, but once they do, the information they require is there.

For derived variables, no traditional documentation exists. There is no clear statement of how the variable was computed other than a computer program which is likely to be physically and intellectually inaccessible to the secondary user. This is particularly problematic because derived variables represent the best summary of the analysts who were in the best position to understand the overall context. Thus the secondary user is often faced with the choice of using a derived variable he doesn't fully understand or deriving one himself which frequently requires a large investment if it is to be done successfully.

In the Medical Outcomes Study we focus on the documentation of derived variables and use existing sources of documentation for raw variables. This compromise is a budgetary one. While it would be nice to have a uniform summary of raw data that would help the secondary user, it is expensive to produce because there are plenty of raw variables. On the other hand, derived data are more heavily used and have no viable alternative sources of documentation.

Nevertheless, the Language of Data is based on a documentation scheme that works for both raw and derived data. This document will also describe how raw variable documentation would be done.

Data and Description

The idea behind a database documentation system is to address both levels of need, which may coexist in a particular user of the data. As a vector of numbers, a variable does not mean anything in itself, but acquires meaning from the environment that surrounds it. The first task is to document that environment, and the second task is to supply essential information for finding and using the variable.

First, we want to describe the *context* of the variable. How does it fit conceptually into the overall study? What is the sample population to which the variable refers?

By *content* of the variable we mean the conditions or qualifications that describe the datum beyond just its numerical value. To what does the variable refer? When does the measurement apply? Where? How? According to whom? As a simple example, while the numbers representing "income" may be readily accessed, the information regarding how, when, why and where the information was collected may or may not be readily available. This documentation schema *systematically* records this contextual information for each variable. These facts are usually known to the survey designers but may be unclear to secondary users.

Finally, we *identify* the location of the variable in the computer system and give other operational information needed for day-to-day use of the data. Identification elements are directed to analysts and other primary users.

In addition, once the data have been collected and the analyst works with them, he/she acquires an understanding of biases or unusual characteristics. All too often in the past, this information has been carried around in a person's head or buried in a mountain of paper. Similarly, when creating derived variables, information regarding the treatment of, or experiences with the data has rarely been captured so that others may understand one's intentions, orientation, or explanations of the variables. In this documentation system, we attempt to correct this by *systematically* recording information so that the context of the variable is computerized along with its numerical value.

General documentation approach

The steps in documentation can be thought of as two–fold. First, there is the actual documentation of the data, which is carried out on a full–screen word processor. Secondly, this information is reformatted into different types of reports by means of a rather sophisticated reporting program. This paper focuses on the first half of the process, serving principally as a working manual for documentation.

The first step consists of identifying and recording entries for the essential elements needed to describe the variable. An *element* or *descriptor* is a basic piece of information that answers what, when, where, why or how issues for each variable. At Rand we refer to the categories of descriptive elements as a *template* or *descriptor set* and sometimes refer to the process of documentation as "templating". The importance of this term is that we systematically record the same information for every variable in the study.

After all the data have been documented, the descriptive information is transmitted to a computer–based reporting system for locating variables and displaying their context. This system is analogous to a library system for finding books. Over the long term, we hope that LOD will have the resources to develop a computer reporting system that can compute symbolically and linguistically the contextual information as well as one now does numerically the numbers.

II. Introduction to elements

The categories of descriptive information which we use to document the variable are grouped into *context, content,* and *identification* elements.

The *context* elements are *subject* and *topic. Topic* in turn is organized into *main topic* and *subtopic* in a hierarchical (tree) structure. In a study such as the MOS, the *subject* describes the set of cases to which the variable applies. The *topic* describes where the variable fits into the MOS Concept Outline.

Next there are five major content elements: *observer, matter, function, space,* and *time,* and two associated content elements: *aspect* and *domain.*

Finally, the identification elements (which include categories developed at Rand) provide historical and access information about the variable. They are *variable name, location, source, prior source, status, creation data, analyst, and footnotes.*

Context Elements

(Indicate how each variable fits into the overall study)[2]

SUBJECT: The set of cases to which the contents of the variable refer. (All patients, all clinicians, patients with diabetes, etc.)

TOPIC: The categories under which the contents of the study are organized[3]

MAIN TOPIC: In the MOS, the top-level study concept under which subtopics are organized (Tracer Conditions, Style of Care, Provider Characteristics, etc.)

SUBTOPICS: In MOS, the study concepts under which groups of related variables are organized (see the MOS "Concept Keys")

Content Elements

(Descriptive structure of each variable)

OBSERVER: The person supplying the values of the input variables; in a survey, the respondent category (doctors, patients)

MATTER: The objects involved in the event discussed

FUNCTION: The nature of the event, the activity or state being discussed

SPACE: Where the event occurred or the location to which it applies (physicians' offices, at home)

TIME: When the event occurred (1986, 1 month prior to screener fielding, patient's lifetime); not synonymous with data-collection date

ASPECT: The specific characteristic being observed

DOMAIN: The nature of the values, a description of the units and/or range of measurement (1=, 2=; 1,...,100)

[2] SOURCE and OBSERVATION DATA have an important bearing on the context and may become context elements for the secondary user (to be included here).

[3] LOD structure includes two extensions not covered here. One is classified description that starts at the topic level, as a guide to the levels below. The other is an extension of the topic structure to the local topic for each variable, represented by one of the content elements in the descriptive structure. For a discussion of these extensions see Dolby et al. (1986) and Clark (1985).

Identification Elements

(Access and historical information)

VARIABLE ID: An 8-character alphanumeric code name for the variable

DATA LOCATION: The data set in which the variable is stored

ANALYST: The analyst, the author of the variable

PROG SOURCE: A pointer to the program defining the variable

PRIOR SOURCE: A pointer to the input variables (e.g. SD1MI01-SD1MI05) and/or documents used to define a derived variable (MOS memos, journals)

VAR CREATN DATE: Survey date or date the program was run to create a derived variable

VAR STATUS: Administrative indicator of the variable's stage of development: PROPOSED (no data yet), PRELIMINARY, SEMI-PERMANENT, PERMANENT

FOOTNOTES:

1. Variable construction (scoring, input variables, missing data rule)

2. Analyst's comments (reason for creation, experiences with variables)

3. Documentor's comments (notes on series of/relations among variables)

4. Administrative variable noting age of template information

III Templates: elaborated definitions and working rules for raw and derived variables

Context Elements

1. SUBJECT: The set of cases to which the contents of the variable refer. (All patients, all clinicians, patients with diabetes, etc.)

   In the MOS concept outline, this is actually the highest classification level for the variables.

2. MAIN TOPIC: is a top-level survey concept under which subtopics are organized.

   A variable should be classified according to the topic and subtopic which best describe that variable. These should be chosen from an approved list.

3. SUBTOPIC: is a major survey concept; related items (variables) grouped together measure a particular concept.

   The main topic / subtopic structure defines the variable's *raison d'etre* from the study's viewpoint. For example "depression" is a subtopic of "mental health" which in turn is a subtopic of "general health". One should be careful in choosing an entry; occasionally a variable will measure something other than what appears on the surface (e.g. subtopic: Socially Desirable Response Set).

   In the MOS survey, these concepts are laid out in the Project Overview in the Figure called, "Conceptual Framework and Key Study Variables."

Once one has a working knowledge of how the various concepts fit into the study's overall organization, one may add the content elements.

Content Elements

1. OBSERVER: person through whose eyes the real-world phenomenon is viewed; in the case of a derived variable, the person who specified the value of the input variables

   In MOS, this will usually be the patient but may be the patient's physician, an MOS laboratory (for a blood test), or other.

2. MATTER: the object(s) observed; whatever objects the data apply to.

   This identifies who or what the question, derived variable, or concept is about. Matter MUST be a physical object (as opposed to a state-of-being).

   In MOS, this will be the patient, a clinician, or a patient-clinican dyad. In some cases you may encounter possessives, such a "(patient's) primary physician". For MOS variables, the stub (the objects in the sample) will always be part of matter.

3. FUNCTION: is the event happening to the object or the state-of-being of the object.

(Read this carefully) Where matter is the object to which the event is happening, function is the event itself. Where matter is the object to which the state-of-being refers, function is that state-of-being itself.

CAUTION: The word "function" is a technical word in classified description and (in a different meaning) a topic of the MOS (e.g. physical functioning, etc.). Don't confuse topics of MOS with the content categories.

4. SPACE: where the event occurred (e.g. physicians's offices, at home) or where the topic applies (e.g. continental U.S., universal).

   If there is no specific place, the answer is often "universal." Also, don't make inferences regarding space because it can be something different from what one would assume. For example, if a patient had a baby, it would be a logical assumption that the event had occurred in a hospital. This might hold true for 98% of cases, but it could also be that 1) baby was delivered at home by a midwife, or 2) baby was delivered at a clinic, or even 3) mama didn't make it to the hospital on time and baby was delivered in a taxi. The only way one could assume the birth took place in a hospital would be if one were in the hospital sampling women who had just given birth there, or if the hospital were specifically mentioned in the question.

5. TIME: when the events described occurred (e.g., 1986, the first year of participation, any time before Sept. 1, 1985); the time of the datum.

   Most often the value of TIME is the date of observation (i.e., the data are current as of the time the questionnaire is filled out). In MOS, the same instrument is sometimes fielded at multiple times. The various fieldings become discrere variables in the database and must be distinguished by their variable IDs. For raw variables, the first three characters of the name are reserved for the source instrument and fielding time (e.g. AA1DRG01, AB1DRG01).

   Sometimes TIME will differ from the date of observation. For instance, in 1986, a questionnaire might be eliciting information about 1985 income. In this case, observation date would be 1986 and TIME would be 1985.

   EXAMPLE: (for occupation or education) Time of observation
   EXAMPLE: (for 1974 income) 1974

   Sometimes the time of observation is another variable in the dataset. If so, the variable name should be given.

   Note: The archivist should check at the beginning of a series of questions for instructions regarding time. Time is frequently indicated there rather than being repeated for each question in the series.

6. ASPECT—the specific characteristic being observed.

   This should be an elaboration of matter, or function, or one of the other descriptors. Examples: frequency of, annual cost, patient's evaluation of ...)

The entry should describe the elaboration as well as that which is being elaborated (the antecedent). The antecedent specifies one of the other elements in parentheses followed by the entry for that element.

Aspect has the form "xxx of (y) zzz" where:

— "xxx" is frequently amount, degree, type or name corresponding to whether the measurement is continuous, ordinal, categorical, or an identity.

— "y" (in capital letters) is most frequently F for function or M for matter, and occasionally T for time or S for space.

— "zzz", the referent, is the entry for matter or function.

For example: "Amount of (F) Mental Health Distress", or "Subspecialty of (M) the physician."

```
"Amount of (F) Mental Health Distress"
    ^       ^   ^
  Aspect      Function
```

— DOMAIN: is a description of units and/or range of possible values.

For example: mm of mercury, Mental Health Index (MHI) scale points 0(poor)–100(good), A=excellent B=good, –2=don't know.

The data domain presumed and reported by the analyst should be distinguished from the actual domain observed in the data. Unless otherwise stated, we are referring to the domain presumed by the question.

For example, a questionnaire may have response choices that are printed on the questionnaire but are never used by any respondents in a given administration of the questionnaire. We might never observe an American Indian in the study even though we have a response category for American Indian. The existence of the possibility of such a response affects the interpretation of the rest of the choices (e.g., we are safe in assuming that Caucasian does not mean American Indian).

Likewise, a respondent may write his/her own answer in the margin of the questionnaire, and the survey coder may be tempted to treat this as a new code. He/she should not do this. Since the write–in response is not one of the given choices, it will not be chosen by a typical respondent. Consequently, the write–in response does not reflect on the choices of the typical respondent. For example, a person might write in the response "transvestite" in response to a question about gender. Some other transvestite might respond "male." So we can't make inferences about the spontaneous category since it was not considered by other respondents. The given categories are the ones that we want to record.

The stated domain will be used for automatic compatibility checks—a standard method of detecting errors in the data.

SYNTAX: The entry should be machine parasable and not a graphic picture of the answer layout. It is presumed that the description adequately represents the choices available to the respondent. Lengthy data codes (e.g., FIPS county codes) will not be listed in the template format library. Instead there will be an example of the format, then a reference to hardcopy translation.

Unlike any of the other facets, the type of value is recognizable by the syntax used. Conventions are as follows:

Syntax used:	*Examples:*
Discrete Variables	
-- Categorical responses specific sets of integers	*1=poor, 2=fair, 3=good, 4,excellent*
-- Boolean logic	*0=no/false, 1=yes/truee*
Continuous Variables	
-- Rational (grainy, measured to limited number of decimal points) NBS format	*0 (.01) 999.99 meters*
-- Integer range (special case of above)	*1(1)5 OR* *1,2,3,....500 units*
-- Real (without grain information (varying number of decimals after decimal point; usually computed)	*[1;infinity] years* *-or-* *[-inf;+inf] std units*
-- Date types, usually in the SAS form MDY date (MDY) (see SAS manual for all forms)	

Identification Elements

1. VARIABLE ID: is an 8-character alphanumeric name reserved for this variable.

   In MOS, variable IDs are limited to 8 characters in order to fit SAS programming conventions. Two or three characters are reserved to indicate the dataset to which the variable belongs.

   For raw variables, names begin with a 2-character instrument code (Moser's, see /a/p/reference/instr.codes) and a 1-field version/administration number. The remaining 5-characters are assigned by analysts. This portion of the name comes from the "measures keys" tables analysts maintain and is supposed to be unique (e.g., PP1HLT01 is the same question as SP1HLT01) within the study.

   For derived variables, the first 6-digits are the name, followed by (where appropriate) a 2-digit code indicating the source of the input variables. The 2-digit code is stored in /a/p/reference/derived.suffix.

2. DATA LOCATION: is a pointer to the dataset in which the variable is stored.

   In MOS, raw variables will be stored, by instrument, as a SAS dataset on WYLBUR. The same applies to derived variable datasets, except that they will be combined into either a patient or clinician database (not instrument specific). Details are worked out in accordance with programmer needs.

3. SOURCE: is the origin of the data—theoretically the source document(s) defining the variable.

   This may become part of the content elements at a later date, when the source is fixed as "Rand" by virtue of publication of study results.

   In the meantime it is very important to track information internally within the project. There are two senses in which this must happen:

   PROG SOURCE: is the computer program in which the variable is defined (needed to gain access to the unambiguous definition).

   PRIOR SOURCE: is the documents used to define a variable (MOS memos, papers, and previous studies). For raw variables, it is most typically an MOS memo describing the variable. For derived variables, the prior source is an original source (or sources) where the data used to construct the derived variable were defined. This will usually refer to a questionnaire or (in complex cases) several questionnaires (from the present study) and will have, tagged onto the end, citations to any previous study.

Table 1

Raw Variables	**Derived Variables**
Source—Typically, a MOS memo. The source could be a previous study if that study stands alone to define the variable.	Source—Location of programming statements used to define the Variable.
Prior Source—If a MOS memo draws from a previous study for variable definitions, reference that study, or other outside source, here.	Prior Source—The Source (and if there is one the Prior Source) from the raw variable.

SYNTAX: See APPENDIX D for specifics.

4. ANALYST: For derived data only, the name of the analyst creating the variable. The analyst is considered an author and it is the anlyst's view that produces the description. We therefore record the context appropriate to this description.

5. OBSERVATION DATE/VAR CREATN DATE: For raw data, the survey date; for derived data, the date the program was run to create a derived variable.

   The date on which a derived variable is created we call the "variable creation date" to distinguish it from the date of collection of the raw variable, the "observation date".

   Even in the case of a simple scale, the propriety of a given decision to include certain items and exclude others is apt to become dated.

   In filling out entries for observation date (and time), one pitfall the archivist needs to avoid is omitting instructions regarding a series of questions. For example, if there is a series of questions for which the patient needs to think back over the past month, (e.g., "For the next series of questions, please consider your feelings during the past month."), be sure to take that instruction into account for all the variables it applies to.

6. VARIABLE STATUS: Administrative indicator of the variable's stage of development: PROPOSED (no data yet), PRELIMINARY, SEMI-PERMANENT, and PERMANENT. The purpose of this is to encourage sharing of well-documented information early in the project.

   Only derived variables need this indicator. (Raw variables have been finalized by the time researchers see them. They had to be in order to be fielded. Thus, upon arrival of raw data, their status is known.)

Proposed item — analyst has conceived the variable. (The variable has a code or programming statements, but it hasn't been created yet.)

Preliminary item — analyst has created the variable, but the meaning can still change.

Semi-permanent item — analyst has verified that this variable measures the concept precisely. In practice, a variable is semi-permanent if the analyst does not expect to change the formula.

Permanent item — has not been revised for at least six months.

7. FOOTNOTES: contain information that does not fit neatly into one of the other prescribed categories or is too voluminous (such as the question text), or information the analyst or programmer wishes to supply explaining the origin of the variable. In documenting MOS variables, we emphasize that the latter, the origin of the variable, is important information that all too often in the past has remained in the heads of the analysts, thus depriving other users of a full understanding of the variable.

   The structure of footnotes is free-form and may be enhanced to meet whatever needs the analyst finds appropriate. However, in our experience the following entries have been considered valuable.

   a. For raw variables, the verbatim text of the survey question. For derived variables, a description of the variable construction (by means of: scoring description, input variable listing, missing data rule).

   b. Analysts comments: For raw/derived variables, a record of analysts' experiences with the variable. These should be collected by the system and appended to the documentation. For derived variables, notes regarding how/why the variable is constructed. Frequently includes the reason for the variable's creation.

   c. Documentor's comments frequently tie together related variables (e.g., This is a series of 3 utilization variables increasing in complexity: PUTIL3 is derived from PUTIL2 which in turn is derived from PUTIL1).

   d. Documentor's administrative variable indicating age of the information contained in the template and whether analyst has reviewed the template.

PART 2 - DOCUMENTATION MANUAL

I. Parallelism Among The Study's Topics/Concepts or Get To Know Your Data

The most difficult aspect of this documentation system is choosing the right level of specificity for entries in the *content* descriptors. If the level is too general, there will be no visible distinctions among large numbers of variables. If it's too specific, both documentor and readers are swamped with detail. One can arrive at the right level only by considering a set of variables.

In order to maintain a structure of parallelism with any degree of accuracy, one must identify a hierarchical summary of the study's major concepts. For MOS, this was provided in a summary table in the Project's Overview Document (Ware, 1985). The framework of a topic classification is obtained from this document

An archivist/documentor should conduct a thorough review of the data collection instruments to become familiar with the items in each. (This is especially important for MOS instruments which usually reflect several different measures.) Next, it is imperative to identify the major study concepts each item in the instruments reflects. Connecting specific items to concepts is essential in order to classify variables in a reasonably consistent fashion.

Overall Strategy

As stated above, probably the most difficult part of data classification is to choose the correct level of specificity. This corresponds to the right "discourse level" of a conversation. It is important that parallelism (of the level of specificity) be maintained when documenting.

If a question pertains to an event, it is relatively easy to choose entries: matter refers to the object(s) involved in the event and function specifies the event If the question pertains to a state-of-being, the matter is usually clear but the function is not

For example, a question such as "During the past week how often did you feel blue?" could be construed as a measure of health, a measure of mental health (a subtopic of health), a question about depression (a subtopic of mental health), or a question about "blueness." Depression would be the best function description, and "Amount of (F) Depression" would be the corresponding aspect Without the knowledge that this question really refers to depression, you might describe the function as "blueness." The other choices (mental health and health) are poorer because they lead to complicated ASPECT descriptor entries.

Varying Number Of Levels In The Hierarchy

Even with knowledge of the topic structure, there will be times when it is difficult to determine the right level of specificity. One aid that we at Rand have found helpful is to make an outline of the instrument being documented. This lays out graphically the structure below the subject/main topic/subtropic level.

We have found that sometimes it is not possible to use the best description of the variable (as described in the aspect) and still have the hierarchy "fit" logically. There is no getting around the fact that the number of levels between topics and items varies.

Consider, for example, a variable describing physician's date of graduation from medical school. It is part of a series that includes sociodemographic (e.g. age) and other education items. The items appear to be parallel in the questionnaire, but to make them parallel in the documentation, we would have to work education into the descriptors along with the additional specifier (date of graduation). If we didn't, there would be no logical connection between demographics and year of graduation (from what?). Our solution: add education as a sub-subtopic.

VARIABLE ID:	CGRADYR
DATA LOCATION:	R.R5500.A4195.CAMADV
PROG SOURCE:	/a/p/programs/ama.pgm
PRIOR SOURCE:	AMA Physician Masterfile (1985), Rand Data Facility DB 303, distributed by American Medical Association, obtained for Rand: 1986
STATUS:	Preliminary
Var CREATN DATE:	04/11/86
ANALYST:	Bill Rogers
OBSERVER:	Clinician
SUBJECT:	All Clinician's enrolled in the MOS Panel Study
TOPIC:	Clinician Characteristics
SUBTOPIC:	Demographics/Socioeconomic status
SUB-SUBTOPIC:	Education
MATTER:	Clinician
FUNCTION:	Graduation
ASPECT:	Year of (F) Graduation
SPACE:	Universal
TIME:	Universal
DOMAIN:	Date (MDY)
FOOTNOTES:	(1) SCORING: Raw item recode INPUT VARIABLES: BS1GRDYR MISSING DATA RULE: No imputation applied (4) 10/15/86 Template reviewed by Bill Rogers

II. The Technicalities Of Making Templates

We have found that the best technique for documenting a large dataset is to fill in as many common entries (e.g. data location) as possible, then copy the template using a full-screen editor from variable to variable and fill in the entries as required. Most raw data entries change little from variable to variable. For both raw and derived data this saves repetitive typing.

Strictly speaking, the term "templating" describes this activity.

One should distinguish between input format and display format. The input format is designed for ease of entry and its ability to be manipulated by string processing programs such as "awk". This is the input format:

Unreported Template

In its unreported form, the template looks like:

```
VARIABLE ID:
DATA LOCATION:
SOURCE:
PRIOR SOURCE:
STATUS:                 (derived only)
OBS DATE:               (or VAR CREATN DATE for derived variables)
SUBJECT:
TOPIC:
SUBTOPIC:
MATTER:
FUNCTION:
SPACE:
TIME:
ASPECT:
DOMAIN:
FOOTNOTES:
```

Notice that there is a difference between raw and derived variable templates. The derived variables template has three more elements than the raw template—STATUS, ANALYST and PROG(ram) SOURCE.

Syntax Conventions

Certain syntax conventions must be maintained so that the data may be parsed (separated and read) by a computer program which will report the data in various forms.

The syntax conventions are:

1. Type in entries after the colon.

2. The text may wrap around to the next line, but should not start in column 1.

3. Leave a blank line between variables.

4. Separate fields by commas (there may be more than one entry per facet).

5. Use special syntax for a particular facet where specified, e.g., for domain

6. Type dates as: mm/dd/yy

7. Orthography requirements:

a. Use the same capitalization and punctuation for entries as one would in English (e.g. capitalize important words).

b. For names, use the standard conventions, e.g. /a/p/programs/ama.pgm, R.R550.A4195,CAMADV, etc.

c. Capitalize all letters in names of variables.

8. If there is no entry for a facet, type "None", so we'll know it has been considered.

9. When there is more than one entry for a facet, separate them by semi-colons.

. *Dealing With Uncertainties*

If you don't know an entry at all, DON'T GUESS. Leave it blank if you are completely baffled and ask for help. If you have a preference but are unsure, start with "(?)". Classifications may be revised later by the analyst who invented the questions or variable (and who presumably has a better perspective on the intent).

Special Instructions For Derived Variables

When making derived variable templates note the name of the input raw variables (in parentheses) wherever possible, For example, note the variable names in parenthesis in the following template:

VARIABLE ID:	PELIGGRP
DATA LOCATION:	R.R5500.A4195.PDERIVED.SAS
PROG SOURCE:	/a/p/programs/derived5
PRIOR SOURCE:	MOS memo123
VAR CREATN DATE:	Screener date (PSCRND)
ANALYST:	Bill Rogers
OBSERVER:	Patient, Clinician
SUBJECT:	All Screened Patients
TOPIC:	Survey Administration
SUBTOPIC:	MOS Panel Study
MATTER:	Patient
FUNCTION:	Eligibility
SPACE:	Universal
TIME:	Screener date (PSCRND)
ASPECT:	Eligibility of (M) patient
DOMAIN:	0 = No hypertension; 1 = Hypertension
FOOTNOTES:	(1) SCORING: Look at both patient's, and clinician's reports of hypertension; INPUT VARIABLES: SP1HYPO1, SD1HYPO2 MISSING DATA RULE: Recoded to missing if either input variable value is missing.

(4) 10/15/86 Template reviewed by Bill Rogers

Also, please notice that derived variables frequently have more than one entry for certain facets. For example, a derived variable may combine the viewpoint of both the clinician and the patient, so both would be entries for the observer facet.

III. Practice and Initiative--Raw Variables

Now we will make a first attempt at descriptive classification at the raw variable level. Specifically, how are the elements connected to reality?

Step- By- Step Raw Variable Classification Example

Helpful Hints

Let's begin by looking at a set of requirements for templating, which need to be kept under consideration in order for the pieces of the puzzle to fit. For example, consider the following issues, in order:

- - the aspect must originate from (or reflect) the domain
- - the aspect must refer back to either matter, function, (or in a few rare cases, to time)
- - function must directly connect with subtopic (or sub-subtopic)
- - within MOS, subtopic and topic fit in with the overview concepts list (Overview, Conceptual Framework and Key study variables) with topic above it.

Keeping these requirements in mind, in the following order, let's determine the entries for MATTER, DOMAIN, ASPECT, FUNCTION, TOPIC, SUBTOPIC, SUB-SUBTOPIC (where necessary), and then the remaining elements.

Specific Example

For a detailed example, let's consider a question that asks: "How many different drugs are you taking for your high blood pressure?" Document the question by doing the following:

1st) Record matter:

MATTER: Patient

Choose patient because we know the question is asked in regards to the patient.

2nd) Record domain:

DOMAIN: 1=none, 2=one, 3=two or more

This is taken directly from the questionnaire. We do this right away since we know that the aspect has to reflect the nature or quality of the response set.

3rd) Record the first half of aspect:

ASPECT: Number of ....something

Choose this since all of the above in domain indicate a specific quantity, a number.

4th) Record the second half of aspect — the antecedent (the thing the aspect refers to):

ASPECT: Number of (F) Drugs taken for Hypertension

"Drugs taken" must be the answer because that is what "Number of" refers to. We complete the phrase by adding "for Hypertension" since we know that is true and relevant here.

Notice that aspect can be very wordy. It is the element that pulls the other elements together.

Sometimes the antecedent refers to another element, commonly the matter. We know this is not the case here because "Number of (M) patient" is not only incorrect, but also illogical.

5th) Record function:

FUNCTION: Drug Usage

Since the second part of the Aspect must refer back to another element (usually matter or function) and the value of matter is already "patient", function is almost predestined to be Drug Usage (a more formal way of saying "taking drugs").

Now one might consider filling in some of the higher-level entries which indicate how this variable fits into the overall scheme of the study, making any adjustments needed along the way to make the lower- and higher-level entries meet in a reasonable, logical fashion.

6th) Record main topic (highest major study concept):

MAIN TOPIC: Patient Characteristics

This variable could have been created to measure Utilization under the topic of "Process of Care", but the analyst creating the variable is the MOS M.D. who is responsible for describing the patient, so "Patient Characteristics" is the best choice.

This variable can also occur as a patient outcome measure but because it is asked at the beginning of the study, we know it is part of initial "Patient Characteristics". Toward the end of the study, the documentor should check with the analyst to determine whether or not the variable has been repeated, and if so, add "Patient Outcomes of Care" as a second entry for this descriptor element.

7th) Record subtopic (major study concept):

SUBTOPIC: Disease Severity

The only way to know for sure which subtopic is correct is to personally ask the analyst what his/her intent is in asking the question. This entry could have easily be mistaken for a measure of utilization instead of disease severity.

8th) Create sub–subtopic to make a logical connection between SUBTOPIC and FUNCTION:

SUB–SUBTOPIC: Hypertension

As it stands right now, the distinction between "Disease Severity" and "Drug Usage" is too weak to stand by itself and does not properly reflect the content of the question. Here is where we need to make adjustments by adding another level of specificity to the template. For the SUB–SUBTOPIC, choose hypertension since that is what the drugs are taken for and how the analyst is measuring the disease severity. Now the link between these three levels is very logical and clear to all analysts.

9th) SUBJECT—identify the pool of people to whom the variable applies

SUB–SUBTOPIC: 1/2 of screening patients

We choose the above since we know that 1/2 the patients screened filled out SP2 and the other half filled out SP1.

Now that the hard part is done, let's round things out by filling in the easier, more obvious entries:

10th) LOCATION — R.R5500.A4195.PDERIVED.SAS

dataset in which variable data resides; supplied by analyst/programmer

10th) VARIABLE ID — SP2PHYPO5

supplied by programmer or analyst

11th) LOCATION — /a/p/datasets/SP2

dateset where variable data resides: supplied by programmer

12th) PROG SOURCE — /a/p/programs/pdv.part1

13th) PRIOR SOURCE — Shelly Greenfield, "Determining Disease Severity", MOS memo 515, 10/10/85

MOS memo cited first, follows; see Appendix D for citation conventions

This could also be a journal article, another research study or for MOS, the raw variable documentation the questionnaire designers created, called "MOS Concept Keys".

14th) OBSERVER — Patient

We know this becasue it is the patient who is filling out the form

15th) OBS DATE — Screener Week

This is the date of data collection.

16th) SPACE — Universal

Number of drugs taken is constant regardless of where the respondent is.

17th) TIME — Screening Date

The time frame is the same as the date of observation by default since the question is asked in the present tense.

Sometimes the archivist has to look at the beginning of a series of questions for directions on time.

18th) FOOTNOTES — (1) How many different drugs are you taking for your high blood pressure?

Question text goes here (no graphics).

(2) Analyst assumes that worse cases require more drugs.

Analyst assumptions are important information that the documentor should be careful to record.

(4) 10/15/86 Template reviewed by Steve Rein

An Administrative variable indicating the age of the information in the template is goes here.

As mentioned above, there is not always a single right answer for an entry. However, you should apply a uniform viewpoint across the questions so that the dimensions are consistently used. Parallelism is very important.

The resulting template is:

VARIABLE ID:	SP2HYPO5
DATA LOCATION:	/a/p/datasets/SP2
PROG SOURCE:	/a/p/programs/pdv.part1
PRIOR SOURCE:	MOS memo 515
OBSERVER:	Patient
OBSERVATION DATE:	Screener week
SUBJECT:	1/2 of screening patients
TOPIC:	Patient Characteristic
SUBTOPIC:	Disease severity
SUB-SUBTOPIC:	Hypertension
MATTER:	Patient
FUNCTION:	Drug Usage
SPACE:	Universal
TIME:	Screener week
ASPECT:	Number of (F) drugs taken for hypertension
DOMAIN:	1 = None; 2 = One; 3 = Two or more
FOOTNOTES:	(1) How many different drugs are you taking for your high blood pressure? (3) Assumes that worse cases require more drugs to control hypertension. (4) 10/15/86 Template reviewed by Steve Rein

For a graphic example of what happens when template entries are filled in without the analyst's intent being verified, see Appendix C.

IV. Practice and Initiation--Derived Variables

Next let's learn how descriptive documentation works at the derived variable level. To understand this one needs to know how:

a) a derived variable is defined (compared with raw variables);
b) item-coded derived variables are similar to raw variables;
c) to make a derived template;
d) derived variables have special documentation concerns
e) a "status" line indicates variable stability
f) derived variables develop in the analysis process

A. Definition of Derived Variables

A derived variable is one that:

1. is not present in raw data and has been created by some kind of computation.

2. is based on an assumption made by the analyst with regards to the meaning of the variable, and is derived from the data plus the analyst's logic.

Raw data from an outside source are considered derived data for the purposes of the study being documented because the data were not generated by the home study instrument.

Derived variables vary greatly in their complexity. Simpler derivations are more numerous and require less sophisticated handling. Conversely, more complex derivations occur less frequently and require sophisticated handling.

Let's look at four cases of derivation, from most simple to most complex:

1. ITEM RECODING has two major functions:

   a. to transform the data to a more computational form (e.g. reverse scaling to align with other variables, change alphabetic characters to numeric, etc.)

   b. reduce error response (e.g., resolve multiple punches, data inconsistencies, out-of-range values, etc.)

   Commonly, simple derivation of variables is tagged onto the raw variable data processing routine.

   While these two actions are both considered simple item recoding, one is a ***substantive*** change and the other is a change in ***form***. They need to be treated in distinct ways. See "Item Recoding and Similarities to Raw Variables" below for more information.

2. SCALES are a product of a systematic combination of individual item scores into a summary score[4].

3. COMPLEX COMPUTATIONS involve formulae that use substitution in the case of missing data (i.e., if the date of a visit is missing from a patient file, it may be picked up from the clinician file).

4. SUPER COMPLEX COMPUTATIONS involve predicting values and making reference to multiple observations in a file. They are so complex because the predicted value depends on the values of the rest of the dataset and sometimes requires use of regressions[5] (e.g., predicting age from income).

B. Item Recoding And Similarities To Raw Variables

A recoded item is harder to classify than other derived variable types because it changes variables in two different ways: 1) ***substantive*** changes and 2) ***form*** changes. Both of these are called item recoding because they are both relatively simple changes, which involve only one variable, limited programming (usually a single line) and assumptions about the variable.

[4] A scaled response is different from a scaled variable, and refers to a graded response set (e.g., excellent, good, fair, poor)

[5] A regression is a model which explains the behavior of some data given the behavior of other data.

The following situation is an example of a ***substantive*** change. If the analyst wants to make assumptions about response possibilities in order to limit errors, he may define age as 1) having the upper bounds of 100 (age value = 0–100), and 2) having the formula (PSCRND – PBIRTHD)/365. This would get rid of many respondent errors (e.g., all 4–digit numbers, alphabetic characters, etc.). It also introduces an assumption that no one was over 100 years old and that there were no decade errors, that people did not write in their age, etc.

On the other hand, *form* changes are actually disguised raw variables. Since the essence of the information is the same, and only the way that information is recorded is different, form–changed variables are very similar to raw variables. Mostly, this type of re–recording of information is done to standardize a measurement (e.g. make all scales conform to values of 0–100, or reverse the direction of a single scale to alighn with the other scales) for use in a scaled derived variable.

Substance–changed variables call for different classification than form–changed variables. While the analysts process these two actions at the same time, we need to document them differently. Recoded variables with no substantive changes need to be templated as raw variables and substantively changed variables need to be templated as derived variables.

C. Creating Derived Variable Templates

An archivist needs to keep the above differences in mind in order to choose whether to use a raw variable template or derived variable template for documentation. (See APPENDIX B for examples of raw and derived variables). The raw template is so similar to that of the lowest–level (item recoding) derived template, that for convenience sake, they will henceforth be referred to together. In order to make a derived template one may start with a raw variable template, and make the following changes:

a) Add PROG SOURCE, STATUS, and ANALYST to the template.
b) Change OBSERVATION DATE to VAR CREATION DATE.
c) Record variable construction information (SCORING, INPUT VARIABLES, and MISSING DATA RULE) in the footnotes.

D. Special Concerns Of Derived Data Documentation

At the data processing level, a major issue in archiving derived datasets is keeping the data in a similar stage of processing. Analysts sometimes lose interest mid–stream which: 1) ignores data which can be different from those already analysed and can cause significant variations in statistical outcomes 2) creates unevenly processed data. An archivist, in tracking the data, should look out for and note this problem for both the analysts' benefit and that of future users.

Another problem specific to the documentation of derived variables concerns capturing data in the most accurate state. For raw data, that may be either (1) when the questionnaire author writes the questionnaire, or (2) when analysis occurs and the question is re–read with a more critical eye. Interesting facts about the raw data may come to light as derived variables are created.

For derived data, there is a similar problem. The concept being derived from a given set of data may evolve over time, or the intensity of knowledge may disappear with time (people may forget assumptions made at the beginning of data collection), leading to mistaken classifications. The problem is reflected in processing: programs may be written which refer to data in one state or another, and it is difficult to tell whether the program should refer to older data or revised data.

In general, derived variables are in flux longer than raw variables and there is no equivalent to the questionnaire text to finalize the meaning of derived variables. Raw variables are fixed as of the date of fielding, whereas derived variables are always subject to change, making them more elusive to understand and document accurately.

E. Status Line To Indicate Variable Stability

For the MOS study, we decided to create an indicator to control the state of flux derived variables are so frequently in. The purpose of the "STATUS" indicator is to flag the underdeveloped derived variables so that more current meanings will be recorded later. To learn to choose which value the STATUS line will be, let's look at the analytic process and how derived variables are created.

F. Derived Variable Development In The Analytic Process

In social science surveys, *measures* are the instruments or tools used to represent concepts.

Typical stages of development measures go through are:

a) Analytic history and technical literature illustrate what research has been done in other projects.

1) *Proposal* stage - these are ideas about *what* should be measured, and thoughts about *how* those ideas should be measured; in the MOS these are spelled out in the Overview of project and in the myriad memos on measures.)

2) *Preliminary* stage - concepts are broken down into specific items and values (item wordings and groupings); both of the latter are subject to change. Tentative programming code is written and run, creating the first version of the variable.

3) *Semi-permanent* stage - items accepted, values subject to change, adjustments made to computations. This is the point at which the variable becomes static. It has been subjected to and passed measurement analysis. (For MOS variables, this would typically be standard psycho-metric criteria.)

4) *Permanent* stage - Not only is the variable static, but no one has even looked at it for six months. It has been computed for all cases in the study database.

A sample application of this process is:
1) Researcher proposes studying income; writes measures memo
2) Economist says wages is a more appropriate measure than income; writes formula for wages
3) After reviewing the first version of the variable, researcher adjusts definition of wages or makes corrections to programming
4) No further changes made to data or definition

When documenting MOS variables, we found the "PROPOSED" value the least valuable since variables at this stage are often too tentative to document. The other indicators are valuable as a mechanism for finalizing a variable's definition with the most recent information.

V. Step-By-STEP DERIVED VARIABLE CLASSIFICATION EXAMPLE

Let's keep in mind again, some criteria that need to be met in the course of templating:

— the aspect must originate from the domain
— the aspect must refer back to either matter, function, (or in a few rare cases, to time)
— function must be directly linked to subtopic
— within MOS, subtopic and topic fit in the overview concepts list (Overview, Conceptual Framework and Key study variables)

So, let us determine, first, matter, domain, aspect, function, topic, and subtopic, then the remaining elements, in that order.

For a detailed example, let us consider a question that asks: "What was your household income, before taxes, in 1985?" Document by doing the following:

1st) Record matter:

MATTER: Patient

Choose patient because we know that is who the question is about.

2nd) Record domain:

DOMAIN: [1:infinity]

This is the range specified in the questionnaire response set.

3rd) Record the first part of aspect;

ASPECT: Log of ... something

Choose this since the numbers reported in the domain are being reformatted into a log for each income level. Archivist wouldn't necessarily know this and may have to clarify when interviewing analyst about derived variable meaning.

4th) Record the second half of the aspect — the antecedent
(the thing the aspect refers to):
ASPECT: Log of (F) Income

This can be surmised by default since a "log of patient" is nonsense.

5th) Record function:

FUNCTION: Income

Choose income because we are talking about that particular characteristic of the patient and this was already determined by the antecedent of the Aspect.

6th) Record topic (top-level study concept):

TOPIC: Patient characteristics

This measure is a patient characteristic that is part of the initial description of the patient.

7th) Record subtopic (major study concept):

SUBTOPIC: Demographics/socioeconomic status

This subtopic is fairly obvious. Most researchers consider income a demographic measure.

8th) SUBJECT—identify the pool of people to whom the variable applies

SUB-SUBTOPIC: All screening patients

We choose the above since we know that this question was asked on both forms of the screener questionnaire.

Now to round things out, let's fill in the easier, more obvious entries:

9th) VARIABLE ID — PINCLOG

supplied by analyst, prefaced with a "P" to note a patient derived variable.

10th) LOCATION — R.R5500.A4195.PDERIVED.SAS

dataset in which variable data resides; supplied by analyst/programmer

11th) PROG SOURCE — /a/p/programs/derived5

location and name of program that created this derived variable

12th) PRIOR SOURCE — MOS memo cited first, a prior study's definition (e.g., HIE) follows; see APPENDIX D for citation conventions

13th) STATUS — Preliminary

Choose this entry since this is the first pass at the variable; its definition will probably be adjusted somewhat before it achieves semi-permanent status.

14th) VAR CREATN DATE — 10/10/85

This is the date the program ran that created the variable.

15th) ANALYST — Ron Hays

This is the analyst who created the variable; often a programmer's name is added afterwards.

16th) OBSERVER — Patient

The analyst is now the person whose view is being presented.

17th) SPACE — Universal

The respondent's income, which is the basis for this derived variable, will be the same regardless of where the respondent is

18th) TIME — 1985

The date of the raw variable is 1985 and there is nothing in the derived formula that alters that fact.

19th) FOOTNOTES — (1) SCORING: Log income (PPIPAT10):
log (minimum value + maximum value)/2, or
log 100000 INPUT VARIABLES: PPIPAT10
MISSING DATA RULE: No missing value imputations
(2) Var used to initially describe study patients
(3) One of three (PINC, PADJINC) income variables
(4) 10/15/86 Template reviewed by Ron Hays

(1) Variable construction description
(2) Analyst comments/reason for variable creation
(3) Other variables in the series
(4) Recency of template information

The resulting template:

VARIABLE ID:	PINCLOG
DATA LOCATION:	R.R5500.A4195.PDERIVED.SAS
PROG SOURCE:	/a/p/programs/derived5
PRIOR SOURCE:	MOS memo510, RAND HIE study
STATUS:	Preliminary

VAR CREATN DATE:	10/10/85
ANALYST:	Ron Hays
OBSERVER:	Patient
SUBJECT:	All screening patients
TOPIC:	Patient characteristics
SUBTOPIC:	Demographics/Socioeconomic status
MATTER:	Patient
FUNCTION:	Income
SPACE:	Universal
TIME:	1985
ASPECT:	Log of (F) Income
DOMAIN:	[1:infinity]
FOOTNOTES:	(1) SCORING: log income: log (lower limit + upper limit)/2, or log 100000 INPUT VARIABLES: PPIPAT10 MISSING DATA RULE: No imputation for missing values (2) Var used to initially describe study patients (3) One of three (PINC, PADJINC) income variables (4) 10/15/86 Template reviewed by Ron Hays

PART 3 - REPORTING AND MANIPULATING TEMPLATE INFORMATION

I. OVERVIEW

Resistence To Template Documentation

A critical question a documentor must ask him/herself after the templates have been so thoughtfully and laboriously created, is how useful the templates are. First experiences with MOS staff illustrated significant resistance to the templates. Even the most open-minded analyst called template terminology "voodoo mumbo-jumbo" and the number of facets grouped into one body was considered to make them difficult to read.

Overcoming Resistance

At Rand, we dealt with the first problem simply by teaching the analysts the terminology and meaning of the variables (since these are based on years of work by other individuals in the Language of Data group, we felt it our duty to preserve the labels they used for their descriptive elements), and renaming other less essential facets to fit MOS vernacular.

A Report Aa A Solution To Readability

As for the question of readability, we devised a report format that was much better suited to visual inspection. This benefited not only the primary analyst, but also satisfied criteria for comprehension of the public codebook.

A Report As A Means Of Augmenting Template Information With Information From Other Sources

Besides readability, the report allowed us to incorporate other sources of variable information into the codebook. For example, an analyst likes to see what raw variables comprise a derived variable. An easy way to produce such information is to copy a brief description of the variable from a data dictionary, rather than repetitiously retype the text. Another important source of information about the variable is statistics. This will be further discussed below.

II. Adding Dynamic Information (Statistics) To The Report

By now one would think that everything one could possibly want to know about the variables has been recorded. Not true. There is a category of dynamic information that we do not record in the template, but that study analysts need. Mainly, these are descriptive statistics. Because this information often changes as the dataset becomes more complete we do not update it regularly, but find it more practical to include it just before it will be viewed — in the reporting stage.

Reported Statistics

The statistics we include in our report are: frequencies — including percent, cumulative frequencies, cumulative percents, and means — including mean, standard deviation, minimum value, and maximum value.

Another element included in the statistics section is a reliability indicator. Reliability is a correlation of the true value of a variable and the analyst's measurement of the variable. If these two values are exactly the same (too idealistic to occur frequently) the correlation would be "1." A correlation of about .82 is more likely; most correlations between true and measured variables are between 0 and 1.

Contextual Information On Statistics

MOS analysts also found information regarding the creation of these figures most helpful so we included the date these statistics were created, the number of observations in the input dataset and the percentage of the dataset represented by missing observations. Information on the point at which an observation becomes missing (e.g., if 4 responses of the 8 input variables are missing, the derived variable is not calculated) is stored in the program creating the derived variable.

III. SAMPLE REPORT

```
VARIABLE ID:          PCSTYLE
STUDY CONTEXT INFO

SUBJECT:              All Panel patients
TOPIC:                Style of Care
SUBTOPIC:             Interpersonal Style
SUB-SUBTOPIC:         Level of Patient Participation
VARIABLE MEANING      OBSERVER:
                      Patient
MATTER:               Patient/Clinician dyad
FUNCTION:             Clinician's willingness to share responsibility
SPACE:                Universal
TIME:                 Screening Date (PSCRND (PATSCRND)
ASPECT:               Patient's Opinion of (F) Clinician's willingness to share responsibility
DOMAIN:               [0;100=Dyad shares responsibility]
VARIABLE CONSTRUCTION SCORING:
                      Average of available items that have been recoded (where necessary) to fit
                      direction of scale; final scale transformed to 0-100
INPUT VARIABLES:      SP1PST01 doctor CHOICE, ask to help make decision?
                      SP1PST02 doctor INFOEX-answer questions politely?
COMMENTS:             1st in a series of 5 delineating...
STATISTICS

MEANS:                MEAN:    83.853   MINIM:    0.00    OBS:      10346
  DATE:               STD:     15.855   MAXIM:  100.00    % MISS:

FREQUENCIES:          VALUES:  FREQS    PERCNT:  CUM FREQ:  CUM PRCT:
 DATE:                   1      10        25        10         25
 OBS:                    2      10        25        20         50
 MISSING:                3      10        25        30         75
                         4      10        25        40        100

EXTERNAL INFORMATION  DATA LOCATION:
                        R.R0062.A4195.PDER2 SASNAME: PDERIVED
PROGRAM SOURCE:         /ma/p/programs/pdv.part2 (SP1 form only)
PRIOR SOURCE:           Memo714, memo664, memo644, memo643,
VARIABLE STATUS:        Semi-permanent
VAR CREATION DATE:      6/86
ANALYST:                Sherrie Kaplan, Ron Hays
ADMIN INDICATOR:        09/22/86 template reviewed by Ron Hays
```

IV. POSSIBILITIES FOR MANIPULATING TEMPLATE INFORMATION

We have been considering several options for implementing this system on a minicomputer or networked microcomputers. It is important that the computers be networked in order that information developed by one user can be easily shared by others.

There are several parts to the proposed system.

1. A display manager.

   A. This could be a text editor, possibly programmed to do interesting things with function keys. The pages would be formatted so that editor Page–Up and Page–Down functions would carry the user from one part of the documentation to another.

   B. A special–purpose viewer that could arrange and compare fields. The Language of Data's Descriptor Manipulator (Franzen, 1986) was an example of this.

   C. A multi–level display manager that worked through several levels: (a) the topic tree, (b) sorted lists of variables meeting given criteria, (c) an elaborated display of information about a given variable, and (d) a text editor for entering user comments.

2. A display generator.

   LOD experience suggests that the physical format of the documentation has a considerable effect on the ability of the user to find the correct information. Our data collection format, while efficient for data entry, does not particularly lend itself to casual viewing because there are too many elements and they all start on the left. Utility programs should be written to convert to more suitable viewing formats.

3. Acquisition of Statistics.

   Utility programs are required to gather the necessary statistics and insert them in appropriate spots. We anticipate unix scripts, or the equivalent, to do this.

4. Database Management.

   The documentation is itself a database, and needs to be accessed as such. Utility programs are required to convert from the input text format to database format. Retrieval operators are also needed for standard relational database purposes.

5. Program Generation

   A system of this kind should help assemble user programs to retrieve selected data from the database.

APPENDIX A

QUICK TEMPLATE REFERENCE

RAW VARIABLE TEMPLATE:

VARIABLE ID: See /a/p/reference/instr.codes for conventions
DATA LOCATION: where the dataset that contains the variable resides
PROG SOURCE: where the program that creates the variable resides
PRIOR SOURCE: MOS memo, bibliographic citation, another study
OBSERVER: person answering questionnaire; respondent
OBSERVATION DATE: (variable containing) date instrument fielding
SUBJECT: Pool of people to whom the variable applies
TOPIC: from MOS Conceptual Framework or MOS Concept Keys
SUBTOPIC: from MOS Conceptual Framework or MOS Concept Keys
SUB-SUBTOPIC: from MOS Conceptual Framework or MOS Concept Keys
MATTER: object under discussion (patient, clinician, visit
FUNCTION: event happening to matter or matter's state-of-being
SPACE: where event took place; For N/A use: Universal
TIME: time the var refers to (1984 income, hlth last week)
ASPECT: specific character of matter of function
DOMAIN: possible value range: ex: yes/no, 1-4, etc
FOOTNOTES: (1) Question text
(2) Analyst Comments/experiences re: variable
(3) Series information here
(4) Template indicator noting recency of information

DERIVED VARIABLE TEMPLATE:

VARIABLE ID: See /a/p/reference/derived.suffix for conventions
DATA LOCATION: where data resides on the computer
PROG SOURCE: source [programs] for data creation
PRIOR SOURCE: MOS memo, bibliographic citation, another study
STATUS: PROPOSED, PRELIMIARY, SEMI-PERMANENT OR PERMANENT;
VAR CREATN DATE: date of dataset creation or MOS memo defining var
ANALYST: Main analyst(s), programmer
OBSERVER: specify the original observer's name here
SUBJECT: Pool of people to whom the variable applies
TOPIC: from MOS Conceptual Framework or MOS Concept Keys
SUBTOPIC: from MOS Conceptual Framework or MOS Concept Keys
MATTER: object under discussion (patient, clinician, visit)
FUNCTION: event happening to matter or matter's state-of-being
SPACE: where event took place; For N/A use: Universal
TIME: time the var refers to (1984 income, hlth last week)
ASPECT: specify character of function or matter

DOMAIN: possible value range; ex: yes/no, 1–4, etc.

FOOTNOTES:
(1) Variab!e construction (SCORING, INPUT VARS, MISSING DATA RULE)
(2) Analyst Comments (assumptions, experiences, reason for creating the variable)
(3) Series information here
(4) Template indicator noting recency of information

APPENDIX B

A RAW VARIABLE TEMPLATE VS. A DERIVED VARIABLE TEMPLATE

VARIABLE ID:	SP4INC10
DATA LOCATION:	/a/p/datasets/SP4
SOURCE:	MOS memo315,
PRIOR SOURCE:	RAND HIE study
OBSERVATION DATE:	Screener week10/10/85
ANALYST:	Ron Hays
OBSERVER:	Patient
SUBJECT:	All screening patients
TOPIC:	Patient characteristics
SUBTOPIC:	Demographics/Socioeconomic status
MATTER:	Patient
FUNCTION:	Income
SPACE:	Universal
TIME:	1985
ASPECT:	Amount of (F) Income
DOMAIN:	1 = $0 to $4999,
	2 = $5000 to $9999,
	3 = $10000 to $14999,
	4 = $15000 to $19999,
	5 = $20000 to $29999,
	6 = $30000 to $49999,
	7 = $50000 to $99999,
	8 = $100000 and up
FOOTNOTES:	(1) What was your total household income before taxes in 1985?
	(2) Income categories are used for the response set instead of asking a specific amount since some people consider the latter an invasion of privacy
	(4) 10/15/86 Template reviewed by Ron Hays

VARIABLE ID:	PINCLOG
LOCATION:	R.R5500.A4195.PDERIVED.SAS
SOURCE:	/a/p/programs/derived5
PRIOR SOURCE:	MOS memo no. 315, RAND HIE study
STATUS:	Prelimiary
VAR CREATN DATE:	10/10/85
ANALYST:	Anita Steward
OBSERVER:	Patient
SUBJECT:	All screening patients
TOPIC:	Patient characteristics
SUBTOPIC:	Demographics
MATTER:	Patient
FUNCTION:	Income

SPACE:	Universal
TIME:	1985
ASPECT:	Log of (F) Income
DOMAIN:	[1:infinity]
FOOTNOTES:	(1) SCORING: log income: log (lower limit + upper limit)/2, or log 10000 INPUT VARIABLES: PPIPAT10 MISSING DATA RULE: No imputation for missing values (2) Var used to initially describe study patients (3) One of three (PINC, PADJINC) income variables (4) 10/15/86 Template reviewed by Ron Hays

APPENDIX C

DESCRIPTIVE ELEMENTS ENTRY PRIORITIES

It is helpful to keep the following issues in the following order:

- the aspect must originate from the domain
the aspect must refer back to either matter, function, or time
function must directly connect with with subtopic (or sub–subtopic)
within the MOS, subtopic must fit in with the overview concepts list

Overview, Conceptual Framework and Key study variables

Below is an example of what can happen when the above is not considered and/or whether the analyst is consulted about the intent of the variable.

e e	Initial Classification	Classification following order
VARIABLE ID:	SP2HYPO5	SP2HYPO5
DATA LOCATION:	S.S8587.A4195.SP2SAS	S.S8587.A4195.SP2SAS
SOURCE:	/a/p/programs/pdv.part1	/a/p/programs/pdv.part1
PRIOR SOURCE:	MOS Memo here	MOS Memo here
OBSERVER:	Screener week	Screener week
OBSERVATION DATE:	Patient	Patient
SUBJECT:	1/2 of screening patients	1/2 of screening patients
	********************	****************************
TOPIC:	* Process of Care *	* Patient characteristics *
SUBTOPIC:	* Treatments *	* Disease Severity *
	********************	****************************
SUB–SUBTOPIC:	(None)	Hypertension
MATTER:	Patient	Patient
	*****************	***************
FUNCTION:	* Hypertension *	* Drug Usage *
	*****************	***************
SPACE:	Universal	Universal
TIME:	Current	Current
	*****************************	******************************
ASPECT:	* Treatment of (F) Hyper– *	* Number of (F) Drugs taken *
	* tension *	* for Hypertension *
	*****************************	******************************
DOMAIN:	1=None, 2=One, 3=Two or more	1=None, 2=One, 3=Two or more

FOOTNOTES:	(1) How many different drugs are you taking for your high blood pressure?	(1) How many different drugs are you taking for your high blood pressure? (3) Assumes that worse cases require more drugs to control hypertension.

In the initial classification, because of the text of the question, the archivist assumed that SP2HYPO2 is about Treatment and Process of Care. However, after verifying with the analyst, the archivist found that SP2HYPO2 is intended to measure patient's disease severity.

APPENDIX D

BIBLIOGRAPHIC CITATIONS

Rand standard bibliographic citations are and memo citations should be in the following form:

MEMOS	REPORTS	JOURNALS	BOOKS
author	author	author	author
title quotes	title quotes	title quotes	title quotes
memo num	report num	volume num	
	publisher/company	publisher/company	publisher
	place	place	place
date	date	date	date
page numbers	page numbers	page numbers	page numbers

In keeping with general syntax, each citation should be separated by commas. Therefore, any commas e.g., ones separating author, title, etc. in the citation should be replaced by semi-colons.

Another type of reference is for datasets, e.g. AMA data. A format could be:

author company
title and date yr of data collection
Rand Data facility database number
publisher/distributor
date data was published

REFERENCES

Dolby, J.L.; Clark, N., 1982: *The Language of Data.* San Jose State University Foundation.

Dolby, J.L., 1983: Meaning from Data, *Proceedings of the AAAS*, May 1983.

Dolby, J.L.; Clark, N.; Rogers, W.H., 1986: The Language of Data: A General Theory of Data, *Proceedings, Eighteenth Symposium on the Interface*, American Statistical Association.

Clark, N., 1985: *Classification Procedures: Classification from Survey Instruments.* Technical Report, San Jose State University Department of Mathematics and Computer Science. For copies contact Nancy Clark, Box R, Sausalito, CA 94966.

Clark, N., 1987: Tables and Graphs as a Form of Exposition. *Scholarly Publishing.* October 1987, in press.

Ware, J. et al, 1985: *A National Study of Medical Care Outcomes: Project Overview.* Contact John Ware, The Rand Corporation, Santa Monica, CA 90406 for copies.