COM 74-10700

MANAGEMENT OF DATA ELEMENTS IN INFORMATION PROCESSING

U.S. DEPARTMENT OF COMMERCE / National Bureau of Standards

FIRST NATIONAL SYMPOSIUM
NATIONAL BUREAU OF STANDARDS
GAITHERSBURG, MARYLAND
1974 JANUARY 24-25

Available by purchase from the

National Technical Information Service,

5285 Port Royal Road, Springfield, Va. 22151.

Price: $9.75 hardcopy; $1.45 microfiche.

COM 74-10700

151

Management of Data Elements in Information Processing

Proceedings of a Symposium Sponsored by the American National Standards Institute and by The National Bureau of Standards

1974 January 24-25,
NBS, Gaithersburg, Maryland

Hazel E. McEwen, Editor

Institute for Computer Sciences and Technology
National Bureau of Standards
Washington, D.C. 20234

U.S. DEPARTMENT OF COMMERCE, Frederick B. Dent, *Secretary*
NATIONAL BUREAU OF STANDARDS, Richard W. Roberts, *Director*

Issued 1974 April

Table of Contents[1]

[1]Papers are sequenced by name of author.

Abstract

Recent technological advances in computers and communications make possible the integration of data systems and the exchange of data among them on an expanding scale. However, the full effect of these advances cannot be realized unless the need for uniform understanding of the common information (data elements) and their expression in data systems is recognized and a means provided to effectively manage this information. The increasing interrelationships among the data systems of Federal, State, and local governments, and with industry and the public add emphasis and dimension to the need for the improved management of data elements in information processing.

These Proceedings are for the first Symposium on the Management of Data Elements in Information Processing held at the National Bureau of Standards on 1974 January 24 and 25. Over 400 representatives of Federal and State governments, industry and universities from 30 states, from Canada, and Sweden were in attendance. 34 speakers discussed data element management in the fields of health care, water resources, state government information systems, transportation, libraries, market research, manufacturing, banking, information retrieval systems, military systems, computer programming and software systems, and motor vehicle registration.

INDEX OF
KEY WORDS AND PHRASES
TO
AUTHORS' NAMES[1]

---

[1]Acknowledgment and appreciation to Debbie Lee, Data Transmission Company for preparation of this Index.

Introduction to the Program of the Symposium on the Management of Data Elements In Information Processing

David V. Savidge, Program Chairman
Manager, Logistics
Data Transmission Company
Vienna, Virginia 22180

Lobachevsky, about 1600 A.D., wrote "God made the integers. The rest of mathematics was invented by man." We can add another truism - Information Processing was invented by man. Before there was mathematics, man collected, stored and used information in the form of numbers.

Pythagoras, about 600 B.C., is reported to have said, "Number rules the Universe." He did this without having the benefit of the "ten wonderful numbers of the Egyptians" to quote Leonardo da Pisa (Fibonacci) about 1200 A.D. Pythagoras' number system consisted of the twenty-five letters of the Greek alphabet plus two additional characters which permitted them to count up to 999 with no more than three letters. The thousands group was represented by the same twenty-seven characters with the addition of a little squiggle to indicate 1000 times the value. This must have been awkward.

The facility afforted by the ten symbols, zero through nine, proved so effective that by the end of World War II they were accepted and understood by all trading nations. This set had become a world-wide de facto standard within eight hundred years of its first introduction to the commerce of Europe.

The 1940's saw the invention, by man, of many tools to expedite information processing. Information processing, using computers, requires the interchange of information between the processing units of a system. Over the past twenty years, we have seen the definition of such a system change from the rigid hardware concept of a single generation from one manufacturer all under the same control to a multi-generation, multi-manufacturer, multi-location and multi-control concept.

A major step in this evolution was the adoption of the American Standard Code for Information Interchange (ASCII). This provided a means for the interchange between heterogeneous units. We are met after much data has been collected and exchanged by those units - some in ASCII and some in Extended Binary Coded Decimal Interchange Code (EBCDIC). We have learned that encoding and decoding characters is not too expensive. We have also learned that effective interchange is only possible if the members of the community of interest attach the same meanings to the same symbols used in the interchange.

We are fortunate in having representatives from a broad spectrum of communities of interest discuss their experiences in making the interchange of information more effective. Some communities represent single functions under the same ownership or control. Some are conglomerates. Some are confederates and some are competitors. All have found an economic need to interchange information.

Each presenter has been asked to describe certain aspects of the community he represents. This was done to make it easier to relate your own problems and experiences to those discussed. With the variety of communities to be presented, you may be able to identify with one or more of them.

We hope you submitted some written questions at the time you picked up your registration package. Additional written questions will be picked up on the aisles after each presentation

as the next presenter asks questions of the preceding speakers. The mechanism of questions before speeches gives each presenter the opportunity to clarify any ideas he may have picked up before making a possibly incorrect assumption.

At the conclusion of all formal presentations in a session, each speaker will respond to as many of the written questions as time permits. The ones not reached orally will be covered in the speaker's supplement which will appear in the proceedings.

Free interchange of information can only occur if there is complete understanding. The format of this symposium is intended to achieve this as much as possible.

NOTE:

A form is provided in the back of these Proceedings for your recommendations and suggestions for future conferences. We welcome your ideas!

Control of Logistics Data in the Department of Defense

T. M. Albert[1]

Logistics Management Institute
Washington, D. C. 20016

Within the Department of Defense (DoD), logistics oriented information interchange is controlled by DoD-wide standard procedures. Five major areas are presently covered: requisitioning, inventory reporting, transportation, system evaluation, and contract administration. Data moves between the Military Services/Agencies via a major world-wide communications network and through a logistics traffic routing and data collecting computer system. Traffic volume exceeds 25 million, 80 column records per month, and many other systems in several organizations are interfaced. A number of data files/bases, and a variety of computer hardware and software in many locations are involved.

Management and control of change of the environment has become a problem and a study is presently ongoing. The paper provides a status report of the effort, describing background, current conditions, and key problem areas. The problem is defined, and basic goals and directions indicated. Data Element Dictionaries (DED) as a key first step of data base control are stressed and a systems structure involving a hierarchy of DED's is suggested.

Key words: Change control; communication networks; data bases; data directory; data element dictionary; logistics; systems interfaces.

1. Introduction

In the best of cases, the management of data elements in information processing is not a simple thing. Even in a small, dedicated environment it is difficult to convince management to treat information as a resource, to manage it, and to build or convert to a system based on data banks.

[1]Project Director

Management prefers to develop programs to produce a specific result.

Consider then, the problems of the management of data and information on an extremely large scale--logistics--throughout the Department of Defense (DoD). More particularly, we will examine the DoD Standard Logistics Data Systems with all the ramifications of the separate interests of the various Services and Agencies, the variety of software and hardware, and the heavy investment in program oriented systems.

At the time this paper was written, the Logistics Management Institute (LMI) was mid-way in the schedule of a project titled "Control of Change Within the DoD Standard Logistics Data Systems." Herein is briefly described, as they pertain to Data Management, the background and existing environment of DoD standard logistics systems, as well as certain preliminary observations by the Institute of key problems, and approaches.

2. Background

In the early 1960's, in the Department of Defense, efforts were begun toward the development and implementation of Military Standard Logistics Data Systems (MILS). DoD indicated that these systems be utilized in the implementation of approved DoD policies in such logistics functional areas as cataloging, inventory management, transportation and movement, storage and distribution, and maintenance. The broad aims of the efforts were to provide compatible methods of logistics information interchange between the Services and Agencies in order to increase DoD-wide cooperation and thereby improve the effectiveness of logistics support at greater efficiency and economy.

Certain specifics of policy are as follows:

"A. Military standard logistics data systems will be designed to:

1. Provide common data languages via standard forms, formats, data elements, codes and rules for their application, to facilitate data interchange and compatibility among users of logistics data.

2. Optimize the use of automatic data processing equipment and digital communications networks for improved logistics operations.

3. Provide a common data base to DoD Components, affected Federal Agencies, Foreign Governments and industrial organizations for use in designing and implementing compatible procedures which (a) involve coding, transmitting, receiving, decoding and using logistics information; and (b) will generally improve operations, customer satisfaction and management control.

B. Approved standard data elements and related features established under DOD Directive 5000.11 will be utilized in the design of new military standard logistics data systems . . ." [1]

The general results of DoD efforts to date are five MILS systems:

- Military Standard Requisitioning and Issue Procedures (MILSTRIP)
- Military Standard Transaction Reporting and Accounting Procedures (MILSTRAP)
- Military Standard Transportation and Movement Procedures (MILSTAMP)
- Military Supply and Transportation Evaluation Procedures (MILSTEP)
- Military Standard Contract Administration Procedures (MILSCAP)

These systems are actually standard procedures for controlling logistics information interchange. They provide a standard system of codes, data elements, formats, policy and procedures for use within and between the Military Services and Agencies.

MILSTRIP, implemented in 1962, provides for the interchange of requisitioning and issue information for most materiel commodities. MILSTRAP, implemented in 1963, concerns inventory accounting information. MILSTAMP, implemented in 1966, provides forms, codes and procedures for the movement of materiel. MILSTEP, implemented in 1968, prescribes reports and methods of data collection to measure supply system performance and transportation effectiveness. MILSCAP, partially implemented in 1971, provides procedures and details for the interchange of contract-related information between and among DoD components and contractors. Other potential candidates for inclusion in MILS are applications such as billing and accounting for materiel sales, procurement, maintenance, and interservice support of weapon systems. With the exception of some aspects of MILSTEP, the MILS are transaction oriented.

To utilize these various procedures advantageously, there was implemented in 1965, a system called the Defense Automatic Addressing System (DAAS). DAAS is a real time, direct access digital computer system with buffered line connections to the Automatic Digital Network (AUTODIN) Switching Centers of the Defense Communications System. The DAAS concept is based on transmitting messages to a single designated point for editing, addressing, routing and retransmission; its reference files include the complete DoD Activity Address Directory.

Messages containing supply related information such as requisitions, supply status, and follow-ups are called documents and are in the form of 80 column records. DAAS performs a number of editing and checking functions, and also stores (for 30 days) an image of each document in a data bank. Valuable statistical reports are produced monthly from this data bank.

Currently only MILSTRIP, MILSTRAP, and MILSCAP documents are processed through DAAS. DAAS computers are in two locations and present average message volume is approximately 25 million documents per month. MILSTAMP documents also may soon be added to DAAS processing. Computer hardware is presently being changed to provide greater capability.

Mention must also be made of the Federal Catalog System (FCS) maintained and operated by the Defense Logistics Services Center (DLSC). The FCS provides for a single uniform catalog system (made up of many automated files)

which, in part, includes the centralized assignment and control of Federal Stock Number identification to more than four million active items of supply, as well as nearly two million items in inactive status. This gigantic data collection and cataloging effort, which began in 1952, replaced many separate systems in DoD with a standard system utilized by everyone involved in the Federal Government's logistics operations.

DLSC is responsible for uniquely naming, identifying, classifying and numbering each item in the catalog system. The information is disseminated via printed catalog, microform, punched card, or magnetic tape, and as a response to unique queries. Several types of materiel-use analysis reports are also produced.

DLSC uses a variety of computers and AUTODIN to provide its services, and present message traffic is between five and six million 80 column records per month.

The Federal Catalog System is well on its way to being further automated in a system called the Defense Integrated Data System (DIDS). DIDS, as planned, will integrate FCS files and utilize an extremely large central data bank, one approaching 15 billion characters of disc storage in size. Computer hardware for DIDS will be Burroughs B6700 systems. The consolidation of such a great amount of logistics data by DIDS along with integrated hardware and software will provide for greatly expanded capability for quick access to extended logistics information.

It should be understood that the various functions discussed above, of the several systems mentioned, do not fully describe the range of processing and reporting carried out. Further, when the new computer equipment for DAAS is operational, and DIDS is implemented, it is planned to have these systems interface or "talk to" each other.

3. Current Conditions and Operations

Before proceeding further, it must be emphasized that the MILS and their related systems, all things considered, function very well. They perform extremely useful, necessary, and valuable services, and are accepted and counted on by the Military Services and Agencies. They provide a common language with which the Services can "talk to" each other logistically. This is becoming more and more important because of the advent of single management of items of supply or weapon systems. For example, the Defense Supply Agency manages the procurement, storage, and distribution of many items which are used by all the Military Services.

However, as their use has increased, the MILS have become complicated through many changes and adjustments. Procedures which could be considered marginal to the specific purpose of a MILS have been added as activities to be covered. For example, MILSTRIP is defined as the requisitioning and issue system and provides for processing transactions related thereto. Consideration has been given to letting it cover the process of excess inventory reporting which is not directly related to MILSTRIP fundamentals.

The field on the MILSTRIP document which identifies the transaction to be handled (the purpose of the particular punched card) is called the Document

Identifier Code. To further illustrate complication, the field is three positions long and has over 150 possible combinations. Other fields have larger possible combinations.

As a very simple example of how the MILS function, consider the need of an Air Force installation in Europe for a truck motor carburetor. If the item is not in stock locally, a MILSTRIP, 80 column card, requisition is created. The requisition includes among other things, identification of the item, quantity, the requisitioner, the source address, and priority. The document is sent via AUTODIN to DAAS in the United States. DAAS, in real time, edits the requisition--checks the item number and address, and sends back any error messages. In some cases, the requisition is transmitted on to the address specified; in others, it is rerouted to one or more different addresses of source of supply. In our example, an Army installation would receive the requisition, since the Army manages vehicle parts.

The recipient of the requisition is usually an Inventory Control Point (ICP) which receives the requisition on a punched card, magnetic tape, or a hard copy print out. The information is then handled manually or is entered into the local ADP system. The process of releasing the carburetor from a warehouse and notifying the requisitioner of the order status is generally handled by additional MILSTRIP documents but in some cases MILSTRAP documents would be necessary.

When the item is ready to be picked up to be shipped, a MILSTAMP document is created to notify the DoD functions responsible for transportation. Other MILSTAMP documents are created at various points in the carburetor's journey to the requisitioner. The MILS are, of course, geared to handle many more functions specific to any supply and distribution operation.

On an after-the-fact basis, certain items of information are derived from the MILSTRIP and MILSTAMP documents and summarized to provide a measure of supply and shipping performance for top management.

As mentioned, the MILS procedures documentation specify the codes and methods to be used. Presumably these standard codes would also be used in each Service's unique supply-oriented application systems. In fact, in many cases, they are. However, past the first level of MILS systems interfaces to the Services we find more and more changes and adjustments to codes and methods. Each Service does not use the MILS in the same way, and each Service implements them through their own internal documents.

4. Administration

DoD policy places MILS administration in the Defense Supply Agency (DSA) with the exception of MILSTAMP which is administered by the Army. In DSA there is a small group called MILS Systems Administrators who are competent and individually at a reasonable management level. However, organizationally, they are at a relatively low level. Their function is basically to monitor the systems' use, maintain the central documentation, and generally coordinate MILS affairs. Beyond this there is no dedicated entity at any

537-860 O - 74 - 2

level within DoD which has as its sole objective the responsibility for planning and expansion of the MILS. It must be noted here that we are speaking only of the MILS systems and not the wider, overall logistics environment where much planning is currently ongoing.

The MILS Administrators also individually head Focal Point Committees. An example would be the MILSTRAP Focal Point Committee made up of a representative of each using Service/Agency. Periodic meetings are held to discuss and decide on adjustments to the procedures.

5. The LMI Task

In mid-1973, LMI was asked by the Office of the Assistant Secretary of Defense for Installations and Logistics to recommend improvements in the management and control of change to the MILS in order to provide more effective interface and coordination among the Service/Agencies.

How are changes to the MILS accomplished currently? Suppose a responsible person at a Navy ICP decided that a change must be made to a particular MILSTRIP code. He would propose this change in writing; the letter would pass up through command channels to arrive eventually at the MILSTRIP Focal Point Committee. If the committee considered the report worthwhile, a draft of a change proposal would be created and disseminated through the Services via each Service's contact point. Comments on the change would flow back to the committee, who, if it still considered the change worthwhile, would write the final change statement. This final statement would also be disseminated for approval or adjustment through the Services.

Upon final approval, the change is issued to all using parties for implementation. This effort, to arrive at concurrence for a change, can take from one to six months or even longer.

Implementing the change of a code, is the major problem. What must be done, is to make a change to, literally, hundreds of ADP systems to take affect at a single point in time. These systems are written in many different languages for many different kinds of computer hardware. Manpower and computer resources must be scheduled and applied. On-going programming or development may have to be postponed, or the change must be delayed. Time schedules for major projects may have to be adjusted. Change implementation of this type can take from six months to four years.

Basic problem areas from the above can be stated as follows:

- The difficulty involved in the implementation and coordination of changes to various systems;
- The lack of a formal change mechanism;
- The difficulties inherent in change control because of the many interfaces with other Service/Agency systems;
- The need for further standardization efforts;

- Current organizational responsibilities; and
- Measures of system effectiveness.

In its task, LMI was asked to review and consider pertinent DoD policy and organization, the interface relationships among the MILS and the other DoD logistics data systems, and existing techniques for managing change and for data base management. The aspects of policy and organization will not be discussed in this paper beyond the following--in the kind and size of environment we are discussing, control of information, its movement, use and standardization requires one or a few strong points of control. The kind of services being discussed or needed must be carried out efficiently, must not be too fragmented, and must be provided without usurping management prerogative and operational control.

6. Approaches to Improvement

Control of change can be approached by adjustment of membership and organizational level of Focal Point Committees, and setting tighter schedules for concurrence on changes and for implementing changes. For example, a cyclical approach for change implementation could be designated--collect them, for some period, and issue the collection at one time, grouped in a meaningful fashion.

If these are the only efforts undertaken, then we are simply treating the symptoms, not the problems. To properly control change in the MILS, we must control and manage the things that change, i.e., data, and the methods of processing and communicating data. Information, made up of data, must be treated as a resource just as people, equipment and money, and managed accordingly.

7. Some Key Logistics Data Problem Areas

7.1 Data Element Dictionaries

Within DoD, as in the non-government world, there have been a number of efforts toward developing a single large centralized data bank. The usual reasons are given; for example: to do away with redundancy, prevent fragmentation of data resources, to economize through elimination of some hardware, and by making the programmer's job easier. The development and control of such data banks has not been overwhelmingly successful. It may be that the state-of-the-art in data base management software has not developed to the point where the data base can be truly independent of the application programs. Or it may be that the cost of data collection is too much and the task too involved to face. There have been a few partial successes where some systems have been developed grouping a number of files or small data banks.

The logistics community throughout DoD and the Services/Agencies uses many different kinds of hardware, software, and programming languages. It is recognized that more efficient and less costly logistics operations could result from better control and transfer of information.

To overcome the problems, we find a number of separate efforts mounted toward the development of Data Element Dictionaries (DED). There is so much data extant that the only way to attempt to control it is to know what exists and to describe it. In general, the DED's contain some or all of such information as:

data element name

source

code

field length

units (inches, gallons, . . .)

definition

where used (program and/or system)

mnemonic

data interrelationships

how used

In itself, a DED has obvious valuable uses such as a reference guide to programmers, a tool to reduce redundancy, and as an impetus to more standardization. Interfaced with a computer system, it could provide control and editing functions for input data, and a control for changes and updates to the system's data base. It is possible for a DED to exist by itself, but the idea is growing that it should be a prerequisite to the development and use of a central data base and data management system.

A particularly valuable use would be to assess the impact of making a change to a system. It should be possible, if a change must be made to a code, for example, to quickly find those locations, systems, and programs involved; and using this information, to make a start toward developing the necessary resources, required schedules, and cost/benefits of the change.

LMI has found a broad but not specifically authorized structure developing in DoD. Each of the Services has developed one or more forms and levels of DED's. Some have specifically defined data elements at length and others have defined logistics functions; some have begun from the ground up by deciding what data should be in the DED and defining it, and others have used existing reports as a source.

Between the various DED's there is some commonality of data elements particularly as they relate to the MILS, there is some redundancy, and, in some cases, the same code is being used in two different ways. But at least a very solid beginning has been made. A firm effort is also being made by DoD toward directing more standardization of data.

Within DSA, there has been set up the Logistics Data Element Standardization and Management Office (LOGDESMO). This office, among other things, has been engaged in developing a DED of MILS related data, and has

been registering, gradually, the standard data elements established in each of the Services and their individual DED's. In the process, relationships have been established between MILS data and programs and systems in the Services. What is developing is a Data Element Directory. Presently the information concerning the data elements is produced as computer print out and on microfiche.

As a future possibility, the DED might be made available through an interactive storage and retrieval system with the capability for unique query, browsing, and analysis of data interrelationships. This might provide information on whether new data bases were required, or how a new file might easily be put together, or how better to use existing data bases.

Because of the size of DoD and each Military Service's individual mission and interests, a desired goal would be a hierarchy of DED's. In each Service there might be two levels of DED's; one a Directory, and the other, one or a series of detail DED's. At LOGDESMO, there could be a Directory to each of the Service's Directories, as well as a detail DED of MILS data and top management required data.

Such an approach will require standardization of methods for development, maintenance, and use of the DED's. Considering the present situation, this will probably come through evolution. However, it will be necessary to establish, in each service and at LOGDESMO, a management function with the power to properly control activities.

7.2 Hardware/Software Standardization

Without question, any large-scale system or group of such systems would be much easier and economical to work with if the same kind of hardware were used throughout. Unfortunately, considering the size of the logistics community, the money presently invested, and national policy against monopoly, this is unlikely to occur.

Efforts have begun toward some standardization of programming languages. COBOL is being used predominantly.

An interesting project which is to begin soon will be to update the "inventory" of automated data systems in DoD. Utilizing standard methods, nomenclature, and forms, the project should provide a catalog of data systems including a description of hardware and software system interrelationships, groupings by function managed, data bases used, etc. Such information should be valuable for, among other things, providing a basis from which to proceed toward some level of system integration.

7.3 80 Column Record Formats

Most DoD ADP systems and communications operations use an 80 column record format. Such a fixed record makes for coding problems in trying to make one punched card cover a variety of transactions; it causes redundancy in data input in the case of multiple card records; it necessitates training in some cases for personnel in the field to code input; in other words,

it is a constraint.

For a number of years, the 80-column card format was the accepted and easiest approach. However, with today's greater hardware and software capability, we have the means to provide more flexibility. Of course, although the means exist, an immediate effort to change from the 80 column orientation is not economically or physically feasible.

Change will come about gradually--probably as hardware is replaced, as more mini-computers and intelligent terminals come into use, and as communications methods are updated. To reach toward more flexibility, however, goals must be set now.

One of the goals we must aim for is to allow for more alphabetic input/output, i.e., natural language words. The MILS systems, as in the case of many other systems, are quite involved and provide a problem at the user interface. Flexibility in record format can allow for alphabetic handling.

7.4 Data Bases

Two aspects of data bases must be addressed: size and location, and management.

In the DoD environment, a single large data base in one central location is not an acceptable situation. The nature of DoD's mission requires planning for contingencies of a type which necessitates backup and dispersion of resources. In the transaction oriented world of logistics, this means a number of active files accessible through a communications network. In other words, distributed data bases, with the requirements of extensive standardization and control.

Data base management is still an evolving art. There are a number of data base management systems (DBMS) software packages available today, none of which are completely general, i.e., complete independence between the data and the application programs. We may never arrive at this state. However, existing techniques which may be tailored, coupled with Data Element Dictionaries/Directories may well provide us with the tools necessary for management of data bases for some time to come.

7.5 Networks

The state-of-the-art of digital communications has arrived at the point where extensive computer networks can be supported. "A computer network can be defined as an interconnected group of host computers that automatically communicate with one another and that can share such resources as programs, data bases, memory space and long-haul links."[2] The concept presents interesting future possibilities for logistics in DoD.

The network developed by the Advanced Research Projects Agency (ARPA) in DoD is considered the pioneering effort. It connects about 40 geographically distributed DoD and University computer complexes made up of many different

hardware/software combinations. At each node, in this heterogeneous distributed computer network, there is a hardware/software interface between the host computer and the communications network. Very simply, the hardware/software interface breaks a message into 1000 bit packets, chooses the best network routing via various nodes at that moment, and sends it along. Each packet is error checked along the way, and regardless of each packet's route, they are all put together properly at the other end. If one or even several network links are not operating, the message can still get through. Several commercial organizations are considering the implementation of Value-Added Networks (VAN) which are based on the ARPA network technique of "packet switching."

Basically, the ARPA network is a number of computers, each with its own software, data bases, and message-switching interface to a communications network. To go slightly beyond the state-of-the-art, research is currently being done by Honeywell on the possibility of moving the data bases, which are to be accessible to the network, into the network. That is, the data base would be interfaced to the message switcher and could be accessed without involving the host computer.

8. Summary

Since LMI is only midway into its project, final recommendations naturally have not been formulated. The topics discussed above have concerned both immediate and future considerations. We can, however, define the problem.

In the past several years, the MILS, which are standard procedures, formats, and codes, have been established to provide for compatible logistics information transfer between the Services/Agencies within DoD. Currently these procedures involve order processing, inventory control, transportation, and contracting, and are in common and extensive use. Each Service/Agency has its own variety of hardware/software/logistics systems and each uses the MILS data somewhat differently. The non-integration of hardware and software, and the lack of common, strong, clear, central direction for use of the MILS results in two problems: 1) the redundant use of computers and manpower with its attendant high cost and energy requirements; 2) the difficulty of changing the MILS or keeping them current because of the many systems interfaces.

These problems cannot be resolved quickly or easily. In DoD, they can probably best be overcome through a series of goal-directed iterations rather than a well-ordered sequence of steps. Goals must be set for the near-term, mid-term and long-term.

Near-term goals must certainly aim at the continued standardization of data elements, the development of data element dictionaries and directories, and information resource control. These are necessary as a basis for later movement toward greater system integration. Some function should be established to provide data dictionary/data base administration. A hierarchy of DED's driving, controlling, or monitoring data bases appears to be a fruitful goal.

Mid-term goals may involve the idea of distributed data bases based on more standardized data, possibly using an existing system as a vehicle or model.

Longer-term goals may involve a dedicated network.

The justification for all this, hopefully, will be the resulting more efficient, more cost effective control of DoD inventory in storage and in pipeline, allowing DoD to better carry out its basic mission.

9. References

[1] Department of Defense Directive 4000.25, Administration of Military Standard Logistics Data Systems, March 23, 1971.

[2] Doll, D. R., "Computer Networking--the Giant Step in Data Communications," Data Communications Systems (New York: McGraw-Hill, September 1973), Vol. 2, No. 1.

End User Data Control

R. B. Batman

Sperry Univac
King of Prussia, Pennsylvania

Information processing has progressed from punched cards thru on-line disk storage.

As the data processing departments evolved through these phases, new procedures such as large data bases and data management have emerged to make efficient use of computers.

Another result of this progress has been a limiting effect on the end user's control of the data that he creates and depends on. In an on-line data base we have the opportunity to return control to the user.

A integrated central data base interfaces with distributed local data bases. The local data base may contain local operating information in a manufacturing plant or census and medical summary information in a satellite hospital.

The purpose of a local or satellite data base is to isolate the data for which there is local responsibility and provide a secure, easy-to-use, terminal oriented interface to this data.

Univac manufacturing has such a system for product definition where a central master data base is maintained in St. Paul and nightly transmission of local requirements for the other plants occurs. In addition there is a in-plant operations system with its data base in Roseville, Minn.

A successful hospital group has a med-scale central system for administrative and remote batch terminals in remote hospitals. In addition they have a satellite for communications, data acquisition, and operational reporting. In this way the remote hospital has local operational control of data required for hour by hour operations. The key benefit in local data bases is the direct control of the data by the end user. He benefits from access to his data, and the total system benefits from a more timely, accurate data base.

Key words: Data base; distributed data base; local data base; end-user; interactive data entry; satellite system.

1. Purpose of Information Processing

Information processing has a purpose - to provide information to people who manage, operate, or perform functions in an organization. In order to do this, information processing systems:

- Hold information in data bases
- Answer inquiries
- Perform computations and associations to prepare reports
- Receive and process information

These functions are performed to serve the end user, the person who needs the information.

2. Loss of Control by the End-User

Manual systems permitted the end user, the shipping clerk or her supervisor, to keep what information they wished and manipulate it at will. Since few people were involved it is easy for the clerk to retrieve information as requested and translate the information to the form a requestor might like. Standards and control of data was minimal since the end-user knew the encoding and could always translate.

The rapid growth in the information required to operate our organizations made automation; and in fact, computer systems, mandatory.

Unfortunately, the trend to automation also caused a trend to centralization with the related standardization and road-blocks to end-user service.

Centralization is required in some cases. In others, its desirable because it eliminates redundancy, improves service and reduces costs and, whenever centralization is employed, it madates standard control of data definitions and usage.

However, the end user should not be slighted. Lets review the chronological evolution of information systems, how controls were developed and how the end-user lost his control of his information. Then we'll look at how control and local data bases can coexist; and how the end-user can be supported.

Manual Systems - the end-user was the data processor and he had complete control directly or through clerks reporting to him.

Punched Card Systems - Moderate form of centralization; some economies in unit record utilization for many functions. Card files were centralized but still accessible for exceptions. Simple special runs were feasible.

Magnetic Tape Systems - Centralization in its worst form. Large tape files defying non-standard inquiries or special reports. Machine efficiency dictated partitioned and structured data bases which imposed difficult error correction and reconciliation on the end user.

Disc Systems - Allowed flexibility in processing, permitted pre-planned inquiries. Centralization continues but data management and report generation systems provide some ability to react to end-user needs. But, the data processing department still generates most output for the end-user.

Interactive Remote Processing - End-user oriented. Provides systems design and implementation capability to end-user at his terminal. Provides local data base for the end-user to create, manipulate, and generate reports. The local data base provides input to the central data base for those applications requiring it.

3. Central Data Base versus End-User Interaction

A Central Data Base evolves from a need by a central authority to control the information and disperse it to the end-user. It also is the solution of choice where many end-users need to share a body of information such as: regional medical records, airline reservations lists, or universal product definitions.

The procedures for creating a central data base would fill a book, and some of them will be addressed by other papers at this Symposium. Most of these procedures require central data definition, constraints of access and processing, and elaborate programming requirements. It is only by this approach that the diversities of interest can be controlled and partially satisfied.

On the other hand: the end-user, the person who needs information to function, loses his ability to access the data he needs. His primary method of access is through inquiries or reports prepared by the programming groups. In the interest of standardization and in order to save money to pay for the system, locally maintained files are absorbed into the central system. This results in two extremes of service to the end-user, neither of which is desirable. On one hand the end-user gets manmouth volumes of data in the form of printed reports and inquiry capabilities for those items that can justify programming costs. On the other hand he gets limited general purpose inquiry capability that requires pre-defining his files or data sets to the query processor and most likely a programmer oriented language.

Although the central data base has advantages, it has some weaknesses:

A. Complex control procedures.
B. Costly audit and recovery.
C. Data Management System overhead.
D. Less end-user control.

The first three of these are cost trade-offs that can be cost-evaluated by an organization.

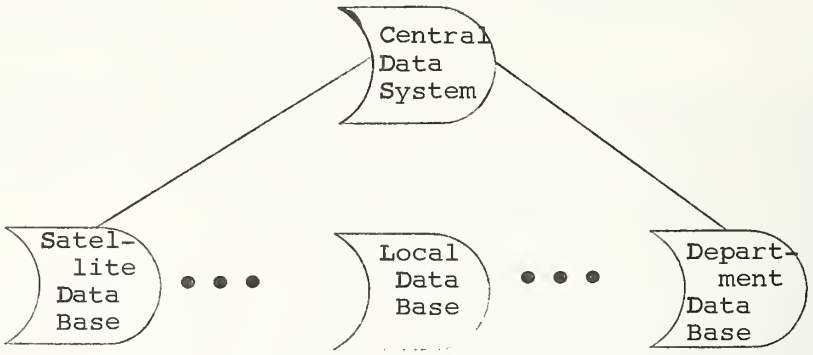
End-user control is more intangible, but more relevant in terms of making the power of a data base available to operational personnel who do the work of an organization.

4. Distributed Data Base

In order to satisfy both needs - that for a central data base and

that for end-user control - a distributed data base is in order.

A distributed data base consists of a central data base and many local or remote data bases as shown.



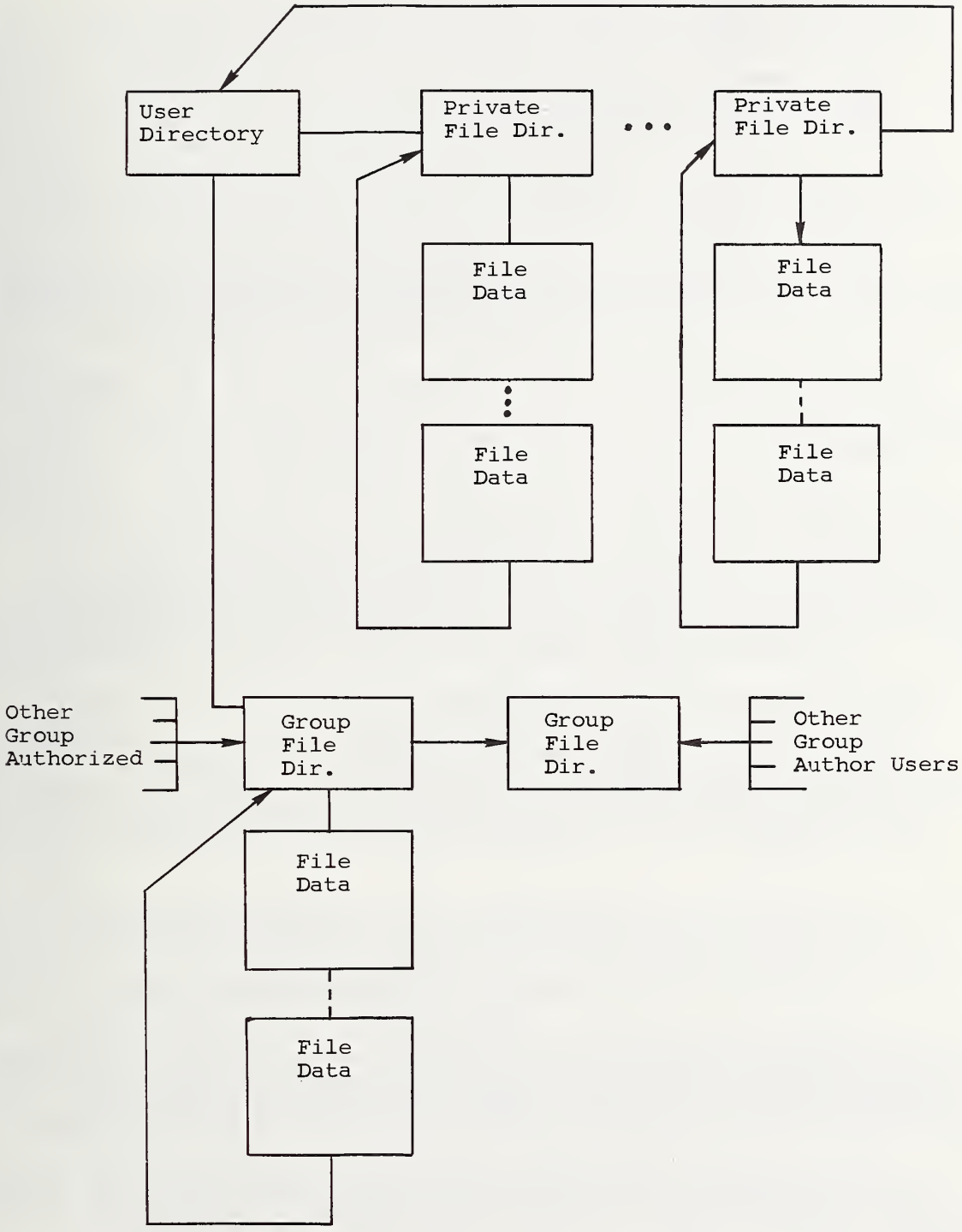
The terms satellite, local, or department in this figure imply a variety of local data bases with different uses.

The term sub-schema in DBTG language applies to these local sub-sets of the entire universe of the data base.

The central data base holds those data sets that need to be central as discussed before - authoritative control, shared simultaneous access, security control, and standardization. The local data base holds those data sets that relate to the end-users.

The central data base is housed in a central system, probably a medium to large system of 262K words (or 1M bytes) of processor storage and 20M bytes or more of disc storage. The local data base could also be housed on the central hardware or on a mini or midi satellite system with a communications link to the central site.

The local data base has a structure as follows:



By definition, the local data base is not related to a group or procedural language applications programs. The Remote Processing System is the terminal end-users interface to the local data base.

Functions available to the end-user are:

- Build and Parameterize a new file.
- Search, sort and match files.
- Format reports by setting up the parameters from the terminal or call on pre-stored formats.
- Conversational, frame-driven data entry.

The files in the local data base can be private files with all the uses of a users own small scale system. He can update them, manipulate them and prepare reports.

In addition, these local files can be used for input to the central system. Files for transmission (figuratively or actually) to the central system can be data extracted by the search function from local files. This input file can also be the accumulation of interactive data entry.

When an end-user wants to extract a set of data from the central data base, a report isn't printed. Rather the central applications or report generation routines create a file that is transmitted to the local data base for perusal by the end-user from his terminal.

The end-user has simple direct control of his data base. He can add, delete or modify information by locating it in the file (search) and displaying it on his CRT. Then simple commands allow him to change it. In this way he can walk up to a terminal and display record like a production order. He can then change the quantity to be produced and scrap quantity without writing a form for transmittal to a keypunch room. Later that shift or at shift change, the latest status of all production orders is fed into the central system.

In the local data base the end-user:

- Creates his own files from the terminal
- Controls the structure of the files
- Accesses them at will
- Can perform strings of user specified operations to create mini-systems.

5. A Manufacturing System Example of a Distributed Data Base

5.1 Central Product Definition for Remote Manufacturing

Engineering data control by a large manufacturing company such as Univac is an excellent example of the need for central and local data bases. In this operation components, subassembly, and units are made at numerous locations throughout the country. These diverse manufacturing plants build numerous parts that need to come together for the finished computer system.

Design engineering is not always in the same location where the parts are made. Long term requirements for spares make long lost cousins of the original design engineer, manufacturing engineer and the plant building the spares.

The only way this network of manufacturing can communicate is to set

up responsible locations for each major product. These engineers are responsible to make all engineering changes, control effectivity of changes by serial number or lot and communicate these to the other locations.

Each plant has local needs for bills of material; some routinely and others when spare requirements for older products occur. Yet, the maintenance of the all-important product structure files can be effectively centralized.

The approach was to maintain a single master Parts and Structure file, the master bills of materials, at one location in Philadelphia. Each night maintenance is performed on this central data base.

Although the data base exists and is procedurally maintained at a central site, updates can be transmitted into this site from other locations. After the master update, those bills of material that are allocated for usage at each remote site are transmitted back to that location, if they have changed. These updated bills of materials then replace the old versions in the local files.

The remote sites can also request bills of material for items not usually theirs.

This has been in existance for numerous years using nightly batch transmission. Future enhancements to spread on-line interactive production control from our Roseville Plant to other locations is in process.

5.2 Remote Processing within a Plant Using a Local Data Base

Our Roseville plant, where the Univac 1100 series is manufactured is a complex production line in itself. But, in addition, the production scheduling manager controls the Jackson Minnesota plant production.

In order to provide communications amongst the cost centers in the factory as well as to provide production scheduling a window into what's happening at any time in the plant; a terminal oriented system using a local data base concept was installed. This system places terminals at all production scheduling operations and cost centers throughout the plant. The foremen and schedulers use their terminals to record progress and note deficiencies; thus permitting the entire plant to look at progress from any point.

This information system utilizes a medium speed real-time computer to provide easy-to-use functions for file update and report preparation, via terminal. The entire system is end-user oriented, eliminating programmer oriented syntax for the processing functions. A terminal user can create a file, update it, search it, sort it, calculate on it, etc. Message switching and record switching is included.

Each terminal user has a mode code or password that giveshim access to his local data base. The local data base is actually one of many stored on the system. He is constrained to activity within this set of files, but if a group of users have similar files they can be grouped together. It's thus easy for them to inspect each others progress and trouble reports. Also, the production scheduling manager can search through files for all similar cost centers and pull out all occurrances of a category of problem.

From the management viewpoint, they finally found a way to eliminate the stack of notes, scribbles on the back of envelopes and hip-pocket

notebooks where people used to record information. Data about their progress and problems in their job are entered into the computer and its local data base. Here it is available for local use or forwarding to the central system.

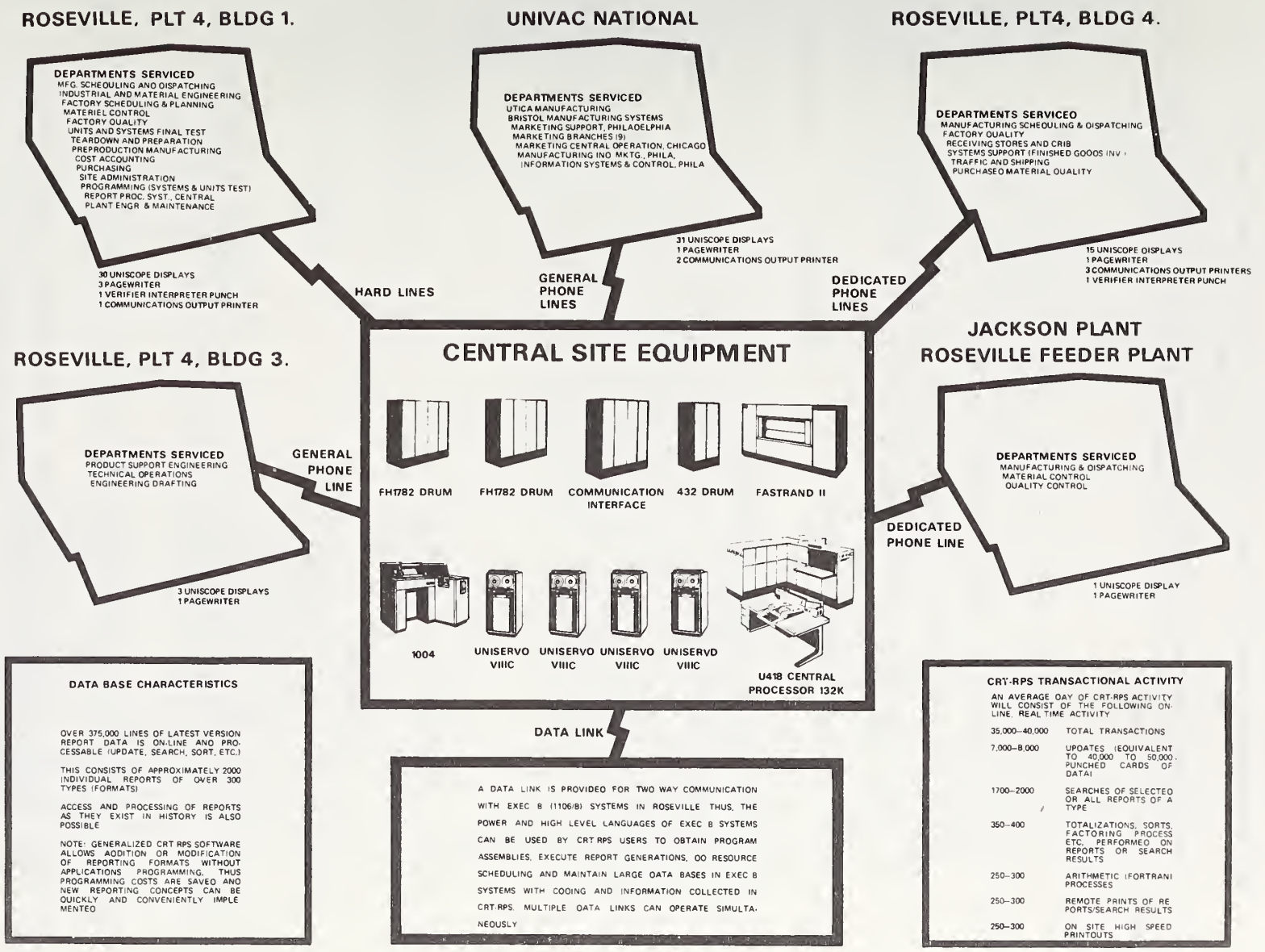
The local data base files are created in two ways. Many of the files are merely local files used by the people on the floor to store records for processing and reporting. This information is only local and is only used by the people in the factory and their management, which also resides in the factory. There are literally hundreds of local files used for many purposes. Many of the ad-hoc as well as regular production reports are merely extracted from these files in a few minutes terminal effort.

Another method of creating these working files relates to how the local data base interfaces to the central system. Production schedules are laid out by a series of programs run on the central system. These programs maintain inventory levels and project requirements of each component for future time periods. Using the product definition files, they explode requirements for all parts of a computer from the highest level assembly down to the simplest IC chip. Production schedules and requirements are fed from the central system to the independent or satellite system and stored in the local data base. It's these local files that serve as the starting point of the production progress reporting system.

After the master production schedule is laid out, the actual job of running the plant and reporting progress begins. Throughout the day, each day, foremen and schedulers post completions, enter delays and causes flag back-ordered materials, and inquire about jobs preceding them in the production flow. Production managers can access these files to evaluate the status of the production floor and what remedial action to take.

The figure on the next page shows the network and usage figures for the satellite system in Roseville.

UNIVAC CRT-REPORT PROCESSING SYSTEM



The important aspect of this system is the end-users ability to set up his own files and develop his own mini-systems.

The entire range of commands available to the end-user in the new version, RPS 1100, has been significantly expanded from the older RPS 418 system and includes:

FUNCTION

Message Send

Enter File

Build File

Help

Exit RPS (Log-off)

Return to Major Command Level

Destroy File

User Definition (Parameterize User)

Form (Parameterize) File

Search

Match

Sort

Compute

Index

Print

Tape Copy

Tutorial Processor

Execute Tutorially Defined Process

TIP Application Selection

To explain the data element control as the end-user sees it, let's examine the build file, parameterize file and search commands.

- Build File

The Build File function allows a new file to be created. Options are:

(1) Create a data-less shell
(2) Selectively copy lines of an existing file
(3) Copy an entire existing file
(4) Combinations of (2) and (3) with multiple files

After creation, the user may choose to parameterize his file. To do so, he must use the 'form file' system function. Initially the new file is formed according to standard file parameters. A standard file has a

line length equal to that of the U-100 screen size (80 or 64 characters per line, dependent upon system generation).

File build parameters (from major command frame) are as follows:

Build File-ID
Create new report specified by File-ID

Build File-ID1, File-ID2
Create new file File-ID1 by copying entire file File-ID2

Build File-ID1, File-ID2: 150-175
Create new file File-ID1 from lines 150 through 175 of File-ID2

Build File-ID1, File-ID2, File-ID3: 250-398; 480-502, File-ID4
Create new file File-ID1 from all of File-ID2,
Lines 250 through 398 and lines 480 through 502 of File-ID3, and
all of File-ID4.

FORM FILE (DEFINE FILE PARAMETERS)

This function is available only to privileged users under most circumstances. It allows a properly trained user to define or update the format and characteristics of a file, and also set certain limitations on access to a shared file. These file parameters may be defined after a file has been "built" with the build file function. To speed parameterization the user may "copy" parameters from another file and use them with or without changes. Certain of the parameters, by their nature, may only be specified at the original parameterization and may not be altered later. Also some parameters are mutually exclusive and choosing one automatically precludes the possibility of choosing the other(s). The result of a form file operation is an updated file directory.

FORM FILE PARAMETERS

Parameter	Possibilites and Rules for Use
Owner/Master User	USERID-owner if private, master user if shared file.
Group Coordinator	Coordinator for a group of files must be previously defined as group coordinator
Trailer Line Code	Any character used to denote lines which extend regular RPS lines
Password	Any character string
Automatic Save	"Y" if desired blank if not-causes file to be saved each time "Save" is run at Computer Center
Automatic Hold Lines	Number of lines to be automatically held at top of screen at entry into file (up to 3 allowed).
Indexed	Inquiry data from pre-existing file directory "Y" = Indexed, N=Not Indexed

Data Key Length	Copied from pre-existing file directory - If length and displacement are displayed, report indexed on data key.
Creation/Update Data	Inquiry data-creater of file and system data of creation also system date/time of last update.
Line Length	Line length in character any value up to 720.
File Data Map	Indicate fields by entering characters-any character is valid.
Extraction Mask Including Editing and Update Protection Attributes	Up to nine display maps may be entered. Each can have from 0 to 2 headers lines. Field editing and protection may be defined by placing appropriate characters under the mapped display fields.

SEARCH

The search function is responsible for searching an entire file and extracting records which match the search criteria. It can also search across more than one file in the same group if the files are formatted the same. The search criteria used may specify one of the following logical conditions.

(1) Extract records on inclusive range
(2) Extract records on exclusive range
(3) Extract records on absolute match

Options are available for future operations. Line extracted may be updated and replaced by specifying the "blend" option at search time and running a match/update operation using the updated results of the search. Search results may be displayed or placed into a file for future handling or both. Also, lines found in a search may be deleted from the searched file.

SEARCH PARAMETERS

(1) Search and Extract Within Range

Entry of two values vertically below a mask field (smaller over greater) causes the creation of a result report composed of copies of all lines containing values falling within the range values in the specified search mask field.

(2) Exclusive Range Search

This is specified by again entering two values vertically in a search mask field, (greater over smaller). The result file will then be a compilation of copies of all lines whose values in the selected field or fields lie outside of the specified range.

(3) Search For <,<or=, >,>, or =

These capabilities are all available by choosing the proper range for a search within range.

(4) Multiple Field Search (Boolean And)

Available between any or all fields in the user's search mask

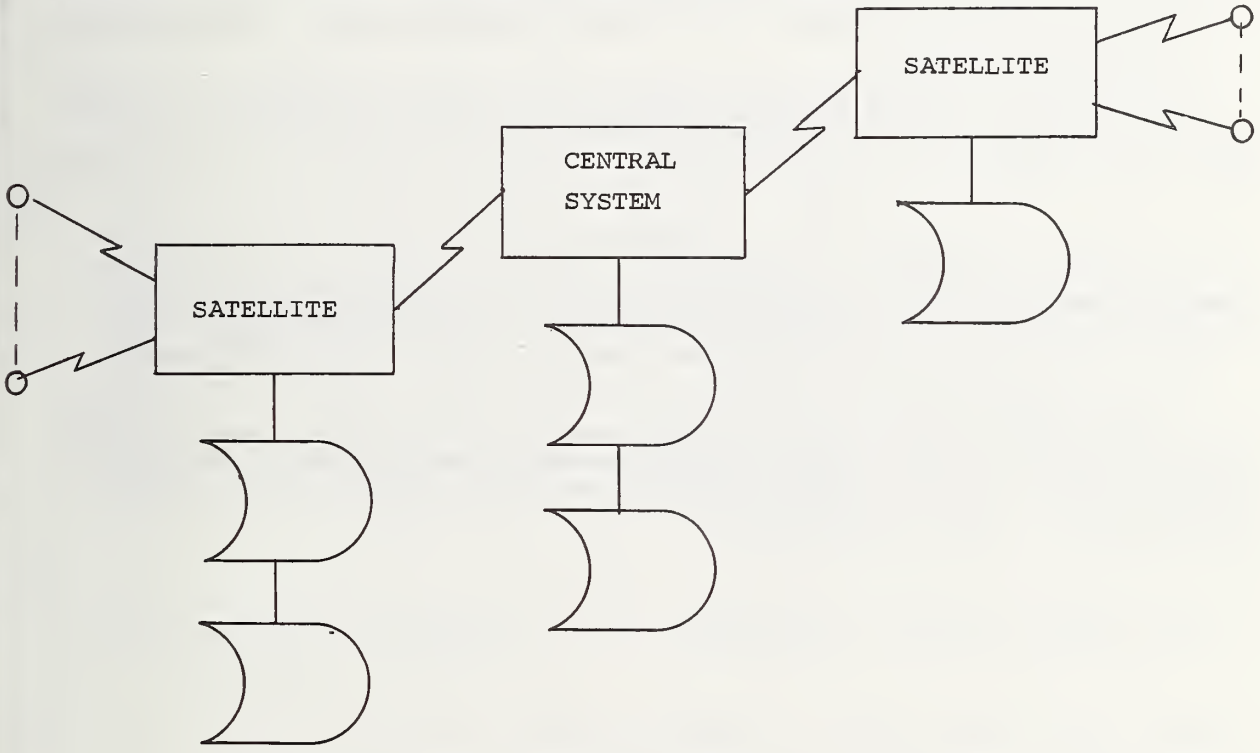
(5) Ommission From Search Criteria

Any character position of any field may be ommitted from search criteria. A partial-search may be indicated by blanking characters in the mask line.

These functions, available to an end-user allows him to control his own data elements. Features are also available that allow him to reformate a file when it is transmitted (or copied) for the central procedural processing systems.

6. A Distributed Health Care System

The best example of a Distributed Health Care System is a regional system with communication links to satellite systems in each hospital or health care unit.



The central system stores the data base for those functions that are regional in scope and for those applications that require the computing power of the central processor.

The central system would handle data sets such as:

Master Patient Register - a network of data sets containing a Master Patient Record and associate summary medical data members in its data sets. Numerous indices by name, disease and operation, as well as specialized number schemes are included.

Waiting List - a data set containing all elective admissions waiting to be scheduled into a hospital in the region.

Patient Accounting - insurance, cost accounting, and accounts receivable data for the region. Centralization expedites control and reduces cost for supporting business office staff.

Central Stores and Pharmacy - maintainence of central inventory orient data sets to expedite locating special items and for centralized inventory management.

There would also be satellite or local data bases for each health care unit. The local data base might reside in a remote satellite processor or merely be a separate data base on the central system. It would handle data such as:

Active Patient Files - a master record for each patient/case, and member data sets for treatments, results, appointments, and charges.

Resource Schedules - a profile of resources, available, appointments, and scheduled patients.

Laboratory Management - queues of test requisitions, interim, and final results, statistics on laboratory equipment.

Local Stores and Pharmacy - maintainence of local stores inventory and pharmacy data sets.

The satellite system would also hold the interactive data entry steps which tutorially guide the terminal operator through a sequence of steps for various functions. These data entry functions simplify input and control the terminal operator assuring accurate requisitions and reporting.

A terminal operator who has signed on to the satellite system could interact with the satellite data base or use the satellite as a store and forward message switches for communications to the central system. He could also cause a file stored in the satellite to be transmitted to the central system or pull a selected data set from the central system.

This configuration or distribution of processing capability provides fast response at remote sites and a reasonable level of data base activity. Systems like this with a tape oriented interface to the central system are operating now. Some limited capability for remote processing exist in concentrators in some systems, but a end-user oriented remote processing satellite is a current development.

ADDENDA

During the seminar there were some questions which I would like to answer here.

Q1: Don't small user files prevent improvements to existing larger systems (ie. less detail in the accounting system)?

A1: Yes and No.
Yes, because an end user will develop mini-systems to serve his specialized needs. The question is should these exceptions be included in the main systems or not?

No, because allowing the end-user to satisfy his local needs provides the following advantages if the systems group decides, after careful study, to incorporate the mini-system into the main system:

- Undersirable mini-systems die as rapidly as they spring up.
- After a mini-system has operated for awhile, the end-user knows what he really wants.
- Conversion of the mini-system (or absorption) can be done more efficiently since it's already operational and better defined.

Q2: How do you measure the value of the system, RPS?

A2: Our Roseville plant using a stand-alone 418III has been able to reduce their expeditors and production scheduling staff while handling increased production. They also have reduced inventory levels and work-order delays due to parts shortage - a unique combination.

Our figures show that the central hardware, support and operations are paid for by savings every 4 months, at our costs. Thats about every 8 months at full commercial prices for equipment.

Q3: How do you reconcile the local Data Base concept with the need to extract across files of information? Is it economically feasible to have the two types of Data Bases?

A3: First let me clarify the local data base. It holds three kinds of data:

- Data which is only used locally or maybe infrequently transferred to the central system.
- Data that is a snapshot of a central data base stored locally for inquiry and short term use.
- Data that is transitory as input or output of a larger system. The interactive data entry approach improves data input while the local data set manipulation and display enhances system output.

The user should be charged for his local data base, based on the file spacehe uses. Then the user will be careful not to uneconomically use

the local data base.

By the way, there are techniques in RPS to extract across files or selectively pull data from one file after another and built a new data set.

Q4: How do you control files being created from an economical point of view? End users, for example, will put up something like a telephone directory while a central manual system may exist.

A4: Firstly, because of the users freedom, we have a System Co-ordinator who advises users and monitors their usage.

Sometimes such redundancies exist, but the user pays from his own budget for his carelessness. If its absurd the Systems Coordinator goes through management channels to stop this misuse.

However, its important to note that many mini-applications can be implemented and used for a year at less cost than merely negotiating with a programming department for a system that may not be implemented.

Data Resource Management[1]

Charles J. Bontempo
& David G. Swanz

IBM Corporation
Federal Systems Division
Gaithersburg, Maryland

Data constitute the sine qua non of information processing. Without this resource, there would be no way to meet the information requirements of management, of problem solvers, and of operations personnel.

Data make up the building blocks of meaningful information. Other system resources are used to transform these basic building blocks into trend, summary, and exception information for management. It is through the analysis of data that problem solvers establish previously unrecognized relationships and formulate and corroborate hypotheses which, hopefully, result in a step forward toward the solution to a problem. At the operations level, constructive data transformation usually consists of the processes of formatting, collating, sorting, and organizing data in a way which facilitates its use for daily operational functions.

It is easy enough to acknowledge the value of data to a successful information processing system. At the same time, however, there is an inexplicable puzzle which arises when we look deeper into the way this prized resource is treated in many data processing centers. All too often, this resource suffers from a form of management neglect. This neglect, though it is not deliberate, is often the source of serious problems for an information processing system.

In order to show how this neglect occurs, consider first our basic attitudes toward other valuable systems resources: people and equipment. In each case, as with our data, we readily acknowledge the value of the resources -- people and equipment -- our concern is translated into pragmatic, constructive management of the resources. Well-established tools, techniques, and procedures are implemented in an effort to monitor and control the allocation and utilization of personnel and equipment.

Unfortunately, we do not follow through with similar management practices with respect to data, even after we recognize the curcial role that this resource plays in the system. Instead, data resources are left virtually unmanaged, while elaborate measurement, monitoring, and control techniques are used to "maximize" the utilization of people and equipment resources.

No data processing center manager would ever allow his people or his equipment to go unmanaged, for obvious reasons. Yet the dangers involved in allowing data to go unmanaged have not been so recognized. The complexities of current data processing systems -- multi-applications environments, on-line processing and dynamic user requirements -- all compound the far-reaching problems resulting from the unmanaged use of data resources.

It is hard to overstate the impact of the development of multi-application systems on data resources. The most obvious effect of this mode of operation is a strong tendency to fragment data resources. Data is introduced on an ad hoc basis to satisfy the needs of applications A,B,C, . . ., or K severally. Each application has as its primary goal the collection and exploitation of those data resources required to satisfy the information requirements of its own area. This fragmentation is compounded for each area when user requirements change rapidly and when requirements must be met by the application in an on-line mode. Applications

[1]Printed with permission from the February 1973 issue of Data Management, published by Data Processing Management Association, 505 Busse Hwy., Park Ridge, Ill. 60068.

area A will thus tend to build, process, and maintain its own set of files without regard for the needs of B or C. B and C, of course, must meet the pressures exercised by their own sets of users, and they will follow suit in their data utilization practices.

Application A will sometimes introduce, process and maintain data without determining first whether it can capitalize on the fact that B and C are using the same data. Once this cycle of duplication begins, it gives rise to further problems. For example, there is now no control over the consistency of data that is used by all three applications areas, and inevitably A, B, and C will generate information which suffers an integrity problem whose source is this very same data inconsistency.

Fragmentation of data resources often results in a form of application isolationism that hampers an application system's responsiveness to change and growth. New information requirements may entail the use of data that is currently unfamiliar to one of these application areas -- say applications area A -- even though this data is, in fact, resident and available for processing in the files of B or C. Since A is unaware of the availability of this data, it will give a misguided response to the request for new information by (1) advising the prospective user that his request cannot be satisfied since the data from which it would be produced is not available, or (2) erroneously estimating the cost of meeting the new requirements since it is assumed that the data is not already available for processing.

Fragmentation hampers the operation in another way. Data that is maintained and processed by application A over a long period of time may acquire de facto but "informal" users from area B. Area A now finds that it has no more reason to process this data and drops it from its maintenance responsibilities. There is no formal record of the use made of this data by B applications users and, as a result, they are not notified of this change in time for them to indicate that their need for the data still exists. Area B users are now faced with the problem of restoring the data processing and maintenance function themselves. Often, this cannot be achieved in time to meet their regular requirements.

The following list provides brief descriptions of situations which serve as indicators of the problem of unmanaged data.

o Identical data elements are distributed over many files. The degree of data redundancy is sometimes not known and, often, management is unaware of any redundancy problem. Sometimes warranted duplication of data is mistaken for redundancy.

o Programmers expend much effort and time in familiarizing themselves with the systems data resources that are required to meet their own program specifications. Often, this is accomplished only through informal "coffee break" exchanges. Since program data definition labels are not standardized, even this technique of exchanging information on data does not always succeed. It is "simpler" for the programmer to build a new file, rather than expend the time and effort required by such a process.

o Management is generally unaware of the extent of the data its system processes. There is no single inventory of data elements and no single source of information on its data resources available. Moreover, it is unable to characterize its data as to timeliness and periodicity (frequency of update). Therefore, management is in a poor position to judge (1) whether or not it can satisfy a new requirement, (2) how much time and effort it will take, and (3) how responsive (timely) the information it provides will be to the prospective user. Sometimes, as a result, requests for new information are needlessly denied.

o Management has no means of monitoring and controlling the users of its data resources. Hence it is unable to notify users of anticipated changes in its data inventory which are likely to affect these users.

Management attention is usually drawn to the problem of unmanaged data only after confusion in day-to-day operations and a loss of efficiency become pronounced. Soon, management develops the uneasy feeling that the effect of this neglect is more costly than they have been willing to admit.

Unfortunately, they grasp for readily available remedies which, although adequate, may not be best in their environment -- and, in some instances, may be more costly than what is actually required to solve their basic problem.

At this point, it is useful to indicate the premise on which the solution presented here is based: the problem relating to data resource utilization is a management problem.

Its solution, therefore, should be based on sound management judgment after examination and evaluation of relevant evidence. Since this is a management problem, we should reasonably expect at least a part of the solution to be some device, tool or procedure which will support the management functions of monitoring, control, and decision-making.

Adoption of this basic approach enables us to avoid the mistake of viewing the problem simply as a technical matter requiring a technical solution.

This mistake is understandable enough, since there is a strong conviction among data processing managers that the solution to any problem connected with data processing is simply either more software or more complex software.

Also, recognition of these problems, if it occurs at all, often causes an overreaction by management. For example, strong lip service is quickly given to the importance of eliminating "data redundancy.' Management eventually comes to see the elimination of redundant data as its main problem.

The danger here is all too clear -- for, actually, the presence of duplicate data does not in itself constitute conclusive evidence of redundancy. The view that it does is based on the first of two non-sequiturs in management's deliberation on these problems.

Certainly, there are circumstances in which it is desirable to maintain and process duplicate data. What is required in order to determine when duplicate data is, in fact, redundant is an unequivocal, clear description of data resources and of the uses to which this data are put. Only then can management make an intelligent trade-off of extra storage and maintenance versus the benefits of duplicate data storage.

Without this evidence, any remedies invoked to eliminate "redundancy" are based on a non-sequitur: any two occurrences of the same data constitute a gratuitous duplication or redundancy of data. But this kind of faulty reasoning can serve merely to compound the original error which is the failure to monitor and control data utilization in a systematic and deliberate way -- in the same way that we monitor and control the use of other resources.

One obvious remedy aimed at the "elimintion of redundancy" is to combine those files which contain overlapping data. A second approach is to avoid the use of special purpose programs which involve building and maintaining "special" subsets of the facility's data resource. Of course, the approach which has the most impact on a system is to incorporate all data resources within the framework of a single data base management system. Indeed, the number of recently developed systems of this sort is concrete evidence of the proportions that the problem of data management has assumed. As with the first two approaches, there is no doubt that the data base management system approach is well worth its costs when its use is accompanied by careful planning, selection, and implementation.

However, the need for careful planning is crucial, for it is all too easy to suppose, prematurely, that any or all of these remedies constitute the best solution to data management problems in a particular operating environment. Even further, it is risky to assume that these remedies are in fact required to achieve the degree of data management that is desirable for a given environment.

The view that a single data base management system (DBMS) is the only solution or the best solution is based on the second non-sequitur involved in management's response to the problem of data resource fragmentation: data resource fragmentation implies a need for a higher degree of actual data integration achievable only by means of a centralized repository of data; i.e., a data base, managed by a complex and elaborate set of programs.

Of course, once management steps out on this path, they are already tacitly committed to the view that integration and centralization are their primary goals; and, that these goals can be achieved only via programming techniques and data structuring concepts (hiearchies, networks, chains, links, lists, rings, etc.).

Now, they must pay the price for this complexity. In addition to lease or purchase charges for the DBMS, they must reeducate programmers and other users, add new systems support, restructure and convert data, and adjust their overall systems flow to allow for the impact of the new DBMS. It is important to stress, as has already been noted, that this investment is often worth

its costs. But, it is not always the only solution nor is it always the best solution.

The guideline espoused here for management in considering solutions to this problem is to regard its primary objective as the ability to monitor and control its data resource, rather than actual integration and centralization of data. With these objectives in mind, management is in a better position to select a tool which will assist in meeting these objectives -- and no more.

Is there a way to develop evidence on the basis of which management (as a part of its planning function) can decide which if any, of these remedies is in fact required; and, if so, which approach is best for its environment?

There is a method for developing such evidence. And it is one of the ironies of this problem, that the implementation of this method can provide management with a tool which, in itself, may be adequate to meet the objectives of sound data management. This tool is the data element directory (DED).

The concept involved here is certainly not new; yet, this tool has not received the consideration it deserves as a means of developing such evidence, even as an alternative to the other remedies already discussed -- especially the data base management systems approach. The DED is supported by a set of programs for building and maintaining a directory file and for producing information which facilitates effective management of the data resources which this file describes.

Data elements are the basic units which the system processes to yield information. Control of data resources can be achieved only when management has sufficient information on the characteristics and use of data elements.

The DED is thus a single, authorative source of information on data elements, their use, and their organization and format. It is a way of monitoring and controlling data resources without actually integrating and centralizing the data itself. Instead, information on data is integrated and centralized in a single file and is available in hard copy form through a set of basic data resource reports.

There is an entry for each data element that can be counted as a part of the facilities total data resources. For any data element, the directory and its associated programs can provide answers to the following questions:

- Is the data element available for processing by our system?
- What is the significance of the data element; or, more generally, what role does it play in a particular application?
- What is its source?
- What is its location?
- How is it used?
- How is it related to other data elements?
- Who uses it?
- Where is it used (by what application, program, report)?
- How often is it used?

Although directory entries can be varied in both content and format to accommodate the special needs of an individual application, it will contain, at a minimum, answers to the above questions. Examples of the many more questions it can accommodate, depending upon the particular needs of the installation, are given in the question list.

The directory is structured into three basic sections to accommodate data element entries for input processing, data base residency and output processing. Entries appear in one or more such sections depending upon the data processing functions in which the data element plays its most important role. A standard DED label is used to identify each element uniquely. Data element entries are cross-referenced when they appear in more than one section.

Important features of the Directory are the alternate name and keyword entries for each element. These features are designed to enable a user to identify correctly the entries in which he is interested although he either 1) lacks knowledge of the correct DED label for the data element, or 2) is not certain that a data element of that kind is a part of the system's data resources.

The DED is designed for use by management, by analysts, and by programmers.

Although the DED's use as a tool to assist management in planning, monitoring, and controlling its data resources has been stressed, its uses go beyond the scope of management's needs. In many instances, it provides information that is useful to all three classes of users -- management, analysts, and programmers. Some of these are noted here.

The DED can be used in planning the allocation and use of other system resources. Management can consult the DED to determine the availability of data that are required to meet new or changed information requirements and can factor this information into planning and estimating development lead times and manpower requirements.

As has already been noted, it can be used by both management and analysts to determine where data duplication amounts to "true" redundancy. As such, it provides crucial evidence in any planning relating to the use of a DBMS. A related use as a planning tool exists in the similar evidence it provides for the development of a common data base that is shared by several applications. If its use points to the need for a common data base and/or a DBMS, the DED can serve as an integral, required resource to support the implementation and operation of these system enhancements. Information supplied by the DED is required in the coding of data description tables that are essential features of most DBMS's. (It is for these very reasons that the DED should not be regarded simply as a substitute for a DBMS.)

One of the most valuable uses of the DED is as a reference guide to programmers. Whenever familarization with some subset of a facility's data resources is required, obviously, the DED can save much of the programmers' time in acquiring this knowledge. Programmer efficiency is thereby increased.

Also, the DED can be used as a control mechanism to preclude the introduction of redundant or inconsistent data elements. It can be coupled with procedures which require that new data elements be checked against the DED to ascertain that they do not duplicate existing elements unnecessarily, and that they are not inconsistent with existing elements.

Since the DED is structured into input, base, and output data elements, it serves as a useful tool for understanding the information flow of a particular applications area.

This by no means exhausts the practical value of this tool. Its utility is a challenge to the imagination of any manager or analyst who realizes the value of effective data administration.

The single most important task in the development of a DED is the collection of information which will comprise the DED file. A systematic collection procedure is used to ensure that such information is comprehensive, accurate, and consistent. This is achieved through the use of a variety of techniques and through periodic, but brief, interaction with applications area specialists at the installation involved. Collection proceeds through a review of documentation, especially input forms, record layouts, report specifications, and actual, sample reports. Informal discussion with installation analysts supplements this comprehensive documentation review as well as a review of bulletins, forms, regulations, etc. which provide insights into the application -- significance of the data element. Report distribution "schedules of users" are reviewed to determine frequency and identity of users.

Analysts who perform this collection task are equipped with worksheets designed to include all of the relevant information required.

Information collected by this means is then analyzed and refined to detect inconsistencies and possible redundancies. Where apparent redundancies exist, the analysts then determine: (1) that the data duplication is not unintentional; (2) that the inclusion of both versions of the same element is warranted. An explanation of the apparent redundancy when it is warranted is documented. Any "true" redundancies are also noted for subsequent review with data processing management.

The next task in DED development is data base creation. This, of course, involves data base design and data preparation. (Both data base design and data preparation forms and procedures can be relatively simple for the DED.)

The final task is probably of greatest interest to management. Report generation

programs are designed to produce the following two sets of reports: verification reports and usage reports.

Verification Reports

- o Discrepancies - Shows duplicate entries (same DED names or alternate names) and cross references to data elements "not found" in the DED data base. Includes update register information, especially unsuccessful update transactions.

- o Completeness - Identifies "missing" characteristic or use information.

- o Summary File Data - Number of entries in DED (current) by processing area; i.e., input, data base, output.

Usage Reports

- o Common Elements - A report of elements that are shared by several applications, and by users. Also, those data elements which originate from the same source. The complete data element entry is printed.

- o Structure Report - For the file or files specified, a listing which identifies data elements comprising records or segments, by record or segment type within each file. The complete data element entry or record is printed.

- o Alphabetic Listing - Data element list alphabetically by data element identifier. The complete entry is printed when a range of data element ID's is specified; e.g., A-J.

- o Key Word Listing - A key word listing with references to every data element ID which carries these keywords in its definition entry.

The purposes of the DED can be achieved, of course, only if (1) management is well schooled on its potential as a tool, (2) the DED is assiduously maintained, and (3) it is used regularly as a device for monitoring the facility's data resources. Responsibility for these functions reside with a data administrator.

One of the attractive features of the DED concept is that the DED is more easily administered by a single person than are other data management methods. In many cases it is unrealistic to suppose that a single person can fulfill all of the data administration functions usually associated with a DBMS; e.g., data structuring and definition, security procedure set-up, and restart/recovery procedures and control.

With the DED, the data administrator monitors all updates (changes, additions and deletions) to the directory file. System procedures specify that all such changes are subject to his approval. New data element definitions for all major applications files and for new files are subject to his review. Supported by the DED, the data administrator serves as a focal point for data control by checking the directory regularly for possible redundancy and inconsistency.

Since he is familiar with the facility's entire data resource, he is in a good position to alert management to significant trends in data resource use -- and perhaps more important, to signs of abuse of this highly valuable system resource.

DED QUESTION LIST

Identification and Definition

What name is regularly used to refer to this data element?

What abbreviations or alternative names are used to refer to this data element? (Include program mnemonics or labels)

How is it defined?

How is it related to other data elements -- logically or functionally?

Usage

What, if any, are the restrictions on the use of this data element?

Origin

With what organization does it originate?

Destination

To whom is it finally transmitted for use?

Integrity

Who has the responsibility for its integrity?

What validity checking and editing are performed?

What degree of accuracy do its values represent?

Is it rated as to reliability?

Characteristics

What are the size and type of data?

Do any of its characters serve as control characters?

What are typical data values?

What are value ranges?

Is it a unit of measure?

Relevent Documents

What documents or forms describe the data element, its use, or procedures relating to either?

Machine Processing

What applications use it?

What programs use it?

Does it have special data processing significance (e.g., hierarchical root node, sort field, retrieval key)?

What is its input media?

What is its position on input and its relative position in the data base?

Input Processing Frequency

How often is it received?

How often is it processed?

Output Processing Characteristics

In what reports does it appear?

How often is it used in regular and ad hoc reports?

What headers identify this element in such reports?

Is it edited (mask, decode) for outputs?

From what other data elements is it generated (data base and output data elements)?

Data Base Characteristics

In what record and file does this element reside?

What special processing significance does it have?

Is it generated from more than one input data element?

The Cost of Information - an Auditors' Viewpoint

Morey J. Chick, CPA

U.S. General Accounting Office
Philadelphia Regional Office
Philadelphia, Pennsylvania 19106

The need for effective information management in business and government is rarely described in terms of dollars and cents. It is recognized that we pay people, and buy equipment and facilities, to collect and process information. We also pay managers and operators for making money decisions based on information they have at hand. There is little, however, in the way of accounting mechanisms for isolating and measuring the total costs of these information processes. Further, there appears to be less available for measuring the total cost of bad decisions made on bad information. The inability to measure these things often make it difficult to support the need for more attention toward improving information management.

Eleven years experience with the General Accounting Office (GAO) can give one an appreciation for some of the things that can go wrong when information is not adequately managed and just how much money it can cost. Problems that cost money crop up in all phases of information processing e.g. collecting, recording, transmitting, processing, printing or displaying, analyzing and interpreting, storing and retrieving. This paper has been adapted from an article presented in "The GAO Review," Summer 1973 edition. The article attempts to highlight with the limited tools available the dollars and cents of some of the information management problems that can exist in all large operations. It was written to demonstrate the need for more effective information management.

Management of data elements is basic to effective information management. This is the case more than ever now that we are in the midst of ever increasing uses of computers in our data processing and decision-making operations.

Key words: Information costs, ADP, duplicate information collection, unneeded data, inadequate processing, unnecessary output, lack of available data, some questions to ask.

1. Introduction

Federal executives and managers, and auditors too, look upon costs as generally being associated with personnel, facilities, equipment, supplies, and services. Because of their tangible nature, these costs in most cases can be readily measured and accounted for in budgets, accounting records, and financial reports.

537-860 O - 74 - 4

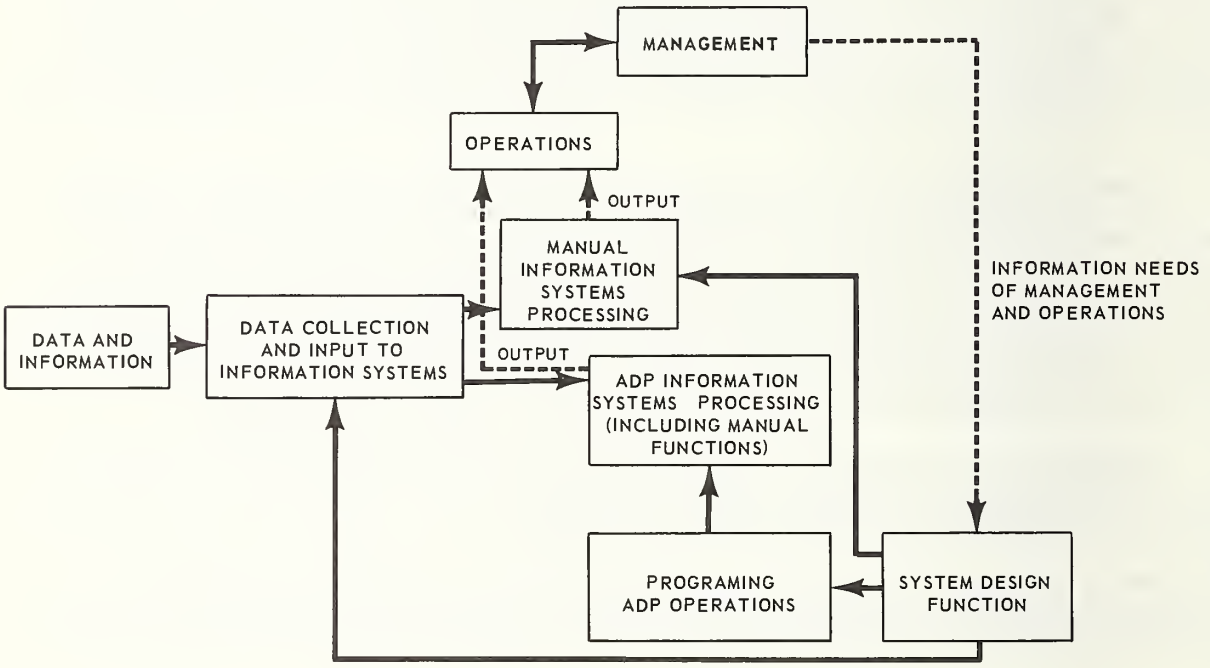
A cost that we often overlook is the cost that flows through, and is a part of, all categories of Federal actions--the cost of information (data).[1] It is very expensive to collect, record, input, and manage the information agencies need to function properly. The costs are incurred regardless of the method of processing--computer or manual.

2. Importance of Information

The extent to which the Federal Government relies on information is widely recognized and does not need to be belabored. Information is needed by all Federal activities, such as inventory control points (ICPs), research and development activities, repair and overhaul facilities, urban area developers, and environmental improvement activities, regardless of their missions.

The chart below shows the flow of information.

FLOW OF INFORMATION



---

[1] Normally, a distinction is made between the terms "data" and "information." Data most often relates to unorganized, sometimes unrecognizable, bits and pieces of facts. Information represents the organized, intelligible, and meaningful results once the data is processed. In this article the terms are considered to be interchangeable.

3. Information Costs--Significance and Measurement Problems

It is virtually impossible to measure the total costs of information, because they are buried in the accounts of (1) programs, (2) salaries, (3) other personnel costs, (4) operation and administration, and (5) other overhead items. They are buried even in contract payments. There is no single place in the Government that can account for the total costs of information. True, the General Services Administration (GSA) publishes an annual summary of Federal ADP activities. The June 1972 summary reported Federal expenditures for ADP at slightly more than $2.3 billion. This figure, however, represented only a small part of the annual costs of information handling and processing. It did not include expenses incurred for

--collection and recording of information;

--employment of indirect support personnel;

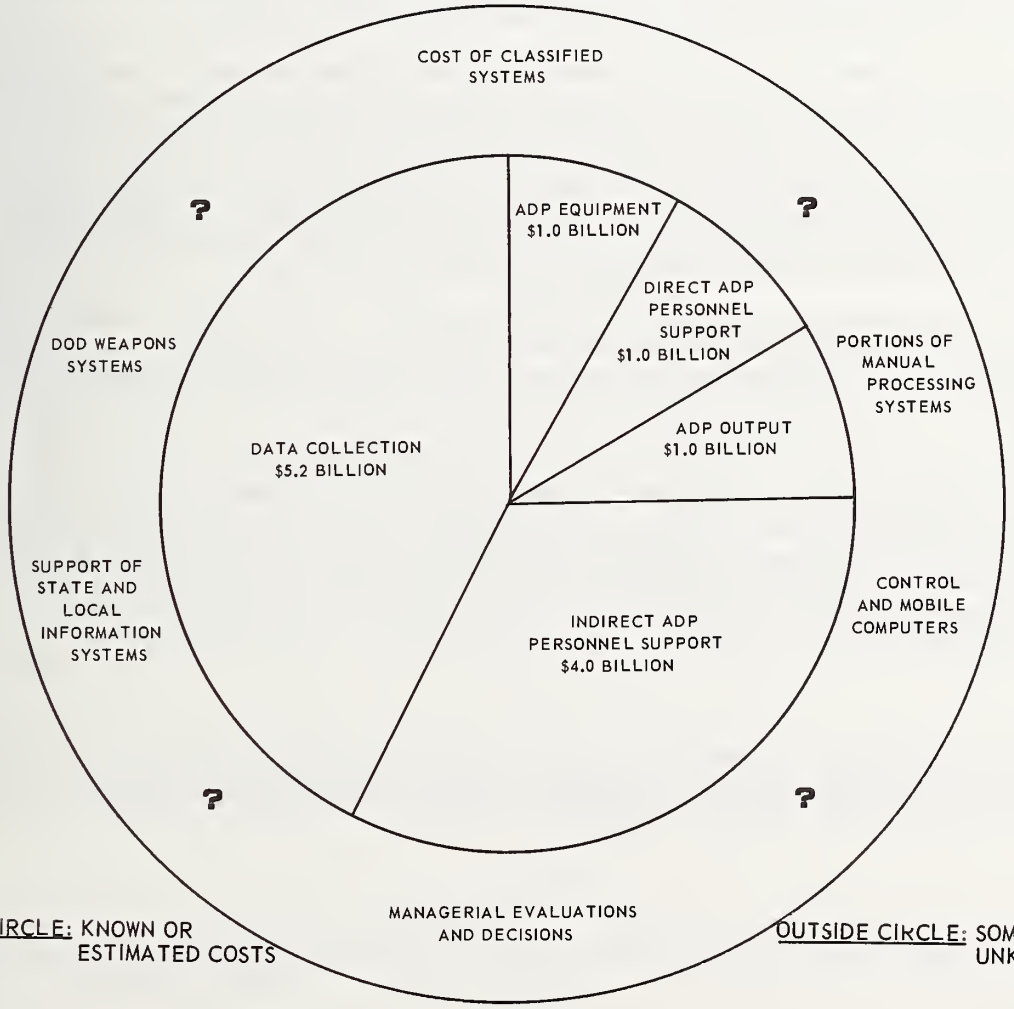
--non-ADP information processing systems;

--all aspects of "special management classification" computers; and

--managerial evaluation, interpretation, and processing of computer output.

In 1970 a Federal agency studied Government information costs. Its comprehensive study showed that the costs of information were significantly more than the $2.3 billion in ADP costs that GSA reported. The study showed that it cost at least $12.2 billion a year to operate Federal information systems. The major costs are shown below.

FEDERAL INFORMATION COSTS



INSIDE CIRCLE: KNOWN OR ESTIMATED COSTS

OUTSIDE CIRCLE: SOME TYPES OF UNKNOWN COSTS

The study showed that the information costs in the Federal Government were enormous.

In considering these enormous costs, we cannot afford to overlook the potential that exists for reducing them. There are numerous ways that data can be mismanaged and misused. Failure to use data also can cost the Government money.

Some of the ways that data activities can increase operational costs follow.

--Collecting data already available.

--Collecting unneeded data.

--Inadequately processing pertinent data.

--Creating unnecessary output.

--Failing to obtain and use needed data.

--Failing to act on pertinent data.

The examples in the following sections illustrate how operating costs were increased by various data management problems.

3.1 Collecting Data Already Available

The $5.2 billion annual cost estimated for collecting data did not include the very high cost of the data input process--e.g., keypunching and verification.

One agency estimated that 25 percent of the data collected by Federal agencies was already in the computer files of other agencies. For example, consider how many agencies are responsible for managing programs dealing with our environment. According to testimony presented before the Subcommittee on Investigations and Oversight of the House Committee on Public Works,[2] no fewer than 14 agencies had jurisdiction, by law or by special expertise, over water quality and water pollution. Consider also the subjects of crime, economic indicators, welfare, chemicals, narcotics control, and air pollution. Responsibilities for collecting data on these and other subjects cross many agency lines.

There is some sharing of information between agencies in these areas. However, about 97 percent of this sharing consists of exchanges of hardcopy forms, which often require reinput of the data into computer-based systems, possibly using different layouts, formats, and codes.

In many cases, however, information is not shared and unnecessary duplication results. On December 8, 1971, GAO issued a report illustrating the significance of this point. The report, "Coordinating Deep-Ocean Geophysical Survey Would Save Money" (B-133188), showed that the Government could save in excess of $20 million if one agency were to collect geophysical data and share it with another. Prior to the review, both agencies had planned to employ vessels in the same area and to independently collect the same or similar information.

---

[2]Hearings on Water Pollution Control Legislation 1971, held on May 25 and 26; June 2, 3, 8, 9, 10, 15, 17, 22, and 24; and July 7, 1971.

3.2 Collecting Unneeded Data

When information systems are designed, analysts, together with management and operational personnel, determine their data requirements. Documentation for the system design should identify, among other things, what data is needed and why and how it is to be processed and used. Hopefully there is a definite need for obtaining the given elements of data that the system is designed to collect.

Tremendous volumes of data are collected in Federal data gathering and processing operations. It has been estimated that more than 3.5 billion data collection documents are generated yearly in Federal data gathering.

To the extent that unneeded data is collected, unnecessary costs are incurred. Added costs for storing and processing are incurred if the unneeded data is entered into information systems.

When information systems are redesigned or replaced, some of the previous data requirements may no longer exist. However, collecting data may be perpetuated because of inadequate systems analysis or design or through management oversight.

A recent project of the Office of Management and Budget (OMB) to improve reporting and to reduce paperwork resulted in reported savings of $270 million.[3] Much of that amount was saved by eliminating reports or information in them that was not needed.

Paperwork studies, such as this are not normally a recurring internal agency operation. They do help to reduce overall paperwork in the Government, but in the long run information requirements appear to develop faster than obsolete requirements are eliminated.

OMB plans to take corrective actions. All agency officials and auditors should seek ways to eliminate the collection of unneeded data.

3.3 Inadequately Processing Pertinent Data

Collecting needed data and injecting it into Federal information systems does not guarantee that computer processing and output will give the managers and operating personnel the information they need for making the right decisions.

Processing relates to the procedures for analyzing data--compiling, combining, calculating, rearranging, structuring, sorting,and interpreting--to produce information needed for decision-making. In ADP systems, computer programs tell the central processing unit how to process input data. Inadequate or erroneous criteria and/or errors in the computer programs can result in uneconomical decisions.

In August 1970 GAO reported on an audit made at a Federal installation on its efforts to reclaim usable parts, from excess equipment. Its computer system was to screen and identify usable parts which normally were on excess equipment and which could be used to support other equipment still in use. The program was written to screen the parts and to output lists of the needed parts. Using these lists, Government personnel would reclaim the needed parts and ship them to designated locations in the agency's supply system.

Inadequate criteria for and programing errors in the computer program processing the information affected the computer's determinations. As a result the installation failed to reclaim more than $410,000 worth of needed and usable parts, as seen in a test of two equipment reclamation projects, and instead purchased more than $250,000 worth of these parts.

---

[3] "Report on the Government-Wide Project to Improve Federal Reporting and Reduce Related Paperwork" (June 12, 1972).

The inadequate criteria and programing errors resulted in excluding needed and usable parts from the output lists and in eliminating from the screening process

--parts that contained components that could deteriorate (even though the components could be replaced);

--parts for which requirements were shown in the tens of thousands of units, since the computer was programed to screen only the last four digits of the requirements quantities;

--parts that were coded as being slow moving even though recent recorded trends showed significant increased usage; and

--parts that had not been used in overhauling the type of equipment no longer needed even though they had been used in overhauling other equipment types.

The data input was current, accurate, and complete and information needed to make the proper decisions was in the files. However, inadequate processing resulted in unnecessary costs to the Government. The report to the Congress (B-157373) was issued on August 6, 1970.

3.4 Creating Unnecessary Output

Estimates of Federal computer output costs run at $1 billion annually for 90,000 periodically printed reports. Printed output which has no use or which duplicates information in other printed output results in unnecessary costs to the Government.

For example, GAO reviewed the "credit returns" program of selected Government supply centers. This program was but a small part of a center's overall operations in terms of (1) the percentage of ADP time used and (2) the personnel employed. Nevertheless, the ADP systems at the centers were printing reports which were not needed. At one center about 90 percent of the printed reports either were not used or duplicated information contained in other reports. Eliminating most of the credit return reports and consolidating others would save the Government almost $50,000 a year at that center. These savings were substantial, considering the scope and cost of the program within the center. Similar situations existed at other centers. The GAO report (B-161766) was issued on June 27, 1967.

3.5 Failing to Obtain and Use Needed Data

The Government incurs unnecessary costs in addition to those included in the previously mentioned $12.2 billion, because agencies make management and operating decisions without the benefit of pertinent data available from other agencies.

A recent GAO review of the Government's procurement of drugs illustrates the above point. The agencies involved independently bought a wide range of the same drug products without exchanging information on prices, manufacturer contracting preferences and pricing policies,and inventories available for interdepartmental use. GAO's tests showed that the Government had spent about $800,000 unnecessarily because such information had not been exchanged. Each agency had information that would have benefited the others and would have saved the Government money.

One agency purchased annually about $250,000 worth of certain drug products from one manufacturer. Although the manufacturer had told the agency that it did not offer discounts on these products, it was at the same time discounting them by as much as 30 percent on another agency's contracts. Because it did not have this information, the agency could not take such alternative actions as (1) ordering the products from the second agency, (2) adding its requirements to another agency's contracts, (3) further negotiating prices with the manufacturer, and (4) purchasing alternative products from other sources.

In testimony before the Monopoly Subcommittee of the Senate Select Committee on Small Business on May 10, 1972,[4] the Comptroller General recommended that procurement and requirements information be shared between agencies.

3.6 Failing to Act on Pertinent Data

The failure of operating and management personnel to act on available pertinent data is another aspect of costs associated with acts of omission. If needed information is economically collected and adequately processed and summarized but is not acted upon, the effort is wasted and potential savings may be lost.

One Federal office had a longstanding policy of screening its inventory of excess repair parts for those parts which can be furnished to contractors manufacturing major equipment. When the Government furnishes parts to its contractors, the Government can negotiate lower contract prices or price reductions in existing contracts and thus realize savings.

The office had implemented this policy until 1966 when it stopped because it thought that no potential existed for offering excess parts to Government contractors. The office did not make periodic followup reviews to determine whether such potential had developed.

The office collected and maintained sufficient and adequate data for making such reviews. Its system contained such elements of information on each item (repair parts and equipment) of supply as (1) requirements data, (2) inventory balances, (3) applications (the major equipment in which each part was used), (4) outstanding equipment contracts and purchase requisitions, and (5) planned production over several years.

Because the office had not examined and related this data, it failed to recognize the potential that existed. GAO identified $2.7 million worth of excess parts that could have been used as Government-furnished parts on existing fiscal years 1969 and 1970 production contracts with attendant reduced or lower prices. The report (B-146727) was issued on December 11, 1969.

We presented these facts to agency officials during our audit, but it was too late to use all but about $130,000 worth of the parts on these contracts. The agency, however, resumed active screening and has since used $434,000 worth of excess parts on fiscal year 1971 production contracts.

3.7 Other Information Costs

Using inaccurate information or failing to use current information can result in uneconomical or erroneous decisions.

---

[4] Competitive Problems in the Drug Industry, vol. 5. GAO issued a report, B-164031(2) on "How to Improve the Procurement and Supply of Drugs in the Federal Government" on December 6, 1973.

Unnecessary information handling and processing, whether by the computer or by manual means, is expensive and inefficient. It may also delay processing more important information.

Both GAO and the executive agencies often have given attention to the design of information. Duplications inherent in the independent design of the same or similar type of systems are costly. Wherever possible, products of already expended design efforts should be used.

Inadequately designed systems represent wasted effort and are often the result of (1) poor planning, (2) inadequate feasibility studies, and (3) a lack of communication between managers, systems analysts, programers, and operating personnel.

To emphasize the need for improved systems design, consider that it costs from $20,000 to $1.5 million, exclusive of hardware costs, to design a computer-based information system, depending on its complexity. Consider also that an estimated 1,000 new computer-based information systems are being designed in the Federal Government each year.

4. Concluding Remarks

In December 1971 a task force commissioned by OMB to improve ADP systems analysis and programing capabilities in the Government recommended that data be managed as a resource.[5] Such resources as equipment, supplies, facilities and personnel cost money, and the Government has provided the criteria, apparatus, and means for managing them. The task force said that data also costs money and cited the $2.3 billion cost for ADP reported by GSA for fiscal year 1971. The task force concluded that, since data represents a substantial investment, it should be considered a resource in the economic sense and resource management principles should therefore be applied.

Considering that Federal information costs are several times $2.3 billion, I believe the recommendation of the task force to be conceptually sound. Implementing such a recommendation, however, may be difficult at present because there are no (1) established accounting systems for information, (2) procedures of identifying and allocating their costs, and (3) apparatus and means for managing information as we manage physical resources. It appears logical that the Government move in the direction of managing data as a resource by studying its feasibility.

Until this is done, there are still ways to improve information-processing operations and to reduce their costs. Thoroughly analyzing and evaluating each processing step seems to be an appropriate way to start improving operations and reducing costs. Some questions for which we should seek answers follow.

---

[5] "Office of Management and Budget Project to Improve the ADP Systems Analysis and Computer Programming Capability of the Federal Government," December 17, 1971.

Information-processing steps	Some questions to ask
Collecting	Is all the needed information being collected? Has needed information been collected elsewhere? Can this information be obtained in machine-sensible form for input into the system? Is the information current, accurate, and complete? Is unneeded information being collected? How is information being collected? Is there a better way?
Recording	Who records the information? Where? How often? How is recording accuracy determined?
Transmitting	Where is the information sent? Should it be sent there? How is it sent? Is there a better way to send it? Should it be sent elsewhere?
Processing	What are the objectives of processing? Is the processing logical? Is it efficient? Are there processing steps that are not needed? Should there be additional processing steps? Is processing being performed as planned?
Output	What is the nature of the output? What data does it show? Is it used? How is it used? Is it shown elsewhere?
Analyzing and interpreting	What criteria exist? What instructions are given to personnel? Are the criteria and instructions adequate? Are they being followed?
Storing and retrieving	What means of storage and retrieval are used? How is information retrieved? Should it be stored? How long should it be stored? Is information that should be stored being disposed of?
System design	How is the system being designed? Is a feasibility study being made? Are there existing systems of the same type available? To what extent are they being used in designing this system? How is the design being coordinated with management and operating personnel? Are all of their needs being considered?

ADDENDUM TO PROCEEDINGS

Below are replies to questions I received at the symposium. Several questions were similar and I have therefore combined them where possible. If you are not satisfied with a reply or wish additional information, I invite you to contact me to discuss the subject further.

QUESTION: What is the impact of the value of the data to your discussion of information costs? If the data has little or no value, what sense does cost reduction make?

REPLY : It is my contention that we should not spend money to obtain information that has little or no value to us. The expensive data collection process should be discontinued as soon as it is determined that certain data has no use in our decision making operation unless it is required for other reasons, e.g. by law. To continue the process would be to perpetuate unnecessary costs.

I also believe that a continuing active effort is needed to evaluate the data being collected in order to keep abreast of the "value" of that data to the collecting organization.

QUESTION: What were the basic problems in the Federal data standards program and when will GAO's report on the subject be issued? How can I obtain a copy of the report when it is issued?

REPLY : Because of the current status of report processing, I would rather not specifically comment on the existing problems until the report is formally issued. A reply to our draft report is due from the Secretary of Commerce and the report could be issued a few weeks after it is received.

Generally, some of the subjects discussed about the Federal program deal with such matters as approaches and philosophies to data standardization, semantics, cost benefits, resources and priorities, coordination and guidance, the role of the program coordinators, etc.

When the report is issued, copies may be obtained by sending a check or money order for $1 to the U.S. General Accounting Office, Room 6417, 441 G. Street N.W., Washington, D.C. 20548. When ordering the report, use the B-number, date, and title (if available) to describe the report. If you are not aware of these facts, contact me in the Philadelphia Regional Office (215-597-4330) and I will be glad to assist you in obtaining the report.

QUESTION: Have you personally encountered any Federal agency who now is operating a good, effective data management program (not merely a data standardization program)?

REPLY : It is a very large government and I've been exposed, on a detail level, to only a relatively small portion of it. My experience, therefore, should in no way be considered all inclusive. No organization is perfect in their data management activities. Some are better than others.

Several agencies of the Department of Defense are probably more advanced than most. I've observed that the Naval Command System Support Activity (NAVCOSSACT) and the Logistics Data Element Standardization and Management Office (LOGDESMO) appear to be on the threshold of going beyond data standardization to developing effective data management tools.

More emphasis must be placed and guidance given on data management activities in Government and industry and not simply on data standardization alone.

QUESTION: Should data standardization be a function separate from system design, development, implementation and maintenance?

REPLY : I believe that data standardization should be an integral part of these functions and not separated from them. The data is at the very core of the system and its source, processing, and use must be considered at least as important as the equipment, people, programing language, logic, documentation, etc.

QUESTION: Did the General Accounting Office report on Federal drug purchasing present an estimate of the cost or other effort that would be required to achieve the data standards that GAO believed was needed? How would this compare to potential benefits?

REPLY : The basic standard recommended by GAO, the National Drug Code, already has been developed and currently represents an under utilized asset. Other standards, of course, would have to be developed to effect the full exchange of data that we believe is required. Much of this data is already contained in computer systems of the respective agencies.

GAO never intended to attempt to measure the full benefits that would be attained as a result of our recommendations. The $867,000 in avoidable costs reported by GAO was based on a test of only $13 million in purchases over a 2-year period. No projection was made. The Federal Government's total drug purchases for the same 2-year period, excluding Medicare, Medicaid, etc., was about $550 million. Total avoidable costs, in my personal opinion, are many times the $867,000 reported. Since they are of a recurring nature, I believe that developing standards for exchanges recommended by GAO will be well worth the cost.

QUESTION: Regarding pharmaceuticals, doesn't the Government use Federal Stock Numbers (FSNs) pursuant to the Federal Cataloging and Standardization Program, (41 CFR 101-29 or 41 CFR 101-30)? If not, are drug products exempt from Federal cataloging?

REPLY : FSNs are used to identify pharmaceuticals managed centrally by the Department of Defense, Veterans Administration, and/or the Department of Health, Education, and Welfare. FSNs are also assigned to many pharmaceuticals not managed centrally, but the central manager's catalogs do not list these items. The hospitals and clinics purchasing items directly from vendors were apparently not made aware of these FSN assignments since their records and purchase orders identified the drugs by (1) description i.e. brand, size, strength doses, etc., (2) manufacturer's number, and/or (3) locally assigned numbers. In addition, other drug products are not assigned FSNs because they are new, or because they do not appear on a Federal Supply Schedule or other Government contract.

On the Connections Between Data and Things in the Real World

Perry Crawford, Jr.
Advanced Systems Development
International Business Machines Corporation
Yorktown Heights, New York 10598

1974 marks the tenth anniversary of the beginning of sustained effort to establish on a broad basis a capability to interchange data among independent data systems. In 1964 the Department of Defense formally established its program on data standardization; steps were taken to extend data standardization to other government agencies; steps were taken to organize what is now the American National Standards Institutes Committee X3L8 on Representations of Data Elements.

The policy statement and directive formally establishing the DOD Data Standardization Program contained definitions of technical terms that had become well established in DOD; data element, data item, data chain and data use identifier were among these. When the X3L8 program began in 1965, the DOD terms were naturally proposed as a basic technical vocabulary; the minutes of the first meeting record the DOD terms; they record also the beginnings of controversy concerning these terms and the views of data and standardization of data that the terms undertake to express.

From the beginning of the X3L8 program vigorous efforts were made to formulate views of data and definitions of data terms that would meet needs of the program and be accepted by all participants. In 1966, a "Data Science Task Group" was established to help accomplish this result. The chief output of the Task Group was a formulation of views of data and definition of data terms that stood as an alternative to the DOD formulation. The proposals of the Task Group were not accepted by the membership of X3L8; the views of data and definitions of data terms used in the X3L8 program are essentially those of the original DOD proposals.

However, there are those who saw merit in the proposals of the Task Group when they were first made in 1967, and have retained interest in them. Dave Savidge is one of these and he has asked me to put before this symposium a review of these proposals and the problems that they present.

The task force proposals can most easily be reviewed by using the diagram (foil) used in the original documentation and presentation of the proposals. Here are depicted two approaches to establishing connections between the strings of characters that we call data and the things in the real world that the data represent.

On the one hand, we can start with the strings of characters that may be contained in a particular field of a particular record in a data system; we can ascertain the individual strings used and what they mean (data items); we can ascertain the set of strings used (data element); and we can ascertain the use of the strings in the particular field in terms of what information is conveyed (data use identifier).

We can, on the other hand, start with the things in the real world concerning which we wish to be informed. We can establish sets of things such that by designating a particular member of a set and a particular time we can convey information; we can establish sets of identifiers that allow members of the set of things to be designated in various situations.

The first approach came to be known as inside-out; the second as outside-in. The outside-in approach had in 1967 and has today strong proponents; however the established approaches to data processing and to views and definitions of data are based on the inside-out approach. Whatever its merits, the outside-in approach was theoretical and unproven; programs of national and international data standardization had to proceed with practical approaches of proven workability.

However, it is not only in the context of data standardization the outside-in approaches have presented themselves; from the beginning of the modern computer era to the present day, outside-in approaches have been pursued toward a number of objectives.

In the 1950's energetic efforts were made to extend languages used to define business activities to permit computerization; decision tables and compilers for translating problem statements containing decision tables into programs were a result of these efforts.

The Language Structure Group of CODASYL was organized at the same time as the COBOL Committee to establish a basis for the development of machine independent problem defining languages. The LSG undertook to develop views of data as representations of "objects and events in the real world" and to develop an algebra for working with information and data in ways independent of machine procedures.

The Air Force in the early 1960's defined a requirement for a problem statement method that would "imply" the programming required to treat the problem.

In the "TAG" technique, the data content of inputs and outputs can be defined -- to a point -- in terms that are machine independent. The ISDOS project at the University of Michigan has, since the late 1960's, been working on extensions of the TAG approach to provide a complete definition of information requirements -- inputs and outputs -- in machine independent terms.

In the late 1960's work on data management and data bases has focused increasingly on achieving data independence. The approaches being pursued by GUIDE are in terms of definitions of the "entities" represented by the data that are independent of the specific form of the data.

In the 1970's work has resumed on "automatic programming" as "automatic programming" was understood in the 1950's -- generation of outputs from definitions of outputs. Work is underway in the Automatic Programming Department of MIT's Project MAC on techniques for eliciting from functional people "knowledge domains" and definitions of reports and for generating the reports from the stored "knowledge" and report defintions.

In all of these areas, approaches having much in common with the outside-in approach of the Data Science Task Group have been pursued energetically by capable, experienced and motivated people, yet the promise held out by the outside-in approach has not been realized.

We have to ask why is this? What is missing in our attempts to bring the outside-in approach to workable status?

In the time that I have, I would like to propose three things that, if they are not missing in our attempts, are not given the kind of consideration they need: The handling of time; the handling of the distinction between identifiers and the subjects they identify; the handling of the distinction between problem definition and problem solution.

First the handling of time. The work of the CODASYL Language Structure Group was an attempt to formalize and extend work that had been going on for many years in the area of business languages and problem statement. In its 1962 report the Group presented results based on two primitive rules:

. Each property has a set of values associated with it

. There is one and only one value associated with each property of each entity.

Based on these rules, the Group formulated postulates and an "Information Algebra" based on modern algebra and set theory.

I propose that one of the key things missing from the approach of the language Structure Group is, in the second rule, the phrase "at a given time":

> There is one and only one value associated with each property of each entity at a given time.

I propose that an adequate treatment of data requires that the question "When?" or "How long?" that apply to data values is attended to systematically from the very outset. In the approach of the Language Structure Group, as in data processing generally, time is attended to in terms of file update; I propose that this is not adequate.

The approach to time in data processing was brought into question by Christopher Strachey in his contribution to the issue of Scientific American devoted to information in 1966. Strachey emphasized that:

> "Programming presents us with certain new questions that are not present, or at least not important, in any other branch of mathematics. Mathematics in general does not recognize the existence of variables... that is values varying over a period of time... In programming on the other hand we deal with time-varying variables by the very nature of the process."

In concluding, Strachey proposes that we will need to develop new concepts in order to get a firm grasp of the situation. He proposes that the way to develop the concepts is through research of meaning, which introduces the second area in which I propose the need for new approaches.

I referred to this area earlier in terms of handling the distinction between identifiers and the subjects that they identify; but this is the area of meaning and semantics.

It is not just in data processing that the contrast between outside-in approaches and inside-out approaches present themselves; in each of us, in each of our waking moments this contrast is presented: whether, on the one hand, as an individual, we give first priority to the direct perception of objects and events in our environment and a lower priority to the choice of names and other labels for the objects and events, or whether we switch these priorities.

Most of us, most of the time in our every day perceptions give first priority to the label; it is an act requiring specific attention to attend in specific terms to what the label refers. I propose that this every day approach is counterpart to the inside-out approach in data processing.

I think that most of us, when we stop to think about it, would accept that the "natural" or "logical" approach in our perceptions is objects and events first and labels second; yet we recognize that from the beginning of our schooling the emphasis has been on the labels and

537-860 O - 74 - 5

their proper handling.

I propose that development of the outside-in approach in data processing depends upon development in each of us of an improved ability to separate the objects and events from the labels and to handle questions concerning the meaning of labels in systematic ways.

In an article reviewing his long experience with language and information, Anthony Ottinger had this to say:

> "All of our understanding of the mechanics of language built up from real or fancied building blocks seems to stop dead before the question of meaning...it almost seems as if the perception of meaning were primary and everything else a consequence of understanding meaning. If this were true, linguistics would have to be built anew, because all of our linguistic theories...seem to start off with the notion of building blocks...and we put these together by rules at various levels... Finally, we say we have...this string of symbols and there is meaning in them."

I propose that Ottinger is here affirming that the established approach in linguistics is inside-out; he affirms also the possibility of an alternative approach:

> "When we try to turn this around...we seem to hit a dead end..."

The third area in which I propose new approaches are needed I spoke of as handling the distinction between problem definition and problem solution.

The report of the Language Structure Group affirmed the importance of separating the statement of the problem from the algorithm for treating the problem; in the other areas of data processing reviewed earlier where outside-in approaches have been and are being attempted -- implicit programming, TAG and ISDOS, Project MAC -- the essential task is defining means for stating problems independently of means eventually selected for treating and solving problems.

Again, each of us, when we stop to think about it, acknowledges the fundamental importance of arriving at the best definition we can of the problem with minimum implications for how the problem may eventually be solved. However, when we don't think about it -- when we react to the pressures of a problem situation -- most of us tend to jump to solutions, even to define problems in terms of proposed solutions.

I propose to you that approaches to problems that put proposed solutions ahead of first rate problem definitions qualify also as versions of inside-out approaches. I propose that developing our ability to put problem definition first -- to take outside in approaches to all of the kinds of problem situations that confront us is of transcendent importance.

One of the ways we talk about the separation of problem definitions from problem solutions in everyday affairs to to talk about the separation of definitions of what is to be accomplished from definitions of how to best accomplish what is defined for accomplishment. In "The Human Use of Human Beings" Norbert Weiner was expressing his grave concern that the new information systems would not in fact secure the human benefits that were possible; his concerns come to a focus in the last chapters where on three occasions he speaks of the difficulty of separating the what from the how; on the first of these he says:

> "Our papers have been making a great deal of American "know-how" ever since we had the misfortune to discover the atomic bomb. There is one quality more important than know-how and we can't accuse the United States of any undue amount of it. This is "know-what" by which we determine not only how to accomplish our

purposes, but what our purposes are to be."

In all human realms the definition of purposes -- of the what--is crucial; in the realm of information systems to extend human abilities the definition of the what independently of the how is crucial to some higher power.

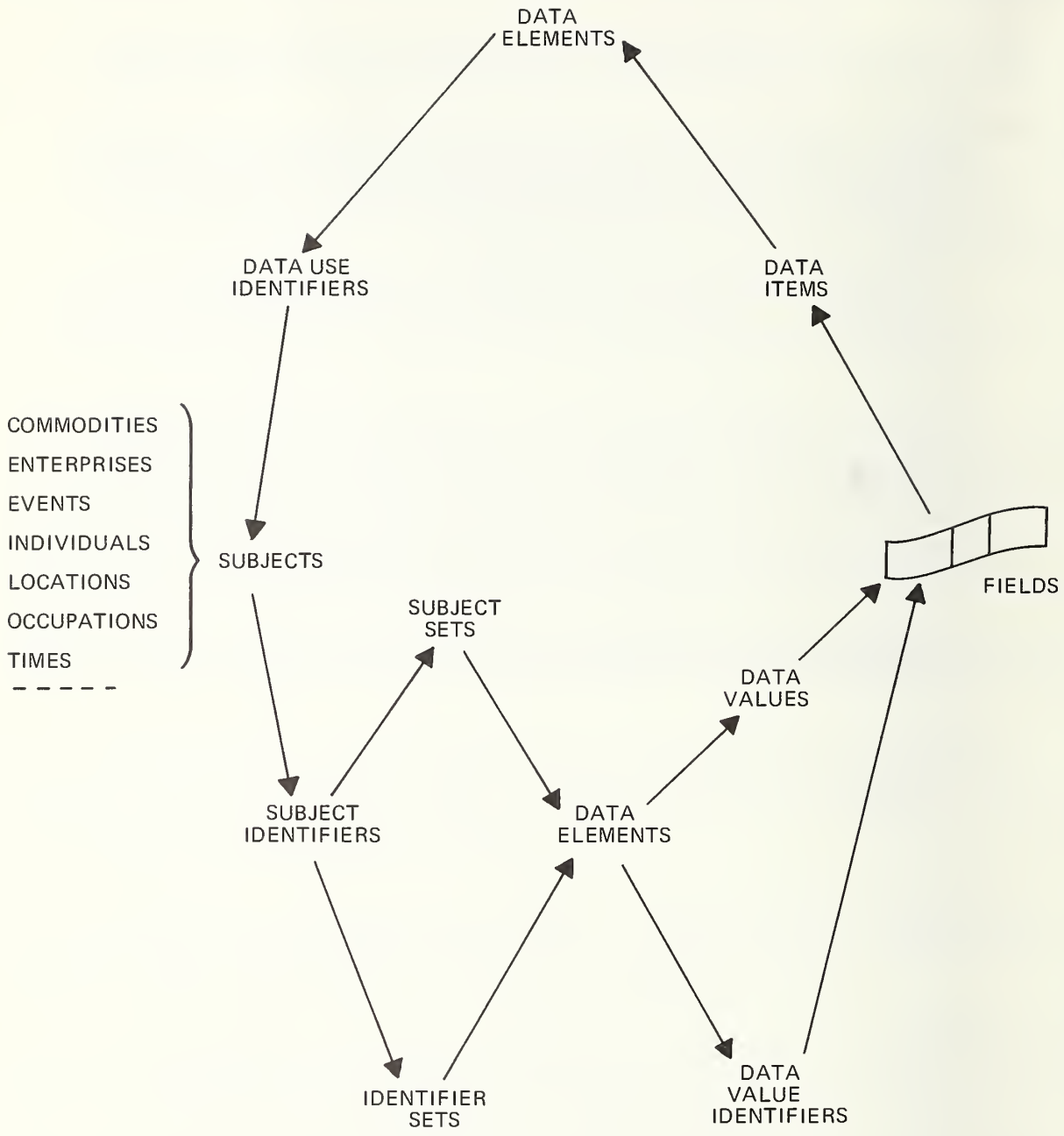
- At one level in information systems, the definition of the what is the starting point for designing the codes and sets of codes that identify subjects and constitute the data to be processed and interchanged:

- At a higher level, the definition of the what is the starting point for designing the messages that are to be interchanged among information systems and delivered by information systems as outputs in ways that are independent of how the messages are handled inside the systems.

- At the level of the end user, the definition of the what independently of the how is the basis for designing more productive and more efficient systems for the conduct of human affairs and for making best use of information system support of design.

We have reviewed briefly some of the background of the pursuit of outside-in approaches in data processing; we have noted some places where these approaches show promise and have been and are being pursued; we have considered some of the reasons why progress in the development of outside-in approaches has not been more rapid than it has. We have suggested -- with the backing of some distinguished students of the information field -- that these reasons include the need to develop new mathematical approaches, the need to develop new approaches to linguistics, and the need to move from a "know-how" to a "know-what" orientation.

To satisfy these needs is a big order; I propose to you that we are in fact called upon to satisfy them and that by concerted action by our profession we can do it.


DATA ELEMENTS
DATA USE IDENTIFIERS
DATA ITEMS
COMMODITIES
ENTERPRISES
EVENTS
INDIVIDUALS
LOCATIONS
OCCUPATIONS
TIMES
SUBJECTS
FIELDS
SUBJECT SETS
DATA VALUES
SUBJECT IDENTIFIERS
DATA ELEMENTS
IDENTIFIER SETS
DATA VALUE IDENTIFIERS

REFERENCES

1. Perry Crawford, Jr., "An Approach to Data Standardization". X3.8.1/25 (X3.8/54) 66-12-30.

2. CODASYL Language Structure Group Phase I Report: "An Information Algebra", ACM Communications, April 1962.

3. Christopher Strachey, "System Analysis and Programming", Scientific American, September, 1966.

4. Anthony Ottinger, "Language and Information", American Documentation, July, 1968.

5. Norbert Weiner, "The Human Use of Human Beings", Houghton-Mifflin, 1950.

US Army Materiel Command (AMC)
Progression From Reports Control
To Data Element Management

Edith F. Curd

Reports Management Branch
Directorate for Management Information Systems
US Army Materiel Command

AMC's first attempt at controlling costs in the reporting area began by establishing monitorship of all reporting requirements. This monitorship consisted of requiring a written justification from the requestor for any new report to be prepared and an annual review of that report to insure its continued necessity. As automation became more prevalent, it was recognized that if maximum savings and efficiency were to be realized, automated products must also be included in the program. The extension of reports management to ADP products greatly increased the savings realized from the program, especially in the area of products from standard AMC programs, where the same computer programs are run at multiple computer sites located at various AMC field installations. An analysis of the savings from management of the ADP products, however, brought on the realization that the majority of the costs incurred was not in the production of the products but in the gathering and maintenance costs of the data elements contained in the products. This realization resulted in the final emergence of the AMC philosophy of data element management. Tools utilized in this program are:

a. The AMC Data Element Dictionary (DED) System which is being adopted this year as the DED standard system for Department of the Army.

b. Standardization of data elements (DE) in AMC.

c. Matrix analysis of DE concept for identification of redundant/overlapping ADP products.

d. DE Management Base files (DED, DE Characteristics, Products Characteristics) for DE management.

e. Standardization of DE in AMC standard systems programming modules.

Key words: Reports control; data element management; Data Element Dictionary (DED); data element standardization; data element matrix analysis; data element management base files; data element characteristics; products characteristics; programming modules; Central System Design Agency (CSDA); Army Materiel Command (AMC); Key Word In Context (KWIC).

1. Introduction

The US Army Materiel Command (AMC) is primarily responsible for the acquisition and maintenance of the US Army's equipment. In order to properly manage this voluminous operation, information is consolidated and/or extracted from our base level operating systems data to provide a basis for managerial decisions. Like all reporting systems, however, our reports are inclined to progress from "necessary" to "nice to have" to "unnecessary" with a resultant spiralling of correlated costs unless controls for essentiality are maintained.

2. Historical Reports Control Procedures

AMC's first attempts at controlling costs of reports began by monitorship of all reporting requirements. This consisted of requiring written justification from requestors for all new reports, as well as annual reviews for all recurring reports to insure necessity for continuance. As this program originally was conceived by Department of the Army (DA) prior to the era of automation, it was addressed primarily towards reports prepared manually.

3. Current Reports/ADP Products Management

As automation became more prevalent, it was recognized that if maximum savings and efficiency were to be realized, automated products must also be included in the program. Rationale for this decision could be based on volume alone, if for no other reason. DA, and subsequently AMC, extended its reports control program to include automatic data processing products, with the exception of such items as program tapes or debugging listings. The extension of reports management to ADP products greatly increased the savings realized from the program, especially in the area of products from standard AMC programs.

4. AMC Standard Computer Systems

4.1 System-Wide Project for Electronic Equipment at Depots Extended (SPEEDEX)

To illustrate, AMC has a standard depot system (SPEEDEX) which was designed and programmed by one AMC Central System Design Agency (CSDA); the system produces approximately 1000 output products. These products actually are being produced on 11 separate computers at 11 separate depots, for a total of approximately 11,000 products. When one product can be eliminated from the system, it results in savings of machine processing and personnel handling at 11 computer sites.

4.2 AMC Logistics Program Hardcore Automated (ALPHA)

A second AMC standard system (ALPHA), one for the major subordinate commands (MSC), was designed and programmed by a different AMC CSDA. This system currently produces approximately 500 products and is to be run at seven MSC's for a total of 3500 products. The number of products is scheduled for expansion.

4.3 Test, Evaluation, Analysis, and Management Uniformity Plan (TEAMUP)

A third AMC standard system (TEAMUP), one for test and evaluation activities, was designed and programmed by a third AMC CSDA. It generates approximately 600 products at eight installations for a total of 4800 products.

5. Data Element Management

An analysis of the savings from management of the ADP products, however, brought on the realization that the majority of the costs incurred was not in the production of the products but in the gathering and maintenance costs of the data elements contained in the products. This realization resulted in the final emergence of the AMC philosophy of data element management.

Our data element management program is still in its infancy, even though much progress has been made, primarily in the area of constructing our tools for the program.

5.1 Standardization of Data Elements (DE)

One of our first steps was to standardize our data elements (DE) in our reports and ADP products. Our regulations carry the stipulations that all DE's must either be standardized or submitted for standardization prior to approval of preparation of a report or ADP product. To date, approximately 9000 AMC DE's have been standardized.

5.2 AMC Data Element Dictionary (DED)

Our most significant accomplishment has been the establishment of an AMC Data Element Dictionary (DED). The DED contains the 9000 AMC standardized DE's; more are being added as standardization is completed.

Computer checks for duplication of DE's are accomplished prior to entry in the DED by comparison of mneumonics and/or titles. A manual check can be made through our Key Word In Context (KWIC) system where each word in the title is listed alphabetically. We are currently investigating generic analysis search techiques to more efficiently locate standardized DE's and to prevent redundant standardization.

The DED is updated monthly and microfilm copies distributed throughout AMC. On a semi-annual basis, printed copies are produced and disseminated.

Department of the Army (DA) has recently selected the AMC DED System as the DA standard DEDS. It is scheduled for proliferation in DA during this fiscal year.

5.3 Matrix Analysis of DE's

One of our DE Management Program's milestones which is still in the conceptual stage is the development of a DE matrix analysis listing. A listing such as shown in Table 1 below could eliminate unnecessary data processing design, programming and machine expenditure. When a new report is needed, input to the computer of the required DE's would produce the identification of existing reports which contain the DE's or the fact no such report exists.

Table 1. Report "X" DE Matrix Analysis

	Report A	Report G	Report F	Report I	Report J	No Report
DE 1	X		X			
DE 2	X			X		
DE 3	X	X	X			
DE 4	X				X	
Etc.						X

5.4 DE Management Base Files

In order to produce the above matrix analysis, two master files are being developed. One, the DE Characteristic File, contains all data which is homogenous to a DE regardles of the file location of that DE or of the various reports in which it is contained. Examples of the characteristics are costs of gathering, maintaining and extracting the DE, number of times accessed and identification of reports in which the DE appears. The second master file is the Reports Characteristic File, which contains characteristics of the reports produced, including identification of the DE's in each report. This file can be used to automatically schedule production of reports as well as review schedules.

5.5 DE Standardization in Programming Modules

Some work has been done by AMC in DE standardization in programming modules, but we still have more progress to make. Out ultimate goal is that DE identification in our files will correlate to our standardized titles. This will facilitate ease of identification and communication between our data processing and functional personnel.

6. Summary

AMC's timeframe to completely phase from its Reports Management to Data Element Management is five years. We do not feel it will be an easy task, however, it is anticipated the results will enable AMC to maximize efficient data flow while minimizing cost of information through standardization of data elements.[1]

1/ Observations and conclusions presented in this paper concerning data standards do not necessarily relate to the DOD Standardization Program.

1. Question: Can a copy of the AMC Data Element Dictionary be obtained? If so, how?

Answer: Copies of the AMC Data Element Dictionary can be obtained by calling AUTOVON 284-9051/2 (Washington area 274-9051/2), or by writing HQ US Army Materiel Command (USAMC) ATTN: AMCMS-IR, 5001 Eisenhower Avenue, Alexandria, VA 22304.

2. Question: Is your dictionary of electronic data elements available through the Defense Documentation Center (DDC)? If yes, please indicate order number (AD Number).

Answer: The dictionary is not available through the Defense Documentation Center but can be obtained by contacting the above address.

3. Question: Would like more information on the design and implementation of your data dictionary. How are element descriptions input; how are reports produced, who has access to the DED; is there on-line interactive support, etc.

Answer: The AMC DED system has been adopted as the Department of Army standard DED system. Documentation is available and can be obtained by contacting the address contained in the answer to Question 1 above. The element descriptions input are placed in card format. Various reports are produced from this system, a description of which is contained in the documentation. All personnel within the AMC complex have access to the DED and are required to use this prior to obtaining of the approvals of a new reporting requirement. Currently, there is no on-line interactive support.

4. Question: Are your 9000 AMC data elements standard throughout the US Army? What percentage are DOD standards? How many are suitable for use in other DOD Departments or Agencies?

Answer: Our 9000 AMC data elements have not been submitted to Department of Army as candidates for standardization. We are now in the process of evaluating the AMC standard data elements for impact if submitted to DA for standardization. It is always a slow process to standardize data elements; the higher level of standardization, the longer the process. As a result, AMC took the initiative in standardizing its own data elements within AMC. We will not be sure how many are suitable for use in other DOD Departments until they have been submitted to Department of Army as candidates for standardization.

5. Question: What if one of your reports produces hours worked by employee number and someone also wants hours worked by employee number and project number; does he get a new report? Does your DED have this?

Answer: In order to produce the first report the data element "hours worked by employee number" must be contained in the data bank. In order to produce the second data element "hours worked by employee number and project number", the second data element must be added to the data base if it is not already in existence. If it is already in existence and is not currently contained on a recurring report, the individual would get a new report. If the second data element was not contained in a data bank, it is projected an economic analysis based on the probable cost of gathering and maintaining the data for that data element in the data bank would occur. Our DED has a list of the different data elements currently in our reports and data bank.

6. Question: How do you cost the benefits resulting from a data element standardization action? What are your elements of expenses used to obtain a cost figure?

Answer: There has been no analysis of cost benefits resulting from data element standardization within AMC. This has been a required program directed by Department of Army. It has also been necessary in order to successfully implement our standard AMC systems.

7. Question: How are you handling the problem of nonuniform input cost data? That is, breaking out recurring vs. nonrecurring, or RDTE vs. production, also variations between different suppliers, cost tracking procedures and learning curve serialization problems.

Answer: AMC has not yet developed costing procedures for our Data Element Management Program. We have felt that it was necessary to identify the cost of producing our various reports control symbol (RCS) reports and our ADP reports first. We currently are not breaking out recurring vs. nonrecurring costs on manual reports, however, we are on ADP reports. On ADP reports we do require developmental (one-time) costs to be furnished, such as systems design manhours and dollar cost, programming manhours and dollar cost, and computer test time and associated dollar costs. On a recurring basis, we require the programmer and systems analyst maintenance cost, supplies cost, and the recurring computer cost to be furnished. We intend to include in the future the cost of handling the report in the functional area. At present, however, we are only identifying developmental costs for ADP reports and recurring costs on both manual and ADP reports. I have been given five years to develop a data element costing procedure and I think it will take at least that long to do so.

8. Question: What is the size of your staff working on development of data element control? Do you have separate functional subgroups for: (1) selection/development of a DED System, (2) standardization of data element representation, and (3) definition of control procedures relative to data administration?

Answer: Currently there are ten personnel in the Reports Management Branch; the majority are working on our current Reports Management Program; this balance will change as our full Data Element Program emerges. The personnel of this branch consist of both computer specialists and management analysts. I personally feel, and this is not necessarily an AMC sentiment, that eventually there may develop a separate series for Data Element Managers. I do not have separate functional subgroups, but utilize my people on a team concept, changing construction of the teams as the need arises.

9. Question: How many programmers, analysts, etc., will be required to complete your job in five years?

Answer: I am not sure how many will be required, however, in all probability, I will be given no additional resources to accomplish the job. I do have the prerogative, however, of having my detail systems design and programming work done at one of AMC's Central Systems Design Agencies (CSDA), as the need arises.

10. Question: Are your Data Element Management computer programs available to other government agencies? Did you investigate commercially available packages, which ones and why were they not selected? I'm assuming your system is an AMC original. How did you begin data element information -- existing reports, source documents, or references in computer system? Any tips would be sincerely appreciated as we are not seeing the value and need of a data element management system and are working toward installing one.

Answer: Any programs that we have developed are certainly available to other government agencies. We have investigated commercially available packages but to date have been unable to locate any that are applicable to the requirements of our Data Element Management Program. In fact, if you know of any, I would appreciate your furnishing me this information. As far as I know, our system is an AMC original. We began building our data element information from our existing reports. We are now building a Reports Characteristic File which gives us the characteristics of all our reports and ADP products produced to include identification of the data elements within these reports. Subsequent to the finalization of our Reports Characteristic File, our Data Element Characteristic File will be designed. I will be glad to keep you informed of our progress and furnish you such documentation as you may desire, if you will contact the address furnished in the answer to Question 1 above.

11. Question: Could I get an advance copy of the USAMC paper entitled "Progression from Reports Control to Data Element Management"? Also, any format information on collecting data for a Data Element Dictionary. How is the DED used in relation to the DBMS? Is it a directory to data as well as a reference dictionary? Does it control access to DBMS

by proponents and users? Could the DED system be converted for use on 6500 and if so, what would be involved?

Answer: An advance copy was furnished as requested. Currently, the only tie-in between the DED and DBMS is the requirement for the programmer to use the name which is assigned in the DED. More discipline, however, needs to be applied in this area. The DED is not a directory to data, but merely a reference dictionary, giving the standardized title of a data element and showing in what systems it is used. It can also identify the reports in which the data are contained. The DED system could be converted for use on 6500, I assume, as it is written in COBOL. Impact would be the work effort involved in conversion of an IBM 360 COBOL series of programs to a 6500, if the 6500 is capable of accepting COBOL.

12. Question: Do you intend, after completion of your initial data element dictionary and report reduction activities, to investigate the possibility of constructing one or more integrated data bases to further reduce ADP costs?

Answer: Yes, this is one of our objectives in our data element management program. The first step, of necessity, must be to more or less "catalog" the data elements which we currently have and where they are located. Our logical step will be then to attempt more integrated data bases than we currently have in order to reduce the gathering and maintaining of data in the data bases.

13. Question: a. What are the main phases planned to implement the use of DED names at the systems programming levels? (Tools?) b. Why "matrix analysis" before that implementation, which could permit automated matrix analysis?

Answer: Implementing and enforcing the use of DED names at the systems and programming levels have been considered the second phase of our Data Element Management Program. AMC considers the implementation of the use of Data Element Dictionary names in our reports to be of first importance. The rationale is the fact the Data Element Management Program is evolving from our Reports Control Program, therfore, we are attempting to purify the Reports Control Program prior to moving into the systems and programming areas.

14. Question: Would you please forward documentation of the content, structure and use of your Data Element Management system?

Answer: I will forward the documentation which I currently have. AMC is only in the first year of our projected five year program in this area; more complete documentation will be available as we are further along in our program.

The Use of a Data Base Management System For Standards Analysis

Sherron L. Eberle, L. D. England, Bernard H. Schiff[1]

Office of Information Services
Office of the Governor
Austin, Texas 78711

A data base management system (System 2000[2]) is being utilized for analysis of recipient-oriented data elements and representations used by a group of Texas state agencies. The data base contains over 700 defined data elements found in computer files and paper forms used by the agencies. In addition, about 30 sets of current data element standards are stored. The data base and the DBMS permit detailed analysis of present standards usage and potential impact of new or revised standards prior to the selection of interagency standards. In addition, the DBMS facilitates the ready generation of reports, listings, data inventories, and current editions of an interagency standards dictionary.

Key Words: DBMS; data base management system; data dictionary; data elements and representations; data element standardization; standards analysis; state government data; Texas state agency data elements.

1. Introduction

The Office of Information Services, a group of computer and information specialists within the Texas Governor's Office, has statutory responsibility for developing interagency information systems and improving utilization of EDP within Texas state government. In accordance with this legislative mandate, the Health and Human Resources Division of the Office of Information Services staffs a data standards project in cooperation with the major health, social, employment and education agencies within Texas state government. Primary goal of the project is the adoption of standard formats for recipient-oriented data elements and representations to be used by participating agencies in storing and exchanging computerized recipient data. The incidence of identified data exchanges among participants in this project lends strong support to the view that governmental information systems today must be able to operate effectively in an environment that is increasingly characterized by both inter-agency and inter-governmental activities and programs.

2. Methodological Approach

In order to assure selection of the most appropriate standards for interagency use,

---

[1] Systems Analysts.

[2] Product of MRI Systems Corporation

it was decided at the beginning of the project that in-depth analysis of existing standards would be essential to the establishment of a set of standard data elements and representations for use by participating agencies. During the data collection phase of the project, it became evident that thorough analysis would require voluminous data gathering from state and federal agencies and national associations, each promulgating standards seemingly independent of one another. Also in evidence was the fact that while standards exist in numerous agencies or functional areas, the data processing shops assumed to be subject to a set of standards may or may not be using the standards. It was therefore decided that file layouts, in addition to formal standards, must be gathered and analyzed to determine what data element standards are actually being used by participating state agencies.

Additionally, the decision was made to analyze relevant paper forms used by state agencies in collecting data about the people they serve. Several benefits were expected from this action. First, in most cases a far greater percentage of an agency's information about its clients is captured on paper forms than is stored in computer files. Therefore, analysis of forms seems essential for accurate assessment of which elements are used most frequently and are hence likely candidates for standardization. Furthermore, data which is stored in computer files is usually first collected on paper forms. Thus, when data standards are adopted which necessitate revisions in computerized files, changes must often also be made in the paper forms which serve as the source documents. By documenting linkages between computerized files and the original data collection forms, we are attempting both to measure the anticipated impact of proposed standards and also to facilitate system-wide conversion once standards have been adopted. Additional benefits are expected to accrue as agency forms undergo periodic revisions designed to improve their use for direct data entry.

As one can see, the more analysis to be done during the standards project, the greater the volume of information that would be needed and the more difficult it would be to organize this information. Two decisions were made at this point: 1) A data base management system should be used to manage the voluminous data and 2) the scope of the standard data elements and representations should primarily center around client services provided by the state agencies rather than any internal agency operation (e.g., payroll, accounting).

System 2000 as a data base manager has been used with much success in the standards project. The data base manager allows for varying reports to be generated at will on specific data elements, a set of standards, a certain agency's forms, the entire data base, etc. Data is easily entered and modified, as necessary. The data base allows for structural changes and the addition of fields containing information initially by-passed but later discovered to be needed in the analysis.

An auxillary file contains data element definitions and these, when accessed with the main files, provide an automated data dictionary.

The approach we have taken with respect to the data dictionary appears to contrast somewhat with what we have seen utilized in other standards projects. Rather than produce a data dictionary which provides a separate citation for each uniquely named element with an accompanying definition which may be specifically tied to particular program regulations or legislative definitions, we have produced a document which attempts to provide a synthesis across agencies. The data dictionary is dynamic, for it is at this stage primarily a working document to be used in establishing common definitions across agencies for data elements which are candidates for standardization.

In selecting definitions for elements which appear in the data dictionary, our objective was to produce definitions which could serve as a starting point for further refinements resulting from interagency discussion. Wherever possible, definitions restricted to a particular agency or program's requirements are avoided. For example, one of our most frequently cited elements is known to project participants as "Applicant/Recipient Name." For our purposes, we are interested in documenting the frequency of use of this element among the various agencies. We do not care what eligibility requirements must be met by an applicant or recipient served by the Department of Public Welfare as compared to the Rehabilitation Commission, for example. Nor do we care that the Department of Mental Health and Mental Retardation calls the people it serves "patients" or "students," while

the Welfare Department refers to "recipients," the Rehabilitation Commission talks about "clients," and so forth. What we wish to document is what piece of data we are describing and in what format it is carried by the various agencies using it. Our definition for "Applicant/Recipient Name" is at this point quite elementary: "name of applicant/recipient as it appears on an agency's records." As this element becomes involved in the standardization process, we would expect the definition to be refined. For example, it may be specified that the element refers only to applicants/recipients who are individuals as opposed to organizational applicants/recipients. Synonyms used for the element name in various agencies may have to be noted. And, of course, the standardization process will lead to the addition of such format-related information as: "Applicant/Recipient Name consists of the person's last name, followed by his first name and middle initial."

3. Administrative Approach

As previously noted, the goal of the data standards project is the adoption of standard formats for recipient-oriented data elements and representations used by state health and human resources agencies in storing and transferring computerized data. The achievement of this goal would facilitate interagency data transfer, facilitate data processing operations within agencies, economize in the use of computer software and hardware, and minimize system conversion problems.

A primary administrative feature of the project has been the formation of a viable Interagency Standards Task Force under the auspices of the Interagency Health and Human Resources Council, an organization of agency commissioners and executive directors. The Task Force, which meets on a periodic basis to establish policy and adopt standard data elements, is composed of data processing management personnel from the following Texas agencies:

Cooperating Agency	Current Hardware
State Department of Public Welfare	IBM 370/155
Texas Department of Mental Health and Mental Retardation	IBM 370/145
Texas State Department of Health	UNIVAC 1106
State Commission for the Blind	IBM 360/40
Texas Rehabilitation Commission	BUR 3500
Texas Industrial Accident Board	IBM 360/40
Texas Employment Commission	IBM 370/158
Texas Education Agency	IBM 360/65
Coordinating Board, Texas College and University System	IBM 360/40

The Standards Task Force has met several times, completed organizational activities, developed task force procedures, and is currently examining a set of proposed data standards. The approval process adopted by the task force is shown in Figure 1.

4. Technical Approach

An innovative aspect of this data standards project is that detailed information has been collected so as to permit thorough analysis on which to base the selection of each standard data element and its representation. Recipient-oriented paper forms, computer files, and promulgated standards have been collected from the participating state agencies. In addition, sets of data standards have been collected from related

537-860 O - 74 - 6

state, federal, and national organizations and agencies which impact members of the Task Force.

Detailed descriptions of data elements and representations as they appear in each form, file, or standard have been entered into the data base using a FORTRAN program with System 2000 procedural language interface. Data values are stored by means of the data base structure shown in Figure 2. The hierarchical structure enhances use of the data base in relating data elements to various files, forms, and/or standards.

Queries and updates are entered directly by the DBMS using a remote keyboard/terminal printer. A separate ISAM file stores element definitions and is accessed as needed. Figure 3 shows the overall computer processing functions.

Reports from the data base may be generated in two ways: 1) First the DBMS provides remote access to the data base via a series of commands which initiate selected counts and listings. Data obtained by the DBMS access method is used for maintenance purposes primarily but also to answer specific questions regarding proposed standards. 2) The second method involves use of a high level language, such as FORTRAN or COBOL, which is linked to the DBMS at compilation time by a procedural language interface feature. A series of COBOL report programs are being written which provide various types of listings and inventories. An example is the data dictionary as illustrated in Figure 4.

The complete report series is as follows:

Report #1, Preliminary Data Standards Element Inventory

Description: This report is a list of data elements showing the media (forms, files and standards) on which each data element appears. The report also describes the data element (length, justification, type, etc.) as it appears on each of the media.

Expected Usage : This report will be used primarily by OIS for initial selection of standards and for initial correction of the data base.

Report #2, Data Standards Element Dictionary

Description: This report is a list of data elements and their definitions. It also indicates which agencies are using which data elements.

Expected Usage : This report will be used by the Standards Task Force members to establish a common definition for each element selected for standardization.

Report #3, Data Standards Media Listing

Description: This report is a list by agency of the media (forms, files and standards) in the standards data base.

Expected Usage : This report will be used by the Standards Task Force and OIS to verify the inclusion of applicable media (forms, files and standards) in the standards data base.

Report #4, Data Standards Element Inventory

Description: This report is similar to Report #1, Preliminary Data Standards Element Inventory, except this report includes element definitions and is of a more usable size (8½ x 11).

Exepcted Usage : This report will provide information for the Standards Task Force

on data elements prior to selection of a standard.

Report #5, Data Standards Media Inventory

Description: This report lists each medium (form, file or standard) with the data elements that comprise that medium.

Expected Usage : The report will be used by the Standards Task Force and OIS to correct and update the Standards Data Base.

Report #6, Data Standards Composite Elements

Description: This report lists those elements which consist of related data items that can be treated as a group or individually. Both the composite data element and its component data elements are listed. An example of a composite element is INDIVIDUAL NAME with its components LAST NAME, FIRST NAME, MIDDLE INITIAL.

Report #7, Data Standards Table Report

Description: This report lists those data elements which appear on forms in a tabular arrangement.

Expected Usage : This report will be used primarily by OIS in its data standards analysis.

5. Project Benefits

One of the difficulties in maintaining a consistently high level of morale and activity in a standards project is that many of the benefits of standardization are realized only in the long run. This is true, for example, of many of the potential benefits noted in Section 3, such as "economize in the use of computer software and hardware". However, we have been pleased to note that certain unexpected benefits have already become evident in our standards project. First we have become aware that the standards data base is increasingly referenced by staff personnel working on other projects within our organization. For instance, the standards data is proving useful in developing a Texas Supply/Demand Information System for Vocational Education, in designing record layouts for state licensing agencies, and in the analysis of minimum data sets for the collection of health manpower and health facilities data.

A second immediate benefit has resulted directly from the capability which we have for dealing with a vast amount of detailed data. That is, utilizing the query feature of System 2000 in conjunction with a manual file of data codes[3], we are able to quickly assemble comprehensive comments regarding standards proposed for adoption by various organizations. Figure 5 is an example of one page taken from a series of comments prepared in response to a set of standards proposed by a group of Texas state administrators. Each of the abbreviations shown on this page represents a set of standards entered in our data base. The standards data has been used similarly in recent weeks to respond to a federal agency's request for comments on a proposed inventory of data elements for a survey questionnaire. The capability to prepare such comments rapidly lends added authority to those comments, while effectively utilizing personnel time.

Another benefit realized from the project has been that of increased communication and cooperation among participating agencies' data processing shops. This benefit is, of course, somewhat intangible, but has resulted primarily from the interaction among data processing managers in the meetings of the Interagency Standards Task Force.

---

[3]The information in this file could easily be entered into System 2000. However, such a conversion is not believed to be cost effective at this point.

Given the benefits already realized from this project, we are convinced that utilization of a data base management system for analyzing data elements in forms, files, and sets of standards is a sound approach to the standardization process. Furthermore, due to the emphasis placed on pre-selection analysis, we are optimistic that the long-range benefits of standardization will follow.

DATA STANDARDS SELECTION PROCEDURES

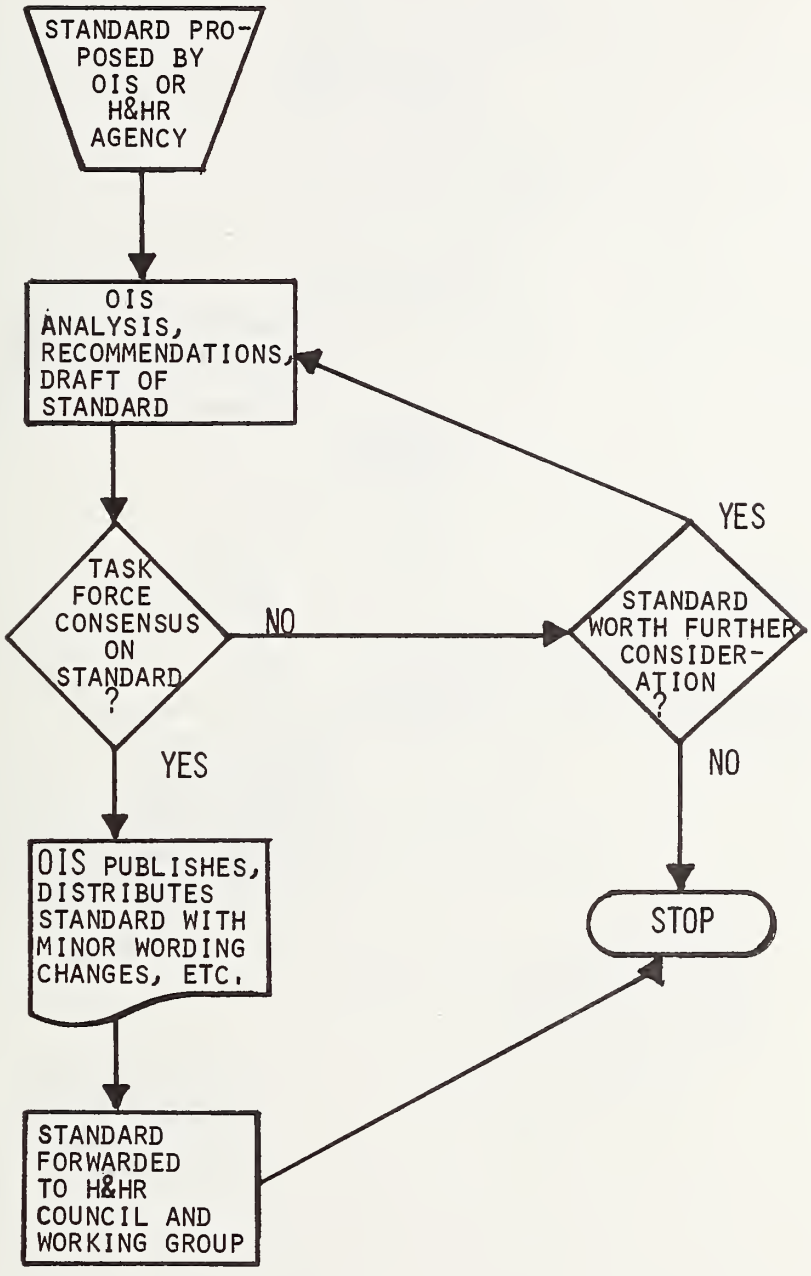


FIGURE 1

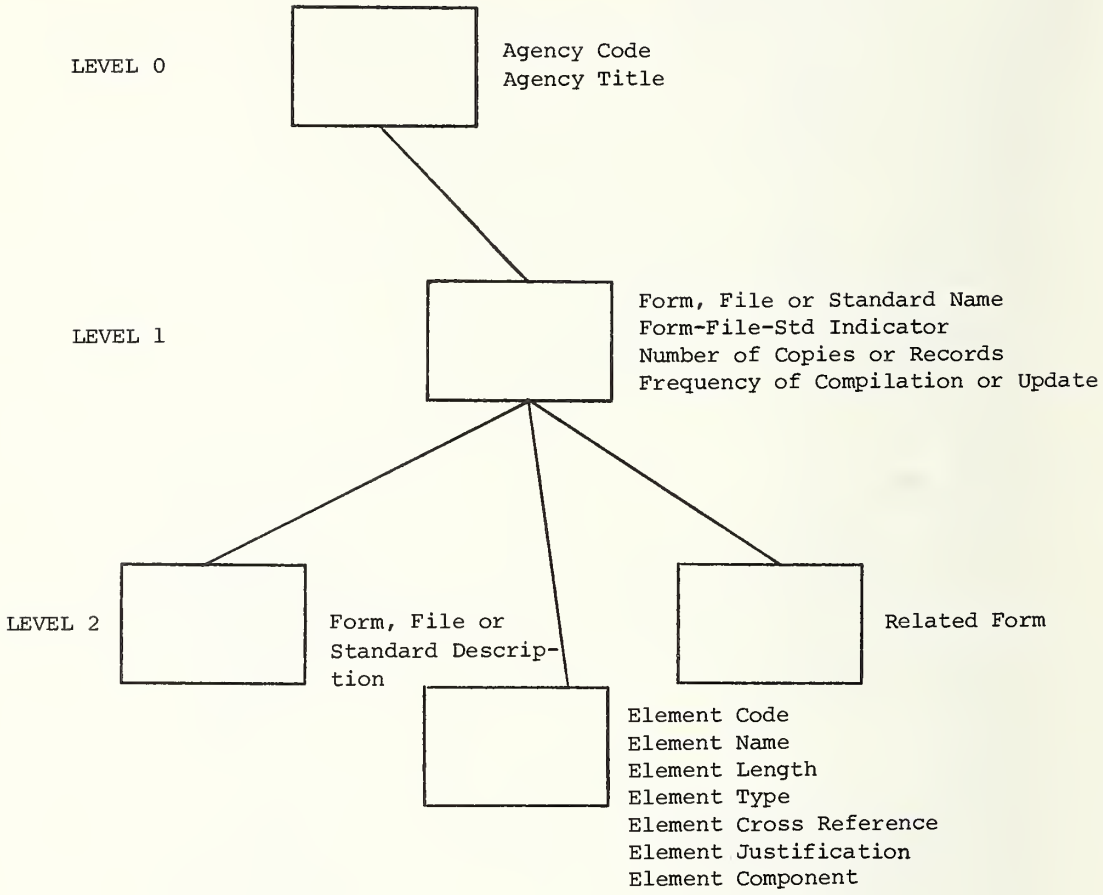

LEVEL 0
Agency Code
Agency Title
LEVEL 1
Form, File or Standard Name
Form-File-Std Indicator
Number of Copies or Records
Frequency of Compilation or Update
LEVEL 2
Form, File or
Standard Descrip-
tion
Related Form
Element Code
Element Name
Element Length
Element Type
Element Cross Reference
Element Justification
Element Component

FIGURE 2

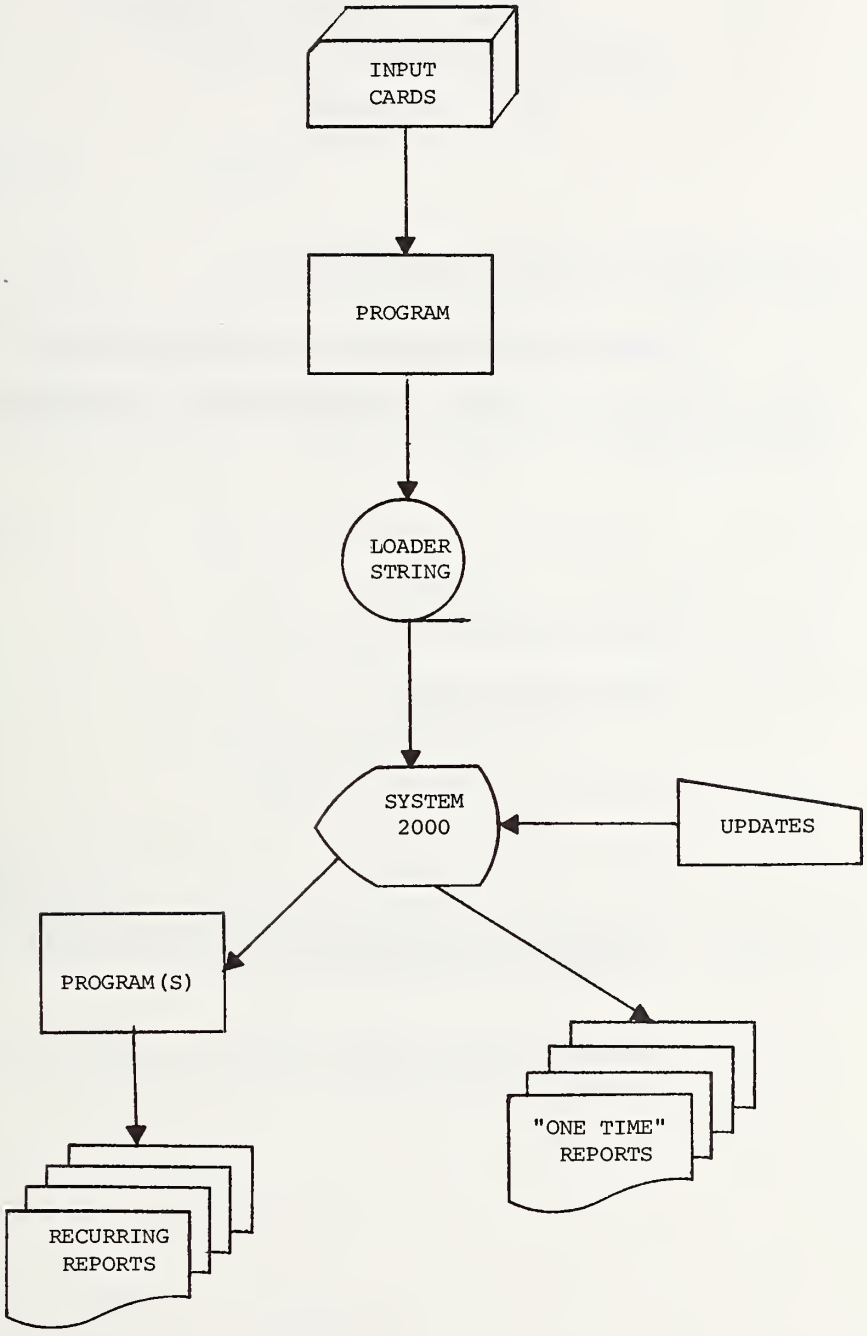

INPUT
CARDS
PROGRAM
LOADER
STRING
SYSTEM
2000
UPDATES
PROGRAM(S)
"ONE TIME"
REPORTS
RECURRING
REPORTS

FIGURE 3

EXAMPLE AND EXPLANATION OF THE DATA DICTIONARY

2350[1] APPLICANT/RECIPIENT'S RELATIVES[2]

NAME OF RELATIVES OF APPLICANT/RECIPIENT,[3]

SEE ELEMENT(S): 2300[4] 2450[4] NUMBER OF OCCURRENCES 8[5]

AGENCY USERS: DPW[6] TRC[6]

1 - DATA ELEMENT NUMBER

2 - DATA ELEMENT NAME

3 - DEFINITION OF DATA ELEMENT

4 - DATA ELEMENT NUMBERS OF SIMILAR DATA ELEMENTS

5 - NUMBER OF TIMES THE DATA ELEMENT APPEARS IN THE STANDARD DATA BASE

6 - ABBREVIATION OF THOSE AGENCIES USING THAT PARTICULAR ELEMENT

FIGURE 4

SAMPLE PAGE FROM COMMENTS
REGARDING PROPOSED STANDARDS

ELEMENT NO: TIPS 7

ELEMENT NAME: COUNTY CODE

CONFORMS WITH: MAIPS, FIPS, OIS, SRS, HEW, USOE, IBM, GSA

AT VARIANCE WITH: DPW, IAB, TCJIS, TSDH, MH/MR, COO, TEA, TRC, SSI

COMMENTS: THE TIPS CODE IS IN CONFORMITY WITH MOST FEDERAL CODES. IN GENERAL, STATE AGENCIES NOT IN CONFORMITY USE ONE OF TWO OTHER TABLES: (1) 1-245, MC'S FIRST AMONG M'S (2) 1-254, MC'S ALPHABETIZED WITHIN M'S

RECOMMENDATION: RECOMMEND ADOPTION IN CONFORMITY WITH FIPS; CONSIDERATION MUST BE GIVEN TO CONVERSION TIME.

FIGURE 5

The States' Model Motorist Data Base Project American National Standards Institute (ANSI) D-20

C. E. Emswiler, Jr. and C. P. Heitzler, Jr.

Commonwealth of Virginia
Division of Automated Data Processing

Each of the States maintains records required for the registration, licensing and control of motor vehicles and drivers. These data systems, 49 or which are automated in whole or in part, make up one of the largest sets of data files in the nation. They perform the primary production functions of revenue collection, vehicle/driver identification and control, and motor vehicle code administration of the individual states. These functions require the retention of motorist data and the interchange of that data between State agencies, individual states and other users of the systems. These systems have been tapped by law enforcement, by commercial users and by some Federal systems. As much as 70 percent of law enforcement data traffic is driver/vehicle related. Commercial users are insurance, credit, employer and statistical collection interests. The Federal systems include law enforcement and traffic safety.

These independent systems, however, have certain weaknesses. They have not satisfactorily responded to the challenges and problems of American mobility (over one trillion miles of travel each year by more than 100 million passenger cars, trucks and buses). There is non-standardization in all elements (forms, terms, methods, usage, local laws and information interchange), although The American Association of Motor Vehicle Administrators (AAMVA), which represents the Motor Vehicle Administrators of the fifty States (and the Canadian Provinces), has been active in the area of standardization. Conviction codes and vehicle identification numbers are two areas of standards accomplishment. The standardization of vehicle titling and registration (ANSI D-19) and establishment of the States' Model Motorist Data Base Standard (ANSI D-20), for which AAMVA is the secretariat, are progressing as active projects.

The ANSI D-20 Project currently involves persons from forty (40) States, and the Federal Government and the private sector. Its objectives include:

1. The standardization of vehicle and motorist related data elements;
2. The automated interchange of information related to drivers and vehicles to effect highway safety and improve motor vehicle administration;
3. The reduction of paper records and transactions processing through access to automated records;
4. The combination of redundant files and communications networks to reduce duplication;

5. The linkage between traffic records and criminal justice systems;
6. The production of local, state and national highway safety information to meet research, evaluation and statistical requirements;
7. Providing the information and processing required to meet the administrative requirements of those responsible for motorist and motorist related information, via standardized State systems instead of Federal systems;
8. The reduction of systems development costs; and
9. Providing for future inclusion of additional data, new information requirements and growth.

Key words: Accidents; ANSI; D-20 project; data elements; data exchange; data files; data interchange; driver history; driver license; fatality analysis; federal; motorist data base; national accident summary; reciprocity; registration/titling; revocations; safety; suspensions; traffic safety; vehicle; vehicle history; vehicle registration.

1. Introduction

The ANSI D-20 Project was an outgrowth of a forward thinking group of concerned individuals, involved in motor vehicle and traffic safety administration at the State and Federal levels of government. This ad hoc group working in conjunction with AAMVA regional data processing workshops, recognized that each State retains the information required to administer its motor vehicle and highway code and its traffic safety programs and that this information composes one of the largest data files in the nation. The size of the data file encompasses both the number of records retained and the number of data elements or information fields that are included. Erosion and duplication of the State motor vehicle and driver automated files was noted as the records in these files also became records in the NCIC, the National Driver Register, local government and other automated systems. An organizational mechanism that would permit the States to work together to improve the situation was necessary. The D-20 project was organized to overcome the weaknesses, deficiencies and information gaps that exist in the individual State systems, which prevent the desired and required interchange of information without the establishment of a federal system.

The three primary facts were immediately identified that precluded acceptable data interchange:

1. The information being retained had no semblance of uniformity or standardization;
2. Desired information was not available; and
3. The information's degree of detail, accuracy and state of being current varied from state to state.

In order to appreciate the scope and goals of the D-20 project, one must understand the need for data exchange at the various levels of government, as well as with the private sector, for interchange is the prime requirement of D-20 although additional benefits can and will be realized. Also, the information areas involved must be identified and related to the individual user requirements.

2. Scope of Project

In general, the information areas within the D-20 environment are those that pertain to what is usually termed "traffic records", which include motorist data (driver and vehicle status and histories), highway, traffic accident, traffic violation and traffic safety program evaluation data.

Access to this data is necessary on two broad scales - Interstate and Intrastate. Interstate interchange involves all levels of Local, State and Federal government and the private sector. Intrastate exchange involves the individual agencies and branches of State government, local government and the private sector.

The requirements for information interchange exist in three general categories, that can be defined in relation to both time and data content. The categories are:

1. Immediate - Data that is required immediately, at the time of a transaction, and is limited to data such as the status of a record or very brief descriptive data.

2. Convenience - Data that is required because of a specific transaction, that may be satisfied either on an immediate transmission or a "convenience" basis. This data transmission is keyed by a specific transaction, however it is not necessary to immediately receive the data pertaining to the transaction and the data that is required is more extensive than that of a simple status check. Currently the convenience method or delayed transmission is used in most systems. However, if the transmission of data can be accomplished in conjunction with an immediate status data transmission, by data transfer or pointer creation, without unacceptably degrading the system or unduly increasing the resources required, it should be.

3. Statistical-- Data that may be generated on an automatic or periodic basis or on an as-requested basis. Included data may take many formats; contain a great variety of data content and extensiveness; and may be transmitted in any of many modes.

Having described the categories of interchange, each may be addressed individually.

Immediate

1. A driver's license transaction of an out-of-state applicant requires identification verification and status information in order to determine the eligibility of the applicant to be licensed.

2. Vehicle and title registration transactions involving an out-of-state vehicle require ownership verification and stolen vehicle, lien or restriction status.

3. The law enforcement community requires immediate access to information that is driver or vehicle related in 70 percent of its data traffic. Such transactions include driver and vehicle identification, ownership, description and status data in connection with crime commission, stolen vehicles, traffic violations, traffic accidents, and abandoned vehicles.

On an interstate basis, with the exception of law enforcement, this type of access and transmission is virtually non-existent. As for intrastate needs, direct inquiry into these information files is prevalent between state government agencies and local government (including law enforcement) and, to a lesser extent, with the private sector, in the area of insurance. In a

very few states the technique of automated, intrastate immediate update and/or status checking at transaction time from remote locations has been developed for driver licensing and/or vehicle registration.

Convenience

1. New driver records, for out-of-state applicants, deemed eligible for licensing, require the transmission of or cross-relation to driver history data from the "old" state, including past accident involvement; traffic violation convictions; license suspensions and revocations; applicable medical history data; rehabilitation programs taken; and financial responsibilities.

2. History data is required on vehicles registered from out-of-state pertaining to inspection failure history, reflecting known defects, and accident involvement history.

3. Upon registering/titling an out-of-state vehicle or licensing an out-of-state driver applicant, or the refusal to do either, the transmission of data notifying the "old" state of the fact of the transaction is necessary, so that required actions and/or records changes can be effected.

4. Data pertaining to a driver's out-of-state traffic violation convictions, indicating any resulting actions or conditions and reflecting the type and conditions of the violation, need to be interchanged.

5. Data pertaining to traffic accidents, occurring out-of-state, involving resident drivers or vehicles, need to be interchanged. These, likewise, must reflect descriptive data and identify any resulting actions or conditions.

6. Vehicle registration verification and owner identification data exchange is required when traffic violation and parking summons are not responded to in the prescribed manner.

7. Data must be exchanged under "reciprocity agreements" existing between states. These include vehicle registration and operation and driver licensing/testing and suspension/revocation.

Interstate and intrastate transfer of almost all of the above data exists currently. However, it is normally accomplished by the transmission of paper documents, either manually or computer produced, or by magnetic tape exchange on a periodic schedule that varies depending on the accumulation rate of individual state processing systems. Recognizing the desirability of immediate exchange of status data, every effort should be made to provide for the immediate (on-line) exchange of "convenience" data in conjunction with and utilizing the same network as that of the immediate status data. This type of data is an almost exclusive need of state and local governments, with little requirement on the Federal level or in the private sector.

Statistical

1. Intrastate, interregional or interstate studies, on either a periodic summary or as requested basis, require data for the purpose of comparisons and program evaluation in the areas of:

   a. Highway safety devices
   b. Driver education programs
   c. Driver testing methods
   d. Emergency medical services

e. Accident investigation
f. Traffic flows
g. Highway planning and construction
h. Countermeasure programs
i. Resource allocation and management
j. Rehabilitation programs

An example is that valid data for comparison for a study in New York City may not be found in the state of New York, but might exist in a similar metropolitan area of comparable size and environment, such as Los Angeles, Chicago or Boston.

2. Vehicle inspection history data pertaining to specific problems common to all vehicles, in all states, of a certain class or type, such as, known defects in "X" style car at "Y" miles, is desirable and necessary.

3. Common roadway engineering methods and problems data, as well as, data pertaining to site location methods, techniques and effectiveness are required.

4. Statistical summaries of traffic violations and convictions, and traffic accidents by type, description, contributing conditions and environment necessary for the National Highway and Traffic Safety Administration (NHTSA) and the National Safety Council, as well as, state and local governmental agencies.

5. Vehicle ovner identification and vehicle description data for recall purposes are needed.

This type of data is widely used by all facets of state and local governments for the execution of their responsibilities in the areas of law enforcement, resource management, program evaluation and planning.

Interstate and interregional exchange is currently accomplished through the Federal data gathering functions such as The National Accident Summary, Fatality Analysis File and The National Driver Register. As standards and conditions change, new requirement areas will emerge that will make new demands for this type of exchange, either periodically or as requested, such as the energy crisis, pollution concern or vehicle recall demands.

Private industry has a continuing need for such data in order to plan for and meet future demands in the areas of manuafcture, sales, insurance and credit.

With the description of the interchange requirement and the general data areas of D-20, one can understand some of the problems and the extent to which this interchange need is currently being met. The major reason and need for interchange on a uniform, standardized basis, is to make the information currently being maintained and exchanged more current, accurate and usable. If these goals can be accomplished, then the interchange process can be made more effective and efficient. The information that is exchanged can be more usable, more accurate and more current, and the processing and transmission of this data can be made in a manner that will actually reduce costs. The purpose of D-20 then is to retain, within the states, the data that they require to execute their responsibilities, eliminating the need and costs of a federal system to accomplish interchange, the by-product of which will be a promotion of highway safety and a general reduction in traffic accidents. With these facts in mind, the previously described ad hoc group set about the organization of the formal D-20 project.

3. Project Establishment

The American National Standards Institute (ANSI) is the only national coordinating organization representing industry, consumers and governments which meets the increasing demand for voluntary standards. ANSI does not develop standards. It makes use of the combined technical talent and experts of its member bodies: the more than 160 technical, professional and trade organizations that comprise the system or federation called the standards institute. The standards developed by these organizations become American National Standards, after the institute determines that they have been developed in accordance with its procedures which include agreement and consensus among interested and affected parties. With the advice and aid from AAMVA, ANSI accepted the "States' Model Motorist Data Base Project". It was assigned to the ANSI Technical Advisory Board (TAB) for highway safety and was identified as the D-20 committee.

On July 13, 1972 at the National Press Club in Washington D-20 was officially organized. One hundred twelve individuals representing 41 states, 42 associations, 11 federal agencies, 3 commercial users, 2 manufacturers, 3 research institutes, 8 hardware vendors and 2 consulting firms were invited. Sixty-three attended with the majority of those not attending desiring to participate as mail members of the parent committee. AAMVA became the secretariat, Charles E. Emswiler, Jr., of Virginia, was appointed Parent Committee Chairman, and Will Wolf, of Washington State, the Vice Chairman.

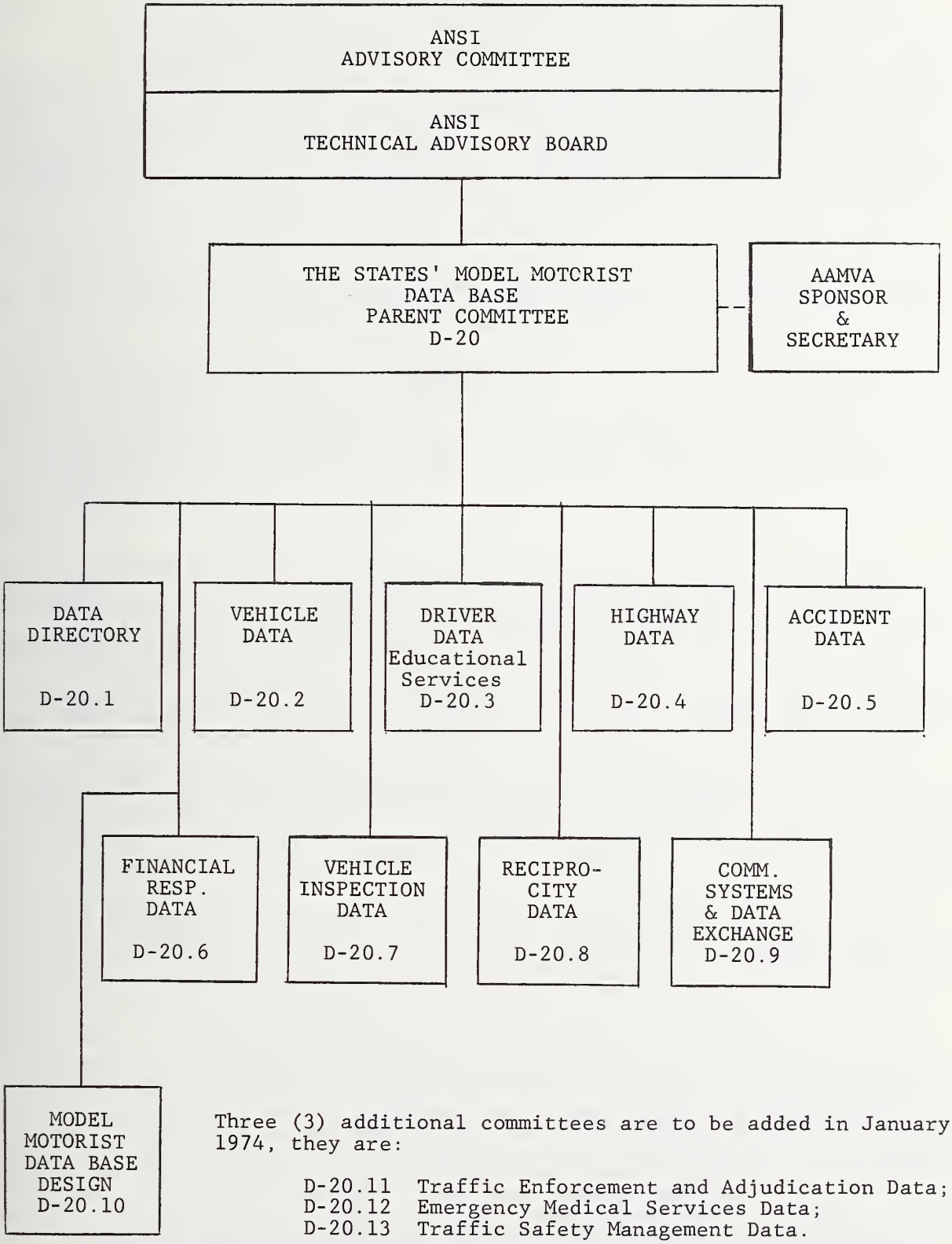
The proposed scope of the committee is: to develop automated data processing procedures acceptable to State vehicle administrative functions in the fields of vehicle registration and certification of ownership, driver licensing, motor vehicle inspection, highway safety and accident statistics and motorist financial responsibility and other reciprocal agreements. The procedures are to be projected for the communication between all states on a systems network enabling message switching and interfacing with other computers as the need arises for an overall coordinated unity and intercommunications.

4. Organization

The D-20 Committee was originally structured with ten technical subcommittees: a data directory committee (D-20.1); seven committees (D-20.2 - D-20.8) dealing with data on the vehicle, the driver, the highway, accidents, financial responsibility, vehicle inspection and reciprocity; and two committees (D-20.9 and D-20.10) to deal with communication systems and data exchange, and the motorist data base design requirements respectively.

As work progressed in the definition of data elements in the subcommittees D-20.2 through D-20.8, additional data element categories were identified and incorporated into the project. Currently additional subcommittees (D-20.11 - D-20.13) are being added to the D-20 organization. These sub-committees cover data pertaining to traffic enforcement and judication, emergency medical services, and traffic safety management programs. In addition, data pertaining to driver education has been identified and incorporated into the D-20.3 committee, concerned with the driver data elements.

D-20 ORGANIZATIONAL CHART



Three (3) additional committees are to be added in January 1974, they are:

D-20.11 Traffic Enforcement and Adjudication Data;
D-20.12 Emergency Medical Services Data;
D-20.13 Traffic Safety Management Data.

537-860 O - 74 - 7

5. Committee Responsibilities and Representation

The Data Directory Technical Committee (D-20.1) is charged with developing the data element standards and for defining and describing all data elements; and for developing a coding system. The D-20.2 through D-20.8 Technical Committees are responsible for defining all data elements relative to the sub-system for which they were created. They are expected to review known or published standards and definitions and to research existing national, state and local governmental agency requirements and identify those which may be useful or needed within state systems; or which may be exchangeable currently or at some future date. Each technical committee must standardize its terminology with the Data Directory Technical Committee.

Forms have been developed for the collection of the data elements and a descriptive pamphlet describing the use of the forms has been produced and distributed for the use by these sub-committees. In addition, a data element check list has been developed for use in communicating with the sub-committees. (Samples of these forms are attached.)

The Communication Systems and Data Exchange Technical Committee, D-20.9, is responsible for defining all data elements which are necessary to satisfy the data exchange requirements of users; researching existing national, state or local governmental agency requirements and consider these during development of the data exchange elements; and defining methods, procedures and facilities for exchange, communication, and security of data. Additionally, they are responsible for defining the interchange formats of those data elements.

D-20.10, the Model Motorist Data Base Design Technical Committee, is responsible for designing a model system or systems to satisfy the processing requirements set forth by the D-20.2 through D-20.8 committees and must provide for the communications and data exchange requirements called for by D-20.9. The cross relationship between the elements within the data base must be defined by the D-20.10 committee. Additionally all data elements that will be required for the proper controls within the data management system, for the security of the system, and for the guarantee of the data privacy, where required, must be added into the data element package by this committee.

The above committees' membership includes representatives of 43 states and a variety of associations, federal agencies, commercial users and manufacturers. The following is a break down: states - 87, associations - 19, federal agencies - 13, commercial users - 9, manufacturers - 2, research institutes - 3, hardware vendors - 4, consultants - 8, for a total of 145 persons.

The Technical Committee Chairmen are the most important members of the D-20 organization in that they are responsible for the productive efforts of the working members of the committees.

6. Progress to Date and Remaining Activities

Much of the initial effort has been devoted to education and public relations. Most motor vehicle administrators were, and are, occupied with meeting the ever increasing production demands of registering, titling, licensing, revenue collecting and related administrative functions. New requirements such as highway safety information were given secondary consideration. In addition, there was little concern on the part of most states that competing, redundant systems were being proposed, established and upgraded. The D-20 secretary, the chairman and others presented these problems to AAMVA members in regional and national conferences, and at workshops. These efforts have awakened AAMVA and most of the states to the fact that

state systems can and should meet the productive and administrative demands of the states, as well as the information needs of Federal agencies and local governments.

Collection and validation of data elements have occupied the data element committees (D-20.2 - D-20.8). The availability of NHTSA traffic records design manual has been of invaluable assistance. Cooperation and assistance of the NHTSA personnel has been excellent. Assistance was given by the National Bureau of Standards personnel in designing the data element description form used to collect and clarify data elements. One major problem has been the high degree of non-standard data presently in use. Most states do not care to contemplate changing to standard data elements. This is especially true with respect to traffic conviction codes.

An important activity was a study of the proposed revised Federal Highway Safety Standards. D-20 technical chairmen contacted all states, reviewed and organized comments, studied the proposed revisions and recommended a major restructuring of the standards. The recommended restructed standards were: purpose, traffic laws and regulations, programs, administrative requirements and information requirements. The recommendation is considered important because the trend has been to consider all motor vehicle and driver records as primary safety records, subject to regulation; the fact is that these records existed as revenue collection and control devices before highway safety was a concern. It is believed that most basic functions can best be served by recognizing the interdependency of the records rather than standardizing them as traffic records.

The major problems of data element collection encountered, thus far, are how to generate an understanding of the total needs of those dependent on this information resource; how to make those involved aware of the capacity of the resources that are available; and how to create an attitude or willingness to change our current individual processes so that the whole may be better satisfied.

Each data element subcommittee has begun its task with an initial discussion as to the depth and brevity of its assignment. Most wrestled with the decision of identifying and defining only those that were thought to be desirable for interchange. The scope of D-20 dictates that the "Standard" procedure must provide for all information within the definition of "Motorist Data Base" currently and in the future, as far as can practically be done. It will be the responsibility, then, of the D-20.1 Data Directory Committee to pursue the product of each data element committee and ensure that its task has been completely accomplished. To comprehend a system that would allow the interchangeability of all the data elements is a difficult task. However, though in all probability this will never be necessary or desired, those establishing the standard must view their task from this point of view if a complete and lasting standard is to be produced. The technical ability to build such a system exists and if we are to strive for a system to satisfy both current and future requirements, we must widen the scope of our vision and fully realize and understand the capabilities that are available to us. This situation, then, brings us to the base problem, creating an atmosphere of change and instilling in the states not only the acceptability of change, but the desire for change. In order to accomplish this we must educate ourselves to the fact that the apparent cost of change can actually effect a savings over the long run by providing a more efficient system of exchange, eliminating current unnecessary processes, reducing development and installation costs and eliminating the need of a central repository for data already maintained in the state systems.

Following the data element definition and description, the task of the Communications Committee will be to identify the current exchange elements and the method of interchange. The data base design team can then define the relationships of the elements and define the management and storage requirements of the data base and the management system.

Activities thus far have accomplished the initial submission of each group's data elements, which are being reviewed by the Data Directory Group, except for the newly established groups and the Driver's Group, which now has to expand its previous work to include Education Data. The Directory Committee is researching dictionary/directory capabilities and methods. The D-20.9 and D-20.10 committees are researching the state of the art and current systems of communications and data base management respectively.

Concurrent with these activities the steering committee, composed of the Chairman, Vice Chairman and the Committee Chairman, has been developing a plan for the future of the Standard and its implementation. It is not the desire of the D-20 project to produce a "paper standard". Rather a viable tool to aid the individual states as well as meet the required data needs of all who interface with the system is needed. In order to have an effective, accepted Standard it must be installed and actually demonstrated if it is to succeed to its ultimate capabilities.

In order to accomplish this overall objective a committee was established to draft the overall plan and goals for the D-20 project. The result was a three part plan that would: (1) define and establish the Data Dictionary/ Directory and its components, (2) design and prepare the specifications and software to implement the exchange of the chosen elements in selected pilot states, and (3) design and write the Data Management System, and create and install the data base, that is hardware independent, in those selected pilot states. The total task was estimated to require 18 to 24 months to complete.

This proposal met with a number of objections as to its time frame, and to the fact that no usable product would be produced until the completion of the total task. However, the objective of producing more than a "paper standard" and the necessity of proving the standard's practicality was agreed upon.

Currently a three phase plan is being produced as a result of the reaction to and comments about the previous plan draft. Basically, the current effort will: (1) produce a paper standard in three parts that will be ready for distribution to the individual states and other interested parties for consensus within 9 months of the start date. The three products of this phase will be (a) the definition and coding structure of all data elements, (b) the definition of all elements to be exchanged and a description of the exchange format of each, and (c) the data base design reflecting the relationships of all data elements including those added in order to supplement and meet the requirements of the data management system, system security and integrity, and data privacy restrictions. This phase will complete the ANSI obligation and provide the paper portion of the standard for interchange.

(2) Phase 2, then, will be an AAMVA sponsored pilot project, in selected states, to actually design and implement the necessary system to demonstrate the data exchange portion of the ANSI D-20 Standard. It will provide the final feasibility determination and possibly the design of a nationwide data exchange communications network. Since it is imperative to link the D-20 System to those of the Law Enforcement Community, which currently operates two nationwide networks, it is desirable to share resources rather than duplicate. In addition, since the Law Enforcement Community has a high rate of dependence on the D-20 information, a single network to meet both needs would be highly desirable from the aspect of resources required. A difficult task in this endeavor will be gaining the agreement of the law enforcement systems to modify in order to gain compatibility and ability to share resources. As stated previously, however, the need to change is not unique to any one group and it will be the most difficult point to establish and the hardest on which to gain concurrence.

(3) Phase 3, like Phase 2, will be an AAMVA sponsored project to establish the data base and data management system portion of the D-20 project. It will parallel phase 2 and accomplish the design, development and installation of a hardware independent Data Management System required to implement the total D-20 standard.

A parallel federal effort of National Highway Traffic Safety Administration, "The Design Manual for States Traffic Records Systems:, will accept this new ANSI Standard as an update to, and a replacement of, the applicable areas of the Design Manual.

This approach has been chosen for two primary reasons: (1) a standard that is unproven and has no implementation aids is not generally acceptable to those who wish to comply with it and (2) in a phased approach, early phases are available for use upon completion, and if subsequent phases prove unfeasible or undesirable the preceding phases accomplished may continue in use.

Upon the completion of Phase 1, ANSI will, through its established mechanisms, maintain the standard on a continuing basis. AAMVA must provide suitable project management for the accomplishment of phases 2 and 3. The State, Federal and private sectors will provide concurrence for phase 1 and personnel for the continuance of phases 2 and 3. Contractual personnel may be required and will be obtained as needed. The steering committee and parent committee will continue to function until the final project completion.

Why does the ANSI D-20 Project feel that this project is so necessary and desirable? The answer is fourfold in terms of feasibility and serves as a summary to he D-20 project description.

1. It is technically feasible to establish a standard data base, including all the necessary data elements. A data management system and a communications system, that can be installed on a variety of manufacturer systems, is a reality in this age that requires software and hardware compatibility.

2. The operational feasibility is reflected in the fact that "traffic records" data bases exist which bring together the elements, such as defined in D-20, under a single data management system in both the centralized and decentralized environments. Regional and national networks currently exist that utilize standard input and inquiry formats, demand the retention of defined elements, and require standardized output formats, as exemplified by the ALECS, NCIC and NLETS systems.

3. Given the fact of technical and operational feasibility, the development of this total package can be justified in the savings of development costs alone, if developed as a hardware independent system and if the software to install and operate the package on each state's equipment is provided. The lack of prior development of a system (hardware and software) to accomplish the D-20 task has possibly been greatly influenced by the limited market potential of such a specialized system. Existing communications networks, not currently being utilized to their fullest, can be shared, thereby eliminating some existing systems and reducing the need of planned systems. Individual transaction costs can be eliminated with the reduction of processes required currently to exchange data in a non-uniform manual mode. The access to existing data and the reuse of this data can greatly reduce the coding, keying and data collection tasks. Data processing and retention redundancy can be greatly reduced. Each reduction can eliminate resource requirements or free resources for other tasks. The degree to which each state desires to adopt each of the products and aids

provided by D-20 will determine how much economic benefit it will derive from the full potential.

4. Political feasibility is the most questionable aspect. This includes the real practicality of "selling the system" and creating the realization of the total need, as related to each individual need. However, the feasibility has been demonstrated in the previously mentioned regional and national networks. This sale will be no simple task, but the final reward will provide satisfaction to all who participate and allow themselves to be beneficiaries of the D-20 system.

   Problems to be overcome are:

   .Resistance to change
   .Resistance to use the products produced by others
   .Resistance to modify state statutes and codes
   .Resistance to apparent (but not actual) control release

   However, there are offsetting arguments and factors:

   .It will be a state and not a federal system
   .Savings can be realized
   .The states will have developed the system, and a concensus will have been gained
   .The degree to which the system is developed and/or utilized is optional

The D-20 project will be successful. Its degree of success will depend on the effort and resources the participating states put forth and their willingness to implement the standard they produce.

AMERICAN NATIONAL STANDARDS INSTITUTE
D.20 COMMITTEE, STATES' MODEL MOTORIST DATA BASE

DATA ELEMENT DESCRIPTION

1. REGISTERING COMMITTEE IDENTIFIER	2. SEQUENCE NO.

3. TYPE OF SUBMISSION ☐ INITIAL ☐ REVISED	A. SEQUENCE NUMBER OF PREVIOUS SUBMISSION	4. PREPARATION DATE	5. DATA ELEMENT NO.

6. DATA ELEMENT NAMES

A. FULL NAME

B. SHORT NAME	C. ABBREVIATION

7. DATA ELEMENT DEFINITION

B. DATA ELEMENT SOURCES

9. DATA ELEMENT USES

10. SYNONYMS

11. TYPE OF DATA ELEMENT ☐ BASIC ☐ COMPOSITE (List component parts)	12. NHTSA DATA ELEMENT ID NO.

13. TYPE OF REPRESENTATION ☐ NAME ☐ ABBREVIATION ☐ CODE ☐ NUMERIC VALUE	14. LENGTH ☐ FIXED ( ) NO. OF CHARACTERS ☐ VARIABLE ( ) MINIMUM ( ) MAXIMUM
15. TYPE OF CHARACTER(S) ☐ NUMERIC ◇ ALPHABETIC ☐ ALPHANUMERIC ☐ SPECIAL	16. OTHER CHARACTERISTICS

17. DESCRIPTION OF DATA ITEMS

NAME OF ITEM	ABBREVIATION (Mnemonic Code)	CODE	DEFINITION

18. SOURCE OF DATA REPRESENTATIONS	19. DATA ELEMENT PRIORITY ☐ REQUIRED ☐ OPTIONAL

INSTRUCTIONS FOR COMPLETING

THE
DATA ELEMENT DESCRIPTION FORM
STATES' MODEL MOTORIST DATA BASE

Prepared by
Data Directory Committee D20.1
American National Standards Institute

INSTRUCTIONS FOR COMPLETING

THE

DATA ELEMENT DESCRIPTION FORM

The Data Element Description Form is used by the D-20 Committee of the American National Standards Institute and its task groups for entering information about a data element contained in the States' Model Motorist Data Base. The information contained on the form is utilized by Task Group D.20.1 in entering the data element in the data directory. The completed forms submitted by the various D-20 Task Groups will be used by D.20.1 to determine elements that have duplicate names or meanings.

The Data Directory resulting from the D-20 standardization work will contain descriptions of all the data elements used in the data base from the various application areas. The directory will be structured to facilitate usage by systems managers, designers, and programmers. Element names can be located from a key word in context (kwic) listing. Cross indexes and categorizations will be provided.

To avoid duplication in the directory and afford precision in the future utilization of the data base, each data element will be provided with a unique, concise name and identifying number.

The following guidelines are provided for completing the Data Element Description Form:

(Items 1 through 4 are for administrative purposes. Items 5 through 12 are used to identify, define, and qualify the data element. Items 13 through 18 provide information concerning the representation of the values (data items) used to record specific facts or conditions of the data element.)

(If additional space is needed, attach sheets identifying the appropriate item(s).)

1. REGISTERING COMMITTEE IDENTIFIER.

Enter the designation of the committee (e.g., D.20.2).

2. SEQUENCE NUMBER.

This number is for reference purposes during the standardization process. This number is consecutively assigned to each form by the preparing committee. (A separate form is used for each data element. Also a new number should be assigned to revised submissions.)

3. TYPE OF SUBMISSION.

Indicate whether this is an initial or revised submission. If revised submission, indicate the sequence number of the previous submission.

4. PREPARATION DATE.

Enter the date that this form was prepared.

5. DATA ELEMENT NUMBER.

Leave this item blank. This number will be assigned later in the standardization process.

6. DATA ELEMENT NAMES.

For entry in the Data Directory, each data element must have a unique, concise name. Shorter names which are not necessarily unique may be used on input forms, outputs, and in other formats where the uniqueness is provided by the context of the application. For example, "Driver's License Expiration Date" would be the complete or full name of the data element as entered in the Data Directory. "Expiration Date" could be the element's name as it might appear on the driver's license or on other forms of input or output. In

certain cases it may be desirable to use abbreviations for the names of data elements. These are provided also in the Data Directory.

6.1 DATA ELEMENT FULL NAME.

Enter the complete concise name of the element (e.g., "Driver's License Expiration Date").

6.2 DATA ELEMENT SHORT NAME.

Enter the name of the element as it would appear on forms where the uniqueness is provided by the context (e.g., "Expiration Date").

6.3 DATA ELEMENT ABBREVIATION.

If the data element name is abbreviated or otherwise shortened in use on forms or as headings of reports, enter the shortened form (e.g., "Expiration Date" could be shortened to "Exp. Date" and Social Security Account Number could be shortened to "SSAN").

7. DATA ELEMENT DEFINITION.

Provide a complete definition of the data element. Cite standard definitions and sources where applicable. Avoid definitions that describe how the element is used (This is provided for in a later item on the description form.) Definitions must provide the information content that is derived from this element. Do not use abbreviations or acroynms in the definition. If techical terms are used, provide explanations of these as deemed necessary to improve understanding.

Indicate if the definition of this element or its items may be subject to local interpretation.

8. DATA ELEMENT SOURCES.

Where possible provide identification of the source or sources of the data. For example: "Date of Birth" is usually provided by the individual concerned; "Vehicle Identification Number" is provided by the vehicle manufacturer; or "Time of Accident" is usually provided by

the investigating officer.

9. DATA ELEMENT USES.

List the actual or intended uses that are or may be made of this data element. For example, in the driver data base, "Date of Birth" may be used to determine "identification", "age groups", and "characterization" of drivers.

10. SYNONYMS.

Provide other names by which this data element was or may have been known prior to its standardization. This will provide a cross reference for bridging between current practices that may not be standard and conversion to the standard.

11. TYPE OF DATA ELEMENT.

Some elements provide only a single fact (basic elements). Others provide information whereby multiple facts can be derived (composite elements). For example "Color of Eyes" is a basic element that provides a single fact (i.e., "Color"). "Date of Birth" is a composite element that in addition to providing the "date", also provides information whereby "year of birth", "month of birth", and "day of month of birth" can be derived. (Note: If the component parts of a composite data element are or are expected to be accessed as independent elements, additional data element description forms should be completed for these in addition to the form for the composite data element. In these cases, identify the name of the composite element and its sequence number.)

Indicate the type of element. If it is a composite element, list its component parts in the order provided.

12. NHTSA DATA ELEMENT IDENTIFICATION NUMBER.

If this data element or its equivalent is listed in the National Highway Traffic Safety Administration "State Traffic Record System Design Manual" indicate the data element number assigned.

13. TYPE OF REPRESENTATION.

Indicate the type or types of representation(s) used to record or document the data items (values) associated with the data element. (Abbreviations are a shortened form of the name of the data item and may or may not be of a fixed length. Codes are fixed length and may or may not be derived from the data item name. When a representation is a shortened fixed form derived from the data item name, e.g., Male = M and Female = F, both abbreviation and code representations should be indicated. Numeric values are representations that convey mathematical or measurement meaning. Numbers that provide for identification, such as serial numbers or Social Security Account Numbers, are indicated as codes, not as numeric values.)

14. LENGTH.

Indicate whether the representation is fixed or variable in length. If fixed, indicate the number of characters. If variable, indicate minimum and maximum number of characters. (If check character(s) are used, these should be counted in the code length.)

If the data element is a composite data element, indicate the position and length of each of its component parts.

15. TYPE OF CHARACTER(S).

Indicate the type(s) of characters used: Numeric (0 through 9), alphabetic (A through Z), alphanumeric (0 through 9 and A through Z) and special characters (such as "+*%/;:" other than 0 through 9 and A through Z).

16. OTHER CHARACTERISTICS.

Describe other features or characteristics of the representation. If a self-checking code(1) is used, describe method employed.

17. LIST OF DATA ITEMS, CODES, ABBREVIATIONS AND DEFINITIONS.

List the names of the data items associated with the data element, their abbreviations or mnemonic codes (optional), and assigned codes (other than mnemonic codes). Definitions of the items should be provided as necessary to provide understandings of the intended meanings. (If all items cannot be listed in the space provided on the form, attach additional pages or references that provide the information needed. If another American National Standard or other authoritative reference is used as a basis for the representation, enter "See Item 18".)

If the data items of this data element are numeric values, use this space to describe their characteristics. This should include the following: (1) Position of sign, if used; (2) Position of assumed or explicit decimal point; (3) Rounding rules applied; and (4) Range of permissible or allowable values, if applicable.

18. SOURCE OF DATA REPRESENTATIONS.

Identify the source or reference that provides the representations or codes used. If another American National Standard is cited, indicate its number and title, e.g., "X3.30-1972, Representation of Calendar Date". If other source(s) are used, indicate number (if assigned), title, and address where copies may be obtained, e.g., "Federal Information Processing Standard 8-2, Standard Metropolitan Statistical Areas, Superintendent of Documents, Washington, D.C. 20402, Price 30 cents." (A copy of the source document should be provided with the Data Element Description Form

(1) A self-checking code is a code that is appended to another code to provide for validity checking. A self-checking code is derived mathematically from the characteristics of the base code.

when it is submitted for inclusion in the Data Directory.)

19. DATA ELEMENT PRIORITY

Indicate the criticality of the data element to the States' regulatory, administrative or safety functions. If the data element is not critical but desirable, check the optional box.

Completed forms should be forwarded to the D.20.1 Chairman for further processing. The address is:

Mr. A. Dewey Jordan
National Highway Traffic Safety Admin.
400 7th Street, S.W.
Washington, D.C. 20590

FINAL DRAFT

AMERICAN NATIONAL STANDARDS INSTITUTE 0.20 COMMITTEE, STATES' MODEL MOTORIST DATA BASE DATA ELEMENT DESCRIPTION FORM CHECK LIST	REGISTERING COMM. IDENTIFIER	SEQUENCE NO.	REVIEW DATE

1. REGISTERING COMMITTEE IDENTIFIER
   - ☐ NOT ENTERED
   - ☐ ENTERED IMPROPERLY

2. SEQUENCE NO.
   - ☐ NOT ENTERED
   - ☐ ENTERED IMPROPERLY - EACH SUBMISSION SHOULD BE SEPARATELY NUMBERED. IF REFERENCE IS MADE TO A PREVIOUS SUBMISSION, IT SHOULD BE IDENTIFIED IN 3.A.

3. TYPE OF SUBMISSION
   - A. SEQUENCE NUMBER OF PREVIOUS SUBMISSION
     - ☐ NOT ENTERED
     - ☐ IF REVISION, 3.A. NEEDS TO BE COMPLETED

4. PREPARATION DATE
   - ☐ NOT ENTERED
   - ☐ NOT COMPLETE

5. DATA ELEMENT NO.
   - ☐ SHOULD BE LEFT BLANK

6. DATA ELEMENT NAMES
   - A. FULL NAME
     - ☐ NOT ENTERED
     - ☐ SUGGESTED CHANGE: _____
   - B. SHORT NAME
     - ☐ SUGGESTED CHANGE: _____
   - C. ABBREVIATION
     - ☐ SUGGESTED CHANGE: _____

7. DATA ELEMENT DEFINITION
   - ☐ NOT ENTERED
   - ☐ INSUFFICIENT FOR ADEQUATE INTERPRETATION
   - ☐ ABBREVIATIONS & ACRONYMS ARE NOT TO BE USED
   - ☐ EXPLAIN TECHNICAL TERMS INDICATED BELOW:
     _____
     _____
     _____
     _____
   - ☐ LOCAL INTERPRETATION NOT INDICATED

8. DATA ELEMENT SOURCES
   - ☐ NOT ENTERED
   - ☐ ENTERED IMPROPERLY, EXPLAIN: _____
     _____

9. DATA ELEMENT USES
   - ☐ NOT ENTERED

10. SYNONYMS
    - ☐ SUGGESTED SYNONYM: _____
      _____

11. TYPE OF DATA ELEMENT
    - ☐ NOT ENTERED
    - ☐ ENTERED IMPROPERLY, EXPLAIN: _____
      _____

12. NHTSA DATA ELEMENT ID NO.
    - ☐ CHANGE TO: _____

13. TYPE OF REPRESENTATION
    - ☐ NOT ENTERED
    - ☐ ENTERED IMPROPERLY, EXPLAIN: _____
      _____

14. LENGTH
    - ☐ NOT ENTERED
    - ☐ ENTERED IMPROPERLY, EXPLAIN: _____
      _____

15. TYPE OF CHARACTER(S)
    - ☐ NOT ENTERED
    - ☐ ENTERED IMPROPERLY, EXPLAIN: _____
      _____

16. OTHER CHARACTERISTICS
- ☐ NOT ENTERED
- ☐ OTHER, EXPLAIN: _____

17. DESCRIPTION OF DATA ITEMS
- ☐ NOT ENTERED
- ☐ OTHER, EXPLAIN: _____

18. SOURCE OF DATA REPRESENTATIONS
- ☐ NOT ENTERED
- ☐ ENTERED IMPROPERLY, EXPLAIN: _____

19. DATA ELEMENT PRIORITY
- ☐ NOT ENTERED

OTHER COMMENTS:

- ☐ THIS DATA ELEMENT IS A COMPOSITE. THE PARTS THAT MAKE UP THE ELEMENT SHOULD HAVE SEPARATE FORMS COMPLETED, IF THESE ARE TO BE ACCESSED OR ADDRESSED INDEPENDENTLY.
- ☐ ITEMS CORRECTED BY THE ANSI D-201 COMMITTEE: _____
- ☐ OTHER: _____
- ☐ ADDITIONAL COMMENTS: _____

REVIEWED BY: _____

ADDENDUM

The numerous questions received from the audience address three general areas. They are (1) the privacy and security of the information and how it will be used, (2) the need and justification for the D-20 project and (3) the progress thus far, and the problems encountered in data element definition and data base/data exchange design.

I would like to address these in reverse order. Progress thus far, as explained in the paper, has been primarily in the area of data element definition. Research is being conducted in the areas of data management systems and existing communications networks, which has indicated that a few transportable software systems are a reality and standardized data element systems are currently operating. The control and processing of the data elements will require the use of some data dictionary/directory facility. No phases of the D-20 project have been implemented as of yet.

The question of need and justification should be addressed from two points of view: (a) the service level desired and (b) the long range potential, rather than the short range costs. The "need" of the system is indicated by the number of participants, which includes representatives of 43 states. If this project is desired then it should be developed with the widest scope and designed for its maximum benefit. This should be done in a manner that will produce a model that is both modular and transferrable, so that it may be implemented to the degree that meets the requirements of the users. If each state chooses to develop on-line systems for immediate data access within their state, then the effort and cost to link those states into an interstate system would be small, in comparison to the total expended by the states in developing their own systems. Likewise, a single development cost that will provide a product each state can use would reduce the need of individual state expenditure to create that product. Any change of conversion will necessarily demand a short term cost. However, if viewed from the perspective of long range benefit and the more effective, efficient use of our resources and knowledge, the short term costs can be offset. Immediate information is as valuable as one wants to consider the value of human lives, safe highways and wasted resources.

Finally, the protection of this data and the determination of its use will be controlled by Legislation, on both state and federal levels, and by individual State and agency policy. However, as stated in the text of the paper, the data management system must address the security of the data from unauthorized access and use; must protect the integrity of the data from erroneous change or deletion; insure completeness and accuracy; and provide for, as required by law, the right of review.

Information System Data Coding Guidelines

M. J. GILLIGAN

Western Electric Company, Inc.
Information Systems Engineering
Newark, New Jersey 07102

Choice of the codes used to represent informational values is an important part of the design of an information system. Each informational value used in a system should be represented by an optimum number of codes, each code being most efficient for its specific use within the system. By "efficient" is meant suitable for accomplishing a task accurately and quickly.

In some applications it may be desirable to use the same codes at every place and for every purpose throughout an information system: for data input and output (man-machine interfaces), internal storage, data processing and data transmission. In other applications it may be better to represent specific informational values by different codes at different places or for different usages, depending upon the information encoded, the type of usage, and the nature of the users. For example, in a specific application, mnemonic alphabetic codes may be best for human-oriented data input and output, while it may be preferable to represent the same values by sequential numeric codes for internal machine data processing.

The choice of what to do in any given application should include consideration of the probably increased accuracy and reliability of a system that uses human-oriented input and output codes as well as the one-time cost of developing conversion routines required if the input and output codes are to be represented by different codes for machine use. In any case an understanding of the various types of character-string codes, code schemes and related matters should be valuable to developers of information systems.

Key Words: Check digits; codes; data standards; information load; information systems; item identification; mnemonic codes; system design.

1. INTRODUCTION -

1.1 Purpose

The purpose of this document is to present descriptions of the kinds of data codes that can be used in information systems and to present some guidelines regarding the selection and development of codes for use in information systems developed for use by Western Electric Company.

An information system can be defined as a set of methods, procedures and physical devices (paper forms, cards, machines, etc.) that are designed, selected, executed and operated for the purpose of acquiring, recording, storing, processing, transmitting and displaying information. The information, or "data" as it is commonly called, may be recorded, etc., in full natural language - that is, as "text" - or, more often, is recorded (etc.) in some condensed, more concise form, such as abbreviations or codes. In fact, it appears that most information systems actually are systems for recording (etc.) codes. The selection and/or development of the codes by means of which information will be recorded and/or input, stored, processed, transmitted and displayed (output), therefore, is an integral and important part of the design and development of every information system.

In any modern information system that uses digital computer equipment to store, process and transmit data there are usually many code-using tasks, some performed exclusively by machines, some performed by humans and some performed by both humans and machines working together at a human-machine interface. The information codes that are "best" (appropriate, efficient, easier to select, etc.) for one task may not be "best" for another task. For example, if one of a system's sets of codes is to be sorted or rank-ordered by computer program then the system programmers may insist that the codes be numeric, even if the human users of the code set - i.e., the clerks or other operators who will originate the system's input - would perform more effectively and efficiently with alphabetic codes. Conversely, system users who want an immediately - interpretable output may insist upon using "abbreviations" as codes for system input and machine processing as well as for output regardless whether the abbreviations are at all appropriate for the first two tasks.

The ISE Data Standards Department believes that the codes used in information systems should be appropriate to the tasks for which they are used. The present document consists essentially of two parts: some recommendations regarding what kinds of codes should be used for various tasks in information systems (Section 1.3); and an explanation of various types of character-string codes and of

some special considerations regarding use of such codes (Sections 1.4 through 6.25).

We hope that this information and opinion will be of interest and value to the Company's information systems personnel and all others concerned with the development and use of effective information systems.

1.2 Scope

By a "code" we mean a graphic symbol or string of adjacent (juxtaposed, catenated) graphic symbols that stand for and can be used in place of a natural-language word or phrase or quantitative value. By graphic symbols we mean the following typographical symbols or characters: the letters of the English alphabet, the Arabic numerals, punctuation marks, plus certain other similar symbols, which are listed in Section 1.3 of this document. In this document and for the present purpose we do not include symbols such as various crosses, stars, astronomical, biological, chemical, mathematical, physical, musical, etc. symbols.

Codes can be described or categorized as numeric, alphabetic, alphanumeric, alpha-numeric-special, etc., according to the typographical characters of which they are composed. Such categorization is helpful in describing codes, and becomes important in naming code types or data types for data validation, editing, etc. during the use of file-management or data-management systems and in data definition sections of COBOL, FORTRAN, etc. programs. Therefore, the terms alphabetic, numeric, etc. should be defined accurately and precisely, as in Section 1.3 of this document.

By "coding" we mean the development or generation of codes as defined above. By "coding" we do not mean the writing of computer-language instructions or commands ("programming") and we do not mean the encoding of characters or character strings themselves into binary bit-strings (e.g., as EBCDIC 8-bit codes, USASCII 7-bit codes, excess-6 binary codes, etc.), Morse code, etc.

Nor by coding do we mean the cryptographic encoding of information (for purposes of secrecy, etc.). Such cryptographic coding may resemble the coding we seek to discuss, but the purposes of Western Electric information system developers are quite contrary to these of cryptography; we seek to encode information merely so that it can be recorded, stored, transmitted, retrieved, and decoded easily, efficiently and effectively.

1.3 Summary of Recommendations for Code Design and Coding Procedures

Design or selection of code sets for use in information systems is an integral and important part of the system design process. We recommend therefore, that information system users, designers and developers familiarize themselves with basic concepts of information code design as presented in the present document and consider the guidelines presented in the following paragraphs of this section during the early stages of information system design. Recommendations regarding various typical information-encoding situations are summarized below.

a. For classification schemes, where the various subcategories of a general body of informationn must be classified and assigned identification codes, and typically not all of the possible subcategories and specifically known initially and in some cases not even the number of hierarchical levels that eventually will be needed is initially known, we recommend the use of blocked sequential numeric code schemes, as described in Section 3.0. If the future needs of the code scheme are not fully known at the time the scheme must be defined then the fixed-length form of the Decimal Classification Code scheme, also described in Section 3.0, can be used, with lower-level digit positions reserved for future expansion.

   We recommend that numeric codes used for classification codes be stcred, processed and transmitted within EDP systems in the formats or modes that are best for those purposes, but we strongly recommend that codes longer than 5 digits in external graphic presentations (manual forms and records, EDP input forms, and all output presentations) be "chunked" as discussed in Section 6.22.

b. For item identification schemes, where it is necessary to set up a coding scheme under which identification codes will be assigned to tangible items (e.g., piece parts) or intangible items (e.g., orders) over a period of time, we recommend the use of sequentially-assigned serial numbers as described in Section 3.0. The serial numbers can be divided intc blocks if necessary. We do not recommend the use of alphabetic or alphanumeric "numbers" for item identification applications.

   We recommend that numeric codes used for item identification be stored, processed and transmitted within information systems in the formats or modes that are best for those purposes, but we strongly recommend that codes longer than 5 digits in external graphic presentations

(manual forms and records, EDP input forms, and all output presentations) be "chunked" as discussed in Section 6.22.

c. Limited-scale code schemes, where a relatively small number of concepts, facts, categories of information, items, etc. are to be coded for purposes of recording, entry into automatic data processing systems, storage, automatic data processing and/or transmission, data display, report generation, etc., are discussed in the following paragraphs.

The best solutions to such data encoding situations cannot be simply and concisely described as though by formula. The usual approach to such data encoding problems is to choose one type of codes and thus one code set for a set of entities and to use that one code set in all code-using stages of an information system. Thus, for example, a numeric code set is chosen and serial-number codes are assigned because "the codes have to be sorted by machine and the computer cannot sort on alphabetic codes;" a "mnemonic" alphabetic code is specified because it is "human-oriented," as though nothing else mattered; or, instead of codes in the ordinary sense, "abbreviations" or, honestly enough, "text" is specified for "input, internal storage and output," because "we have to be able to understand the output and we do not want to waste time looking up codes."

The solutions to the apparent dilemmas implied by the examples cited above lie not merely in code design but in information system design. It may be necessary to encode each informational concept (or set of concepts) by different codes (code sets) at different stages of an information system, and it is our recommendation that this be done unless it is economically unjustifiable.

For the manual recording of information, including filling out of forms, direct keyed data entry, etc., where a relatively small code set (up to about 50 codes) will be entered repeatedly by the same operators, and where the task can be structured so that the operators can learn the codes, we recommend the use of alphabetic mnemonic codes (equal-length abbreviations as described in Section 4.4.). Such mnemonic alphabetic codes should be 3, 4 or (no more than) 5 letters long. If for automatic data processing purposes it is desirable to represent the information thus encoded by numeric codes, then the translation into numerical codes ought to be accomplished by the computer program, not by the persons entering the codes.

For manual recording (input) of information, where the number of codes is large (more than about 50 codes) or where the codes will be entered infrequently (say, only once or twice, for example as when establishing an initial data base) and where, therefore, there is no need or no benefit to be derived from learning of the codes, we recommend the use of sequential numeric codes. Such codes should be satisfactory alsc for automatic data processing purposes, but if it is necessary that they be translated such translation should be done by computer program.

For output (display, report generation, etc.) we recommend the use of human-oriented alphabetic codes, abbreviations, or even full-length natural-language text. If the information to be presented is stored internally via numeric codes then the translation to alphabetic human-oriented output form should be accomplished by the computer prcgram prior to output.

The recommendations presented above represent in condensed form a "philosophy" of information code selection and usage for information system design and development. The Corporate Data Standards Organization will welcome any opportunity to discuss these recommendations and will consult constructively on any pertinent code design or usage application.

1.4 Data Types

Various sources name, describe and attempt to define various character sets, "alphabets" or "data types." Differences in names and definitions are due to differences in training and experience of the persons who devise and publish lists of data types and the needs of the specific applications for which such data types are defined. For example, Western Electric Data Standard 10143, "Character-Code Data Type," which was prepared to support an input-data auditing and editing function, defines and assigns one-letter codes to four character sets, as indicated in the next table.

Character-Code Data Types

Code	Name	Description
X	Alpha-Numeric-Special	All printable characters (Note 1)
A	Alphabetic	Alphabetic characters A through Z (Note 2)
N	Numeric	Numeric Characters 0 (zero) through 9
S	Special	All printable characters in set X except Alphabetic (Set A) and Numeric (Set N). (Notes 1, 3)

Note 1: These descriptions do not explicitly define "all printable character" or all "special" graphic characters, as by naming or listing them. This was done deliberately because in electronic data processing the various character sets used have different numbers of special characters. For example IBM's EBCDIC (256-code) character set has twenty-seven "Special Graphic Characters," including the unique code for "blank space" while the USASCII 128-code character set allows for the representation of thirty-four such "graphic characters" as does USASCII's 95-character graphic subset, while USASCII's 64-character graphic subset includes only twenty-eight special graphic characters. The characters actually available for input and output (display, printout, etc.) depend upon the physical device. For example, varous IBM print chains have different subsets of the full EBCDIC character set, and some print chains may have extraordinary graphic characters assigned to certain EBCDIC codes, for example the British pound Sterling sign, £, in place of the American # symbol. Similarly for automatic typewriters, CRT display matrices, etc.

Note 2: Alphabetic includes only the twenty-six upper case (capital) letters; the lower case (small) letters are not included, nor is the "blank space" character.

Note 3: Although many binary codes, when interpreted by an automatic data processing output device, will cause no visible character to be printed or displayed, EBCDIC, ASCII, etc. each assign a specific unique binary code to represent the "blank" or "space" character. This blank space character code is considered to be a "special graphic character."

Other sources define many more character sets, some of which are subsets of one another. For example, Federal Information Processing Standards Publication (FIPS PUB) 20, "Guidelines for Describing Information Interchange Formats", lists the following "character types", where "character type" is defined as "An indication of the type of characters or bytes to represent a value (i.e., alphabetic, numeric, pure alphabetic, pure numeric, binary, packed numeric, etc.)."

FIPS PUB 20 Character Types

Name	Description
Alphabetic	A representation which is expressed using only letters and punctuation symbols.
Pure Alphabetic	A representation which is expressed using only letters.
Alphanumeric	A representation which is expressed using letters, numbers, and punctuation symbols.
Pure Alphanumeric	A representation which is expressed using only letters and numbers.
Numeric	A representation which is expressed using only numbers and selected mathematical punctuation symbols.
Pure Numeric	A representation which is expressed using only numbers.
Packed Numeric	A representation of numeric values that compresses each character representation in such a way that the original value can be recovered, e.g., in an eight-bit byte, two numeric characters can be represented by two four-bit units.
Binary	A representation of numbers which is expressed using only the numbers 0 and 1. E.g., 5 is expressed as 101.

We recommend that codes for Western Electric Information Systems be composed of (1) certainly no other than the 95 Graphic characters (including the explicit blank space) of USASCII, or (2) preferably only the twenty-six upper-case English-alphabet letters, the ten Arabic numerals, plus - where a separator is needed - the hyphen (-).

The reason for the first recommendation is that its observance will ensure that codes and code sets developed for use in information systems will be usable no matter what manufacturer's machines are used to implement the information system, so long as the machines implement at least the basic 95 graphic characters of USASCII. The reason for the second recommendation is that principles for development and use of human-oriented codes recommend that "special" symbols not be used in codes, except that the hyphen should be used to separate (connect) the parts of a long code or code chain.

Therefore, for the purposes of this document we define three data types or character sets into which we categorize character-string codes, as follows:

Name	Description
Alphabetic	The twenty-six upper-case English-alphabet letters A through Z.
Numeric	The ten Arabic decimal numerals 0 (zero) through 9.
Alphanumeric	The combination of the 26 upper-case letters A through Z and the ten numerals 0 through 9.

1.5 Terminology

A group of one or more codes is called a set of codes or a "code set." The "American National Standard Vocabulary for Information Processing" defines the term "code value" as "one element of a code set." We call an individual code a "value" or a "code value" and a set of codes, a "value set." This seems to be by analogy to the way in which specific numbers - e.g., 685 or 6785487 - are often called "values" whether or not they represent magnitudes or quantities. For example, the nine-digit number 247389731 can represent: the magnitude two hundred forty-seven million, three hundred eighty-nine thousand, seven hundred thirty-one; or it can represent a U.S. Social Security Number; or it can represent a Western Electric COMCODE. In the first case the number would probably be called a "value," in the second, probably a code, and in the third case, quite literally it is called a code.

Similarly, given a 3-character alphabetic code set ACK, DIS, RET, and SKP, each of the three-character strings is a code, but each could also be called a value. In particular, in electronic data processing we often speak of fields' being populated with codes or with data values.

Considerable experience dealing with this problem in semantics has led us (the Data Standards Organization) to adopt the position that it is not especially useful to seek or insist upon extremely precise and/or mutually exclusive definitions of the terms "code" vs. "value;" we consider them to be effectively synonomous in the context of the work with which this document is concerned. Similarly, terms such as designator, designation, abbreviation, level, category, class, classification, number, type, identity, identifier, identification, "ID," etc. are used as synonyms. This appears to arise from the facts that some codes do "designate" or "identify" or do represent class or type or category or level, or are numbers or abbreviations, and in such situations quite often the generic word (name) "Code" is replaced by another, more specific word that names what the code does or represents or is. This is normal in human communication and it is not useful to try to "prohibit" such language. However, we do hope and expect that readers and users of this document will realize that a code is a code is a code, even if it is called an "identifier" or a "type" or a "number" or an "ID," etc.

Another question that appears often in discussions of codes and coding is the question whether, for example, a string of characters "is" two codes or one code when the first (left-most) few characters stand for one thing or one aspect of an informational concept and the remaining (right-most) few characters stand for another. For example, consider a five-character code format, where the first three characters indicate "physical device identity" and the last two indicate "mode of operation." Some workers would argue that the codes for physical device identities would be one code set while the codes for the modes of operation made up another code set, and might insist that the two code sets be listed in two separate Data Standards, linked by a third, "chain" Data Standard. This question is not significantly affected by whether the two segments are presented graphically (displayed, printed out, etc.) with a graphic separator (e.g., hyphen) between them.

After considerable discussion regarding such problems we have developed the following policy: in general, a character string represents one code (and thus is documented by one Data Standard) if it is normally recorded and/or input to an information system at one time as an entire string (even if such a system input character string is assembled or composed out of separate parts prior to its

being recorded or input) and if the parts are not used separately elsewhere.

On the other hand, if a group of two or more codes (each consisting of one or more characters), which are recorded and/or input separately, thereafter are processed or transmitted together or output (displayed) in such a manner as to appear to be one character string, then the resulting longer string is called a code chain and is documented via Western Electric Data Standards as a chain. For example, three elements are used to make up the usual calendar date: day, month and year. Each element can be coded and can be used separately, so each element has its own code set. In fact, in this case each element can have more than one code set; see the following illustrative table. Therefore, there would be one or more Data Standards for each element, documenting its code set(s), plus an overall chain Data Standard documenting each of the possible code chains that it has been judged desirable to document as Western Electric Data Standards.

Codes for Various Elements of Calendar Dates

ELEMENTS:	DAY OF WEEK		DAY OF MONTH (1)	DAY OF YEAR (2)
CODES:	ALPHA	NUMERIC	NUMERIC	NUMERIC
	MON	1	01	001
	TUE	2	02	002
	WED	3	03	003
	THU	4	(etc.)	(etc.)
	FRI	5	29	364
	SAT	6	30	365
	SUN	7	31	366

ELEMENTS:	MONTH OF YEAR (3)	(4)	WEEK OF MONTH (5)	WEEK OF YEAR (6)
CODES:	ALPHA	NUMERIC	NUMERIC	NUMERIC
	JAN	01	1	01
	FEB	02	2	02
	MAR	03	3	03
	(etc.)	(etc.)	4	(etc.)
	OCT	10	5	51
	NOV	11		52
	DEC	12		53

ELEMENT:	YEAR (7)			
	1	2	3	4 digits
CODES:	2	72	972	1972
	(etc.)			

Western Electric Data Standards:

Note	Number	Name	Format	Example
(1)	10069	Day of Month	NN	31
(2)	10011	Day of Year	NNN	366
(3)	10068	Month-Mnemonic	AAA	AUG
(4)	10067	Month-Numeric	NN	12
(5)	10201	Week of Month	N	5
(6)	10202	Week of Year	NN	53
(7)	10010	Year	NN	99

2.0 Types of Codes

For the purposes of this document we defined codes according to their "data types" - i.e., according to the character sets from which their characters are selected.

The resulting categories are, again:

Numeric,
Alphabetic,
Alphanumeric

Codes and code sets can also be categorized or described according to whether they manifest certain characteristics (features, attributes), either by design or de facto, as listed below. These various attributes are not necessarily mutually exclusive; some codes or code sets may manifest more than one of these characteristics.

a. Blocked (grouped) or positionally-significant, including dependent and non-dependent, "decimal" (fractional), exponential, high-order (first-digit) low-order (final digits), etc., codes;

b. Sequenced, including collating codes and serial codes;

c. Non-sequenced, including "hashed" and other "random" codes;

d. Algorithmic (generated according to a set of rules, i.e., a mechanistic algorithm), including error-detecting or self-checking codes and certain kinds of abbreviations.

e. Mnemonic codes, i.e., codes specifically designed to be easy to remember, including alphabetic abbreviations, acronyms, etc., and some numeric codes.

3.0 Numeric Codes

Numeric codes are codes made up of the ten Arabic numerals (numeric digits) 0 (zero) through 9.

3.1 Sequential Numeric Codes

A sequenced or sequential set of numeric codes is one whose codes are arranged in numeric integer sequence ascending in value as though the codes were magnitudes (representing quantities). For example, a set of persons' Social Security Numbers could be arranged in sequential order as though they were truly "numbers" - i.e., representaticns of quantities - rather than merely almost-

random codes. A simpler sequenced set of numeric codes would be, for example, the code set 1, 2, 3, 4. Thus sequencing (sequenced, sequential) is a property of a code set, not of any one code by itself.

3.11 Types of Sequential Numeric Codes

If all possible values in such a sequenced code set are used - or at least reserved for use - then the code set is said to be serial, for example, the code set above, 1, 2, 3, 4. A non-serial sequenced code set would be one having unreserved gaps, for example, the complete code set 01, 02, 05, 06, 09, 12. A sequenced numeric code set, be it serial or not, is also a collating code set. This is familiar to programmers who are aware of the collating sequence of EBCDIC or ASCII binary codes for letters and numerals, as illustrated on programmers' reference cards and the like. For example, a sequential (but not serial), collating, 2-digit, numeric code set for the twenty-six letters of the alphabet plus the blank space character is illustrated by the table below. This code set, of course, is merely "for example."

A Hypothetical Sequential Numeric Code Set

Symbol	Code	Symbol	Code
(blank)	01	M	52
A	02	N	53
B	03	O	61
C	21	P	62
D	22	Q	63
E	23	R	71
F	31	S	72
G	32	T	73
H	33	U	81
I	41	V	82
J	42	W	83
K	43	X	91
L	51	Y	92
		Z	93

Sequenced numeric code sets are usually applied to sets of entities which themselves have been arranged in some useful

sequence, e.g., in alphabetical order for natural-language words or word phrases, such as a list of the states of the United States, or a company's vendors, or employees' last names, etc. Another type of sequence is chronological sequence, the natural sequence of codes assigned sequentially over a period of time.

Other sequences into which lists of entities can be arranged before being encoded are hierarchical classification sequences, where subdivisions of major categories are grouped under their major categories. Several more complicated versions of sequential code sets can be developed to provide numeric codes for hierarchical classification schemes, with as many levels of embeddedness as required. They are described in later paragraphs of this section.

The process of deciding how to arrange (categorize, classify) the entities to be coded is a part of information system design that should be dealt with before the coding scheme is chosen. This aspect of system design is known as taxonomy, systematics, or, simply, classification.

A problem that can arise in use of simple sequential numeric code sets, as heretofore described, all of whose values have been assigned to entities, is that if it then becomes desirable to insert an entity into the list it may be impossible to assign to it a code appropriate to or in accordance with its "natural" (i.e., alphabetical, hierarchical) sequence position. Fortunately certain more sophisticated sequential numeric code sets can be devised to provide for such contingencies, and they are described in Section 3.12 of this document.

Random numeric codes also could be used to solve (avoid) the sequential-code problem mentioned above but they present human-factors problems such that they are not recommended for human use - i.e., for use in information systems where they must be selected, transcribed or translated by humans. Nevertheless, for the sake of completeness random numeric codes are discussed briefly in Section 3.3 of this document.

However a numeric code set is generated, and whether it is serial, sequential, non-sequential or random, a rule that is generally agreed upon is that the code value zero (0 or 00 or 000, etc., depending on the length of the code) should not be used to stand for any entity in the list to be coded. Instead, the zero value should be reserved to indicate the fact that the field (or box or line, on a paper form) has not yet been populated with a meaningful code. Some coding authorities recommend that the highest value (9 or 99 or 999, etc.) also be reserved to indicate the "last" code or the end of a sequential code set. The latter suggestion is not always easy to follow but if it is possible to do

537-860 O - 74 - 9

so and if such an indicator code is needed then 9 (or 99 or 999, etc.) seems a reasonable choice for that function.

The next level of sophistication of sequenced numeric codes are blocked sequential codes. A simple block sequential code scheme divides a sequential code set into a specific number of blocks, representing equal-level categories, by assigning one or more digit positions to that purpose. Normally the highest-order (leftmost) digit or digits are used for such block codes, because the codes can then be most easily automatically sorted, when necessary, as though they were quantitative numbers rather than non-quantitative codes.

For example, consider the following block code scheme:

Codes	Entities Coded
1000 thru 1999	Meats
2000 thru 2999	Produce (Fruits, vegetables, etc.)
3000 thru 3999	Dairy Products
4000 thru 4999	Groceries
5000 thru 5999	Bakery Products
6000 thru 6999	Frozen Foods
7000 thru 7999	Beverages
8000 thru 8999	Cleaning and Paper Products
9000 thru 9999	General Merchandise

This scheme provides 9,000 serial-number codes, 1000 through 9999, for a supermarket's merchandise items. However, the first (leftmost, high-order) digit is used to divide or block the 9,000 codes into 9 high-level categories of 1,000 serial-number codes (000 thru 999) each. This scheme does not reduce the number of serial codes available; it merely provides specific significance to the high-order digit position. The 1,000 codes within each of the 9 equal-level categories remain undifferentiated as to significance; they are merely serial numbers and would normally be serially (sequentially) assigned to the specific entities to be coded.

Notice that the digit 0 (zero) is not used in the high-order position. If zero were used then a tenth high-level category could thus be encoded, allowing 999 more 4-digit codes (0001 through 0999, exlcuding the code 0000). Thus approximately 10% of the otherwise-possible 10,000 codes are not used. Why? Because this code set illustrates a coding prinicple, that codes to be used by humans should not begin with zeroes. The reason for this restraint is concern that high order zeroes might be discarded by a person transcribing (re-writing, key-punching, etc.) such a code, thus changing the code 0123, for example, into the code 123 and thus

potentially into the code 1230, if the code 123 became left-justified in a system that filled on the right with zeroes. Thus 10% of the 10^n possible codes in an n-decimal-digit numeric code set are given up in order to avoid having a zero in the high-order digit position. In other words this restraint or condition means that 90% or $9 \times 10^{n-1}$ codes will be actually available for use out of the 10^n possible codes that a n-decimal-digit numeric code scheme can provide. We recommend that this constraint be accepted.

If it is necessary to encode more than nine categories at any level in a block code then two alternatives exist:

a. Use of symbols other than numerals - i.e., use of letters. This would change a numeric code into an alphanumeric code, which is not recommended, as discussed in Section 5.0.

b. Use of two or more digit positions to encode the categories at a given level, as in the code scheme described below.

Codes	Entities Coded
100000 thru 109999	Diodes
110000 thru 119999	Transistors
120000 thru 129999	Electron Tubes
(etc.)	(etc.)
980000 thru 989999	Resistors
990000 thru 999999	Capacitors

This scheme provides 900,000 (100000 thru 999999) serial-number codes for an electronics-supply stockroom's stock items. The two high order digits are used to provide 90 (10 thru 99, inclusive) categories or blocks of 10,000 (0000 thru 9999) serial number codes each. This is an order-of-magnitude (power-of-ten) increase in the number of categories encoded, from 9 categories using one high-order digit to 90 using two high order digits. Simultaneously the number of serial-numbers available for each category has also been increased and also by a factor of ten, from 1,000 to 10,000 per category.

This approach can be extended indefinitely, with 1, 2, 3 or more high-order digits being used to designate 9, 90, 900, etc., equal-level categories, and as many trailing digits used as are needed to provide 10, 100, 1000, 10,000, etc. equal-level serial-number codes, to be assigned to the lowest-level entities to be coded.

Whether one, two or more high-order digits are used to define the blocks, what these block codes amount to are combinations of (a) positionally-significant equal-level category-block codes with (b) simple sequential or serial codes. Thus, such code schemes are two-level code schemes.

The order-of-magnitude jumps in the high-level categories designated by such high-order digit block codes may result in a jump from "too-few" codes to "too-many" codes. That is, for example, if 25 or 50 equal-level categories are needed, then 9 are not enough and 90 are much more than enough. In such cases adjacent blocks can be operationally combined, as it were, as illustrated by the following:

Codes	Entities Coded
10000 thru 19999	Resistors
20000 thru 29999	Capacitors
30000 thru 39999	Inductors
40000 thru 44999	Electron Tubes
45000 thru 47999	Tube Sockets
48000 thru 48999	Tube Caps
49000 thru 49999	Tube Shields
50000 thru 69999	Transistors
80000 thru 89999	Diodes
90000 thru 99999	Other

In this scheme blocks of 10,000 codes are assigned to Resistors, Capacitors, etc., but one block of 10,000 codes has been subdivided and assigned as follows:

5000 codes to Electron Tubes,
3000 codes to Tube Sockets,
1000 codes to Tube Caps, and
1000 codes to Tube Shields,

because an entire block of 10,000 codes was not needed for Electron Tubes but codes as shown were needed for the other three, related categories. On the other hand, 20,000 consecutive codes are assigned to Transistors and 20,000 to Diodes, as shown above.

Allocations of equal-level blocks in the blocked sequential code schemes discussed above, if done carefully after an adequate determination of the categories needed, may provide an adequate code scheme. However, the result will be only a two-level scheme consisting of sequentially-arranged blocks of serial numbers, as stated earlier.

If blocking, categorization, classification, etc., is needed at levels below the high-order level, then an extension or general case of the high-order blocked sequential code scheme can be used which is called the dependent block, hierarchical or exponential classification code scheme. In these schemes successive digit positions (or adjacent pairs, triads, etc.), from the highest-order digit towards the low-order digits, indicate successively-lower embedded hierarchical levels of classification. Thus the name hierarchial code scheme.

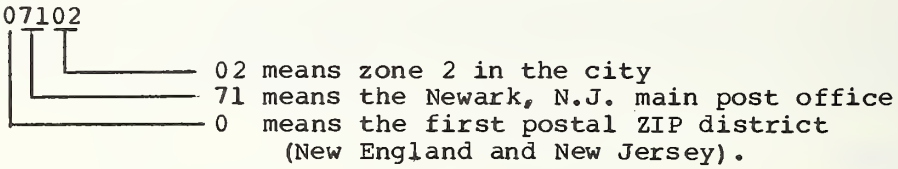
For example, consider again the last coding scheme, for electronic components, described above. It is seen that the "40000 block" (the numbers 40000 thru 49999) has been allocated to 4 categories, as follows:

Codes	Entities
40000 thru 49999	Electron Tubes
45000 thru 47999	Tube Sockets
48000 thru 48999	Tube Caps
49000 thru 49999	Tube Shields

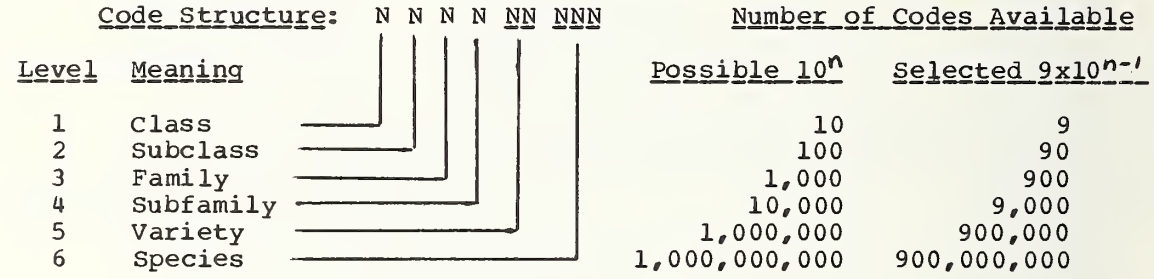
This can be interpreted to mean that a "4" in the high-order digit position means Electron Tubes and related parts, and that a 0, 1, 2, 3 or 4 in the second digit position means Electron Tubes specifically if the first digit is a 4. Similarly a 5, 6, or 7 in the second position means Tube Sockets if the first digit is a 4, an 8 in the second position means Tube Caps when there is a 4 in the first position, and a second-level 9 means Tube Shields dependent upon a first-level 4. Thus the name dependent block code for such schemes.

Similarly, consider the high-level blocks allocated to Diodes and to Transistors in the same code scheme. Whether each 20,000-code group is considered to be one 20,000-code block or two adjacent 10,000-code blocks would appear to be merely a matter of semantics and thus merely academic or inconsequential. That is not so. If the 20,000 codes in each group are assigned sequentially beginning with the lowest number (50000 for Transistors, 70000 for Diodes) then each group is a single 20,000-code block. However if, for example, codes are assigned to different types of Diodes beginning at different numbers within the Diode block (e.g., beginning at 50000, 55000, 60000 and 65000, for 4 different kinds of Diodes) then, as for Electron Tubes and related parts, a second, dependent level of blocking has been developed, because here a 0, 1, 2, 3 or 4 in the second digit position means Diode type I (for example) dependent upon a 5 in position 1, but means Diode type III dependent upon a 6 in position 1.

Another, familiar dependent blocked sequential code is the U.S. Postal Service's ZIP Code, a 5-digit numeric code for post offices. In this code the leftmost digit indicates one of ten national postal districts, the next two digits indicate one of 100 major post offices or Sectional Centers within each district, and the rightmost two digits are assigned sequentially from 00 thru 99 to branch post offices, zones within a city, etc., served by a major post office or to the smaller post offices served by a Sectional Center. Thus for example, in the Zip Code



The general case of decimal numeric dependent hierarchical block classification codes is called the exponential coding scheme because the number of codes available increases exponentially as powers of ten with addition of digit positions to the code. Consider the next illustration:



If all digit positions are allowed to take all values 0 (zero) through 9 then this 9-decimal-digit numeric code scheme provides 10^9 or 1,000,000,000 codes. If the recommended restraint is imposed, that codes beginning with a zero not be used, then the high-order digit position, representing Class, can take the values 1 thru 9 and the code scheme provides 9×10^8 or 900,000,000 codes.

This code scheme provides a six-level hierarchical classification scheme, with 9 classes, 90 Subclasses (10 in each Class), etc., as illustrated above. The user of such a code set would arrange his highest-level categories (classes) in whatever order suits his purposes, and would arrange the 10 Subclasses within each class appropriately, etc. After all the known items to

be coded were arranged (categorized, classified) appropriately then the user could begin assigning codes.

3.12 Assignment of Sequential Numeric Codes

The simplest way to assign codes, after an initial ordering, is to assign them serially at each level, that is, to leave no gaps. For example, if at the beginning there were 5 Classes they would be assigned codes 0 through 5. If Class 1 had 4 Subclasses they would be assigned codes 0 through 3 in the second digit position. If Class 2 had 7 Subclasses they would be assigned codes 0 through 6 in position 2. Et cetera, so that Class 5, Subclass 7, Family 0, Subfamily 3, Variety 82, and first Species would be coded 570382000, and the 524th Species in the same Variety would be coded 570382523.

A consequence of such serial assignment of codes is that if it becomes necessary to assign codes to new entities after the initial entity list has been coded then there is no possibility of inserting entities into the code set at or at least near their proper place as determined by alphabetical order, taxonomical classification sequence, or the like. This problem is solved with varying degrees of sophistication by the techniques described in the following paragraphs.

The next level of sophistication in assigning sequential codes is to arbitrarily skip a constant number of codes between codes assigned, i.e., to establish a constant-value increment to codes assigned, at any level or all levels, subject of course to the constraint that such skipping will not result in running out of code numbers at any level before all entities at that level have been coded. For example, if a given level is assigned one digit position (which can take the values 1 thru 9) and there are 3 categories to be coded at that level then codes can be assigned as follows:

Increment	Codes Assigned
1	1,2,3
2	1,3,5
3	1,4,7
4	1,5,9

If, for example, a level is allocated 2 digits that can take values 00 through 99 (thus,100 values) and there are 25 categories to be coded increments can be used as follows:

Increment	Codes Assigned
1	00, 01, 02, ..., 24
2	00, 02, 04, ..., 48
3	00, 03, 06, 09, ..., 72
4	00, 04, 08, ..., 96

What increment should be chosen? That depends on the needs of the code set user. If it is known that no entities will have to be coded after the initial set or if it does not matter if new entities are added at the end of the list rather than being inserted, then an increment of 1 is the obvious choice. If it is expected that there will be many new entities to be coded after the initial set and if it is important that they be inserted into their proper place in the code set rather than merely being added to its end, then the largest increment, which allows the most opportunities to do so, should be chosen.

It is possible to compute directly the largest constant increment that will disperse the entities to be coded evenly and most widely over the available serial numbers. This can be done by techniques of modular arithmetic which are indicated briefly below. Consider the following example: Given a 4-digit numeric code format and the constraint that only codes 1000 thru 9950 should be used, with 600 items on the initial list to be coded, what increment should be used to disperse them evenly across the available mapping space (code set)? The number of available codes are 8951, computed as:

```
 9950
-1000
 8950
+   1
 8951
```

The increment then can be computed as the integer portion of the quotient 8951/600 = 14.92, or 14. Using this increment the codes initially assigned would be 1000, 1014, 1028, 1042, 1056, etc.

If it seemed reasonable, a lower increment could be chosen instead, such as 12 or, very useful, 10. If an integer increment between 5 and 10 were computed it might seem reasonable to chose 5 as an increment instead. A computed increment of 23 would suggest actual use of the value 20. In other words, whatever increment is computed, choice of the next lower integral multiple of 10 for decimal numeric codes might make assignment of codes easier and should produce codes that might seem more reasonable to users who are not aware of these code-assignment processes.

The next level of sophistication in assigning sequential codes would involve varying the increment between codes assigned to entities on the initial list according to an expected need for varying numbers of spaces for new items, depending upon some characteristics of the items. For example, if it were necessary to assign sequential codes to persons' surnames, as in a list of employees, so that new persons' names could be inserted in their proper place, it would appear desirable to use larger increments between the initial Browns, Joneses, Smiths, etc., for example, than between the Zbniewski's and Zenders - unless most of your employees are Polish and German. So, in order to apply such a technique successfully, it is necessary to have accurate statistical measures of the distribution of the entities to be coded. Computer programs have been written to apply such techniques to the structuring and coding of name files. The initial set of codes assigned by such a programmed coding scheme are called predictive codes. However, it has been stated that variations in frequencies-of-occurrence of various names or even initial letters are so great, depending on the source, that initial results were unsatisfactory. It may be possible, with sufficient effort, to apply such techniques successfully to coding keys for name files but I cannot report further regarding this approach at this time.

A classification and coding scheme that has a feature to allow easy insertion and coding of new, not-specifically-foreseen items is the so-called "decimal" classification scheme, as exemplified by the Dewey Decimal System used by libraries to classify books. This is actually a conventional blocked-dependent hierarchical classification scheme, in which the first three digits are used to designate 10 primary categories, 100 secondary categories and 1,000 teritary categories, using codes 000 through 999. After the third digit position in each code a decimal point or period is customarily written as a separator, as illustrated below:

Codes	Entities (Categories)
300.	Sociology
400.	Philology
500.	Natural Science
510.	Mathematics
520.	Astronomy
530.	Physics
531.	Mechanics

The essential feature of this coding scheme is the way in which codes are assigned after the decimal point (after the third digit position). Codes are assigned to the initial entity set (categories, items or the like) serially (sequentially with an

increment of 1) in a continuation of the hierarchical dependent classification scheme, using as many digit positions after the decimal point as are needed to accomodate the number of categories and levels of embeddedness needed to encode the initial list. Thereafter, if it becomes necessary to insert an item into the initial list at any point, it must be considered to be a subdivision of an existing category, and another digit position must be added to the existing category's code to accomodate the new item's code. This is illustrated by the addition of "Pulley, Compound" on the last line of the table below.

Codes	Entities
530.	Physics
531.	Mechanics
531.1	Machines
531.11	Level and Balance
531.12	Wheel and Axle
531.13	Cord and Catenary
531.14	Pulley
531.141	Pulley, Compound

This example also illustrates that the period or "decimal point" is used to separate the left-hand, fixed, "integer" portion of the code from the right-hand, variable-length, "fractional" part. Thus the name decimal classification system. It is interesting that the decimal point also serves to separate the long numeric code into two, visually-separate subcodes, thus "chunking" it as recommended by modern authorities on human factors aspects of character-string code use.

The "fractional" part of a decimal-type classification code can get very long as several levels of embeddedness of categories are successively indicated, and the resulting long string of digits becomes difficult to read, transcribe (manually), etc. without the introduction of errors. One technique for alleviating this problem is to extend the use of the decimal point to blocking out successive levels of embeddedness, thus continuing the "chunking" of the code chain. Thus, for example, a very special type of pulley might be coded as 531.141.1234.1155. Even with chunking, such long codes seem somewhat long for human use in manual or semi-automatic information system.

The decimal method of coding is designed to be used for identifying data in manual information systems where the quantity of items to be coded cannot be limited to any specific anticipated volume. It is particularly well suited for classifying and filing abstracts of written material because it is able to handle an indefinite number of items (lower-level categories) as they are added to any given classification.

The decimal code works well in most of the cases for which it was designed. However, decimal codes are not in general use in modern information systems because they vary in length and the decimal point can be a nuisance. Both aspects of the decimal system can be altered, however, by allocating fixed-length subfields to each "chunk," eliminating the explicit decimal point(s), and left-justifying the code values within each subfield, filling on the right with zeroes, as illustrated below:

Code	Subject
531000	Mechanics
531100	Machines
531110	Level and Balance
531120	Wheel and Axle
531130	Cord and Catenary
531140	Pulley
531141	Pulley, Compound

In this example, the decimal code has been converted to a six-digit fixed-field block classification code. The organization of the decimal code is retained, but the degree of expandability has been limited to ten subdivisions for each machine class.

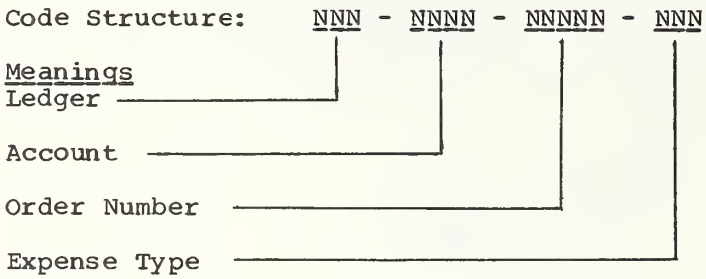
In general we would recommend using a well-planned block sequential code scheme for item or information classification applications. However, if the special advantages of decimal classification code schemes - flexibility, expandability, etc. - are desirable for a specific coding application in a modern information system such a scheme could be implemented with relatively little trouble using the codes left-justified in fixed-length fields, as illustrated above. The fields should be defined long enough, however, to accomodate the longest codes likely to be developed during the life of the information system.

3.2 Group-block Non-dependent Numeric Codes

In the previous section we described essentially only one type of numeric code, serial numbers, but considered them from various points of view and with different names, all of which merely describe various features of such codes.

There is another type of numeric codes, called group-block or non-dependent codes. Their essential difference is that a value in a given digit positicn (or group of adjacent digit positions) represents one, specific, unique meaning independent of the values and their meanings of the higher-order digits in the code (or the lower-order digits, for that matter).

Financial Accounting system codes are typical examples of non-dependent group-block codes. Consider the following illustration:



Under this scheme the code 830-5402-62015-062 would mean Ledger 830, Account 5402, Order No. 62015, and Expense Type 062. (What "Ledger 830" means is another matter.) The point is that "5402" in the 4th through 7th digit positions would mean, for example, "Ace Printing Co." on every ledger from ledger 000 (or, preferably, Ledger 100) through Ledger 999, and "62015" in positions 8 thru 11 means Order No. 62015 (probably a serial number) on every Ledger and no matter what Account Number appears in positions 4 thru 7. Similarily for Expense Type. As a matter of fact, such a code structure is actually a chain of four independent codes, as described in Section 1.4. We have called it a code rather than a code chain because it is conventional to do so, even though the term "code chain" is technically more accurate.

The separation (connection) of the four codes by hyphens also illustrates a few important points.

a. The essential code (code chain) structure is 15 decimal digits. The separators (hyphens) do not represent essential information for data input, internal storage, processing or automatic (mechanized) data transmission, and thus are redundant, unnecessary for those purposes.

b. However, the hyphens are very useful for efficient graphic presentation (output, display) to humans. The hyphens graphically "chunk" the code chain into short strings that can be read and processed (perceived, remembered, etc.) more easily by the people that must somewhere in the information system use the data that the machines are moving. Such chunking seems to be virtually necessary for

reducing errors in human code handling, as described further in Section 6.2.

c. The 4 components codes themselves (Ledger, Account, Order Number, and Expense Type) are probably each some version of the serial-number codes described in Section 3.1. However, the individual codes can be other than serial-number codes, as discussed later in this Section.

Another example of a group-block non-dependent code is that presented in Western Electric Data Standard No. 10178, Rating Defect Code. These 6-digit codes are actually chains made up of two catenated (chained) 3-digit codes, "Characteristic or Location" and "Defect Classification." There are 221 codes of the first type, numbered sequentially 003 thru 663, with an increment of 3, and 147 codes of the second type, numbered 003 through 441, also with an increment of 3. Typical codes are as follows:

Characteristic or Location		Defect Classification	
Code	Meaning	Code	Meaning
003	Adaptor	003	Above Maximum
006	Alignment	006	Base
009	Angle	009	Below Minimum
012	Apparatus	012	Bent
(etc.)		(etc.)	
660	Wire Wrap	438	Wrinkles(ed)
663	Wrap(s)(ed)	441	Wrong

Thus, it is seen that the meanings or entity lists were arranged alphabetically - rather than according to some other classification scheme - before being coded with increment-of-3 sequence numbers.

Six-digit codes (code chains) would be generated by looking up the words to be coded in the two alphabetical lists, selecting one 3-digit code from each list, and writing them (or perhaps, directly keying them into an input or recording device) in the proper sequence. Thus, in order to encode a "Bent Apparatus" one would look up "Apparatus" in the first list and find it coded as 012, then look up "Bent" in the second list and find it coded as 012 (a coincidence), then record or enter the code 012012.

Since these 6-digit codes each report or measure two aspects of an information item, the "Characteristic or Location" of a defect and the "Defect Classification," such a code could be called a 2-dimensional code and a graphic scheme could be devised such as 2-dimensional matrix (an array) to aid in generating the 6-digit codes. However, for this code scheme a 221 by 147 array would

result, and since many of the resultant 6-digit codes would be meaningless (e.g., code 375147 would mean "Nylon Illegible" and code 033408 would mean "Bottom Too Much") there would not be much point to constructing a graphic array for this code scheme. However, a matrix presentation could reasonably be applied to the coding schemes described in the next few paragraphs, which would then be called matrix codes.

A two-dimensional non-dependent group block code that could be illustrated as as a matrix code is the "Experience Category" code used in the Corporate Personnel System and documented by Western Electric Data Standard No. 10042. This is a 2-digit code in which the first digit designates Bell System Experience by a "1" and Non-Bell System Experience by a "2." The second digit can take values 1 thru 6 as follows:

Value	Meaning
1	Non-Professional Engineering Experience
2	Professional Engineering Experience
3	Non-Professional Information System Experience
4	Professional Information System Experience
5	Non-Professional Accounting Experience
6	Professional Accounting Experience

This code scheme can also be presented in matrix form, as illustrated by the next table.

	Experience					
	Engineering		Information System		Accounting	
	Non-Prof. (1)	Prof. (2)	Non-Prof. (3)	Prof. (4)	Non-Prof. (5)	Prof. (6)
Bell System (1)	11	12	13	14	15	16
Non-Bell (2)	21	22	23	24	25	26

The same information could have been coded more explicitly but less concisely by a 3-digit non-dependent block code scheme of the format ABC, where:

digit A designates Bell System (=1) or non-Bell (=2)
digit B designates Non-Professional (=1) or Professional (=2)
digit C designates Engineering (=1), Information Systems (=2), and Accounting (=3).

The resulting 3-dimensional codes can be illustrated by a matrix on 2-dimensional paper as follows:

		Experience					
		Non-Professional (=1)			Professional (=2)		
		Eng. (1)	Info. Sys. (2)	Acct'g. (3)	Eng. (1)	Info. Sys. (2)	Acct'g. (3)
Bell System	(1)	111	112	113	121	122	123
Non-Bell	(2)	211	212	213	221	222	223

The above two coding schemes illustrate among other things that group-block non-dependent codes exhibit the characteristic of positional significance. That is, a given digit value has significance (meaning) dependent upon its position in the digit string, not upon the values of other digits in the string.

The two coding schemes illustrated above also demonstrate that essentially the same information can be coded in two or three dimensions, and that conciseness is traded-off vs explicitness or decoding simplicity. Only a code user and/or system designer can decide what is more important in a given case. If conciseness seems most important - and it is our opinion that conciseness is not most important; reduction of errors via ease of understanding, coding and decoding is more important - then the twelve categories (2 x 2 x 3 = 12, whatever way they are coded) could have been coded more concisely by a 1-character alphanumeric code set: 1, 2, 3, ..., 9, A, B, C. However, such alphanumeric code sets are not recommended for reasons explained later in Section 5.0.

Another group-block code is the "Character of Service" code presented in Western Electric Data Standard No. 10018. Used by the Corporate Personnel System (CPS), this code or code chain consists of 4 decimal digits, in the format or sequence ABCD, where

digit A designates an evaluation of work,
digit B designates an evaluation of ability,
digit C designates an evaluation of conduct, and
digit D designates an evaluation of attendance;
and each digit position can take one of four values,
where a 1 denotes "outstanding"
a 2 denotes "very good,"
a 3 denotes "satisfactory," and
a 4 denotes "unsatisfactory."

Thus, 4 x 4 or 16 possible codes can be produced by this scheme. Under this coding scheme the code 1321 means outstanding work, satisfactory ability, very good conduct and outstanding attendance.

It is apparent that these factors were judged to be related but independent of one another, and that it was deemed to be appropriate that they be coded together via a non-dependent group-block code.

Since this coding scheme reports measurements of four attributes it is a 4-dimensional group-block code and could be illustrated by a 4-dimensional matrix. As for the 3-dimensional matrix illustrated before, this beccmes complicated to do on 2-dimensional paper, so will not be done here. Also, per the previous coding scheme, these 16 values could have been designated by more concise, though less readily decoded, code sets, as follows: 2-digit codes 01 through 16 or 1-character alphabetic codes A through R (excluding I and O).

Matrix codes can be considered to be a simple type of algorithmic codes, i.e., codes generated by mechanistic application of a set of rules to a set or string of input by a person or a machine. Other algorithmic codes or coding schemes are considered later in this document, in Section 6.

A special case of group-block numeric codes is that of telephone numbers. Complete telephone numbers consist of ten decimal numeric digits: 3 for Numbering Plan Area (NPA) code ("area code"), 3 for Central Office Code (COC) (exchange) and 4 for Telephone Line Number. (Reference Western Electric Data Standards 10070 through 10074.) The NPA code and the COC code are assigned not sequentially, randomly or mnemonically, but rather dependent upon electrical systems engineering (switching) considerations. The 4-digit line numbers can be considered to be sequential numeric

codes. The essential complete telephone number, then, consists of a chain of ten digits, and can be recorded, input, stored, etc. in that form. However, for human factors reasons it is common to record, etc., and especially to output, display, or present the code chain graphically in the format of three codes separated from (and connected to) one another. For human information-processing factors reasons we recommend that such codes be presented in "chunks" separated by the recommended separator, the hyphen, as illustrated by the following examples.

NPA-COC-LINE

201-468-6000
212-555-1212

Closing our remarks on group-block non-dependent numeric codes we wish to point out that each group or block in such a code chain can be not only a sequential number but also could be a random number or a mnemonic numeric code. The latter two numeric code types are described in Section 3.3 and 3.4, respectively. In section 3.3 we point out that there does not appear to be any reason for using random numeric codes in Western Electric information systems, and here we wish to point out that there seems to be even less reason for using random numbers in chains. Therefore, we recommend that random numeric codes not be used in code chains in Western Electric information systems.

Regarding the use of mnemonic numeric codes in chains (group blocks), we point out in Section 3.4 that mnemonic numeric codes can be useful, and we observe here that therefore they can be useful chained in groups. However, we do recommend that if mnemonic numeric codes are used in chains that the individual codes be separated graphically (visually) by means of the recommended graphic separator, the hyphen. Thus, for example, the mnemonic telephone number 212-555-1212 (written thusly, not as (212) KL 5-1212) or the mnemonic dimensional analysis code chain 36-24-36.

3.3 Random Numeric Codes

Truly random numbers are numbers that are generated and made available (either as a prepared list or as needed, in real time) in a way such that they satisfy certain mathematical statistical tests of randomness. Techniques for generating random numbers are outside the scope of this document. However, we wish to mention that truly random (and thus non-sequential) code sets for various entity sets can be generated by algorithms such as sequential entry into a certified table of random numbers.

There does not seem at this writing to be any reason for using random numbers as identifiers (identifying codes) in information

537-860 O - 74 - 10

systems. For a numeric code of any length (any number of digits) there are as many random numbers as serial numbers. (In fact they are the same numbers, merely listed in different sequences.) Reasonable care in choice of code length, blocking, and assignment of sequential codes, with provision for handling "overflow," should suffice to provide an adequate sequential coding scheme for any application.

The techniques known as "hashing," which produce pseudo-random codes and which are used in compiling computer source-language programs into machine-executable modules may be applicable to generation of information system codes as defined in Section 1.2 but such applications are beyond the scope of the first edition of this document.

Computer programs have also been devised that will scan words or phrases, i.e., character strings representing natural-language words and phrases, and on the basis of the words detected assign a pre-determined numeric code to each phrase. This scheme is deterministic rather than stochastic; its purpose is to map into a specific code all possible phrases that have the meaning assigned to that code, whereas the purpose of "hashing" is to generate a unique, different code for every non-identical character string encountered.

3.4 Mnemonic Numeric Codes

In Webster's Third New International Dictionary, perhaps the most generally accepted published American authority on the meanings of words, "mnemonic" is defined as: "1. assisting or intended to assist memory; 2. of or relating to memory." Thus even a string tied around one's finger can be called "a mnemonic" (if it works). But, confining our discussion to codes as defined in Section 1.2, let us consider mnemonic codes, in particular mnemonic numeric codes.

The term "mnemonic" is usually applied only to alphabetic codes that resemble in some way the natural-language words they represent. Thus, abbreviations, acronyms, etc. are called "mnemonics." But in a larger sense any code can be considered to be menemonic if in some way it helps a person to remember the code when presented with the need to remember it; or to remember the natural-language word or the like for which the code stands, when presented with the code; or if in some way the code itself is easier to learn by rote and thus is easier to remember than some other code that would otherwise be as useful or meaningful.

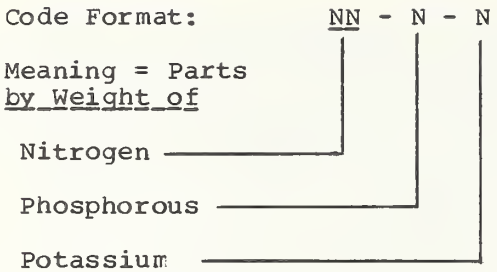
Thus we offer the opinion that numeric codes <u>can be</u> mnemonic and that some numeric codes <u>are</u> mnemonic.

For example, consider telephone numbers. They are codes and some are easier to remember than others. (For the purpose of this specific discussion we shall discuss all-numeric telephone numbers only.)) The 3-digit "area codes"or NPA codes do not seem to be especially mnemonic, considered as a code set, but the individual area codes 202, 212, 303, 404, 505, 515, and 707 might be easier to remember than other because of their symmetry, a mnemonic property. Individual telephone exchange codes or Central Office Codes (COC), also 3 digits long, might also be somewhat mnemonic because of symmetry, or because of sequence (e.g., 123, 234, 456, etc.) or because 2 of the 3 digits are identical (e.g., exchanges 299, 399, 233) or because all three digits are identical (e.g., exchange 555). The four-digit line or extension numbers can manifest such properties as sequence (e.g., 1234, 2368), half sequence (e.g., 1122, 5566) mixed sequence (e.g., 1212, 8989, 2434) repeated digits (e.g., 9900, 9989), being multiples of 1000 (2000, 5000) or of 100 (e.g., 9900,2300) or of 50 (e.g., 9950, 2750), or where digits or digit pairs appear to be multiples of one another (e.g., 1020, 1224, 2040, 4488, etc.)or where the 4 digits can be considered to represent memorable years (e.g., 1776, 1949, 1972, 1984).

Complete local telephone numbers (exchange+line) evidently can thus be designed to be mnemonic. Consider for example the "Universal Directory Assistance Number," 555-1212, or the "New York City Report" (for traffic and transportation conditions) number, 999-1234, or a typical first number for a PBX: 222-2345. It has been stated that the fire-police-ambulance emergency telephone number, 911, used in cities such as New York, Washington, Seattle, Omaha and Denver, "was selected after computer tests found that it was an easy number to remember." We can reasonably infer then that certain similar numbers are also easy to remember, namely:

Number (Code)	Meaning
411	Directory Assistance
611	Repair Service
811	Business Office
911	Emergency

Another way in which numeric codes can be made somewhat mnemonic to make the digit groups "significant," i.e., by giving them quantitative weight. This is demonstrated by group-block non-dependent codes which encode quantitative information. For example, consider the garden fertilizer analysis code whose format is illustrated by the next table.



Thus the code 10-3-2 would indicate a fertilizer mix containing 10 parts Nitrogen, 3 parts Phosphorous and 2 parts Potassium by weight. This type of code is a descriptive code rather than an identification code but apparently is adequate for its intended use.

Another illustration would be the alphanumeric partial code set for automobile tires illustrated below.

Code	Meaning
T67013B1	Tube-type 6.70 x 13 Blackwall, First-line
T69013B1	Tube-type 6.90 x 13 Blackwall, First-line
T71013B1	Tube-type 7.10 x 13 Blackwall, First-line
N73515W2	Tubeless* 7.35 x 15 Whitewall, Second-line
N77515W2	Tubeless* 7.75 x 15 Whitewall, Second-line
N82515W2	Tubeless* 8.25 x 15 Whitewall, Second-line

(*N=) "No tube" = tubeless.

The above code chains mix alphabetic mnemonic codes with quantitative-mnemonic numeric codes, and thus are not pure numeric codes, but do, we hope, illustrate the notion that in an adequate context numeric codes can be mnemonic.

Summarizing our comments on mnemonic numeric codes, we believe we have demonstrated that numeric codes can have mnemonic qualities and that some numeric codes <u>are</u> mnemonic. To use mnemonic numeric codes may be an adequate solution to a coding problem, but we make the general recommendation that this course not be taken unless there is a constraint (as with telephone numbers) that the codes must be numeric. If the mnemonic requirement is sufficiently important then alphabetic mnemonic codes (Section 4.4) are demonstrably superior; if a logical classification scheme, to which codes can be added or inserted, is required, then a numeric or blocked alphanumeric coding scheme is recommended.

4.0 Alphabetic Codes

Alphabetic codes are codes whose characters are taken from the twenty-six standard English-language upper-case letters A through Z.

4.1 Sequential Alphabetic Codes

If the assumption is accepted that "alphabetic order"(A,B,C,...,Z) is as firm and meaningful as numerical order (0,1,2,...,9), we can say that the 26 English-alphabet letters form their own symbol system, a hexavigesimal symbol system. If for human-factors (error-reduction) reasons the decision is made not to use the symbols 0 (oh) and I (eye) then the remaining 24 symbols form a quadravigesimal symbol system.

It is also necessary to agree on a convention for the weighting or order-of-significance scheme for alphabetic sequence or serial "numbers" of length longer than one letter, and the obvious choice is to imitate the customary scheme for decimal numbers, i.e., high-order towards low-order sequence is from left to right. Thus AABCD is analogous to 11234 and is 1 unit larger than AABCC.

Parenthetically, it is requested that readers do not confuse this sequential alphabetic code scheme with "Roman Numerals," which use the letters M, D, C, L, X, V and I in a scheme that will not be explained in this document. We do not recommend the use of "Roman Numerals" in Western Electric Information Systems.

Assuming the above conventions are accepted then the various characteristics of decimal numeric sequential codes apply also to hexavigesimal or quadravigesimal alphabetic sequential codes, including weight or positional significance and therefore, dependence, the collating property, dependent blocking, hierarchical classification, and the exponential property. Therefore, the alphabetical-order sequence of a 4-letter alphabetic sequential code scheme would be as illustrated in the next table.

A Sequential Alphabetical Code Set

AAAA
AAAB
AAAC
(etc.)
AAAZ
AABA
AABB
(etc.)
AABZ
AACA
AACB
(etc.)
ZZZY
ZZZZ

Although we recommend in Section 3.11 that the numeric code zero ("all zeroes," e.g., 0 or 00 or 000, etc.) or numeric codes beginning with a zero (e.g., 03215) should not be used, we recommend here that the alphabetical serial codes A or AA or AAA, etc., and other codes beginning with A such as AABCD be used in their proper sequence where applicable. There is no generally accepted convention that "all A's" is the zero value for alphabetic codes and it would be artifical, arbitrary and undoubtedly unsuccessful to attempt to impose such a convention. Instead we recommend general acceptance of the data-processing convention that "all blanks" be considered the zero-value for fields designed to be populated by alphabetic (or alphanumeric) codes.

Since in EDP machines letters are actually encoded in binary numeric collating codes, computer programs can successfully use alphabetic sequential codes of great lengths. By encoding 26 values per character position (and 26^2 per 2 character positions, etc.) they can provide many more values in a given code length than decimal numeric codes.

However, there are several reasons why sequential alphabetical codes are not widely used in information systems. One is technical and scientifically measurable. The reason is that the "information load" of alphabetic codes is much greater than that of numeric codes of the same length, and that if the information load of a code exceeds a certain value the human error rate in handling such codes begins to increase rapidly, almost exponentially.

Information load and its consequences are discussed at length in Section 6.0 of this document. Suffice it to say here that the information load of pure alphabetic codes is 4.70 per character position, or 18.8 for a 4-letter code and 23.5 for a 5-letter code. The information load for a pure numeric code is 3.32 per digit or only 19.92 for a 6-digit code. Experiments have shown that the error rates in human recording and transcribing of codes begins to increase when the information load of the codes exceeds the value 20.

Therefore, alphabetic sequence codes are not recommended for use in manual or EDP information systems except where their use by humans is limited as follows:

a. transcribing (rewriting, keying from a copy, etc.) of alpha codes no longer than 5 letters;

b. encoding (looking up the code, given the entity) or decoding (looking up the entity, given the code) of sequential alphabetic codes no longer than 4 letters in a table in which the codes are listed sequentially and the entity list is arranged alphabetically or according to a logical classification scheme.

Another reason is economic, though not so readily quantifiable. In information systems implemented using modern EDP equipment it is questionable whether any machine storage space or processing time would actually be saved by use of alphabetic sequential classification codes. Numeric codes can be "packed" or converted to binary or floating point representations if saving of storage space is especially important, and data-processing programs (storage, searching, retrieval, sorting, merging, etc.) seem to be principally, if not exclusively, oriented towards numeric identification and classification codes rather than alphabetic codes.

Finally, alphabetic code schemes for codes longer than one letter allow the production of codes that are identical to or reminiscent of natural-language words or phrases. This feature is exploited in the design of mnemonic alphabetic codes, as discussed in Section 4.4. However, as a consequence of this feature, sequentially-assigned alphabetic serial codes can inadvertently manifest pseudo-mnemonic characteristics that are unintended, unwanted and perhaps distracting to the code-using task. For example, four-letter alphabetic sequential codes, expecially if somehow automatically assigned without a screening procedure, can produce four-letter natural-language words - of all kinds.

For all the above reasons it is recommended that sequentially-assigned alphabetic codes not be used in information systems where

they are to be used by humans. For the same reasons the use of randomly-assigned alphabetic codes is not recommended. However, the use of carefully-designed mnemonic alphabetic codes is recommended, for certain applications, as discussed in Section 4.4.

4.2 Group-block Non-dependent Alphabetic Codes

All the properties of group-block non-dependent decimal numeric codes or code chains apply to alphabetic codes. For example, the "Experience Category" code and "Character of Service" group block codes discussed in Section 3 could have been coded using alphabetic sequential codes. Thus, for "Character of Service,"the 4-character code could have been described as having the format 1234, where

letter 1 designates an evaluation of work
letter 2 designates an evaluation of ability
letter 3 designates an evaluation of conduct, and
letter 4 designates an evaluation of attendance,

and the letter A designates "outstanding,"
the letter B designates "very good,"
the letter C designates "satisfactory,"and
the letter D designates "unsatisfactory."

Thus, the code ACBA would indicate outstanding work, satisfactory ability, very good conduct and outstanding attendance.

Other similar, but more complicated group-block alphabetic codes schemes could be illustrated here but will not be, because there does not appear to be any significant advantage to use of sequential alphabetical codes in non-dependent group-block combinations. As stated in section 4.1, sequential numerical codes can solve virtaully any problem to which sequential alphabetical codes might be applied.

If two or more mnemonic alphabetic codes (discussed in Section 4.4) are strung together into a code chain, with or without separators, such a chain could also be considered to be a group-block non-dependent code chain. In fact, if random alphabetic codes (section 4.3) were strung into a chain the result could be called a group-block code. However, we do not recommend the use of random alphabetic codes in any combination. Chaining of mnemonic alphabetic codes will be discussed in Section 4.4.

4.3 Random Alphabetic Codes

Random alphabetic codes of any length could be generated by algorithms similar to or derived from those by which random numeric codes are generated. (See Section 3.3.) Random alphabetic codes

share the disadvantages of sequential alphabetic codes (large information load per character, possible inadvertent generation of objectionable "words"), but not their advantage (sequence). Truly random alphabetic codes, like truly random numeric codes, might be useful in cryptographic applications, but that is outside the scope of this document.

As for random numeric codes, there does not seem to be any reason for using random alphabetic codes in information systems and so we do not recommend their use in Western Electric information systems.

4.4 Mnemonic Alphabetic Codes

This category of codes is as important and significant to information systems design and operation as sequential numeric codes. Various studies have ascertained that use of mnemonic codes can reduce human error rates significantly (by 50% or more) in operations such as recognizing, recording, transcribing, etc. coded information. Consequently, many authorities recommend the specification and design of "human-oriented" mnemonic alphabetic codes above all other types for information systems where the codes must be used by humans.

In general we share this approval of alphabetic mnemonic codes but we observe that the determination of what codes are truly mnemonic, and to what degree, in any specific using environment is not necessarily a simple matter and so we advise that the specification and design of mnemonic alphabetic codes for each information system be approached with some background knowledge and a determination to evaluate objectively and thoroughly the codes' designs and actual codes produced in view of the needs of the specific information system.

In Section 3.4, Mnemonic Numeric Codes, we stated that "mnemonic" means "assisting or intended to assist memory." Human memory characteristics and capabilities have been and continue to be studies by means of techniques of experimental psychology, and human (mental) processing of character-string codes have been the subject of some of this study. For the purposes of this document it can be stated that human memory can be divided into short-term or immediate memory and long-term or permanent memory. Immediate memory is the memory we use, for example, when we remember a telephone number or other code for a few seconds while we write it down or dial it, and then forget it. Permanent memory is the memory we use when we remember a code or anything else for longer than a few seconds, for example for several minutes up to a lifetime. There is some evidence also for the existence of a "medium-term" memory which stores for a few minutes information that has passed through the short-term memory and which, if it is

properly selected and reinforced, can pass into the permanent memory. For the purposes of this document we can consider such a medium-term memory to be an input channel or stage of permanent memory.

The mental mechanisms or techniques that we use, subconsciously or deliberately, to aid us in remembering new information include organization and familiarity. Organization includes such techniques as grouping message elements or ideas into physically- or logically- related groups or into a logical sequence, including alphabetical order. Familiarity includes deliberate memorization, by rote or by some associative technique, and also includes that non-deliberate memorization that seems to come about naturally when we encounter the same idea, word, number, code, or sequence of numerals or of letters, repeatedly.

Familiarity, therefore, enhances learning, and for the purposes of the present document we infer that if codes can be organized so that they can easily become familiar they can be learned more accurately than might otherwise be the case and thus, it seems obvious, can be used more accurately (with lower error rates) than might otherwise happen, with consequent economic and social benefits - that is, with lower fiscal costs and less of the aggravation that is caused by errors in information system data.

The effects of organization, familiarity and learning on immdiate memory seem to be small in comparison with their potential benefits to permanent memory. Nevertheless, code schemes and specific codes intended to be used in short-term memory situations (human code transmission tasks, such as recording, table look-up and transcription, etc.) can be designed to take advantage of such menmonic properties as symmetry, sequence, etc. For example, the specific codes of a 2-letter, not-otherwise-mnemonic, alphabetic code set could be assigned so that the codes AA, BB, CC, DD, etc., and perhaps natural-language digrams such as TH, CH, SH, OU, EA, etc., were assigned to the more important meanings or those likely to be encountered with greater frequency then others.

Similarly, for numeric codes the properties of uniformity, symmetry, sequence, etc. could be exploited, as discussed in more detail in Section 3.4, Mnemonic Numeric Codes. The form of organization known as "chunking" also has a significant beneficial effect on human transmission of codes. Chunking is discussed at length in Section 6.21 of this document.

The benefits of organization and familiarity are demonstrated much more noticeably, however, when these factors are applied to development of codes for tasks involving long-term or "permanent" memory, that is, code-using tasks which require or can demonstrably benefit from learning of the codes in a code set. The most obvious

example of this is natural language itself. Although for the purposes of this document we defined a code as any substitute for a natural-language word or phrase, psychologists and linguists define natural language itself as a system of codes whose use is described by syntax. The benefits of organization, familiarity and learning with regards to the codes of natural language (words) seem obvious: we must learn at least a basic set of a natural language's words and their meanings (semantics) and acquire an instinctive understanding of how to use them (the rules of syntax) if we are to communicate with our fellows. Similarly, if the synthetic language that is comprised by a code set and the rules for how to use it are made easy to understand and to learn then we can expect, other factors being equal (motivation, working conditions, etc.), that the code set will be used more effectively (more accurately, faster, with lower error rate, etc.) than might otherwise be the case. We hope that this seems reasonable to the reader, for it is merely a simply-stated summary of the conclusions drawn by experts who have carefully investigated the use of mnemonic codes versus non-mnemonic codes in code-using tasks involving learning and who have found that alphabetic mnemonic codes are remarkably superior for such tasks.

Alphabetic codes which are intended to be "mnemonic" may actually be more or less mnemonic for any individual human code user depending on whether he can easily learn (and thus remember) them and whether he has already implicitly learned them as part of his education and other life experience. Many mnemonic alphabetic codes are abbreviations (such as RCD for "received," M for "male," Y for "yes," MR for "Mister," etc.) and others are acronyms (such as FBI, USAF, PBA, RADAR, etc.) with which many - but not all - adult Americans may be more or less familiar prior to and independently of their being encountered as part of a code set in a designed information system. That is to say, the learning of mnemonic codes such as more-or-less commonly used abbreviations or acronyms may have taken place before a specific information system is designed to use such codes. This may be helpful or harmful to the code-using tasks. It can be beneficial if what has already been learned can be incorporated into the requirements of the information system - that is, if the information system can be designed to take advantage of the codes the users already know. On the other hand if the information system designers choose or design codes that in some way conflict with the established learning of the system's users (operators) then that established learning will probably have a harmful rather than a helpful effect on the operation of the information system. For example, if one-letter codes for the polarity of electric battery terminals are being chosen (and if the symbols + and - cannot be used) an electrochemical engineer, nominally the system's "user," might choose A for "anode" and C for "cathode," which might be contrary to the experience and learning of the system's actual operators,

who might better understand, remember and correctly use the codes P for "positive" and N for "negative." Similarly, a military officer controlling the design of an information system to be used by civilian employees might insist upon use of the codes A for "affirmative" and N for "negative" in place of Y for "yes" and N for "no." Use of the code A instead of Y in this code set would please the officer in charge but might also result in a poorer-performing information system.

There has been extensive research into the factors that make alphabetic codes mnemonic, such as their inclusion of specific single letters and groups of two, three, etc. letters, their pronouncebility, and other factors, all of which can be summarized as follows: The more that alphabetic codes resemble natural-language words, the easier they are to remember correctly. Much of the above may seem to the reader so obvious as to be trivial. However, it is interesting that scientific research into these matters has produced essentially the same conclusions that "common sense" might have, and that is somehow reassuring. Unfortunately, however, it has been demonstrated that the converse is not always true. That is, "common-sense" code design, as illustrated by hundreds of code sets reviewed by the Corporate Data Standards Department as part of our task of producing Western Electric Data Standards, does not consistently result in codes and code sets that are at once concise and as truly mnemonic as they might be. If anything has been demonstrated by this experience it is that one man's "common sense" does not necessarily agree with another's and that "mnemonic" alphabetic codes selected or developed using the guidance of only "common sense" are not in all cases satisfactory according to sound, methodical principles of code design. Therefore, we cite in the following portions of this section some principles regarding the development of mnemonic alphabetic codes that should assist code set developers in the generation of truly menmonic alphabetic codes.

Virtually all mnemonic alphabetic codes are abbreviations of one kind or another. They include:

a. abbreviations in the traditional sense, including:

   1. contractions - shortened representations for natural-language words, formed by eliminating some of the letters of the word, often the vowels and sometimes in accordance with a rule or algorithm, for example, "MR" for "mister," "LTD" for "limited," etc.;

   2. truncations - shortened representations formed by eliminating letters from the end of a word, such as "INC" for incorporated, "BL" for "blue," "NO" for "north," etc.;

3. Codes - such as "NO" for "number," "LB" for "pound," etc.

b. acronyms - formed by selecting the initial letter or letters from a phrase (group of words), e.g., "RSVP" from "respondez s'il vous plait," RADAR from "RAdio Detection And Ranging," LORAN from "LOng RAnge Navigation," etc.;

c. Mnemonic alphabetic codes designed specifically as such for use in information systems. They may be contractions, truncations or may even use entirely different letters than the words for which they stand (though this last case is unlikely).

Traditional abbreviations and acronyms typically are of varying lengths (numbers of letters), from one letter (M for Monsieur) on up to as many as twenty-two letters (ADCOMSUBORDCOMPHIESPAC for "Administrative Command, Amphibious Forces, Pacific Fleet, Subordinate Command"). Specifically-designed mnemonic alphabetic code sets typically have this common characteristic: all the codes are of the same length, the same number of letters.

If it is desired to use traditional abbreviations or acronyms as codes for information systems then we recommend that system designers consult authoritative references for lists of abbreviations and acronyms, for example, such commercially-published dictionaries as Webster's Third New International Dictionary (G. & C. Merriam Co.) or the Acronyms and Initialisms Dictionary, Third Edition (Gale Research Company).

We in the Bell System are fortunate to have available in a group of Bell System Practices a massive collection of common abbreviations and acronyms. BSP 790-100-100, "Standard Abbreviations and Letter Symbols Master List," is a descriptive guide to the complete series of BSP's on abbreviations and should be consulted by anyone who is interested. Closer to home, the Western Electric Data Element Standards manual, CI 95.186, produced by the Information Systems Data Standards Organization includes numerous alphabetic code sets and lists of acronyms. The Data Element Standards manual should be consulted by every system designer.

For those who prefer to generate their own sets of fixed-length abbreviations - i.e., mnemonic alphabetic codes - the procedures should be useful that are described in BSP 790-100-110, "Creating Abbreviations Guidelines." The procedures described in BSP 790-100-110 describe rules for eliminating vowels (first) and consonants (second) from the graphic (printed or written) representations of English-language words in order to shorten them

to form abbreviations (codes) of any desired length. The resulting abbreviations usually consist principally, if not exclusively, of consonants, and if a reasonably large proportion of the original letters remain the codes are usually mnemonic when perceived visually. However, since alphabetic codes (pseudo-words) consisting principally of consonants are not usually easily pronounced audibly like words by English-speaking persons, such consonant-type abbreviations may not be very mnemonic if they must be transmitted by auditory means (by being spoken and heard). Thus it seems that the "consonants" technique for generating codes goes contrary to recent opinions regarding what makes alphabetic codes mnemonic, as described in earlier paragraphs of this section, which were summarized by the simplicism, "the more that alphabetic codes resemble natural-language words the easier they are to remember correctly." Natural-language words in English do contain vowels and thus alphabetic codes (including abbreviations) that are to resemble natural-language words should contain vowels. The experimental findings regarding this phenomenon are extensive and varied. Letter patterns such as CV, CVC, VCC, CVCC, and CCVC (where C = a consonant and V = a vowel) seem to be more successful (more easily remembered correctly) than others, and usage of naturally-occurring digrams such as QU (instead of UQ), TH, CH, SH, CK, etc., where possible, also seems to help. Two remarkably successful mnemonic alphabetic code sets which were carefully designed according to the principles described above are presented in the following tables.

U.S.A.F. Maintenance Data Reporting System
Experimental Mnemonic How-Malfunctioned Codes

	Malfunction Descriptions	Mnemonic Codes	Codes Replaced
	Mechanical		
01	Travel Incorrect	TRI	599
02	Torque Incorrect	TOI	167
03	Tension Incorrect	TEI	664
04	Punctured	PUN	540
05	Pressure Incorrect	PRI	525
	Electrical		
06	Current Incorrect	CUR	029
07	Voltage Incorrect	VOL	169
08	Insulation Breakdown	INS	350
09	Fuse Blown	FUS	472
10	Impedance Incorrect	IMP	816
	General		
11	Improper Handling	IMH	086
12	Launch Damage	LAD	158
13	Lost in Flight	LIF	386
14	Secondary Failure	SEF	602
15	Battle Damage	BAD	731
	Component		
16	Engine Removed	ENR	142
17	Compressor Damaged	COD	380
18	Turbine Damaged	TUD	486
19	Tire Defective	TID	782
20	Bearing Failure	BEF	953

BISP Common Language Code 65E, Trouble-Pair and Binding Post

Descriptive Word or Phrase	Code
Capacity Unbalance	CUB
Cross	CRS
Crosstalk	CTK
Echo	ECH
Ground	GRD
Grounded Tip	GTP
Grounded Ring	GRG
Grounded Tip and Ring	GTR
Induction	IND
Low Insulation	LIN
Noisy	NSY
Open	OPN
Open Tip	OTP
Open Ring	ORG
Resistance Unbalance	RUB
Short	SHT
Signalling	SIG
Transmission	TMS
Transpose	TNS
Unbalance	UBL
Universal Bad Pair	UBP

However a would-be mnemonic alphabetic code set is chosen or generated, we caution that system designers should not automatically and immediately assume that the code set is as fully mnemonic as it might be. They should consult with ISE's Corporate Data Standards department, who are charged with the responsibility of being knowledgable concerning code sets and their usage, and should consider arranging tests of the codes' performance in actual usage.

5.0 Alphanumeric Codes

Alphanumeric codes are character-string codes made up of the alphabetic characters A through Z and the numeric characters 0 through 99. In some situations it may be specified that alphabetic characters not be used that can easily be confused with numerals. Thus, typically the letters I and O are excluded, and in some cases Z, B, G, S, U, V, Q, J and even H are excluded to avoid their being mistaken for numerals or for other letters. (See Section 6.2, Visual and Auditory Perceptual Factors.)

Alphanumeric codes can be categorized as follows:

a. Codes in which some character positions can be populated only with alphabetic characters and other positions can be populated only with numerals. For example, codes of the format AANNN, where A = any alphabetic character and N = any numeric character;

b. Codes in which character positions can be populated by alphabetic and/or numeric characters. For example, codes of the structure XXXX, where X = any alphabetic or numeric character;

c. Codes of format (structure) consisting of mixtures of categories 1 and 2. For example:

Code Structure (Format)	Examples
XXNNNA	12345D, AB123B
AAXXNN	AB1234, ABCD12

It should be apparent that to define a code set as alphanumeric implies that the consequent maximum possible code set can include an all-alphabetic subset and/or an all-numeric subset. The simplest alphanumeric code format, "X," includes the subsets A-Z (all alphabetic) and 0-9 (all numeric).

A code set of the format XXX includes subsets of the following structures:

537-860 O - 74 - 11

AAA (all-alphabetic), (for example LZY), and
NNN (all-numeric) (for example 543)

as well as 25 other combinations of the symbols X, A and N.

Similarly, a code set of the format XXNNNX includes subsets of the formats:

NNNNNN (all-numeric), e.g., 123456
AANNNA, e.g., IZ345G
etc.

Before going any further into descriptions of alphanumeric codes and their applications we must state here our recommendation that alphanumeric codes of any format should not be specified for use in information systems unless there exist factors or considerations that outweigh the higher error rates and consequent higher costs that can be expected due to use of alphanumeric codes instead of numeric codes or mnemonic alphabetic codes. This recommendation is based on published results of experimental investigations that measured and compared error rates experienced in use of various types of codes in typical code-using tasks.

For table look-up of 3-character codes (codes 3 characters long) it was found that the lowest error rates by far were experienced when the looked-up codes were all-numeric (NNN) in format. In one set of experiments the average error rate for the six possible alphanumeric code formats was 8.5 times the error rate for simple all-numeric codes when the look-up table was indexed by sequential numeric codes, and was greater by a factor of approximately 2.8 when the look-up table was indexed by sequentially-ordered alphabetic codes.

It was found that the error rates increased as more character positions became alphabetic (e.g., from NNN to ANN to AAN to AAA) and it was found that the principal types of errors were substitution of a number for a letter or a letter for a number. Other, less obvious effects were noted, which are mentioned elsewhere in this document, and it is apparent also that the results might be interpreted and perhaps explained in terms of their information load (see Section 6.21 of this document), but the point to be made is that short alphanumeric codes tend to cause higher error rates in a simple code-using task than numeric codes of equal length, and thus we recommend that alphanumeric codes be avoided insofar as possible.

Despite our blanket recommendation above we must also describe and comment upon possible and typical uses of alphanumeric codes or code chains:

First, alphabetic prefixes to numeric codes are sometimes used for blocking, that is, for dividing the resulting code set into high-level blocks or categories, because it is believed that the alphabetic prefixes make the blocking more obvious. We wish to comment that although the resultant blocking may be slightly more obvious (e.g., perhaps codes AA1234 and AC1234 may appear more obviously to be in different blocks than codes 111234 and 131234), the potential resulting higher error rate in human use of such mixed codes is good reason to avoid alphabetic-prefix blocking of sequential numeric code sets.

Second, alphabetic prefixes to numeric codes are sometimes used for blocking because alphabetic prefixes allow a larger number of codes (and thus, blocks) in a given number of character positions. Thus in a code set of the format AANNN the prefix AA could be used to designate 26 x 26 or 676 high-level categories, or 15 x 15 = 225 categories, if I, O, Z, B, G, Q, SS, J, U, V, and H are excluded, as they should be. However, the numeric prefix of format NN can be used to designate 90 (10 thru 99) categories and the prefix NNN, only one digit longer, can be used to designate 900 (100 thru 999) categories, if so many are needed. Therefore, we recommend that numeric prefixes be used, not alphabetic prefixes, for blocking sequential numeric codes, in order to avoid the potentially higher error rates that can be expected in use of alphanumeric codes.

Third, alphabetic suffixes to numeric codes are sometimes used to indicate insertions of a new code or subdivision of an existing code. For example, if a toll highway's interchanges are numbered 1, 2, 3... and a new interchange is built between interchanges 8 and 9 it may be better (in terms of costs and ease of understanding in a total system context) to designate the new interchange as 8A rather than to renumber all the interchanges. Similarly, if a spur is built onto such a turnpike from its exit 14 it may be better to designate the spur's exits as 14A, 14B, 14C, etc. than to renumber all the exits or to number the new exits 19, 20, 21, etc. Similarly, if the highway is so altered that a second exit 16 is built several miles west of the original exit 16 it may seem best to designate the new exit as exit 16W and the original exit 16 as exit 16E - or perhaps exits 16A and 16B would have been better. (The examples refer to the New Jersey Turnpike.) For such unforseen developments alphabetic suffixes to numeric codes serve admirably. However, for well-planned information systems we recommend that insertion or subdivision of sequential numeric codes be provided-for in initial design of sequential numeric code formats, by using increments such as 10, 100, etc. between initially-assigned codes, so as to allow with no problems all the insertions and subdivisions of codes that can reasonably be foreseen.

Another type of "alphanumeric code"that might be considered is a code chain in which alphabetic characters appear in the midst of numerals, for example, codes of the format NNAANN. It is unlikely that such codes would be any type of sequential code; it is more likely that they would be non-dependent group-block codes or code chains. We do not recommend the use of such codes; we recommend that sequential numeric codes or mnemonic alphabetic codes be used wherever possible. However, if such codes must for some good reason be used, we recommend that even though the code chains might be input to and stored and transmitted within an EDP system in contiguous formats such as that illustrated above, they should be recorded and presented graphically with the recommended graphic separator, the hyphen, separating (connecting)the code elements, in the following format (for example): NN-AA-NN. Such a presentation, by "chunking" the code chain (See Section 6.21) helps to overcome the potential error-causing effects mentioned earlier in this section.

6.0 Special Considerations

In the following subsections several topics are discussed that pertain to all types of character-string codes.

6.1 Algorithmic Codes

In general, an algorithm can be defined as a set of rules or procedures describing how to solve a problem or how to execute a logical or mathematical operation. Since the term algorithm came to be used to denote a computer-program algorithm it may have acquire some more specific definitions, but it is not our intention to discuss them here. Rather, we wish to point out that in a general sense virtually all codes are algorithmic codes insofar as they are substitutes for natural-language words, etc. and have been generated or derived in some way in accordance with some set of rules or procedures.

For the purposes of the present section, however, we shall consider only algorithmic codes in a more narrow sense, meaning codes that are derived by performing some thoroughly-defined logical and/or mathematical operations on a character-string representation of the word, phrase, etc. to be coded. In this sense the numeric hash codes generated by hashing are algorithmic codes, and many abbreviations, acronyms and the like, as discussed in Section 4.4 are also algorithmic codes. The matrix codes discussed in Section 3.2 can also be called algorithmic codes, as use of matrix representations to display code elements and to assemble them into chains can be considered to be an algorithmic process.

In this section we wish to describe two more types of algorithmic codes, letter-value codes and self-checking codes, that can be useful for information systems applications.

6.11 Letter-value codes

Letter-value codes are formed by assigning specific, predetermined, numeric values to individual alphabetic letters and to other, specifically-designated letter combinations according to a set of rules, so as to derive numeric codes from alphabetic character strings. Such algorithmic coding schemes are useful, for example, for providing blocking categories and sorting and retrieval "keys" for alphabetically-ordered entity lists such as name lists. Various types of letter-value coding schemes can be devised to serve specific purposes. Any scheme's effectiveness ought to be tested on a representative sample of the entity list to be coded. In a typical letter-value code scheme letters are assigned numeric values after, for example, they have been arranged in sequence ordered according to their frequency of occurrence in typical English text, or after they have been grouped according to their phonetic characteristics, etc.

A typical letter-value coding scheme is described as follows: Each derived code is a chain, 4 characters long, consisting of the first letter of the name being coded followed by a 3-digit numeric code. A letter-value code is derived by applying the rules below to the name.

a. The first letter is "saved," then all letters including the first are operated upon according to the following set of rules.

b. Names with internal spaces or other punctuation are packed to eliminate all but the 26 upper-case letters. Thus D'Arcy becomes DARCY, Scott-Smith becomes SCOTTSMITH, and Van Der Laurens becomes VANDERLAURENS.

c. Vowels (A,E,I,O,U and Y) in positions other than the first character of the name are given a value of zero, i.e., are "dropped."

d. Consonants H and W are given a value of zero.

e. The other consonants are given values in accordance with the table below:

Value	Letter	Phonetic Category
0	A,E,I,O,U,Y,H,W	Vowels, etc.
1	B,F,P,V	Labials
2	C,G,J,K,Q,S,X,Z	Gutterals and sibilants
3	D,T	Dentals
4	L	Long liquid
5	M,N	Nasals
6	R	Short liquid

Note that consonants that sound alike are given the same numeric value.

f. Two or more adjacent letters having the same numeric value according to the table are coded as though they were ne letter. Thus LL = L = 4, SC = 2, MN = S, CKS = 2, WHAY = 0, etc.

g. The code is complete when the end of the name has been reached or 3 digits have been generated.

Examples of codes produced by application of the above rules are as follows:

Rugge → RGG → RG → R200

Scott → SCTT → SCT → S300
Stock → STCK → S320

Bookkeeper → BKKR → BKR → B260

Anderson → ANDRSN → A536

This system overcomes the problem of variations in name spellings and provides a "blocking factor" for organizing a name file. It is useful for coding related items of limited number, since synonyms (duplicate code values for different items) will occur fairly soon. This system permits a rough sort of the data items in alphabetical sequence. Code values can be derived from tables in a computer program for so-called "automatic" coding.

The above rules were extensively re-written to make them appear more algorithmic than they seemed in their original form but it is apparent that they would have to be even more rigorously stated to be considered a mechanistic algorithm suitable for implementation as a computer program.

6.12 Self-checking Codes

The most frequently-occurring types of errors in human use of codes, i.e., recording, transcribing, etc. of character-string codes, are:

a. Inversion, or transposition of adjacent digits (e.g., 623 becomes 263 or 632, etc.);

b. double transposition errors (e.g., 1234 becomes 1432);

c. transition errors, including duplication of adjacent digits (e.g., 623 becomes 622 or 123, etc.);

d. random errors, (e.g., 1234 becomes 2283, etc.).

An effective technique for detecting such errors is the use of self-checking or error-detecting codes. A self-checking code is one that includes a check character as part of the character string. The check character is generated by operating on a root code according to a set of rules (an algorithm). The root code can be any type of code heretofore described but usually is a serial or blocked sequential numeric code. The self-checking code formed by appending the check character (check digit) to the right end of the root code has the property that when another algorithm is applied to the self-checking code the result inidicates whether the code has been transcribed, etc., perfectly or whether an error has been introduced. The action to be taken when an error has been detected in a self-checking code is a matter of system design.

It is also possible to contrive error-correcting codes, which are of especial usefulness in the design of high-volume, high-speed automatic data transmission systems, but they are outside the scope of this document.

Common self-checking codes are the "Modulus 10," "Modulus 11," and "Double Add 10" types. They are described elsewhere and so the calculation of check characters will not be illustrated here. Self-checking codes have advantages and disadvantages. The obvious advantage is that they will detect errors - not all errors, but most errors. In particular, codes generated by the Modulus 10 method will detect:

100% of all transition errors,
97.8% of single transposition errors,
90% of random errors, and
no double transposition errors.

The disadvantages of error-detecting codes include the costs of: calculating them initially (and correctly, therefore very

carefully); inputting, storing, transmitting, etc. the extra character; and applying the checking algorithm to test for errors. There is some opinion that the costs of self-checking codes outweigh their benefits in most business data-processing applications. The decision whether to use error-detecting (self-checking) codes in a business data-processing system should be made based upon an economic analysis of the costs, not merely on the desire for technical perfection. One Western Electric Company code scheme that uses the Modulus 10 check digit scheme is the Comcode Identification System.

6.2 Visual and Auditory Perceptual Factors

The following sections discuss some factors that affect the accuracy with which codes are visually and audibly perceived.

6.21 Information Load

The concepts and techniques of the modern quantitative discipline of information theory have been applied to problems of measuring human information-processing capabilities (including recognition and use of character-string codes) as part of the discipline of experimental psychology. The concept that the information content of a message can be measured quantitatively has been used to formulate predictions of the difficulty of human information-processing tasks and to estimate the quantitative informational value of accomplished tasks.

Information measurement techniques can be useful to designers of information systems that involve human information-processing tasks in several ways:

a. Code formats can be designed and specific codes developed that should be efficient within a total information system design context including consideration of human-factors aspects of information processing as well as computer system factors, resulting in information acquisition and processing with lower error rates and thus greater value at lower cost than might otherwise be the case.

b. Error rates and consequent costs likely to be experienced in use of developed code sets can be at least roughly predicted during early phases of information system design and development and if the predictions indicate trouble ahead then possible alternatives can be explored in good time.

c. The discipline of estimating in a concrete way <u>via</u> scientific code design the amount and variety of information that an information system is expected to

acquire, process, store and deliver should be a positive influential factor in total good system design.

The basic concept of information theory is that information can be measured quantitatively at its most elementary level in terms of dichotomy, the dividing into two alternatives or dyads represented by the binary digits or "bits" 0 (zero) and 1 (one). An informational situation that has two possible values (such as "yes" and "no" or "on" and "off") can be represented or coded by the two mutually-exclusive binary digits 0 and 1, each of which occupy one bit position. An information situation having four possible values can be coded in two bit positions and a set of 8 values can be coded in 3 bit positions, as illustrated in the tables below.

Four-value situation

Binary Code	Value
00	1
01	2
10	3
11	4

Eight-value situation

Binary Code	Value
000	1
001	2
010	3
011	4
100	5
101	6
110	7
111	8

The number of bits required to code sets of values can be presented, then, in another way as follows:

Number of values	bits needed for binary encoding
2	1
4	2
8	3
16	4
32	5
64	6
(etc.)	(etc.)

It is seen, therefore, that the number of bits, n, needed to encode a number of values, V, can be indicated as the power to which the binary system base, 2, must be raised:

$$2^n = V.$$

The inverse of the above exponential formula is the logarithmic formula

$$\log_2 V = n.$$

That is, the exact number of bits, n, needed to encode a number of values, V, is given in decimal-number notation by the logarithm to the base 2 of V.

Measurement of the informational content of value sets, including messages, via logarithms to the base 2, therefore, is a basic technique of information theory, and has been applied to character-string codes as described below.

The potential information content, measured in bits, of a character position in a character-string code that can be populated equi-probably by any of N possible codes is given by $\log_2 N$. Such a value, called entropy in basic quantitative information theory, is also called the "information load" of the character position, or the "character load." Hereafter, in this document we shall refer to this measure merely as "information load" and we shall assume equiprobability of occurrence for all characters of a defined character set.

For a character position that can be populated equiprobably by the ten Arabic decimal numerals 0 (zero) through 9 the information load of that character position is quantified as $\log_2 10 = 3.32193$ or 3.322. For a single "alphabetic" character position, which can be populated equiprobably (let us assume) by the 26 English-language alphabetic letters, the information load would be $\log_2 26 = 4.70044$ or 4.70. Information loads per character position for various alphabets (character sets) are given by the next table.

Information Loads for Some Character Sets

Usual Name	Members	Number of Members	Information Load per Character position
Numeric	0 thru 9	10	3.322
Alphabetic	A thru Z	26	4.70
Alphanumeric	A thru Z and 0 thru 9	36	5.17
	A thru Z except I and O	24	4.585
	A thru Z and 0 thru 9, except I and O	34	5.09
	A thru Z except I,O,Z,B,G,S, U,V,Y,J and H	15	3.91
	A thru Z and 0 thru 9, except I,O,Z,B,G,S U,V,Y,J and H	25	4.64

The information load of a character-string code format - that is, a code format or structure defined as a string of characters, such as four numeric digits (NNNN) or three alphabetic letters (AAA) or two alpha concatenated with three numeric (AANNN) - is defined as the sum of the information loads of the individual character positions in the code. Thus, for example, the information load of a character-string code of format 4 alphabetic characters would be 4.70 + 4.70 + 4.70 + 4.70 = 18.80 bits, and the information load of a code of format 6 decimal digits would be 6 x 3.322 = 19.932 or 19.9 bits.

The information load of a code of format XXNNNA, where A = any letter A thru Z, N = any digit 0 thru 9, and X = any letter and/or digit, would be 2 x 5.17 + 3 x 3.322 + 4.70 = 25.006 or 25.0.

What is the practical significance and usefulness of computing information loads for character-string code formats, as

above? It has been found in studies of human transmission of coded information (writing down or copying codes, tasks requiring use of short-term memory for unfamiliar character strings) that the error rate is negligible up to a certain point and then begins to increase rapidly, almost exponentially. The point differs for alphabetic codes and for numeric codes and also differs depending upon whether the codes are perceived aurally (by hearing them) or visually (by reading them). (The subjects of the studies were adult employees of the Dutch postal system, ranging from workers to the Director.)

Considering only whether the codes are alphabetic or numeric, it was found that the error rate was negligible for codes consisting of up to 5 letters or 6 digits. Considered in terms of information load it was found that information loss, as demonstrated by a sharply increasing error rate, was practically zero for codes with information load up to the value 20, and then began to increase significantly for codes whose load exceeded 20, except for alphabetic codes presented visually, for which the error rate became significant for loads over 23.5, i.e., for alphabetic codes longer than 5 letters. For mixed letter-digit codes the effects are similar, though because of the many possible formats and consequent effects of position (where in the code the letters appear with respect to the digits), it is impossible to present a complete range of quantitative values here. The important point to be derived from these studies is that the information load measurement technique, is a simple and concrete means of quantitatively measuring the potential difficulty of human "manipulation" of character-string codes and consequently offers a means of predicting whether a given code format is likely to experience a significant error rate when being used by humans in information systems.

The experiments that produced the results described above have been imitated using different subjects (American college students) and the results were similar though not identical. The error threshhold was found to be about 16 bits for numerical codes and for alphabetic codes perceived aurally (by hearing) and about 19 bits for alphabetic codes presented visually. These figures indicate that 5-digit numeric codes and 4-letter alphabetic codes were the longest codes that could be handled by American young adults consistently with negligible error rates.

It is possible, of course, to question whether the results reported above are of universal validity, but unless we are willing to conduct our own, "better" research, it would seem reasonable to accept the findings described above as being reasonably indicative of human capabilities for such simple code-using tasks and to draw conclusions regarding recommendations for these coding guidelines.

We recommend, therefore, for "simple" tasks such as reading (or hearing) and then writing, dialing, keying, etc., that pure alphabetic codes no longer than 4 lettters or pure numeric codes no longer than 5 digits be used.

Other important factors regarding alphabetic codes are discussed in Section 4.0 and we call the reader's attention in particular to Section 4.4, Mnemonic Alphabetic Codes.

6.22 "Chunking"

For situations where it is necessary to use numeric codes longer than 5 digits we recommend that they be "chunked" or broken into groups of 3 and/or 4 digits. Studies of human transmission (immediate-memory tasks) of numeric codes have repeatedly found that:

a. people naturally tend to break long strings of digits (equivalent to numeric codes, in our terminology) into groups of 3 or 4 digits during simple code-using tasks, and

b. when long numeric codes are presented either aurally or visually in similar simple code-using tasks the lowest error rates and/or fastest performance with equivalent error rates were achieved when the codes were presented in groups of three or four digits.

The psychological mechanism by which long numeric codes are instrinctively broken into groups by users has been called "chunking" by one of the most eminent researchers in the field and the term "chunking" has been applied by extension to the deliberate, planned breaking into groups of long numeric codes.

Typical long numeric codes that are "chunked" in their usual graphic presentations include telephone numbers, Social Security Numbers, some department store charge account numbers, some gasoline company and airline credit-card numbers, etc. Typical long numeric codes that are not usually "chunked" include some other department store, gasoline company and airline credit card numbers, Western Electric COMCODE numbers (9 digits - e.g., 101424653), Western Electric department numbers (6 numeric digits immediately following 2 digits and 2 letters - e.g., 33HF112710), and Western Electric Transportation Commodity Code (10 digits immediately following one letter - e.g., D2899946004). The difficulty in dealing with such codes and consequent probably-greater error rate has been recognized implicitly in two of the Western Electric Company code formats described immediately above; in both COMCODE and Transportation Commodity Code the last (right-most) digit is a Modulus-10 check digit, which serve as an error-

detection aid when the codes are analyzed by an error-detection algorithm. (Refer to Section 6.12, Self-Checking Codes.)

It is a matter of opinion which would have to be investigated and tested analytically whether it would be "better" to help prevent errors in dealing with such codes by "chunking" the root code than to present and require humans to deal with such long codes unchunked and rely upon a check digit merely to detect errors. It is my opinion and recommendation that if only one of the two approaches is to be used, then it should be "chunking." If 100% code integrity is required at any cost then both chunking and a check digit could be used. To use 9, 10, and 11-digit codes unchunked but with check digits seems analogous to deliberately designing and using an unsafe automobile and then paying 10% or 20% extra for bolt-on safety devices: unnecessarily expensive.

The following table illustrates "chunking" recommendations for graphic presentations (input and output formats) for character-string codes.

Number of Characters	Chunked Formats(s)
1	C
2	CC
3	CCC
4	CCCC
5	CC-CCC
6	CCC-CCC
7	CCC-CCCC
8	CCCC-CCCC or CC-CCC-CCC
9	CCC-CCC-CCC
10	CCC-CCC-CCCC
11	CC-CCC-CCC-CCC
12	CCC-CCC-CCC-CCC or CCCC-CCCC-CCCC
(etc.)	(etc.)

6.23 Special Graphic Characters in Codes

Special graphic characters were defined in Section 1.3 as all printable characters in USASCII (United States of America Standard Code for Information Interchange) or EBCDIC (Extended Binary Coded Decimal Interchange Code) except the twenty-six upper-case English-alphabet letters A through Z and the ten Arabic numerals 0 through 9. Various versions of USASCII and EBCDIC have from twenty-six to thirty-four such special graphic characters, including various punctuation symbols and other symbols.

We recommend that special graphic symbols (except the hyphen) not be used in character-string codes, for the following reasons:

a. Considered as graphic (visually-perceived) symbols, some are difficult to see (e.g., the period, comma, apostrophe). Others can easily be confused with one another or with letters or numerals. For example, the "slash" (virgule) can be and often is confused visually with the letter I, the numeral 1, the exclamation point !, the simple vertical line used to symbolize "or" in EBCDIC, and perhaps the left or right parenthesis, bracket or brace. The "equals sign" (=) can be confused with the colon (:) and vice-versa, and the underline (_) if used anywhere except actually under another symbol can be mistaken for a hyphen (minus sign) (-) or dash (-).

b. For aural transmission (by hearing) there appears to be disagreement, or at least a lack of general agreement, on the names and even the pronunciation of agreed-upon names of some graphic symbols. For example the symbol /, used ad nauseam by programmers and others associated with computer systems, is properly named the virgule, is called a solidus by others, but is called simply a "slash" or a "slant" by many others. "Slash" can be and has been mistaken for "dash" in aural perception, but we doubt that a very large percentage of our information systems personnel would immediately recognize the words "virgule" or "solidus" and write or key the appropriate symbol. Similarly, some people know the symbol # only as "number sign," others as "hashmark," and others as "pound sign" - though some would write the symbol L when they hear the phrase "pcund sign."

c. Considered in terms of "Information load" per character position (see Section 6.21), adding as many as 34 special graphic symbols to the 36 symbols of the alpha-numeric character set means that the alpha-numeric-special character set denoted by the symbol X (see Section 1.3) will consist of as many as 72 characters, resulting in an information load per character of 6.17 bits. Such a load means that a code defined to have the format XXNNNX, for example, has an information load of 26.32, dangerously high in terms of potential error rate, whereas if the code format AANNNN could effectively serve the same purposes, its lower information load of 22.46, a 15% reduction, suggests a lower potential error rate.

We except the hyphen (-) from our blanket recommendation against the use of special graphic characters in character-string codes. We recommend that hyphens be used to separate (connect) the

parts of "chunked" codes or code chains, for the reasons detailed below.

a. Codes longer than 5 characters should be "chunked," i.e., broken up into smaller segments or "chunks," as explained in Section 6.21. Other « group-block codes, as described in Sections 3.2 and 4.2, "naturally" are organized as small groups (strings) of characters placed adjacent to one another.

b. It is desirable to indicate both aspects of such chunks' or segments' relationship to one another - their separateness and their connectedness - and it seems equally obvious that it is desirable to indicate their separateness and their connectedness simultaneously in the same way (visually by the same graphic symbol).

c. In our natural language (English) the standard symbol used for the purpose described above - to simultaneously connect and separate parts of a word - is the hyphen. This graphic symbol has the visual advantages of being visible enough (more so than the period) but not too visible (not as obtrusive as the asterisk (*), for example) and not as easily confused with other symbols as is the virgule ("slash") (/). The hyphen serves the same purpose as a single blank space in separating parts of a code or code chain but serves, as the blank space cannot, to connect the parts visually (graphically).

Therefore, we recommend that in graphic presentation or representations of codes or code chains longer than 5 characters the code's characters be organized into groups of no less than 3 and no more than 4 characters and that the groups be connected (separated) by hyphens.

6.24 Positional Effects

This topic, in a sense, belongs in the discussion of alphanumeric codes (Section 5.0) because the available information pertains specifically to alphanumeric codes. However, the findings have implications for alpha-numeric-special codes also and so we present this discussion in the context of "special perceptual factors."

First, let us summarize the findings and repeat the recommendation of Section 5.0: alphanumeric codes experience greater error rates in human use than do all-numeric codes of the same length or equivalent mnemonic alphabetic codes. Therefore, we recommend that alphanumeric codes not be used in information systems unless there are other, overriding reasons for using them.

However, if it is necessary to use alphanumeric codes then the information presented in the following paragraphs should be considered.

It has been found in studies of human use of short (3-character) alphanumeric, non-mnemonic codes that the error-rate tends to be higher for codes whose middle character is of a different type than the two outer characters (i.e., for codes of format ANA or NAN versus those of type AAN, ANN, NNA or NAA). More specifically, it was found for codes consisting of 2 letters and one numeral that codes of format ANA had higher error rates than codes of formats NAA or AAN, and for codes of 2 numerals and one letter the format NAN had higher error rates than the formats ANN or NNA. Codes of the format ANA had the highest error rate and all-numeric codes (NNN) had the lowest error rates. We have not found any "proven" simple explanation for the above phenomena, but it appears that in simple, "semi-automatic," immediate-memory code-processing tasks humans experience the phenomenon known as "perceptual set" by which our unconscious or subconscious expectation for the next character is "set" to a specific type such as alphabetic or numeric, and this perceptual set tends to influence us to respond in a predisposed and perhaps erroneous manner to the visual stimuli received during a character-to-character visual scan of a character-string code.

In any case the recommendations derived from the information summarized above are as follows: for simple tasks (those involving only short-term memory and not able to benefit from learning of the codes), if it is necessary to use alphanumeric codes then codes should not be designed that intersperse letters and numerals, and in particular codes that surround a single numeral with letters should be avoided.

Another set of experiments tested alphanumeric codes, all consisting of four letters and two numerals, to determine in what positions the numerals should be placed in order to enhance immediate recall (short-term memory). It had earlier been determined that in alphabetic or numeric character-string codes ranging in length from 5 to 10 characters the penultimate and antepenultimate (2nd and 3rd from last) characters experienced the greatest frequency of errors in human processing. Furthermore, as already described in several places in the present document, it has been established that for short-term memory tasks short numeric codes are easier to remember than alphabetic codes of the same length. Interaction and hypothesised counter-action of the two phenomena were investigated by testing short-term recall of various combinations of 4 letters and 2 numerals. It was found that the two effects did interact and that when the easier-to-remember numerals were placed in positions 4 and 5 of the 6-character codes the resulting code formats were "easiest to remember" in the sense

537-860 O - 74 - 12

that they experienced the lowest error rate by far of the five possible code formats having 2 adjacent numerals. This "best" (lowest) error rate, however, was still 48.8%; the worst error rate was 64.5%. From these findings we derive two recommendations:

a. Do not use alphanumeric codes.

b. If you must use alphanumeric codes then put the numerals in the penultimate and antepenultimate character positions, i.e., use the format AAA...NNA.

6.25 Visual Recognition of Graphic Characters

Elsewhere in this document we have mentioned some of the problems of visual and auditory recognition of graphic characters and symbols that are used in character-string codes, such as the problem of distinguishing visually between I and 1 or between O (oh) and 0 (zero). In this section we discuss these problems more thoroughly and offer some recommendations.

When we consider the problem of visual recognition of graphic characters we must admit that for some people in some circumstances and using some character sets there seems to be no serious problem; they do not make errors. However, many workers in common code-using situations, trying to identify some characters in commonly-used type faces or handwritten or hand-printed characters, do make errors. So the abilities, training, and motivation of the workers, and their working conditions, equipment and materials used, etc., in code-using tasks are factors which may have greater or lesser effects on accurate perception, transcription, keying, etc. Furthermore, the shapes of the characters in different type faces or fonts, as well as the legibility of handwritten or printed characters, also are factors which probably have effects on human code-using accuracy. Therefore, the material presented below must be interpreted with an awareness of the factors mentioned above and their possible consequences.

Examples of letters, numerals and special graphic characters that can be mistaken for one another are given below.

The letter I, the numeral 1 and the virgule, /.
The letters 0 and Q (and sometimes D) and the numeral 0 (zero).
The letter Z and the numeral 2.
The letter G and the numeral 6.
The letter B and the numerals 8 and 3.
The letter S and the numeral 5.
The letter J and the numeral 5.
The letter H and the numeral 4.
The letters U and V.
The letters Y and V.

The letters Y and X.

Letters and numerals have been rank-ordered (separately) according to their increasing potential for being confused with other characters. If avoidance of such error is of overriding importance in a code-assignment problem such as assigning single-character codes to a short list of items, then the codes should be assigned in the order in which they are listed in the next table. (Adjacent to the lower characters, in parentheses, are the other characters with which it is believed they are most likely to be confused.)

Recommended Order Of Assignment

Alphabetic Characters	Numeric Characters
R	3
W	7
M	9
N	8 (B)
F	6 (G)
L	5 (S,J)
T	4 (H)
A	2 (Z)
C	1 (I,/)
E	0 (O,Q,D)
P	
H (4)	
D (O,Q)	
B (8)	
K (X)	
Q (O,0,9)	
V (U,Y)	
J (5,S)	
S (5,J)	
X (Y)	
Y (V,X)	
G (6)	
Z (2)	
U (V)	
I (1,/)	
O (0,Q,D)	

Letters and numerals have also been rank-ordered according to their graphic complexity. There is some opinion that under poor conditions, such as use of dimly-illuminated cathode-ray tube displays, there would probably be more errors made in perceiving the more complex characters. In such situations it is judged better to use the less-complex, more-easily recognized characters insofar as possible.

The following rankings, listed in order of increasing complexity, for alphabetic and numeric characters, has been derived for such use. (Underlining indicates characters judged to be of approximately equal discriminability.)

Alphabetic:

L A J Z T U E P S M V N F W R D C X I K Y B O G H Q

m p d b c u y v w h n z q k g r x j o s f e i t a

Numeric:

1 2 0 7 3 5 6 9 4 8

We recommend that in information system design the type faces or fonts actually to be used be examined by the system designers in order to determine the appearance of the characters and the likelihood of their being confused with one another. The information presented in the earlier paragraphs of this section can be used to guide this process rather than as absolute rules. If it becomes apparent that some characters could be mistaken for one another then perhaps different type faces could be chosen or the system's codes should be designed tc avoid use of easily-confused codes. In any case the letters I and O should not be used, and if letter codes are to be hand-printed at any point then the letters Q and Z certainly should be avoided and the letters B, G, S, U, V, and Y should also be avoided if at all possible.

Further useful information and recommendations regarding this topic can be found in the proposed American National Standard "Character Set for Handprinting," BSR X3.45, available from: CBEMA, 1828 L Street, N.W., Washington, D.C. 20036.

6.26 Auditory Recognition of Letters and Numerals

During design of information systems, in the operation of which codes will be transmitted by auditory means (*via* being spoken by one person and heard by another person), attention should be directed to the phonetic characteristics of the codes. Specifically, some letters and numerals can be mistaken for one another when they are spoken and heard, because they sound like one another even under "normal" conditions. In noisy ambient environments or during transmission over noisy, weak or unclear voice communications channels, or when spoken by people with poor or extraordinary pronunciation, or heard by people with poor hearing, these factors are exaggerated, so that errors can be made in hearing and recording codes.

The following table is a list of letters and numerals that are commonly mistaken for one another in auditory communication. (Please note specifically that this list refers only to the pronunciations of the *names* of the symbols, not the *sounds* that the symbols represent in natural language. Thus, for example, by "U" we mean the sound "you," not the sounds "oo" or "uh.")

Letters and Numerals Often Confused
in Auditory Communication

M and N,
B and V,
P and T,
B and P,
D and T,
C and G,
A and K,
A and H,
K and J,
A and J,
S and X,
Q and U,
I and Y,
C and Z,
B, C, D, E, G, P, T, V and Z (under severely degraded conditions),
the numerals 5 and 9,
and in alphanumeric codes,
the letters A, H, J, and K and the numeral 8,
the letters I and Y and the numerals 5 and 9.

We recommend, therefore, that in design and development of codes that will be exchanged via human hearing that the conflicts indicated above be avoided insofar as possible. This, it seems, would be best accomplished by avoiding alphabetic codes entirely and avoiding use of the digit 9 in numeric codes. If alphabetic codes are to be used then they should be restricted insofar as possible to use of the letters A, B, F, L, M, O, Q, R, W, X, and Y.

Some of the auditory problems discussed above can be alleviated for numeric codes longer than one digit. If such codes are "chunked"into groups of two (preferably) or three digits they can be spoken as though they were quantities instead of merely code strings of numeric characters. Thus, for example, the code "55" can be transmitted as "fifty-five" instead of as "five, five," and the code "12345678" could be transmitted as "twelve, thirty-four, fifty-six, seventy-eight" instead of as "one, two, three, four, five, six, seven, eight." The code "123456789" would, we believe, naturally be chunked and orally transmitted either in 2-digit groups, as above, or as three-digit groups "one twenty-three, four fifty-six, seven eighty-nine." The phonic redundancy inherent in such "chunked" oral transmission of numeric codes helps improve accurate reception (hearing) and thus helps reduce error rates in auditory transmission of numeric codes.

The techniques described above should prove useful also in transmission of the numeric portions of alphanumeric codes or code chains but such techniques should not be used for auditory transmission of alphabetic codes unless the codes have been specifically designed for such use. Voice transmission of "chunks" of alphabetic codes implies, inevitably, the pronunciation of pseudo-words, some of which may be homonyms or almost-homonyms (e.g., the codes TAT and TAD or GOME and GOAN, etc.). Alphabetic codes are discussed in more detail in Section 4.0, but recommendation in the context of the present section is: "Do not design information systems to require transmission of alphabetic codes by voice!"

We wish, nevertheless, to mention the fact that the problem of auditory transmission of alphabetic codes (including natural-language words) under conditions of reduced intelligibility has been encountered since electrical communications were invented, and in marine and aeronautical communications has led to the development and use of the "international phonetic alphabet," in which individual letters are transmitted by voice as the first letters of internationally agreed-upon words that are relatively easily understood even when spoken and heard by persons of different pronunciations under difficult auditory conditions. For example, the code "ARW" would be transmitted as "Alpha Romeo Whiskey." The complete list of the International Phonetic Alphabet can be found in nautical and aeronautical communications reference publications.

7.0 References

A list of books, reports, articles, etc. compiled during preparation of this document is appended.

Information System Data Coding Guidelines

References

Alden, D. G., et al, "Keyboard Design and Operation: A Review of the Major Issues", Human Factors, August 1972, 275-293.

Aume, N. M. and Topmiller, D. A., "An Evaluation of Experimental How-Malfunctioned Codes," Human Factors, 1970, 261-269.

Bell, J. R., "The Quadratic Quotient Method: A Hash Code Eliminating Secondary Clustering," Comm. ACM, February 1970, 107-109.

Bell, J. R. and Kaman, C. H., "The Linear Quotient Hash Code," Comm. ACM, November 1970, 675-677.

Blankenship, A. B., "Memory Span: A Review of the Literature", Psychological Bulletin, 1938, 1-25.

Bonn, T. H., "A Standard for Computer Networks," IEEE Computer magazine, May-June 1971, 10-14.

Brown, J., "Some Facts of the Decay Theory of Immediate Memory", Quarterly Journal of Experimental Psychology, 1958, 12-21.

Cardozo, B. L. and Leopold, F. F., "Human Code Transmission," Ergonomics, 1963, 133-141.

Carson, J. G. H., "Item Identification and Classification in Management Operating Systems," Production and Inventory Management, June 1971, 23-32.

Cherry, C., On Human Communication, 1966, Cambridge, M.I.T. Press.

Chesson, F. W., "Computers and Cryptology," Datamation, January 1973, 62-81.

Conrad, R., "Experimental Psychology in the Field of Telecommunications" Ergonomics, 1960, 289-295.

Conrad, R., "The Location of Figures in Alpha-numeric Codes," Ergonomics, 1962, 403-406.

Conrad, R. and Hille, B. A., "Memory for Long Telephone Numbers", Post Office Telecommunications Journal, 1957, 37-39.

Conrad, R. and Hull, A. J., "Copying Alpha and Numeric Codes by Hand: An Experimental Study" Journal of Applied Psychology, 1967, 444-448.

Crossman, E. R. F. W., "Information and Serial Order in Human Immediate Memory," in Proceedings of 4th London Symposium on Information Theory, 1961.

Crowley, E. T. and Crowley, R. C., Acronyms and Initialisms Dictionary, Third Edition, 1970, Gale Research Company.

Crowley, E. T., New Acronyms and Initialisms, 1971, Detroit, Gale Research Company.

Dreyfuss, H., Symbol Sourcebook: an Authoritative Guide to International Graphic Symbols, 1972, New York, McGraw - Hill Book Company.

Feistel, H., "Cryptography and Computer Privacy," Scientific American, May, 1973, 15-23.

Field, M. M., et al, Guidelines for Constructing Human Performance - Based Codes, Bell Telephone Laboratories Technical Report, 1971.

Fletcher, J. G., "The Octopus Computer Network", Datamation, April 1973, 58-63.

Frink, W. J., "In Coding It's Structure That Counts," Control Engineering, October 1962.

Giddings, B. J., "Alpha-numerics for Raster Displays", Ergonomics, January 1972, 65-72.

Gilbert, E. N., "A Comparison of Signaling Alphabets," Bell System Technical Journal, May 1952, 504-522.

Gilbert, E. N., "Information Theory After Eighteen Years," Science, April 15, 1966, 320-326.

Goldman, S., Information Theory, 1953, New York, Prentice-Hall.

Gombinski, J. and Hyde, W. F., "Classification and Coding," Graphic Science, March 1968.

Gombinski, J., "Industrial Classification and Coding," Engineering Materials and Design, September 1964.

Hall, R. A., Linguistics and Your Language, 1960, Garden City, Doubleday.

Hamming, R. W., "Error Detecting and Error Correcting Codes," Bell System Technical Journal, April 1950, 147-160.

Harris, D. H., et al, "Wire Sorting Performance with Color and Number Coded Wires," Human Factors, April 1964, 127-131.

Hauck, E. J., "Be Kind to Your Data Codes," Journal of Systems Management, December 1972, 8-12.

Hare, V. C., Systems Analysis: A Diagnostic Approach, 1969, New York, Harcourt, Brace and World.

Heron, A. "Immediate Memory in Dialing Performance with and Without Simple Rehearsal," Quarterly Journal of Experimental Psychology, 1962, 94-103.

Hitt, W. D., "An Evaluation of Five Different Abstract Coding Methods - Experiment IV," Human Factors, July 1961, 120-130.

Hodge, M. H. and Field, M. M., Human Coding Processes, University of Georgia, 1970.

Hull, T. E. and Dobell, A. R., "Random Number Generators," SIAM Review, July 1962.

Jackson, M., "Mnemonics," Datamation, April 1967, 26-28.

Jackson, R. L., "Dial 911' Setup Qualified Success," Gannett News Service, July 1972.

Jones, B. W., Modular Arithmetic, 1964, New York, Blaisdell Publishing Company.

Kahn, D., "Modern Cryptology," Scientific American, July 1966, 38-46.

Kahn, D., The Codebreakers, 1967, New York, Macmillan

Klemmer, E. T., "Grouping of Printed Digits For Manual Entry," Human Factors, 1969, 397-400.

Klemmer, E. T., "Grouping of Printed Digits for Telephone Entry," in Proceedings of Fourth International Conference on Human Factors in Telephony, Munich, 1968.

Klemmer, E. T., "Keyboard Entry," Applied Ergonomics, 1971, 2-6.

Klemmer, E. T., "Numerical Error Checking," Journal of Applied Psychology, 1959, 316-320.

Klemmer, E. T. and Stocker, L. P. "Optimum Grouping of Printed Digits," American Psychological Association Proceedings, 1972, 689-690.

Konz, S. et al, "Human Transmission of Numbers and Letters," The Journal of Industrial Engineering, May 1968, 219-224.

Laden H. N. and Gildersleeve, T. R., System Design for Computer Applications, 1963, New York, Wiley.

L'Insalata, B. B., "COBOL Module for the Generation and Verification of a Check-bit Using the Modulus 10 Method," Western Electric Company, 1971.

Little, J. L., "Some Evolving Conventions and Standards for Character Information Coded in Six, Seven and Eight Bits," U.S. Dept. of Commerce, National Bureau of Standards, 1969.

Little, J. L. and Mooers, C. N., "Standards for user procedures and data formats in automated information systems and networks," 1968, Spring Joint Computer Conference, 89-94.

Mackworth, J. F., "The Effects of Display Time Upon the Recall of Digits," Canadian Journal of Psychology, 1962, 48-55.

Maurer, W. D., "An Improved Hash Code for Scatter Storage," Comm. ACM, January 1968, 35-38.

Mayzner, M. S. and Gabriel, R. F., "Information "Chunking" and Short-Term Retention," Journal of Psychology, 1963, 161-164.

McKinsey, D. L., "Equipment Codes and Etymology: the Common Language Problem," August 14, 1967 (privately published).

Meltzer, H. S. and Ickes, H. F., "Information Interchange Between Dissimilar Systems," Modern Data, April 1971, 56-67.

Miller, G. A., "The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information," The Psychological Review, March 1956, 81-97.

Miller, G. A., The Psychology of Communication, 1967, New York, Basic Books; 1969, Baltimore, Pelican Books.

Miller, G. A. and Nicely, P. E., "An analysis of perceptual confusions among some English consonants," Journal of the Acoustical Society of America, 1955, 338-352.

Morris, R., "Scatter Storage Techniques," Comm. ACM, January 1968, 38-44.

Oberly, H. S., "A Comparison of the Spans of Attention and Memory," American Journal of Psychology, 1928, 295-302.

O'Reagan, R. T., "Computer-Assigned Codes from Verbal Responses," Communications of the ACM , June 1972, 455-459.

Owsowitz, S. and Sweetland, A., Factors Affecting Coding Errors, 1965, Santa Monica, The Rand Corporation

Parsons, H. M., "The Scope of Human Factors in Computer-Based Data Processing Systems," Human Factors, 1970, 165-175.

Peace, D. M. S. and Easterby, R. S., "The Evaluation of User Interaction with Computer-Based Management Information Systems," Human Factors, April 1973, 163-177.

Peterson, W. W., Error-Correcting Codes, 2nd Ed., 1972, Cambridge, M.I.T. Press

Pierce, J. R., "Information Theory," Bell Laboratories Record, February 1968, 47-51.

Pierce, J. R., "Real Science for Real Problems," Analog Science Fiction/Science Fact, 1971, 34-56.

Plath, D. W., "The Readability of Segmented and Conventional Numerals," Human Factors, 1970, 493-497.

Pollack, I., "Assimilation of Sequentially Coded Information," American Journal of Psychology, 1953, 421-435.

Radke, C. E., "The Use of Quadratic Residue Research," Comm. ACM, February 1970, 103-105.

RAND Corporation, A Million Random Digits with 100,000 Normal Deviates, The Free Press, 1955.

Rocke, M. G., "Data Codification Principles and Methods," Caterpillar Tractor Company, 1971.

Rocke, M. G., "The Need for Data Code Control," Datamation, September 1973, 105-108.

Severin, F. T. and Rigby, M. K., "Influence of Digit Grouping on Memory for Long Telephone Numbers," Journal of Applied Psychology, 1963, 117-119.

Shannon, C. E., "Communication Theory of Secrecy Systems," Bell System Technical Journal, October 1949, 656-715.

Shannon, C. E., "Prediction and Entropy of Printed English," Bell System Technical Journal, January 1951, 50-64.

Shannon, C. E., and Weaver, W., The Mathematical Theory of Communication, 1948, Bell System Technical Journal; 1949, Urbana, University of Illinois Press.

Shepherd, W., Shepherd's Glossary of Graphic Signs and Symbols, 1971, New York, Dover Publications.

Shurtleff, D., "Design Problems in Visual Displays, Part 1: Classical Factors in the Legibility of Numerals and Capital Letters" 1966, Mitre Corp. Technical Report 20.

Simpson, G. C., "A Comparison of the Legibility of Three Types of Electronic Digital Displays," Ergonomics, July 1971, 497-507.

Sinkov, A., Elementary Cryptanalysis , 1970, New York, Random House.

Smith, S. L. and Goodwin, N. C., "Alphabetic Data Entry Via the Touch-Tone Pad: A Comment," Human Factors, April 1971, 189-190.

Smith, S. L. and Goodwin, N. C., "Computer-Generated Speech and Man-Computer Interaction," Human Factors, April 1970, 215-224.

Sonntag, L., "Designing Human-Oriented Codes," Bell Laboratories Record, February 1971, 43-49.

Stallcup, K. L., "New Growth in Linguistics Produces Clarity, Confusion and Controversy", The New York Times, January 8, 1973, page 90.

Talbot, J. E., "The Human Side of Data Input," Data Processing Magazine, April 1971, 28-35.

Thorpe, C. E. and Rowland, G. E., "The Effect of "Natural" Grouping of Numerals on Short-Term Memory," Human Factors, 1965, 38-44.

Wicklegrew, W. A., "Size of Rehearsal Group and Short-term Memory," Journal of Applied Psychology, 1964, 413-419.

Withington, F. G., "Cosmetic Programming," Datamation, March, 1970, 91-95.

Woodbury, M. A. and Lipkin, M., "Coding of Medical Histories for Computer Analysis," Comm. ACM, October 1972.

Woodward, R. M., "Proximity and Direction of Arrangement in Numeric Displays," Human Factors, August 1972, 337-343.

Wooldridge, D. E., The Machinery of the Brain, 1963, New York, McGraw-Hill

Woznick, A. M., "Item Identification Standards," Western Electric Company, 1971.

--, "A Standard Labeling Code for Food," Business Week, April 7, 1973, pages 71-73.

--, "Character Set for Handprinting" (proposed American National Standard), 1972, Business Equipment Manufacturers Association.

--, Classification and Coding Techniques to Facilitate Accounting Operations, 1959, New York, National Assocation of Accountants.

--, "Coding List for 1970-1973 Cars," Form TNG-3, Consumers Union, 1973.

--, "Comcode Identification System," Western Electric Company.

--, Common Language Coding Guide, Bell Telephone Laboratories, 1970.

--, "Creating Abbreviations, Guidelines," AT&T Co., October 1968.

--, "Federal Item Identification Guides for Supply Cataloging: Section A, Cataloging Handbook H6-1."

--, "Guidelines for Describing Information Interchange Formats," Federal Information Processing Standards Publication 20, 1972 March 1, US Dept. of Commerce, National Bureau of Standards.

--, Handbook for Data Standardization , U.S. Department of Commerce, National Bureau of Standards, 1970.

--, "Human Factor in Data Codes," National Bureau of Standards Technical News Bulletin, March 1970.

--, International Code of Signals, 1969, U.S. Government Printing Office.

--, Legibility of alphanumeric characters and other symbols: Part I, Permuted title index and bibliography, 1964; Part II, Reference Handbook, 1967, U.S. Dept. of Commerce, National Bureau of Standards.

--, "List of Merchandise Classes," Western Electric Company.

--, Manual for Standard Data Elements, DoD-5000.12M, March 1970, Office of the Assistant Secretary of Defense (Comptroller).

--, National ZIP Code Directory, U. S. Postal Service.

--, "Playing the Nickname Game," Business Week, July 29, 1972, Page 37.

--, "Subsets of the Standard Code for Information Interchange," Federal Information Processing Standards Publication 15, 1971 October 1, U.S. Dept. of Commerce, National Bureau of Standards.

--, System/360 Reference Card, IBM Form GX20-1703.

--, "Systems Codes for Classes of Equipment," Western Electric Company.

--, "USA Standard Code for Information Interchange, USAS X3.4-1968, American National Standards Institute.

--, "Vitamins - All the Way From A to K," Food Facts From Rutgers, July-August 1972, Rutgers University College of Agriculture and Environmental Science.

--, "With License Plates Now at Y, a Plan B, or Rather Z, Is Set," New York Times, Dec. 29, 1972.

Addendum to
Information System Data Coding Guidelines

M. J. Gilligan

Western Electric Company, Inc.
Information Systems Engineering
Newark, New Jersey 07102

This coding guidelines document was written primarily to provide a thorough presentation of useful information and recommendations regarding use of character-string codes in business information systems developed for use by Western Electric Company. The primary use of the document is as a reference for courses in Data Base Design, Data Administration and Information System Analysis and Design at Western Electric's Corporate Education Center. The document is also used for guidance and reference by the Company's Data Element Standards Department and by information system users and developers.

There remains much more to be learned and published regarding the topics treated in these coding guidelines; it is apparent that information systems can be made more effective and more economical through attention to design of "good" codes and through standardization of data element codes for information interchange. I hope that readers of this document will contribute their findings and opinions regarding these topics to the next Symposium.

Questions submitted to me at the Symposium, and my answers, are presented below.

QUESTIONS: How is your work being integrated with Arthur Wright's BISP Common Language System? Are these separate activities or is there a standard Bell approach?

ANSWER: Since the different Bell System Companies provide different products and services and have different corporate structures, they establish their own internal data element standards based on their own needs. However, for data elements exchanged between Bell System Companies (including Operating Telephone Companies) the Bell System Data Interchange Language Standardization Committee described by Art Wright establishes Bell System Data Standards. Participation in this Committee's work enables us to learn new coding principles, as does our participation in ANSI Committee X3L8. Otherwise our work and Art Wright's (BIS) are separate activities within separate companies.

QUESTIONS: Mnemonics often present a problem of overlap, especially in a data element requiring numerous data items. Do you agree? How would this change your statement of always using mnemonics for people?

ANSWER: Selection or generation of mnemonic alphabetic codes can be more difficult if a large number of informational values must be encoded for a given data element. (This was illustrated by Art Wright's slide that showed abbreviations of various lengths for closely-related words such as locate, location, locator, located and locating.) However, I did not state that mnemonic codes should always be used for people. We recommend that alphabetic mnemonic codes be used if possible in human code-using tasks where, due to repetitious use, the codes can be learned (memorized). For human code-using tasks where repetitious use of the codes is not likely to occur or where there are too many codes for them to be learned even by repetitious use, we recommend the use of sequential numeric codes. Whether alpha-mnemonic codes or sequential numeric codes should be chosen can be estimated or can be determined experimentally. The article by Aume and Topmiller and the monograph by Osowitz and Sweetland cited in my bibliography provide more infor-

mation on this subject.

QUESTION: If you recommend alphabetic-mnemonic codes, why do we get digital area codes and telephone numbers?

ANSWER: Our recommendations regarding mnemonic alphabetic codes are intended to be applied to data input and output problems in business information systems. They are not intended to apply to the problem of providing designations for telephone line terminations. In terms of data coding for human use, telephone line designations are a type of item identification code, for which we recommend blocked sequential numeric codes, "chunked" into groups of three or four digits. (See Coding Guidelines sections 1.3, 3.2 and 6.22.) However, I believe that choice of numeric characters to designate telephone lines and the prevalence of all-numeric telephone numbers has been influenced by other factors, as described below.

The telephone switching systems in this and most other countries were designed by electrical engineers to operate using modulus-10 (decimal) switching. Thus each character position in the sequence of switching signals ("telephone number") accepted and used by telephone systems can have ten values. Whether these ten values are represented electrically by voltage or current levels, by electromechanical counters, or by electronic circuitry, they must be represented for human use by some set of symbols that can be recognized audibly and visually, and preferably easily. The graphic symbols chosen could be numerals, Latin letters, letters of another alphabet, silhouettes of animals, or any other set of symbols. However, the ten Arabic numerals of the decimal numeric system seem to be most suitable for the purpose. Some of the reasons are as follows. The ten numerals match exactly, on a one-for-one basis, the ten values required electrically by telephone switching systems. Also, the set of Arabic numerals are recognized in countries that use different letter alphabets. This makes international recognition and use of telephone numbers easier. Further, the set of ten Arabic numerals has a lower "information load" than the set of twenty-six Latin letters. (See Section 6.21 of the Coding Guidelines.) This suggests that there will be lower error rates in human use of all-numeric telephone numbers than in identical use of alphabetic or alphanumeric "telephone numbers" of the same length. Studies summarized in Section 5.0 of the Coding Guidelines (those of Aume and Topmiller, Konz et al, and Osowitz and Sweetland) also suggest that alphanumeric codes - such as partially alphabetic telephone numbers - should be avoided, if possible, in order to reduce error rates.

Probably there are other aspects of this matter that could be investigated and discussed, such as the influence of culture (in the anthropological sense) on the use of letter symbols vs numerals in telephone numbers. However, whatever the theoretical possibilities, our American telephone switching system (Bell and non-Bell) is already designed and to a great extent, already installed. It does and must continue to interface (exchange switching signals as well as voice signals) with other countries' telephone systems. To change the electrical switching method (from decimal) would appear practically impossible. Assuming, therefore, that Modulus-10 switching will prevail, the ten Arabic decimal numerals will probably continue to be the most widely-recognized and efficient character set from which the symbols of "telephone numbers" can be selected.

QUESTIONS: In a sense, isn't it frustrating to set up standards without any "clout" to enforce their usage? Do you see any official usage policy in the near future within your corporation? Were any economic analyses done in connection with the standards effort? How do you cost out benefits from a new standard data element?

ANSWERS: Like any technical staff workers sometimes we wonder whether the production workers (information system development staff) appreciate our efforts, but we believe that our "friendly persuasion" approach is and will continue to be the most effective way to establish the use of data element standards within our Company. Past experience indicates that information system developers will use data element standards if the standards are technically good, if they are adequately publicized to potential users,

537-860 O - 74 - 13

and if a willingness, and indeed a need, to seek developers' and users' advice and comments regarding proposed data element standards is shown. I'm not aware of any planned change in our policy. I don't know whether any specific quantitative economic analysis was done in connection with our standards effort, but it is apparent that a qualitative economic analysis was done by our higher management, who authorized our activity and who continue to approve its budget. We do not compute specific dollar values of expected benefits for individual data element standards. We are convinced that the overall effort will, in time, prove to be of benefit to the Company.

Transportation Data Exchange

Edward A. Guilbert[1]

Transportation Data Coordinating Committee
Washington, D. C.

Shippers and carriers, confronted with the increasing costs of paperwork processing, are assessing the technology of computer-to-computer data transmission for cargo documentation and payment. The Transportation Data Coordinating Committee (TDCC) has been established by industry as the national center for development and promotion of uniform terms, standard data elements, codes, formats, and systems interface for data exchange between shippers, carriers, and banking institutions. The TDCC has achieved many of its goals for the ultimate development and adoption of common data languages for computer-to-computer information interchange within the transportation community.

Key words: Cargo; cargo movement; carriers; data exchange; information systems; shippers; transportation community; transportation industry.

"It should be noted that no industry uses computers more extensively than transportation. The transportation industry is well on its way to becoming the largest domestic and international user of data communications. ...It is clear that data exchange standards are an absolute necessity. In-house computers speak eloquently and efficiently to one another but when they attempt to communicate across company lines there is only the inaudible hum of incomprehension. The potential for future transportation applications is exciting but it cannot be accomplished without standardization along the lines proposed by TDCC. We must put aside diverse procedures and opinions and get on with the job. The differences are substantial, but we are going to have to subordinate them to a greater goal: Agreeing on standards and implementing them." (1)[2]

In these words, Mr. Joseph M. Henson, Vice President of IBM, identified the goal of the Transportation Data Coordinating Committee (TDCC), which was formed by a broad community of shippers, carriers, financial

---

[1] President

[2] Figures in parentheses indicate the literature references at the end of this paper.

institutions and other concerned interests. Its mission is to coordinate data elements, standards, and codes in the transportation industry for use in transportation data exchange systems. In pursuing this task, the TDCC directly participates with government and private sectors of the transportation community, as well as with international organizations and associations which are similarly dedicated.

It is not surprising that the transportation industry is forecast to become the largest user of data communications. The industry is second only to agriculture in size and second to none in dynamics. As stated by Mr. Arthur C. Clarke: "Most of the energy expended in the history of the world has been to move things from one place to another." (2) This historical truth is demonstrated by the U. S. transportation community as it responds to a continually expanding and changing market. The industry moves cargo to the market through the coordination of a multiplicity of propulsion systems, management systems, regulatory systems, procedural systems and, most importantly, information systems.

Mr. T. L. Simis, Vice President, American Telephone and Telegraph Company, recently said: "To live effectively is to live with adequate information. To operate a business effectively...is to operate with adequate information, timely information, where you need it." (3) This is a statement of fact most appropriate for the transportation industry where accurate and timely information, delivered to the right places and people, is required to initiate, control, account, pay, audit, and perform many other functions involved in cargo transportation.

In the United States, operational information concerning cargo movement is generated, communicated, and processed by a series of separate but interacting information systems. Individual shippers and carriers, interchanging carriers of a single mode, and the intermodal cargo handling process provide five distinct system environments which function individually and collectively to deliver cargo from point 'A' to point 'B'. These information systems are of central importance but do not contain the total transportation data base. The international port, for example, requires a unique information system to perform its role as the communications interface point between national entities on matters concerning international trade. Further, billing and paying functions, freight forwarder services, auditing services, and government reporting contribute importantly to the total information flow necessary to complete the transportation transaction cycle.

The TDCC has analyzed the data elements in the transportation cycle and has selected the key elements for identification through recommended standard codes. For example:

> Commodity identification: The TDCC has endorsed the structure of the Standard Transportation Commodity Code (4) for data exchange purposes for domestic transportation. A thesaurus is being produced under contract with the Department of Transportation which will identify commodity descriptors that have been harmonized with the Standard Transportation Commodity Code, Standard International Trade Classification (5), Brussels Nomenclature (6), and other classifications required for international trade.

Geographic points: The Standard Point Location Code (7) has been endorsed by TDCC as the uniform geographic identifier for shipment pick-up and delivery.

Carrier identification: The Standard Carrier Alpha Code (8) was selected and endorsed as the uniform carrier designator.

Customer Code: The Dun & Bradstreet Data Universal Numbering System (9) has been endorsed by TDCC as the consignee/consignor identifier.

These and other codes that are being evaluated will provide a transportation data dictionary for the data elements required in information exchange programs.

Currently, the information is being exchanged between shippers and carriers by means of paper documentation involving invoices, bills of lading, waybills, freight bills, etc. It is this paper system which is the target of the TDCC for upgrading to electronic systems applications. The goal is for computer-to-computer interaction, online cathode ray tube terminals in the shipping and freight handling areas, and the application of digital data communications to link shippers, carriers, banks, international ports, appropriate governmental agencies and other transportation services to satisfy transportation data exchange requirements.

Mr. James W. Germany, Vice President, Southern Pacific Transportation Company, said, in reference to these objectives: "With the capabilities of modern computers and communications, it now is technically feasible to have shipments move...from origin to destination with no paper documents at all ...and...to handle all accounting in the same paperless fashion...This paperless future, however, is far out in time...and needs a great deal of thought, planning and effort." (10)

Mr. Germany also suggests that it is too early to speak in terms of hardware and software systems. The number of parties and sites involved or the number of communications paths to be accommodated have yet to be resolved. However, the task is of such a magnitude and the applications requirements so varied that it is reasonable to assume that many forms of data display, processing and communications techniques will have a place in the transportation information environment of the future. Therefore, the task today is to prepare a solid base for tomorrow with a sensitivity to the fact that "we cannot realistically ignore the restraints and controls imposed by precedent." (11)

The primary area of emphasis in the TDCC is towards getting the 'transportation data house' in order. This is a prerequisite to establishing a foundation on which systems can be developed and through which transportation data exchange can occur. While the orientation is essentially towards information concerned with the movement of cargo, the relationship of that information to the internal operations of both carriers and shippers is significant. For example, the shipper is concerned with the distribution service. Therefore, cargo movement information is critical to his logistics management operations. The carrier's information system must encompass cargo movement, rolling stock, schedules, revenue accounting and other phases of

operation. The basic information flow must serve the needs of its users. In this case, it must reveal who is shipping, what is being shipped, when, at what cost, via what means and route, to whom, and to what location.

As cargo movement is generated, communicated and processed, the basic elements of data are utilized to serve a wide spectrum of applications for control, management summary, statistical evaluation and record purposes. These varied use requirements and the close relationship of cargo movement information to the operations of the shippers, carriers, government and other concerned organizations dictates an approach for wide coordination between many parties to achieve standardization for data exchange.

The TDCC effort to define a system of transportation data standards is feasible and practical. The various interests involved in the shipping, transporting, regulatory and accounting aspects of cargo movement require essentially the same information to perform their respective functions. In terms of computers, communications, software and data base design, the information required can be standardized. Terms of reference can be developed. Communications and operational data elements can be identified and defined. Appropriate codes can be developed to convert information to a machine readable form. The TDCC has made significant progress in developing a standardized transportation data base and recommending adoption by the industry. This data standardization effort, when completed and the standards adopted, will enable the transportation community to move with continuity towards a truly modern transportation data exchange capability.

The data concept recognizes problems yet to be resolved concerning the collection, control, and use of interchanged data. Various companies will agree to the transmission of information but not when they believe its collection, processing and collation in 'external' data bases will operate to the detriment or disadvantage of the transmitter. Therefore, protection of data will be critical to the data interchange system. It is entirely reasonable to expect that solutions to requirements for data privacy can and will be achieved.

In summary, the shipper/carrier community has acknowledged the need for an improved data exchange capability. The future holds a feasibility for a 'paperless' transportation information system employing the most advanced display, processing, and communications technology. It is a future wherein transportation information systems environments will be able to exchange data in a timely and economical manner. The TDCC will speed the process through the development of transportation data element standards, a transportation communications guide, and through continued effort to convert undisciplined narrative data to a machine readable system of representative codes.

References

(1) Henson, J. M., 'Role of the Computer in Transportation', Proceedings TDCC 1973 National Forum on Interfacing Transportation Data Systems (1974).

(2) Clarke, A. C., Profiles Of The Future; An Inquiry Into The Limits Of The Possible, Harper & Row, New York (1962), 121pp.

(2) Simis, T. L., 'Role of Communications in Transportation', Proceedings TDCC 1973 National Forum on Interfacing Transportation Data Systems (1974).

(4) Standard Transportation Commodity Code, Association of American Railroads, Washington, D. C.

(5) Standard International Trade Classification, United Nations Publishing Service, New York.

(6) Brussels Nomenclature, Customs Cooperation Council, Brussels, Belgium.

(7) Standard Point Location Code, National Motor Freight Traffic Association, Inc., Washington, D. C.

(8) Standard Carrier Alpha Code, National Motor Freight Traffic Association, Inc., Washington, D. C.

(9) Data Universal Numbering System, Dun & Bradstreet, Inc., New York.

(10) Germany, J. W., 'Improved Data Systems in the Railroads', Proceedings TDCC 1973 National Forum on Interfacing Transportation Data Systems (1974).

(11) Manion, D. L., 'Data Systems Impact on Railroads', Proceedings TDCC 1972 National Forum on Transportation Data Systems (1973).

Cargo Data Interchange System for Transportation (CARDIS)

Murray A. Haber[1]

Office of Assistant Secretary for
Environment, Safety and Consumer Affairs
Department of Transportation
Washington, D. C. 20590

Paperwork threatens the success of the trade it was intended to assist. Documentation entailed in domestic as well as international transactions has reached such magnitude that it delays shipments, boosts costs, imperils profits, discourages expansion, and overburdens industry and government.

In the report "Paperwork or Profits? in International Trade"(1)[2], the reasons for the high costs of exporting goods become very apparent: a total of 125 different types of documents are in regular and special use, representing more than 1,000 separate forms; average shipments involve 46 different documents; U.S. international trade documentation annually costs $6.5 billion or 7.5 percent of the value of all U.S. imports and exports.

In order to overcome these problems, the Department of Transportation has developed the Cargo Data Interchange System (CARDIS), a data processing and transmission system whereby shippers, carriers, government agencies, banks, insurance underwriters, and others engaged in the transportation of goods will enter standardized data elements and codes into a data processing center from remote terminals located in their own premises. Information related to specific shipments will then be transmitted by high-speed methods to other data processing centers either within the U.S. for domestic shipments, or to foreign countries for export shipments. The receiving data centers will produce bills of lading, commercial invoices, manifests, and other documentation necessary to enter, clear, and release goods to consignees. In the case of overseas shipments, the goods could be entered and cleared through Customs before the arrival of the vessel or aircraft.

Test/demonstrations were conducted at Department of Transportation headquarters on March 21, 1973, to show the feasibility of the system. Authority has been granted to the Office of Facilitation to proceed with the further development of the system, and the planning for the many individual studies deemed necessary to find solutions to the many problems that exist, before CARDIS can become an operating reality.

---

[1] Consultant on Documentation and Procedures, Office of Facilitation.

[2] Figures in parenthesis indicate the literature references at the end of this paper.

Key words: Cargo data interchange; commodity description and code; international trade documentation; standardized data elements; transportation documentation; U.S Standard Master.

1. Background

1.1. Magnitude of the Problem

United States exports (excluding military) in calendar year 1969 were $37,331,000,000. They rose to $49,116,000,000 in 1972, an increase of 31.6 percent. Data available so far for calendar year 1973 indicate that exports are proceeding at a rate in excess of $64 billion, or an increase of more than 20 percent over the previous year. The Export-Import Bank of the United States, in projecting for the foreseeable future, anticipates a 15 to 20 percent increase in U.S. exports each year through fiscal year 1976.(2)

To further increase the volume of U.S. exports, and to enhance the U.S. balance of payments, the President recently set up two new export expansion advisory groups:

The President's Export Council, an organization of leading American businessmen, to advise him on ways to increase U.S. sales overseas.

The President's Interagency Committee on Export Expansion, representing thirteen Government departments and agencies, to initiate and coordinate Government programs and policies affecting the U.S. export performance.

1.2. Paperwork Problem

Paperwork threatens the success of the trade it was intended to assist. It is estimated to exceed 800 million documents and 6 billion copies per year. The volume for domestic trade is substantially larger than for exports and imports; however, the paperwork is less complex for domestic trade.

The Department of Transportation, Office of Facilitation, and the National Committee on International Trade Documentation (NCITD), a private, non-profit research organization dedicated to the simplification of international trade procedures, jointly conducted a world-wide study on international transportation paperwork. The resultant report, "Paperwork or Profits? in International Trade" (1) discloses some startling statistics:

- A total of 46 different types of firms and government agencies regularly are involved in international trade. . .
- As many as 28 of these parties may participate in a single export shipment. . .
- A total of 125 types of documents are in regular or special use. . .
- The 125 different types of documents represent more than 1,000 different forms. . .
- A total of 80 types of documents are in regular use as opposed to 45 in special use. . .
- Average shipments involve 46 separate documents, with an average

of over 360 copies per shipment being employed. . .

- U.S. international trade annually creates an estimated 828 million documents and these generate an estimated 6½ billion copies. . .
- Average export and import shipments required 64 man-hours to prepare and process, split on the average of 36½ man-hours for an export shipment and 27½ man-hours for an import shipment. . .
- Total U.S. international trade documentation annually consumes more than a billion man-hours, equivalent to more than 144 million days of work, and equal to 600 thousand work years. . .
- Average documentation cost per international shipment amounts to $351.04, divided $375.77 for exports and $320.58 for imports. . .
- On the basis of current shipping volumes, total U.S. documentation costs aggregate almost 6½ billion dollars a year, and represent 7.5 percent of the value of the total U.S. export and import shipments.

1.3. Cargo Delays

As the volume of trade increases, and as transit time for moving goods decreases, particularly with expanded use of wide-bodied aircraft and high-speed container ships, paperwork is becoming a major limiting factor in shipper-consignee total transaction time.

Delays in the creation and receipt of documents have become a major problem at ports of departure and ports of arrival, as well as at intermediate interchange points in the movement of the cargo. Shipments often accumulate at these points in the movement, awaiting the arrival of documents, and often delay the preparation and processing of additional essential movement documents. Existing manual procedures for preparing and processing documents are usually slow, repetitious, prone to errors, and expensive. Failure to receive documentation in a timely manner often causes cargo to be delayed at piers and airports awaiting loading or pick-up. Such conditions subject the cargo to loss by pilferage, demurrage costs, and possible damage by weather.

1.4. Automation in Use

Some companies, including exporters, importers, and carriers, as well as government agencies, have automated their internal operating systems. Sea Land Service, Inc., and Atlantic Container Lines are but two examples of large ocean container ship companies which are operating automated systems for preparing bills of lading and manifests and transmitting them internationally. These systems are limited, however, since carriers do not have access to commercial invoice data to meet U.S. and foreign entry and clearance requirements. Other examples can be cited: The Ford Motor Company plans to eliminate the preparation of documents and transmit movement data to the Penn Central Railroad on its shipments.

The Bureau of the Census of the U.S. Department of Commerce encourages exporters who meet certain prescribed criteria to discontinue providing individual "Shipper's Export Declarations" for each exported shipment valued at $250.00 or more. These companies are authorized to report on a monthly basis using magnetic tape or punched cards (3). The data thus provided, can

be fed directly into Bureau of the Census computers for incorporation into U.S. trade statistics.

The U.S. Customs Service is automating import entry paperwork to help speed the flow of goods coming into the United States. The total system is known as the Automated Merchandise Processing System (AMPS).(4)

Most of these systems are oriented to "in-house" needs, each serving a limited or specific purpose, and each having its own data requirements, its own coding systems for processing, and its own document formats. The systems thus created do not interface with any other systems to permit the automated interchange of data.

1.5. Automation in Other Countries

Some countries have automated, or are in the process of automating their own Customs cargo entry systems. The United Kingdom's London Airport Cargo EDP System (LACES) is an excellent example of an automated entry system in operation. It has an added capability of maintaining location control of all consignments in the area serviced by the system, and also offers benefits to private users. All carriers, consignees, consignors, freight forwarders and brokers are permitted to have terminals connected on-line with LACES. The US, of course, is developing AMPS. Australia, France, Germany, Japan and the Netherlands have or are planning to install similar automated systems.

2. Progress Thus Far

The transportation of goods has had a growing amount of paperwork or documentation red tape associated with it. This condition has become progressively worse over the years until joint industry/Government attention was concentrated on the problem. Several years ago, industry and Government in the United States launched a full-scale attack on the problem of trade and transportation documentation.

2.1. Paperwork Simplification and Standardization

The program to simplify, standardize, and otherwise improve transportation documentation and related procedures and coding, and to obtain greater use of automatic data processing, was one of the early programs of the Department of Transportation. The many problems associated with the transfer of data on documents from point-to-point as goods move from origin to destination, made it essential for Government and industry to work hand-in-hand on solutions. The joint Government/industry participation was made possible and enhanced by the establishment of the Office of Facilitation in DOT, and the simultaneous creation of two non-profit organizations by the private sector known as the National Committee on International Trade Documentation (NCITD), and the Transportation Data Coordinating Committee (TDCC).

2.2. DOT Mission

The Office of Facilitation was established to help carry out the mandate of the Congress (5) that directed DOT to:

- Facilitate the development and improvement of coordinated transportation service.
- Stimulate technological advances in transportation.
- Provide general leadership in the identification and solution of transportation problems.

- Promote and undertake development, collection, and dissemination of technological, statistical, economic, and other information relevant to domestic and international transportation.

2.3. NCITD and TDCC Missions

NCITD has as its role the simplifying and standardizing of international trade documentation and related procedures.

TDCC was created to develop and coordinate the use of standard codes and to extend the use of ADP through transportation.

NCITD and TDCC, separately funded and staffed by private industry, represent hundreds of shippers, carriers, bankers, insurance underwriters, brokers, agents, and freight forwarders to provide the necessary support to properly represent the private sector.

2.4. Government Support

Considerable support continues to be received from offices of Government with responsibilities in this area. Included are the Office of Management and Budget, U.S. Customs Service, Office of Export Control, Bureau of the Census, Maritime Administration, National Bureau of Standards, Federal Maritime Commission, Interstate Commerce Commission, Civil Aeronautics Board, Department of Defense, General Services Administration, General Accounting Office, U.S. Tariff Commission, and others. This cooperative support has helped DOT increase the effectiveness of the Government's program to simplify transportation paperwork.

2.5. Joint Government/Industry Program

As a result of this broad support for an improvement program, a joint industry/Government attack on transportation paperwork was launched in 1967 by DOT, NCITD, AND TDCC. For purposes of explanation, the progress being made can best be described in three phases:

Phase I. This was the organization phase covering the period from 1967 to 1969. The three organizations were engaged in planning a comprehinsive documentation, coding, and data processing program for transportation. During this period it was necessary to determine what techniques could best be utilized to enable Government and industry to work together and find solutions to common problems. Exercising its role to coordinate facilitation activities in the Government, DOT solicited the assistance of other Federal departments and agencies to obtain complete participation in the analysis, development, and implementation of more efficient and effective data and documentation systems.

Phase II. During this period, 1969 to 1972, programs were launched to improve documentation, procedures, and coding, and to obtain wider use of ADP and data transmission. Alignment of transport and trade documents was accomplished on a world-wide basis. DOT and NCITD provided the leadership to achieve this important improvement. Work was also initiated to develop a standard commodity description and code system for transportation. The DOT and TDCC are currently engaged in this activity.

A case-by-case study of actual documents and procedures required to move goods domestically and internationally was undertaken jointly by DOT and NCITD. The results were published as the joint DOT/NCITD report "Paperwork or Profits? in International Trade", previously mentioned in this paper.

DOT, NCITD, and TDCC later joined forces to explore the feasibility of transmitting the essential data elements to satisfy all requirements from origin to destination with a minimum of paper documents. The high pay-off

potential, in terms of time and dollar savings, brought forth full support from the private sector, Federal agencies, and foreign businesses throughout the world. In fact, many more organizations were willing to devote the time of their experts through NCITD, TDCC, other organizations, foreign governments, and Federal agencies to enhance and extend the benefits of the program. Such wide support was convincing evidence that the program and specific projects would benefit all segments of the transportation community and were geared to maximum results and effectiveness.

Phase III. This phase of the program began in 1973, and will continue through fiscal year 1974 and beyond:

Standard Format. During this period, the long-haul truck carriers agreed to adopt the U.S. Standard Format (fig. 1) for their bills of lading applicable to domestic and international cargo shipments. The air carriers adopted a Shipper's Letter of Instruction based on the standard format, from which the air waybill can be prepared. The ocean carriers have accepted the standard format for ocean bills of lading, and are using or are in the process of printing the newly aligned forms for use. The U.S. Customs Service, Department of Commerce offices, and other government agencies have aligned their trade or transport documents to the standard format.

Government Bills of Lading. A vigorous program is underway to simplify and improve Government Bills of Lading. An 1823 Act has been amended to permit more expeditious payment of Government transportation charges without the consignee's receipt which was required under the Act. The amendment paves the way for eventual Government use of commercial bills of lading, thus permitting the same standard documentation for all shippers, whether Government or private.

Commodity Coding. The commodity description and code area, one of the most complex areas to improve, and where reliable estimates have indicated potential savings of $1.2 billion, is well along toward completion, for the domestic listing. This will enable a single description and code to replace as many as 17 different descriptions for a single commodity often required in a shipment from origin to destination. TDCC has done the development work on this project for DOT. (6)

Bill of Lading. A shipper-prepared bill of lading (fig. 2a and 2b) is now being put into use for maritime shipments. A standard set of terms and conditions which more effectively meet the needs of shippers and carriers has been approved, and appears on the reverse of the form. This new form, with its blank masthead, will eliminate the need for printing, stocking, and using hundreds of different carrier bills of lading, and will be applicable to domestic as well as international cargo, and will enhance the use of through bills of lading.

Joint DOT/NCITD Study. The 1971 joint DOT/NCITD study report (1) contained 28 recommendations which at this point are being implemented. A supplementary report (7), has been jointly prepared to show the progress being made in implementation. The estimated savings from complete implementation of the 28 recommendations will be $4.5 billion.

3. Cargo Data Interchange System (CARDIS)

3.1. Need for Data Interchange

The projected increases in trade and transport, and resultant paperwork problems and delays in the expeditious movement of cargo; the development of independent data processing systems by shippers, carriers, and Government agencies; and the design of Customs entry systems by many governments (AMPS by the U.S., LACES by the UK, SOFIA by France, etc.) emphasized the need for

an overall, fully interfaced automated system for receiving, processing, and transmitting cargo data to serve the needs of the private sector as well as government agencies. The United States Cargo Data Interchange System (CARDIS) is being developed to satisfy this need.

3.2. Test/Demonstration Program

It was agreed by participants in the United States and in the United Kingdom that a series of tests should be initiated to determine the feasibility of the concept, and to provide the groundwork for the eventual design of the operating system. The test/demonstrations would consist of a series of transmissions of data on air shipments from the United States to the UK LACES installation at Heathrow Airport outside London. Subsequent tests could then proceed on air shipments from the United Kingdom to the United States, followed by similar series on ocean shipments. Also, the initial series would be limited to shipments from airport to airport and from ocean port to ocean port. At a later time, the series would be expanded to cover shipments from inland points in one country to inland points in the second country.

a. Participation by Government and Industry

U.S. Government agencies that play a role in trade and transportation agreed to support and cooperate in the project. Included were the Office of Management and Budget, U.S. Customs Service, Office of Export Control, Bureau of the Census, Maritime Administration, National Bureau of Standards, Federal Maritime Commission, Interstate Commerce Commission, Civil Aeronautics Board, Department of Defense, General Services Administration, General Accounting Office, U.S. Tariff Commission, and others.

The National Committee on International Trade Documentation was given the role of simplifying and standardizing trade documentation and related procedures, and to obtain the cooperation of selected shippers and carriers to participate in the tests.

The Transportation Data Coordinating Committee was assigned the task of developing and coordinating the standardized codes, and to obtain the cooperation of its membership in enhancing the tests.

These two industry-funded organizations, representing hundreds of shippers, carriers, bankers, insurance underwriters, brokers, freight forwarders, and agents, provided the necessary support to ensure proper representation by the private sector.

4. CARDIS System Requirements

4.1. Concept

From the start, it was agreed that the system must provide data required by the shipper, the carrier, and the Government agencies involved. Other participants in transport, such as bankers, and insurance underwriters would be brought in at a later phase. The system at the outset would have to provide for bills of lading, commercial invoices, manifests, both outward and inward, export controls and statistics, import data for Customs authorities, etc. The concept required that the system provide a central data bank, accessible on a security and privacy controlled basis, to all participants, with the input and output capability described above.

4.2. Data Interchange Centers

The heart of the planned automated cargo data interchange system, which came to be called CARDIS, would be a data center, or centers. These

centers, operated by the private sector, but under some control of the Federal Government, could receive, store, process, and transmit data on domestic shipments, as well as on exports and imports, through all transportation stages, and would provide audit trails, controls, and statistics for Government as well as industry. For the initial tests, the U.S. Department of Transportation facility known as the Transportation Systems Center, located in Cambridge, Massachusetts, would serve as the data interchange center.

4.3. Standardized Coding

The most economical recording, transmitting, and processing of data would rely on the use of standard code systems. Some of the systems currently in use by industry were developed within individual companies. Consequently, there is no interface between shippers, carriers, and governments. The proposed system would specify only recognized standardized coding schemes to be used by all participants in the system. Many of these codes already exist. Some code systems, such as the DOT Standard Commodity Description and Code System (6), had reached a sufficient stage of development as to permit their use in the tests. Other codes would have to be developed, but all would be standardized and would be part of the overall system.

4.4. Document Formats

Standard documents or new standardized formats would be prescribed for all input and output needs.

Input. A format containing all data element fields would be developed for input into the system. This format would simplify setting up a shipment record, permit inquiry by need-to-know participants, and make possible standard outputs.

Output. Bills of lading and commercial invoices would align with the U.S. Standard Master (fig. 1), the U.K. aligned series, and the ECE layout key. Manifests would conform to the Customs requirements of the countries involved. Efforts would be made to utilize standard documentation for manifesting as prescribed by the International Civil Aviation Organization (ICAO), and the Inter-Governmental Maritime Consultative Organization (IMCO).

4.5. Confidentiality of Data

The system would be so designed that only authorized parties, be they shippers, carriers, or Government agencies, would have access to the data bank to add, delete, change, or read data. In addition, such access would be limited to a need-to-know, and would be individually prescribed for each data element in the record.

4.6. Flow Charts

Studies were made of the documentation used by a number of the participants involved in moving cargo from a port of export to a port of destination. The data elements were analyzed to ascertain the irreducible minimum of information that would satisfy the needs of these participants.

From the information thus obtained, a task force from Government and industry devised the basic flow of information as described in figures 3a, 3b, 3c, and 3d. The system uses a central data bank, labeled DPC (Data Processing Center) to accumulate and process data inputted on-line by shippers or their freight forwarders, carriers, and Government agencies. In turn, the DPC produces, upon demand, bills of lading, manifests, statistical reports, and other messages for transmission to destination port.

At destination, each message is decoded, printed out in proper format, and entered into the LACES System by the broker for processing according to United Kingdom Customs requirements.

4.7. Data Elements

Analysis of information contained on transportation documents used on air shipments from the U.S. to the U.K. indicated that less than 60 data elements provided all the information needed to complete bills of lading, commercial invoices, and manifests. These data could also serve to verify shipments for export control purposes, and could provide input to the statistical programs on exports maintained by the Bureau of the Census. As additional participants would be brought into the system (insurance underwriters, banks, etc.) the list could be augmented to serve their needs as well.

4.8. Input Document

In order to insure the standardization of input data elements into the system, a form (fig. 4) was designed, listing all of the data elements, and showing the maximum number of characters available in the particular data field. In the case of some data elements (i.e., data elements 6 through 12) alternate fields were provided so that either the established code could be entered, or the entire item could be entered in text form. The document thus developed could serve as a shipper's letter of instruction, and could be used for direct sequential input into the data bank from a device as simple as a teletypewriter. Additional input could be made at any time to fill in missing data, or to change existing data.

4.9. Output Document

It was decided from the outset of the project that the U.S. Standard Master format (fig. 1), which is aligned with the ECE Layout Key, would be the basis for the bill of lading and the commercial invoice. For purposes of the test, some modifications in columnar arrangement on the commercial invoice were made. Further studies are underway to determine whether closer alignment can be achieved on that document. The program was designed to use the standardized manifest form prescribed by the United States Customs Service.

5. Test/Demonstration of Air Cargo Shipments from U.S. to U.K.

5.1. DOT Transportation Systems Center

The DOT Transportation Systems Center (TSC) at Cambridge, Massachusetts, designed the automated cargo data processing system program for demonstration purposes. Using teletype terminals acoustically coupled through the telephone network to the Center's computer, and with built-in security measures to protect the data from unauthorized users, the system provides for data entry, editing, reading individual items, transmitting shipment data overseas, and producing hard copies of bills of lading and commercial invoices aligned with the U.S. Standard Master, as well as the cargo manifest.

The data are contained on a Honeywell DDP-516 computer located at TSC with a core memory of 16,000 16-bit words, disk storage, and a data-phone interface capable of controlling three telephone lines. The program contains about 8,500 lines of source codes.

5.2. Participants in the Test/Demonstrations

Airlines. Pan Americal World Airways, Trans World Airways, and British Overseas Airways Corporation participated in the studies and provided

537-860 O - 74 - 14

valuable information which was useful in developing the data base for the system.

Shippers. Four large shippers (IBM, Dow Chemical, International General Electric, and DuPont) provided copies of actual shipping documents, which were used as the basis for setting up the data bank for the test.

5.3. The Test/Demonstration

Prior to the test/demonstration, the shipments were entered into the computer at TSC. The information on the shipments had previously been entered in code form on the Master Record Input Sheet. Four on-line input/output teletype terminals located at DOT Headquarters, Washington, D.C. were represented as the offices of a shipper/forwarder, carrier loading platform, carrier traffic office, and Government agency (Customs, Export Control, Census, etc.). The demonstration to the public was actually held on March 21, 1973.

Demonstrated to the audience of industry, transportation, and Government representatives, were various types of actions and queries between terminals, and to and from the computer located in Cambridge, Mass. These various types of actions included:

a. Goods on one shipment arrived at carrier loading platform at airport. Airline receiving clerk entered transaction number, shipper's name, and number of packages. Data printed out in carrier traffic office.

b. Operator at carrier traffic office entered transaction number and asked for print-out of Field 6 to verify shipper's name.

c. At airline's request, DPC printed out a bill of lading in the airline's traffic office.

d. Airline traffic office assigned shipments to a TWA flight, and entered request for flight manifest by entering flight number, date, and transaction number of each consignment designated for that flight.

e. Computer advised one shipment had improper destination airport entered on record.

f. Carrier traffic office changed destination to correct airport.

g. Computer printed out complete manifest in airline traffice office for carrier use, and on Government terminal for Customs use.

At this point the airline traffic office instructed the computer to transmit to the U.K. the required data related to that specific flight. The message was transmitted to London via New York, using facilities provided by Pan American World Airways.

h. Other queries to the data base followed:

1. To test the security of the system, one airline asked for data on a shipment moving on a competitive airline. The query was rebuffed by the computer.

2. Shipper/forwarder requested status information about his consignment and received an answer within several seconds.

3. The unit price on a commodity was changed. The computer refigured the price extensions and totals, and printed out a new invoice on the shipper/forwarder terminal.

Messages related to the other two flights of the demonstration were transmitted. During the course of the demonstration, acknowledgement of receipt of the transmitted messages was received from SITPRO in the U.K., showing the transaction numbers and the LACES channels of selection, indicating the fact that the U.K. system had accepted the data and was processing it for clearance.

6. Conclusions.

The tests had been undertaken to determine the feasibility of assembling all the data elements needed to meet trade and transport needs, as well as Government requirements, and to transmit these data elements to all parties involved in a transaction from origin to destination. Therefore, there was a greater emphasis placed on the data elements themselves, than on the technique of transmission. Also a primary goal was to transmit the cargo data in an expeditious manner in order to eliminate delays in the movement of cargo associated with the late arrival of documents.

All of the objectives were attained, and the results were very encouraging. The following conclusions were drawn as a result of this first series of tests:

a. Data elements for trade and transportation are very similar, with only slight uniqueness of data for each.

b. Data can be furnished and received by a large number of shippers, carriers, and others, in machine language, eliminating the need for each to convert data from a variety of different documents at each step of the process from origin to destination.

c. Standard codes for data elements are essential to facilitate the exchange of data in an understandable and efficient manner.

d. Standard commodity descriptions and codes are essential, and are by far the most important data elements and codes in the system.

e. Standard documentation formats will facilitate data interchange.

f. When moving from a document system to a data transmission system, solutions to several problems will have to be found. Among them are:

1. Satisfaction of signature requirements.

2. Negotiable documents accomplished by the system.

3. Agreement terms and conditions for transport purposes accomplished without having to transmit minute details on the documents.

g. Delays in the shipment and transshipment of cargo due to the late arrival of documents can be overcome by data transmission and interchange techniques.

h. A number of data interchange centers may be needed at strategic locations around the United States.

i. Different computer hardware can be used to process and transmit data provided they interface with each other.

j. Guidelines and controls will be needed to bring about the required standardization in order to permit data interchange.

k. Security and privacy of data in the system will be essential.

l. Statutory controls may be needed to ensure effectiveness of the system.

m. A joint Government/Industry Coordination Group may be helpful to ensure full representation and participation.

n. A Legal Committee will be essential to study, evaluate, develop, and establish the legal aspects of data transmission and interchange.

o. Government leadership and coordination is essential to ensure that all efforts are directed toward the common goal of data interchange.

p. A program plan to develop a cargo data interchange system should be developed and approved at the earliest possible time.

q. Resources will be required to complete the necessary studies and to develop the details required for an operating system.

r. Many countries are interested in data interchange. Tests with these interested countries are essential in order to perfect a world-wide data interchange system.

7. Future Plans

Now that the initial steps have been successfully taken, the Department of Transportation is proceeding with the further development of CARDIS. Japan, Canada, France, Australia, West Germany, Russia, and Hong Kong have expressed interest in participating in the program. The United Kingdom has offered to continue the test series to include ocean movement of cargo. Japan has also offered to conduct joint U.S./Japan tests for cargo moving between the two countries.

7.1. Funding

The many aspects of the CARDIS program and its relationships to the private sector and the numerous Federal agencies involved, make it essential that the Department of Transportation play a lead role in the development of the operating system. As industry becomes a participating member on the CARDIS development team, and contributes to it, the system will grow substantially. The Department of Transportation resources will be used to develop the basic system, the standardized data elements and codes, the interfaces with shippers, carriers, freight forwarders, banks, insurance underwriters, and Federal agencies. The cost of establishing data centers will, of necessity, be borne by the private sector. The resources contributed by DOT will be reduced over a three or four year period, and the entire cost of CARDIS will then be borne by the users of the system.

7.2. Coordination

The importance of the interfaces with the numerous participants involved in the United States, as well as the alignment of the U.S. system with other countries and international organizations, makes it mandatory that the Department of Transportation serve as the overall CARDIS coordinator. This coordination role will be a continuing function. It will be enhanced by the establishment of a Government/industry group to increase the effectiveness of the coordination. This Government/Industry group will be chaired by the Department of Transportation.

7.3. Industry Expertise

NCITD and TDCC have done much of the research and development work of CARDIS thus far. Together, they represent hundreds of organizations

intimately involved in the movement of goods domestically and internationally. This arrangement of working with these experts has proven to be an economical and effective technique for solving the documentation, procedural, coding, and ADP problems as they arose. DOT plans to continue this arrangement and, in fact, to make greater use of these resources to accomplish the objectives of a fully operational CARDIS in the shortest possible time.

8. References

(1) Paperwork or Profits? in International Trade, DOT/NCITD (1971).

(2) Statement of Condition, Fiscal Year '73, Export-Import Bank of the United States.

(3) Sec. 30.39, Foreign Statistics Regulations, BuCensus, DofC.

(4) AMPS, Automated Merchandise Processing System - Past, Present, Future, U.S. Customs (Nov. 1973).

(5) Public Law 86-670, 89th Congress.

(6) Standard Transportation Commodity Description and Code System Reports, prepared by TDCC for DOT (1973).

(7) Progress Report on Paperwork or Profits? in International Trade, DOT/NCITD (1973).

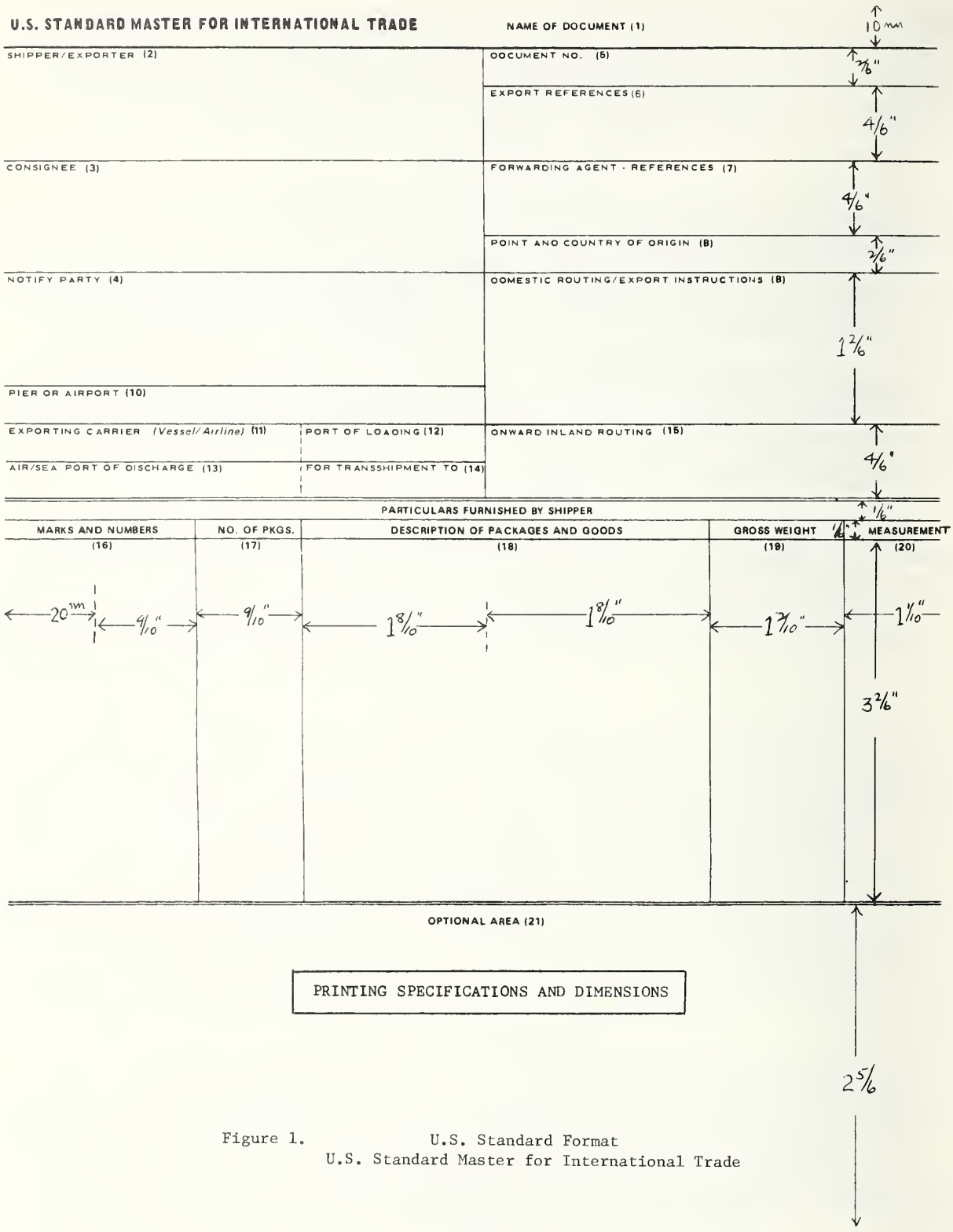


Figure 1. U.S. Standard Format
U.S. Standard Master for International Trade

SHORT FORM BILL OF LADING
(Non-Negotiable Unless Consigned To Order)

NAME OF CARRIER

SHIPPER/EXPORTER (2) (COMPLETE NAME AND ADDRESS)

DOCUMENT NO. (5)

EXPORT REFERENCES (6)

CONSIGNEE (3) (COMPLETE NAME AND ADDRESS)

FORWARDING AGENT - REFERENCES (7) (COMPLETE NAME AND ADDRESS)

POINT AND COUNTRY OF ORIGIN (8)

NOTIFY PARTY (4) (COMPLETE NAME AND ADDRESS)

DOMESTIC ROUTING/EXPORT INSTRUCTIONS (9)

PIER/TERMINAL (10)

VESSEL (11) FLAG

PORT OF LOADING (12)

ONWARD INLAND ROUTING (15)

PORT OF DISCHARGE FROM VESSEL (13)

FOR TRANSSHIPMENT TO (14)

PARTICULARS FURNISHED BY SHIPPER

MARKS AND NUMBERS (16)	NO. OF PKGS. (17)	DESCRIPTION OF PACKAGES AND GOODS (18)	GROSS WEIGHT (19)	MEASUREMENT (20)

FREIGHT AND CHARGES PAYABLE AT

PREPAID ☐ COLLECT ☐

	PREPAID		COLLECT	
TOTAL				

RECEIVED the goods or the containers, vans, trailers, pallet units or other packages said to contain goods herein mentioned, in apparent good order and condition, except as otherwise indicated, to be transported, delivered or transhipped as provided herein. All of the provisions written, printed or stamped on either side hereof are part of this bill of lading contract.

IN WITNESS WHEREOF, the Master or agent of said vessel has signed _____ bills of lading, all of the same tenor and date, one of which being accomplished, the others to stand void.

BY _____
FOR THE MASTER

DATED B/L NO.

Figure 2a. Shipper-Provided Short Form Bill of Lading (Front)

SHIPPER–PROVIDED SHORT FORM BILL OF LADING

(Terms continued from overside)

This short form Bill of Lading is provided by the Shipper and issued for his convenience and at his request instead of the Carrier's regular long/short form Bill of Lading. Copies of the Carrier's regular long/short form Bill of Lading and the clauses presently being stamped or endorsed thereon are available from the Carrier on request and are incorporated in tariffs or classifications on file with the Interstate Commerce Commission or the Federal Maritime Commission.

In using this short form Bill of Lading, the Shipper, Consignee, and Holder hereof agree that all the terms and conditions of the Carrier's regular long/short form Bill of Lading, normally used in the service for which this bill of lading is issued, including any clauses presently being stamped or endorsed thereon filed with the above agencies, are incorporated herein with like force and effect as if they were written at length herein, and all such terms and conditions so incorporated by reference are agreed by Shipper to be binding and to govern the relations, whatever they may be, between all who are or may become parties to this Bill of Lading as fully as if this Bill of Lading had been prepared on the Carrier's regular long/short form Bill of Lading.

As used herein, the term "Carrier" means any and all carriers whether on land or sea on whose modes of conveyance the goods described on the face hereof are carried.

If this Bill of Lading evidences a contract for the carriage of goods by sea to or from ports of the United States, in foreign trade, or provides for routing within the United States, it shall have effect subject to the provisions of the U.S. Carriage of Goods by Sea Act of 1936, and other applicable statutes, to the extent that any such Act or statutes may apply to the transportation contract of any one or more of the carriers involved.

If this Bill of Lading evidences a contract for the carriage of goods by sea or by surface transportation to, from or through countries other than the United States, it shall have effect subject to the provisions of the applicable Acts, statutes or regulations of such countries, to the extent that any such Acts, statutes or regulations may apply to the transportation contract of any one or more of the carriers involved.

The Carrier's regular long/short form Bill of Lading may contain a number of provisions giving the Carrier certain rights and privileges and certain exceptions and immunities from and limitations of liability additional to those provided by the Acts or Laws referred to above and may extend the benefit of its provisions to stevedores and others.

If required by the Carrier, a signed original Bill of Lading, duly endorsed, must be surrendered to the Carrier on delivery of the goods.

All agreements with respect to the above goods are superseded hereby and none of the terms hereof shall be deemed waived except in writing by an authorized agent of the Carrier.

Figure 2b. Shipper-Provided Short Form Bill of Lading (Back)

US - UK CARGO DATA TRANSMISSION TESTS
(US Segment - Shipment via Air or Ocean)

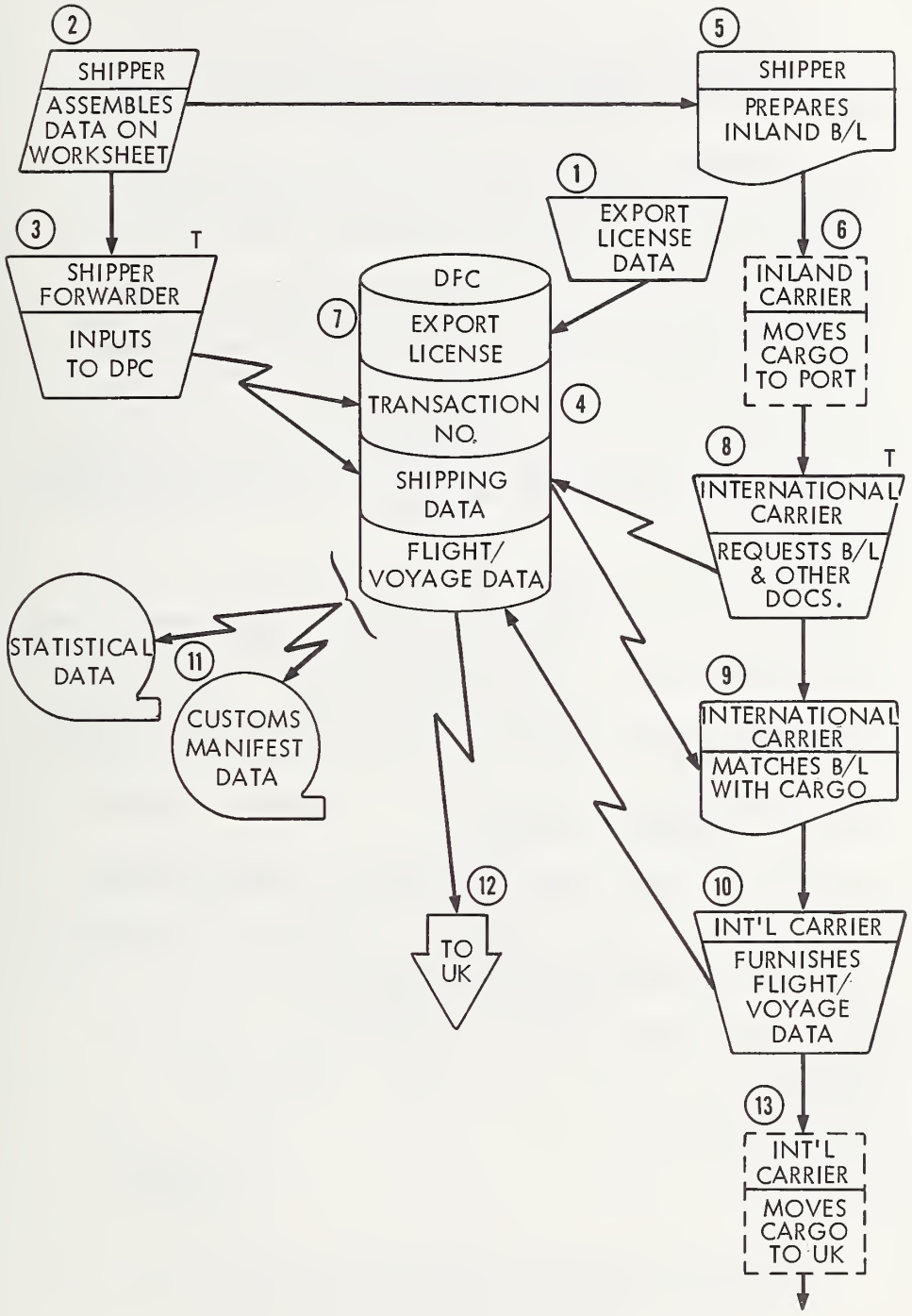


Figure 3a. Chart - US-UK Cargo Data Transmission Tests
(US Segment - Shipment via Air or Ocean)

1. Office of Export Control, upon application by shipper, issues export license and transmits data to the Data Processing Center.

2. Shipper assembles shipping data on work sheet, consisting of bill of lading, Shipper's Export Declaration (stub portion), Commercial Invoice line (delivery terms, net invoice value, and amount insured), and page 2 of Commercial Invoice.

3. Shipper (or forwarder) assigns transaction number and transmits the number and shipping data, assembled in step 2,above,to the Data Processing Center prior to or at time of shipment. Note: Input may be in stages--shipper(or forwarder) enters data as it becomes available.

4. Data Processing Center records transaction number assigned to the shipment and all shipment data. Note: All other reference numbers (if used) are subordinate to the transaction number and are cross-referenced to it.

5. Shipper arranges for shipment and prepares inland bill of lading.

6. Inland carrier picks up and delivers cargo to international carrier.

7. Data Processing Center verifies export license.

8. International carrier requests bill of lading and related documents from Data Processing Center.

9. Data Processing Center transmits bill of lading and related documents to international carrier.

10. International carrier inputs flight/voyage information to Data Processing Center.

11. Data Processing Center produces manifests for Customs, and export statistics for Census and other authorized users.

12. Data Processing Center transmits data to UK--bill of lading and Commercial Invoice.

Figure 3b. Explanation of Process Steps, Figure 3a.

US - UK CARGO DATA TRANSMISSION TESTS
J. F. Kennedy Airport to Heathrow Airport
(Entry into LACES)

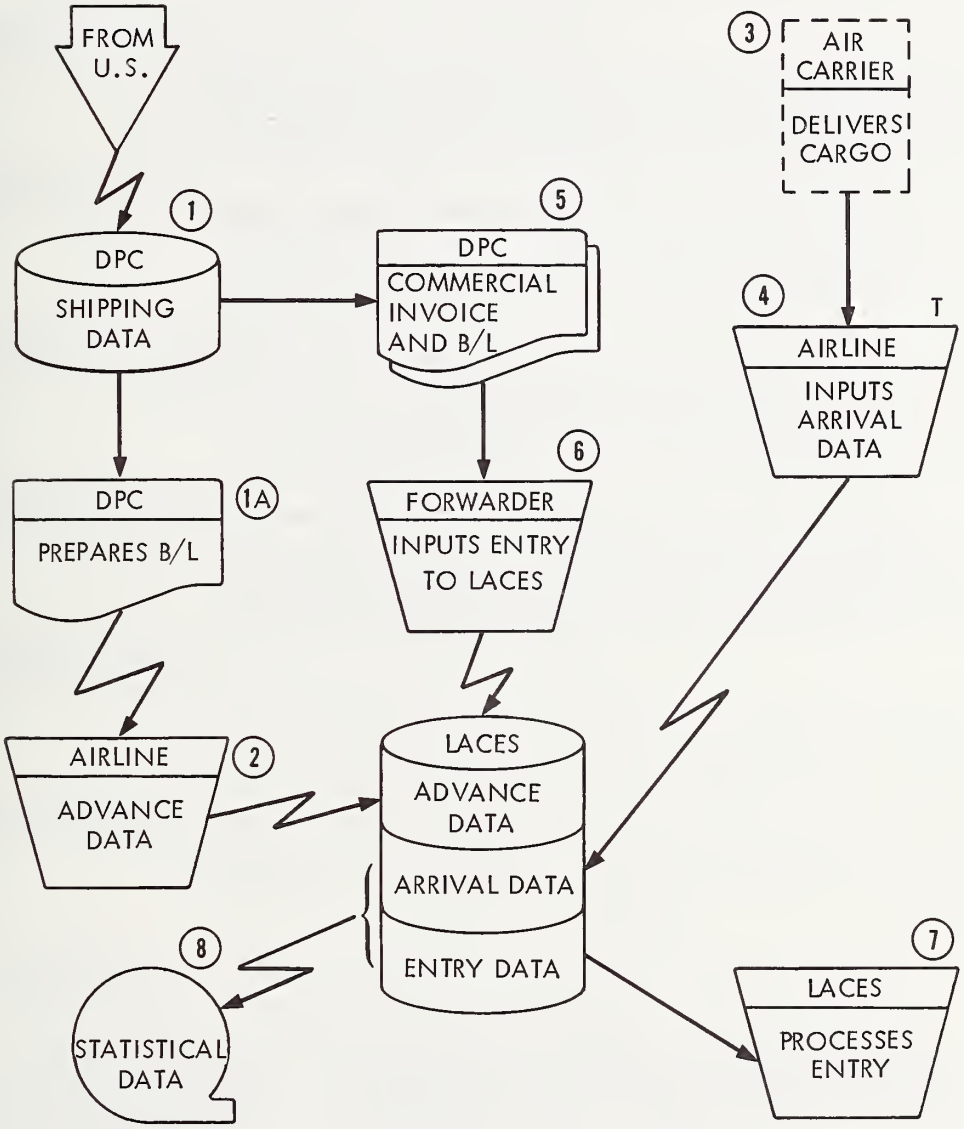


Figure 3c. Chart - US-UK Cargo Data Transmission Tests
J. F. Kennedy Airport to Heathrow Airport
(Entry into LACES)

US - UK CARGO DATA TRANSMISSION TESTS
J.F. Kennedy Airport to Heathrow Airport
(Entry into LACES)

1. Data Processing Center receives shipment data from U.S.

1A. Data Processing Center prepares B/L and transmits to airline as advance notice of shipment.

2. Airline inputs advance shipment data direct to LACES.

3. Airline delivers cargo to shed at Heathrow Airport.

4. Airline inputs arrival and location data into LACES.

5. Data Processing Center produces hard copy bill of lading and Commercial Invoice for forwarder.

6. Forwarder inputs entry documents into LACES.

7. LACES processes entry, computes duty, taxes and fees, and assigns clearance channel.

8. LACES produces necessary reports and statistics.

Figure 3d. Explanation of Process Steps, Figure 3c.

US-UK CARGO DATA TRANSMISSION TESTS - MASTER RECORD INPUT SHEET
AIR SHIPMENT - EASTBOUND

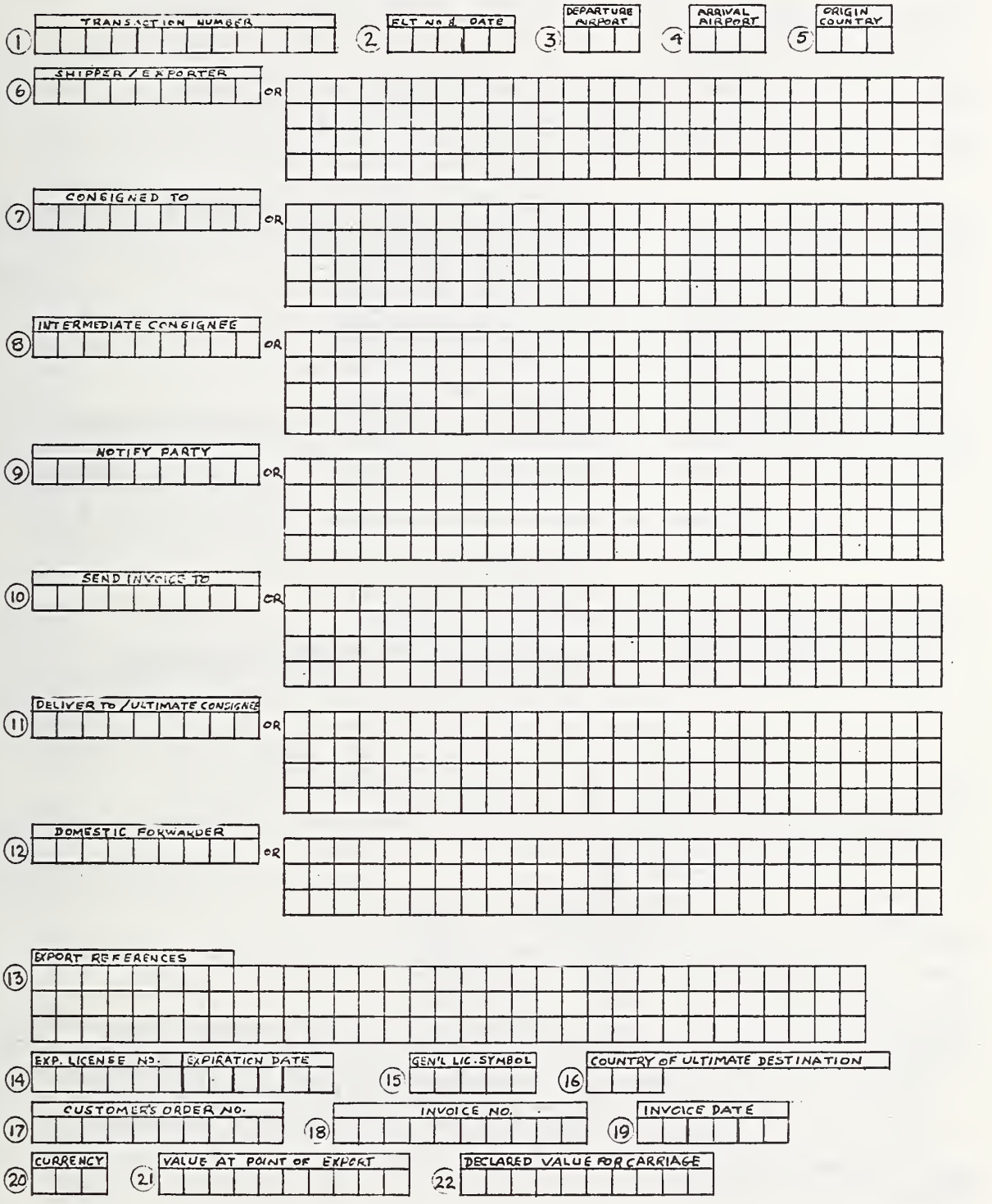
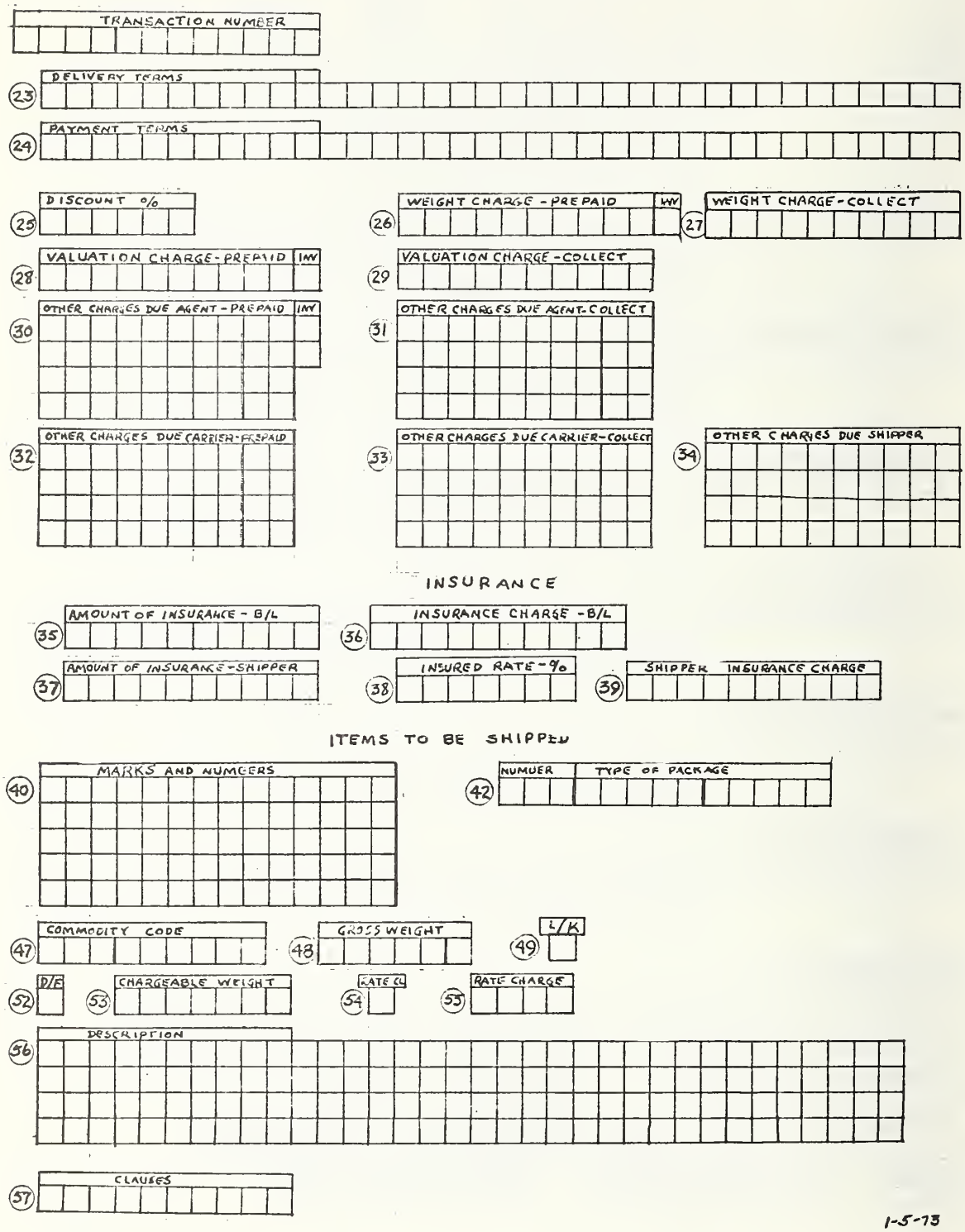


Figure 4. Master Record Input Sheets (3 pages)
(Air Shipment - Eastbound)

US-UK CARGO DATA TRANSMISSION TESTS - MASTER RECORD INPUT



1-5-73

US-UK CARGO DATA TRANSMISSION TESTS
INVOICE DETAIL INPUT

TRANSACTION NUMBER

(65) ITEM (66) GROSS WEIGHT (67) NET WEIGHT (68) UNITS (69) UNIT PRICE (70) DETAIL

(71) INVOICE DESCRIPTOR

(65) ITEM (66) GROSS WEIGHT (67) NET WEIGHT (68) UNITS (69) UNIT PRICE (70) DETAIL

(71) INVOICE DESCRIPTOR

(65) ITEM (66) GROSS WEIGHT (67) NET WEIGHT (68) UNITS (69) UNIT PRICE (70) DETAIL

(71) INVOICE DESCRIPTOR

(65) ITEM (66) GROSS WEIGHT (67) NET WEIGHT (68) UNITS (69) UNIT PRICE (70) DETAIL

(71) INVOICE DESCRIPTOR

(65) ITEM (66) GROSS WEIGHT (67) NET WEIGHT (68) UNITS (69) UNIT PRICE (70) DETAIL

(71) INVOICE DESCRIPTOR

(65) ITEM (66) GROSS WEIGHT (67) NET WEIGHT (68) UNITS (69) UNIT PRICE (70) DETAIL

(71) INVOICE DESCRIPTOR

11-10-72

The Standard Data Element System (STADES) for Controlling Data Elements used in Navy Computer Programs for the Worldwide Military Command and Control System (WWMCCS).

Robert R. Hegland

Naval Command Systems
Support Activity[1]
Washington D. C. 20374

The Standard Data Element System (STADES) for application programs of the Navy Worldwide Military Command and Control System (WWMCCS) is designed to provide the data manager with several necessary tools.

STADES allows system developers to determine the current status of existing and proposed Standard Data Elements and to report the use of both standard and other attribute data elements in their systems. Using the concept of a distributive data base, each of the major Navy organizations involved in the effort is able to have available the same information as the others.

Included in the STADES data base for each application program is information concerning not only their attribute data elements but also information on the system, the files used and their record types.

The program system used to create, maintain and query the STADES data base is the Record Association System (RAS), designed and used by the Naval Command Systems Support Activity (NAVCOSSACT) for five years prior to this application. RAS provides a wide variety of report and display capabilities that allow the system developer to easily find information from other development efforts that may be of benefit in his application.

Key words: Application computer program data elements; attribute data elements; data management; data use identifiers; distributive data base; file descriptions; Naval Command Systems Support Activity; Navy WWMCCS Standard Data Element System; Record Association System; record type descriptions; standard data element; system descriptions.

1. Scope

The Navy is currently involved in developing computer programs for use on the Honeywel 6000 series computers that have been acquired for Navy commands in the JCS Worldwide Military Command and Control System (WWMCCS). The Chief of Naval Operations (Op-91) directed that the Naval Command Systems Support Activity (NAVCOSSACT) develop procedures and a supporting computer program system to ensure that those data elements that have been standardized by the National Bureau of Standards (NBS), the Department of Defense (DOD) and other significant agencies are used in developing these application computer programs. Those

[1]The views expressed in this paper are the author's and do not necessarily reflect those o the U. S. Department of Defense.

537-860 O - 74 - 15

programs contain data ranging from financial management to command and control.

2. Contents of Data Base

The Navy WWMCCS Standard Data Element System (STADES) has been designed by NAVCOSSACT to provide this capability. The data base of STADES contains several different kinds of information.

(1) All data elements standardized by NBS, DOD and DON that have been implemented for use within the Department of the Navy.

(2) Data elements that have been promulgated by certain other activities, commands and organizations involved in using the H-6000 computers, such as the Defense Intelligence Agency.

(3) Attribute data elements (often called data use identifiers) that are used in computer programs being developed to run on the H-6000 computers.

(4) Information on the records, files, subsystems, and systems being developed for the H-6000 computers.

3. Maintenance of Data Base

To maintain this data base, STADES uses the computer programs of the Record Association System (RAS) [1] that was developed by NAVCOSSACT. These programs provide updating and retrieval capabilities as well as a wide variety of specific outputs for use in analysis.

4. Exchange of Data

STADES is currently or will soon be installed at the sites shown in figure 1. Each of these sites develops computer programs whose data elements and other pertinent information must be included in the STADES Data Base. At each site the Local Data Manager updates his local data base using RAS and forwards a transaction tape containing all the changes since the last submission to the Central Data Manager at NAVCOSSACT. These submissions are reviewed, incorporated into the STADES Data Base with errors or questions about the submissions noted by the insertion of error codes, and the new STADES data base is forwarded to all users to become their new master. Receipt of this data base signals each site to prepare a new transaction tape for submission to the Central Data Manager for review. All such exchanges are by magnetic tape.

5. Data Collection and Use

5.1. Scope of Data Collected

As shown in figure 2, the developers of each of the application programs that are documented using STADES, submit information ranging from a description of that system to the attribute data elements that it uses following the instructions contained in a detailed procedures manual [2]. In the STADES Data Base, each such entry is formatted in essentially the same way.

5.2. Format of Entries

Figure 3 shows the generalized picture of a representative entry. The Name of the entry may be the name of the system, file, record type or data element with each of the other card types within the entry providing information about what is named. The A, B, and D card types form the basic, non-repeated part of the entry. The E, F, and J card types provide information about how a particular system uses the data specified in the basic part of the entry. When the entry describes an attribute data element, card type F is included. When another

system that uses the same attribute data element is coded and entered into the data base the developers simply add another set of E, F, and J cards.

5.3. Retrieval and Report Capabilities

The RAS programs produce permuted or keyword indexes for the terms in card types A and B and allow retrieval of entries containing information specified in card types E and J. Included as reference codes in J cards is the unique reference number of the next senior entry, that is record type for attribute data elements, files for record types, etc. The RAS programs provide users with a complete picture of the data in an application system by using this referencing scheme as shown in figure 3. An example of a data element from the STADES data base is shown in figure 4.

6. Use of Standard and Attribute Data Elements

In using STADES, project developers must use any applicable data elements that have been standardized or the attribute data elements used in other systems that have already been entered into the data base. When a project developer can find nothing in the STADES data base that satisfies his needs, he submits a new attribute data element specifying how it will appear in his data base. We do not require that he adhere to specified codes and abbreviations of standard data elements in his input, processing or output reports, but, in so far as possible, his data base must contain the codes, abbreviations, and specified format of those data elements that have been standardized.

7. Benefits of STADES

The uses and benefits of the information contained in the STADES data base cover a wide spectrum.

7.1. Documentation

Outputs from the data base satisfy most of the requirements in the DOD Documentation Standard [3] for the information that must be included in the Data Requirements Document, the Data Base Specifications and the Users and Program Maintenance Manuals. These outputs also reduce the delays inherent in preparing documentation that are caused by typing and proofreading information about attribute data elements.

7.2. Managers

STADES provides a tool to managers for monitoring project development since management can review the entries provided by the project developers as the application computer programs are developed. It also allows early review of the attribute data elements of a program to ensure aderence to standardized data elements and to ensure that good data coding structures have been used. The exchange of programs and data is also encouraged through the use of this system.

7.3. Project Developers

STADES provides developers with tools for data analysis during the analysis phase of program development and, later in development, with information about the data they are using in the application program. It also provides the data codes and abbreviations from existing data elements that have been standardized and from attribute data elements used in other systems in order to avoid reinventing such codes and abbreviations.

7.4. Data Standardization

Review of the contents of the data base will provide personnel involved in establishing data standards with information about data that is currently being used and that should be standardized.

8. Conclusion

STADES is designed to ensure that managers of application program development recognize that data is a very valuable resource that must be treated in a way that will ensure its usefulnes to the greatest possible number of users. It provides benefits to program developers and users alike. Initial preparation of the different entries is rather time consuming and invades the historically sacred area of a programmer and "his" data. This area is now, however, too important to view in any other way than in the broadest possible context. The structure of data in a data base must be visible to and used by the largest possible number of other developers and conform to the format of as many standard data elements as possible. If standards are to be used, program managers must enforce their use; if standards are to be used effectively, programmers must understand the need for their use.

9. References

[1] Naval Command Systems Support Activity Document No. 88T002 UM-01, Record Association System II (RAS), October 1973.

[2] Naval Command Systems Support Activity Document Number 88T001 TN-01, Navy WWMCCS Standard Data Element System (STADES), December 1973.

[3] Department of Defense Manual 4120.17-M, Automated Data System Documentation Standards Manual, Office of the Assistant Secretary of Defense (Comptroller) December 1972.

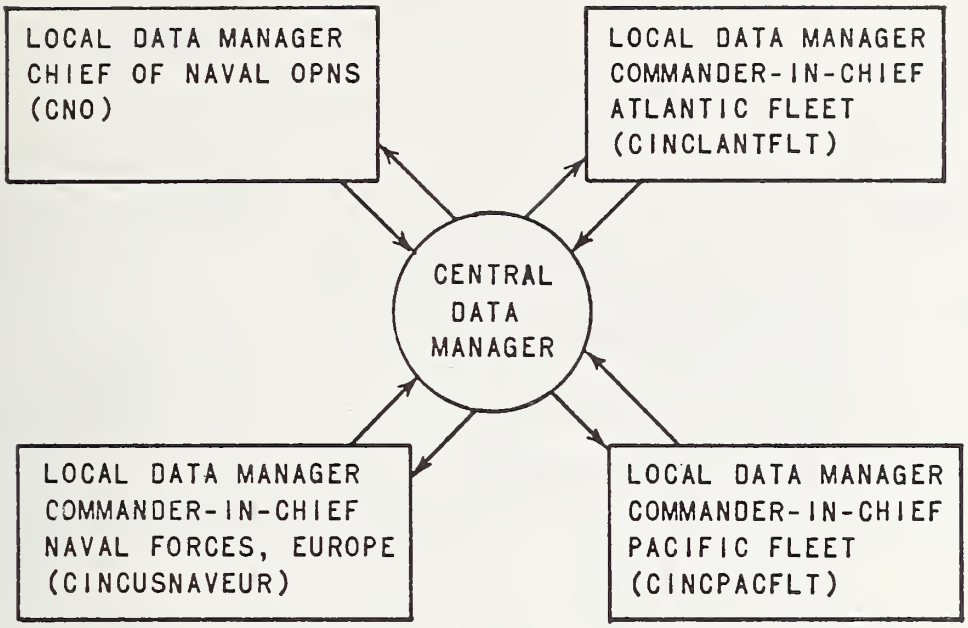


Figure 1

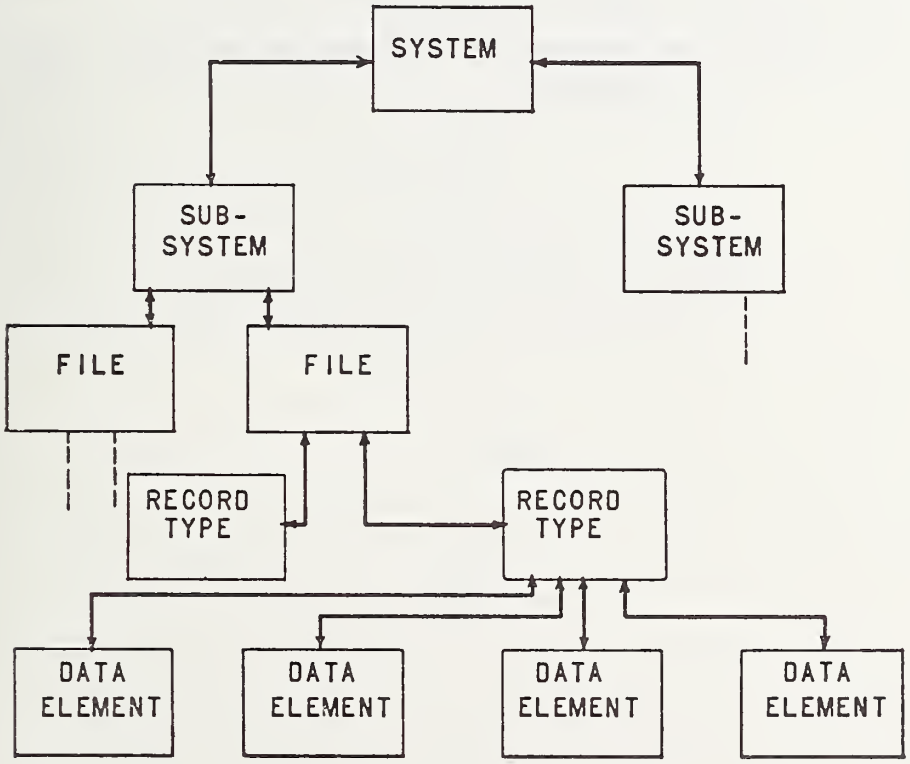


Figure 2

USES:		CARD TYPE	REF #
P	NAME OF ENTRY (SYSTEM, FILE DATA ELEMENT, etc.)	A	XXXXX
P	SYNONYMOUS NAMES	B	XXXXX
	DESCRIPTION	C	XXXXX
	SOURCE OF DESCRIPTION	D	XXXXX
R	SYSTEM #	E	XXXXX
	FORMAT OF DATA ELEMENT	F	
R	REFERENCE CODES SENIOR ENTRY SUBJECT CODES	J	XXXXX
	SYSTEM #	E	XXXXX
	FORMAT OF DATA ELEMENT	F	XXXXX
	REFERENCE CODES	J	XXXXX
	DATA ITEM NAMES, ABBREVIATIONS, CODES, EXPLANATIONS	U	XXXXX

P-INDICATES THAT WORDS IN THIS CARD TYPE MAY BE PERMUTED AND REVIEWED IN A KEYWORD LISTING.

R-INDICATES THAT THE CODES AND NUMBERS ON THESE CARD TYPES MAY BE RETRIEVED.

Figure 3

UNCLASSIFIED

PRINT OF ENTIRE DATABASE

	RIC	C D	S N	MODE IND
**********				
ALIEN STATUS	AL-IEA0			
**********				
SYNONYMOUS NAMES				
1 ALIEN-STAT	AL-IEB0			
DEFINITION				
THE STATUS OF AN ALIEN IN THE UNITED STATES UNDER THE IMMIGRATION	AL-IEC0			
LAWS.	AL-IEC1			
SOURCE OF DEFINITION				
SECNAVINST 5200.20A	AL-IED0			
ASSOC DATA ELE ID / SUBJ ID / DATA INDEX SET/ARA				
STAN / NAVY / APRV	AL-IEJA			
SOURCE OF RECORD USE / CLASS / UNIT OF MEASUREMENT / EFFECTIVE DATE				
U / 710910	AL-IEE			
FORMAT / RANGE OF VALUES				
A / 1	AL-IEF0			
DATA ITEM NAME / ABBREVIATION / CODE				
IMMIGRANT ALIEN / IALN / I	AL-IEU0			
FOREIGN NATIONAL / FORNTL / A	AL-IEU1			
FILIPINO-IMMIGRANT ALIEN / FLPNO / F	AL-IEU2			
CUBAN-IMMIGRANT ALIEN / CUBAN / C	AL-IEU3			
UNKNOWN / UNK / Z	AL-IEU4			
NOT APPLICABLE / NA / Y	AL-IEU5			
CODE / EXPLANATION				
I ANY FOREIGN NATIONAL LAWFULLY ADMITTED TO THE UNITED STATES	AL-IED0	C		
UNDER AN IMMIGRATION VISA FOR PERMANENT RESIDENCE EXCLUDING	AL-IE01			
FILIPINO AND CUBAN IMMIGRANT ALIENS.	AL-IE02			
A ANY PERSON LAWFULLY ADMITTED INTO THE UNITED STATES FOR A	AL-IE03			
TEMPORARY PERIOD OF TIME, OR FOR A SPECIAL LIMITED PURPOSE,	AL-IE04			
WHO IS NOT A U. S. CITIZEN, IMMIGRANT ALIEN OR U. S.	AL-IE05			
NATIONAL AND WHO HAS NOT BEEN ADMITTED TO THE U. S. UNDER AN	AL-IED6			
IMMIGRATION VISA FOR PERMANENT RESIDENCE.	AL-IE07			
F A FOREIGN NATIONAL OF THE PHILIPPINE ISLANDS WHO HAS BEEN	AL-IE08	C		
ACCEPTED FOR ENLISTMENT IN THE ARMED FORCES OF THE UNITED	AL-IE09			
STATES IN ACCORDANCE WITH A TREATY BETWEEN THE U. S. AND THE	AL-IE0A			
PHILIPPINES ENTERED INTO UNDER DATE OF 521213, AS AMENDED	AL-IE0B			
UNDER DATE OF 540621, BUT WHO HAS NOT BEEN LAWFULLY ADMITTED	AL-IE0C			
TO THE U. S. AS AN IMMIGRANT ALIEN FOR PERMANENT RESIDENCE.	AL-IED0			
C FOREIGN NATIONALS OF CUBA WHO MAY UNDER A SPECIAL LAW	AL-IE0E			

UNCLASSIFIED

Figure 4

Q. Are RAS and STADES operational? How long did it take to develop each?

A. Both RAS and STADES are operational. Changes and enhancements will, of course, continue to be made to further assist the users of both systems. RAS has been operational for 5 years but various new capabilities have always been under development. Development time would probably total about 8 man-years. The STADES procedures and two supporting programs that are now included in RAS took about two man-years to develop.

Q. Are the procedures for STADES available to other DOD WWMCCS users?

A. Copies of the RAS and STADES manuals are available to anyone who will write for copies on their letterhead stationary. The RAS programs are available to WWMCCS users if installation is approved by CNO Op-91. Correspondence should be directed to:

Original Copy	Information Copy
Commanding Officer, Naval Command Systems Support Activity (attn. Code 70.3) Washington Navy Yard Washington, D.C. 20374	Director, DON ADP Management (Op-916) Washington, D.C. 20350

Q. How closely does STADES conform to the DOD catalog of Standard Data Elements?

A. The STADES Data Base includes all the information about Standard Data Elements required by DOD and includes all the Standard Data Elements approved by DOD and implemented within the Department of the Navy.

Q. What coordination is underway between the armed services?

A. The DOD program provides for this coordination. By using STADES we hope to identify more elements that should be standardized and expedite this process within the Navy.

Q. How do you eliminate elements that are no longer used?

A. The command having maintenance responsibility for the program must maintain the entries in STADES.

Q. Why do you tolerate synonyms?

A. The STADES Data Base includes synonyms in order to reflect what is actually being used in data systems. While we want to reach the point where one name is used, that certainly is not the case in current systems. We also include as synonymous names the COBOL name of the file, record type or data element; other significant keywords that can be permuted with the name in the AØ Card; and some other information useful to analysts who are searching for information about data.

Q. Couldn't STADES automatically generate file and record type definitions from the related data element definitions?

A. For the file and record definitions, we want more than a duplication of the information in the related data element definitions. What should be included is an overview of all the data contained in the file and record type. We hope to eventually have a program that will compare a COBOL source tape with the entries in STADES to ensure that all the pertinent information has been correctly reported to STADES and to ensure that it is still current.

Q. How do you cost out the benefits derived from standardizing a data element?

A. I have seen no comprehensive costing of standardizing individual data elements. Their benefits really are obvious, particularly in third generation computers with shared data bases, considering the expense of redundant reporting for different systems and the time involved to change from one format to another.

Q. When a new Standard Data Element is adopted, how do you ensure all Navy ADP centers convert to the new standard.

A. As with most other implementations of standards, conversion only takes place when the system is redesigned or has new interfaces. Primary factors to consider in implementing all standards are their long term benefits and the simplifications that the standard offers to the overall ADP community.

FROM DISCORD INTO HARMONY

WILLIAM T. KNOX

My theme is "the challenge for fast, comprehensive standardization of data elements in information processing is herewith placed in the perspective of human communication needs." Or in the vernacular, "why it's going to take lots of hard work and a long time to get there."

As I read the abstracts for this meeting, I was struck by their endless variety. It takes a genius to embrace them with a common element. Being no genius, I failed in the attempt. But as I fell back into the everyday world of my problems with data elements in information processing, I began to sense some of the congruences of our topic with similar problems in other areas of human communication. And to recognize my personal involvement in data element standardization.

I was reminded of the less-than-adequate formats in which I get the most crucial data--NTIS' income and costs. And the confusion caused the buyer and NTIS when an eight-digit order number is transposed or errs. We do have a real, gutsy problem!

But let's not think we are the first to have this problem. From the tower of Babel to the immense library for printed publications to the mag tape files of interest to this group, people have been plagued with discord within the media--lack of media standards, if you will.

Some philosphers have wondered if the discord wasn't a divine implant to keep mankind from becoming too knowledgeable and thereby, too powerful. You recall the Old Testament version: "Behold, the people is one, and they have one language...and now nothing will be restrained from them which they want to do. Let us go down, and there confound their language, that they may not understand one another's speech." And from time to time we're helped along by such events as the burning of the library at Alexandria or the erasures on certain tapes.

Progress has been painfully slow and still limps along, hobbled mainly by the human source of the discord and the absence of broad societal acceptance of the need to work for harmony. In other words, people generally don't expect much better than what they have. Who cared, other than the librarians, that the English-speaking countries finally agreed about 5 years ago on common book cataloging rules--after centuries of different rules and millions of catalog card entries. Has it made any difference to you, as a library user, that there now is an Anglo-American standard?

My challenge to you today is to generate the power and thrust within society to make sure that the need for harmony in processing data bases and other machineable files is--contrary to past history--adequately and quickly met. But learn from the lessons of history, following Alfred North Whitehead's advice on how to avoid extinction.

It takes power to get anything done. And in an economic society, it usually takes economic power. Although I admit that emotional power and the power of logic, also, from time to time, are the mainsprings of action. But economic power is the best power source in our society for something as hard to grasp by the layman as data element standardization.

Let's talk for a few minutes about the human-derived nature of the discord. Our meeting prospectus glibly states that "unless the transmitter and the intended receiver agree on the meanings of words and other symbols or codes, there can be no transfer of information". But getting two people to agree on the precise meaning--all the denotations and connotations--of a word is exactly the problem in much of our day-to-day existence. And sometimes, as in agreements to stop war, action only takes place when the words have <u>different</u> meanings for the warring camps.

In the end, we rely on the flexibility of the human mind to correct errors in format, in phrasing, even in content. For example, in the erratic indexes in books. Not good, but there's nothing better.

But when we turn to mechanical transmission of signals between mechanical devices the problem becomes a challenge to our scientists and engineers. They can move out of the murky mists of meanings into the bonny brilliance of bits. Not entirely, of course, but in most cases--enough to vastly simplify the problem--compared to that faced by the first-grade teacher and her successors.

Can you marshal evidence enough to persuade our society that an effort commensurate with that spent on language standardization is equally worthy on machineable data standardization? Our schools spend about $70 billion a year, and it is fair to guess that 10%, or $7 billion a year is spent on language standardization. And still--and still, everyone is "laying" all over the place. How much must be spent to get the machines to talk to each other unambiguously?

Let me raise yet another question--do we really want the unfettered, unlimited data transmission which is such a logical goal, or are we better off with errors and inconsistencies, with ambiguities and omissions which slow down the transmission? There is a parallel in the verbal communication world. Let me digress for a few minutes. You might call these some splashings from my pool of thoughts.

Within a generation we have moved into an era in which the average citizen suffers, not from scarcity, but from an over-abundance of information and of communication machines. Peter Drucker has called our times "The Age of Discontinuity." Nowhere is this more evident than in information and communication.

We are in trouble, like that experienced by all living beings and even non-living systems, such as corporations, when flooded with unusually large quantities of information. The person (or system) becomes unstable. It may shrink from contact with other information. Its response to new stimuli becomes erratic and sometimes irrational. Wild gyrations occur, and in the extreme case, paralysis of thought and action result. In individuals, we call it a nervous breakdown. We call the behavior of the Stock Exchange "erratic".

We have learned, primarily due to Norbert Wiener, that the effective functioning of all dynamic systems (including people) is critically dependent on the proper balance between 1) the type and quantity of information flowing through the system's communication channels, and 2) the control and response mechanisms. A new profession--control or systems engineering--has been created to handle problems such as these.

Large, expensive, information-dependent systems, such as airplanes, oil refineries, the telephone network, and military command-and-control systems, are carefully designed and engineered to ensure that new, incoming information can be effectively digested and that the correct response will be given. Control instruments automatically avoid over reaction, avoid swinging wildly from one course to another, and allow messages to enter the system only when the system can effectively handle them. Such controls are absolutely necessary for the proper functioning of these systems.

Although people have developed over the millenia of evolution some remarkable sensing, communication, and information processing devices--especially the eye, ear, brain, and nervous system--these have not changed in the past 10,000 years. People are, therefore, trying to live in an over-rich information environment, surrounded by unimagined communication machines, with their own personal information processing, control, and communication capabilities biologically adapted to the pre-civilized eras.

Most people managed--until about 1940--to adapt to the gradual increase in the amount and variety of information sent out via an increasing number and variety of communication machines. They adapted by learning to read and write (few have learned to listen!), by inventing more compact ways of expressing information (e.g., symbols and generalities such as the laws of science), and by making more time available for communication through increased productivity in other areas.

Social institutions were also created whose major function was to provide a means for communication and a generalized response to new information--the church, the school, the university, government, the courts, civic groups, etc. Those people who were not personally capable of effectively operating in the increasingly information-oriented environment were still able to work and live productively by relying heavily on their institutions. People and their institutions thus kept a reasonable balance between the information flow and the control of and response to the information.

The post-World War II explosive growth of both communication machines and information, coupled with increased industrialization, destroyed that balance for many people. Some social institutions, such as the church, have lost their credibility as useful sources of generalized information, and thereby have become less useful as social mechanisms for control and response.

No other society moved so quickly into the communication era, the knowledge-oriented world of the future, as did the U. S. after the 1940's.

The results have not been all good. With the declining influence of accepted social organizations for screening, filtering, condensing, and responding to information, more of the burden has been shifted to the individual. A major reason for the declining influence of these organizations is, of course, due to the inadequacies of their out-dated internal communication system, which cannot respond quickly enough to meet people's expectations. Our criminal justice system is widely recognized to be collapsing due to information overload on the ancient channels of communication in the system. The organizations show the classic symptoms of system instability--under-reacting or over-reacting unpredictably and concentrating on the familiar kinds of information instead of recognizing the intrusive forerunners of the future (how long the established organizations ignored the deteriorating environment!). They also throw up barriers to further communication and further information input; the bearer of bad news has never had an easy lot--frequently losing his head or job.

Passing the burden from the organization to the individual has aggravated his problems. Thrown on their own limited personal communication capabilities, many people have given up trying.

People barely skilled in reading and writing are being shoved into a ubiquitous audio-video technology for communication for which they have no training at all. The unusual, abnormal, and exotic are stressed to attract attention until they seem the norm. Riots and other societal aberrations are pictured while they happen, even in faraway places. The distortion of reality and significance by communication machines employing pictures and sound is not easily grasped; the human's inborn communication system prefers to believe otherwise. It reacts viscerally, fast, emotionally to pictures, color, and sound. Although necessary for man's survival during evolutionary times, these reactions tear down today's complex societies which demand rational, considered actions. The turbulent, destructive Reformation was propelled by the products of the newly invented printing press. The rapid, uncontrolled development and exploitation of today's more potent communication machines makes more turbulent, destructive times likely for our society.

In sum, our cleverness has created communication machines which threaten our survival. Our capabilities to generate information and to move it about in attention-demanding forms have far outstripped our capacities for its effective use.

Did anyone's feet or legs get uncomfortably wet from those splashings? Likely not, because I've hinted that we should not suboptimize the data transmission system. That we should look ahead and make sure that the faster growth, or higher efficiency, or different product qualities which we will help our organizations achieve by data element standardization are, in themselves, worthy on a larger scale of human values.

Some in the rear of this classroom have already raised their hands. If they can restrain themselves for another two minutes, I'll conclude with some summary opinions:

1. Data element standardization is terribly important to the future well-being of our information-dominated society. It's a worthy cause, as far into the misty future as I can judge. If I didn't feel this way, I wouldn't have supported NTIS becoming the central clearinghouse for the new ANSI report numbering system. At our current level of confusion, I see no need for divine intervention to further muddy the waters.

2. Progress in implementing data element standardization is very slow. Don't be discouraged. Some years ago I checked into some of the information processing activities of the military departments, and found that the personnel records of the three military departments had earlier been machine incompatible. But the redoubtable Secretary McNamara had created a task force to make them compatible. The task force reported after a year's study that the records could not, in fact, be made compatible. Mr. McNamara simply ordered that they be standardized within 4 months. And so they were. Except that I've heard recently that they never "really" were, and are not today. It's slow going!

3. Data handling systems will always be non-standard where they interface with human users. We're ornery critters! Our standard NTIS accounting system, for example, is not completely acceptable to some division heads, so they create their own--for <u>their</u> purposes. Standardization of data elements may standardize the management information to an unacceptable degree in the eyes of those very people--the managers--on whom an organization depends for its major initiatives and innovations. We cannot lightly overlook this fact.

4. The most useful approach will continue to be area-by-area, discipline-by-discipline. Those wanting to standardize the physical characteristics of the record have a different

set of motivations from those wanting to standardize the record contents. Those concerned about telecom protocols for data transmission have different problems from those concerned about descriptive cataloging.

5. So rather than one, all-embracing data element standardization effort--which would have to look for a power source interested in all data systems, it will be most fruitful to develop a concentrated power and thrust in limited fields. This should also make it easier to develop the economic rationale which will provide the real power needed. So far the power seems to be based on a logical rationale, in a professional sense.

   It must be given a quantified economic rationale. For example, the advent of networks providing access to bibliographic data files has created a greater reason for data element standardization in that field. But I have yet to see $ values assigned to the benefits that the producer, middleman, and user would get from various degrees of standardization. You are the ones who should be giving the answers to such questions.

6. Finally, I suggest that it will be easier to develop the economic case for data element standardization to the extent that we develop the concept that data--and indeed all information--is property. The more I struggle to advance the application of information technologies, the more critical the "information is property" concept becomes.

The Constitutional Convention recognized the property aspects of information when it established the patent and copyright clause in the Constitution. Information created by the human mind thus was entitled to some of the property rights enjoyed by landholders and bankers. Our private sector has, ever since, utilized the patent and copyright mechanisms to promote the creation of intellectual property and its utilization within society. Bear in mind that intellectual property rights are given in exchange for public access to and use of the property. They are not devices for restricting use of the property to the owner. Nor do they last indefinitely unlike real property rights.

One of the signs of the Age of Discontinuity is the increasing dominance of information--deliberately created information--as the primary consumer good in our society. Economists still prefer to ignore this development; perhaps because they don't know how to handle it in economic theory. They continue to treat information as a free good. Some with whom I have talked have shown awareness of this ever-widening gap between economic theory and practice.

Another sign is the increasing role of government--national, state, and local--as the creator of information. I call to your mind the Census Bureau, Federal Reserve Bank, and the large national laboratories, such as the Bureau of Standards. And nearly all government organizations create useful computer software packages.

Here we come to the crux of the issue. Although in the private sector the individual creators of intellectual property obtain property rights to it, their counterparts in government are usually denied this creative right. And the problem is compounded then by the traditional government position that its information belongs to everyone. Such a posture surely conflicts with the property concept developed for all kinds of informational products in the Constitution. The resolution of this issue will call for the scrapping of many traditional--and sometimes emotionally cherished ideas on the part of government employees and managers. It will equally demand the development of a new set of criteria and procedures recognizing that information is <u>property</u>.

This will then be of great help to you in the basic charge I have laid on you--to develop the economic, property-based rationale for data-element standardization.

Right on!

537-860 O - 74 - 16

Standardization Problems Involved in Interactive Direct Access to Large Data Base Systems Using Remote On-Line Terminals

Robert M. Landau

Science Information Association
3514 Plyers Mill Road
Kensington, Md. 20795

The number of large data bases in the field of science and technology (S&T) being placed in machine readable form has accelerated rapidly in the last few years. There are now over ten million bibliographic records in the field of S&T which are increasing at the rate of over three million records per year. Most of these large data bases started being put into machine readable language in the late 1960s. No significant effort has been made to put these data bases in standard format or have standardized data elements within each record. Further, there has been no serious, concerted national effort to reduce the high redundancy between these large files. Most important, as the new on-line systems are coming into being, there has been little effort to standardize on the English-like languages for the on-line users. A large number of potential on-line users will use such systems only if they are provided an easy-to-use command language. Those involved in creating such languages for proliferating on-line bibliographic search systems should be encouraged to use common commands, conventions and procedures.

Key words: Bibliographic references; data bases; data elements; interactive searching; on-line systems; search strategies; standardization; user training.

1. Introduction

There are approximately 100 data bases in the field of science and technology now commercially available in machine readable form on magnetic tape.[1] Fourteen of the major data bases are from the following organizations: (Numbers are thousands of records increase/year)

1. Chemical Abstracts Service (Chemical Abstracts Condensates) (360)
2. National Agricultural Library (Cataloging and Indexing System) (120)
3. National Technical Information Service (56)
4. Educational Resources Information Center (ERIC) (27)

[1] Schneider, John H.; Gechman, Marvin; Furth, Stephen E., eds. Survey of Commercially Available Computer-readable Bibliographic Data Bases, American Society for Information Science, Wash., D.C., 1973.

5. Engineering Index (COMPENDEX)(85)
6. The Institution of Electrical Engineers (INSPEC)(Information Service in Physics, Electrotechnology, and Control)(150)
7. National Library of Medicine (MEDLARS)(150)
8. National Library of Medicine (TOXICON)(50)
9. Pandex (250)
10. American Institute of Physics (Searchable Physics Information Notes (SPIN)(30)
11. Biosciences Information Service of Biological Abstracts (BA Previews)(240)
12. Science Information Exchange (30)
13. Exerpta Medica (250)
14. Institute for Scientific Information (ISI)(374)

Thirteen of these data bases are increasing approximately 1.8 million records per year. The fourteenth data base, ISI, is increasing about 374,000 more source document records (representing 4 million citations) per year. Another 85 or 90 data bases include approximately another one million records increase per year. Thus the total increase of records is well over three million per year.

The average length of the bibliographic records in these data bases is about 250 characters per record in about eight to fifteen fields. Most of the data bases (with the notable exception of ISI, which uses citation indexing) contain index terms as one of the major fields for subject searching. Four of the fourteen (NTIS, TOXICON, COMPENDEX and SPIN, totaling about 200,000 records increase per year) contain abstracts. The records that include abstracts average around 1,000 characters per record. Thus, it can be estimated that there are approximately 625 million characters per year (2.5 million records x 250 char/record) increase in the bibliographic records, and approximately 375 million characters per year (375* thousand records x 1,000 char/record) for the records that contain abstracts, for an estimated total yearly increase of about one billion characters. The real increase is much less (perhaps as high as 50%) because of duplication of citations in the various data bases.

Retrospective and SDI searches in most of the above-listed data bases are available from over twenty organizations, such as the University of Georgia, Illinois Institute of Technology, Research Institute, North Carolina Science and Technology Research Center, and others. Numbers 1-8 above are now available in large direct random access interactive systems from such organizations as System Development Corporation, Lockheed, Battelle Memorial Institute, Informatics, Inc. and Lehigh University. There are other large systems which are not commercially available operated by a number of government agencies, including the Defense Document Center, Patent Office, NASA, AEC and EPA. There are also a number of large random access on-line systems in many universities; examples of these include: SPIRES (Stanford Physics Information Retrieval System or Stanford Public Information Retrieval System) at Stanford University; SUNY, MIT, Northwestern, and the LEADERMART system at Lehigh University. In addition, IBM has available a software package called STAIRS (Storage and Information Retrieval System) which provides multi-terminal interactive retrieval for large bibliographic data bases. Unfortunately, all these systems have different commands, search strategies and user conventions, which has increased the difficulty of users to learn the various systems' logic, nomenclature and command structure. A project underway at the Stanford University is studying ten of the major on-line bibliographic retrieval systems' logic, nomenclature and command structure.

*(Includes 200,000 records of the 14 data bases listed above plus 175,000 more records from the 85 or so smaller data bases listed in reference 1.)

Described above are a number of large systems, including dozens of large data bases containing tens of millions of records stored in a number of very large, random access, computer systems scattered all over the United States. The information in the various data bases which the author will characterize as the "information bank" is already being made available by local dial-up through communication networks to hundreds of people in hundreds of locations. It is clearly predictable that within the very near future (one to three years), this information will be available to at least a million people through hundreds of thousands of terminals in thousands of organizations - corporate, governmental and educational.

There are four major areas of concern regarding these systems vis-a-vis standards: (1) the standardization of the data elements in terms of content or meaning as well as labels; (2) the standardization of data elements so that reference record redundancy between the various data bases can be easily eliminated;(3) the agreement on standard search techniques and methodoligies; and, most important, (4) the standardization of the search language, nomenclature, conventions and procedures.

There have been a number of studies and efforts made by organizations both within the United States government and the various standards groups to settle on standard data elements. The various efforts have met with indifferent success. One need only to cite the efforts of the Library of Congress, the efforts by a number of COSATI panels, the recommendations of the SATCOM report, etc., etc., etc. The National Bureau of Standards has been assiduously working on this subject with standards groups for many years. Although progress in this area is slow and should be encouraged, it is probably being made at a speed as fast as can be expected in this pluralistic environment with conflicting organizational goals.

A number of studies have been made within recent years about the excessive and expensive overlaps in the secondary services that provide bibliographic information about the major items published in the fields of S&T. It was concluded several years ago that there were significant overlaps between such machine readable data bases as the Chemical Abstracts Condensates, the Biological Abstracts, Engineering Index, etc. Standardization of key data elements in these data bases is essential in order to eliminate redundant records. A major concern in this area is the numerous efforts taking place in various government agencies to, in effect, reorganize a number of the major data bases into yet new groupings to satisfy an operational or organizational need. This trend is important and will not be easily changed; however, it should be pointed out that it may be self-defeating because what a user really wants is all of the references relevant to a query no matter what grouping of data bases the results may be taken from. Therefore, the emphasis should be on the ability to quickly put on-line interactive questions to a series of data bases which would yield them only the relevant references from each data base. The grouping of the references within the data base whether by subject or organizational requirements is really not relevant to the ultimate user. The trade-offs between these two trends ought to be determined and analyzed by those interested in the standards area.

A large number of new search techniques based on interactive search logic has been developed within recent years. Unfortunately, however, most of these experiments and conclusions were based on those procedures and logic best suited to relatively small data bases of five or ten thousand records. Most of the system features and procedures found to optimize performance in these systems do not apply to the large data bases containing hundreds of thousands or millions of records. Although this is an important area for further research and one which standards people should be aware of, there is little that could be gained at this point of development in terms of standardization (in the literal sense) of such procedures. However, a number of conventions are being developed in a very disparate

manner and those groups involved in standardization must provide significant help by attempting to identify and regularize a number of these procedures.

An area seen by the author to be a crucial one is the variety of command languages being developed by a number of groups for the user who is searching interactively in large data banks through remote terminals. It has been discovered that, although a few people can quickly learn such user languages based on English-type commands, most people cannot. There are four levels of learning which the potential user must surmount: physical, logical, intellectual and emotional. The physical level involves the various problems surrounding the actual use of the equipment including how to type, the physical location of the command keys, the setting of the right switches, etc. The logical level involves the understanding and use of the logic of the system to obtain the particular results desired and the understanding of the terms used in the command language. The intellectual level involves the understanding of the structure and the contents of the particular data base being searched. The emotional level is a complex personality factor which involves the cognative acceptance by the user of this new means of access. Before a user can become a competent and effective direct-access inquirer, he must be adequately trained at all four of these levels. In many cases, this is a formidable, if not hopeless, task. However, in other cases it can be achieved literally in a matter of a very few hours. Most learners require a few days to a few weeks to master these four levels. Because of this multi-level problem, it is obvious that, if we were to be able to standardize on a few simple commands, the degree of training and amount of resistance on the part of users would be decreased dramatically. It is felt that until such action is taken, there will not be a mass market for the use of interactive access to large bibliographic data bases in the S&T field. Those involved in standards are, therefore, urged to become more aware of this problem, consider the alternatives and take steps to assure the simplest possible set of user command conventions as is feasible. This is one of the major goals of the Science Information Association.

A Data Manager Looks at the Development of the Colorado Water Data Bank[1]

Robert A. Longenbaugh[2] and Norval E. McMillin[3]

Department of Civil Engineering
Colorado State University
Ft. Collins, Colorado 80521

The Colorado Water Data Bank Project is developing a central computerized data base for capturing, storing, and retrieving all types of water data collected in the State of Colorado. The three-year project includes development of all software programs, operational procedures, program documentation, and user manuals for capturing both current and historic records. The project is funded by the Colorado Division of Water Resources (DWR) and represents a cooperative venture between DWR and Colorado State University.

The first task for the Colorado Water Data Bank Project was to evaluate and choose a Data Base Management System (DBMS) to be used for the project. Following selection of the DBMS, a major effort was required to develop record formats and file structures which were compatible with the DBMS and would still provide efficient and economic storage with maximum retrieval flexibility. External programs, written in COBOL and FORTRAN, interface with the DBMS to perform editing and updating of data, as well as preparing sophisticated reports.

A system was developed for capturing, editing, reformatting, loading and retrieving the desired water data and is identified as the Colorado Water Data Bank System (CWDBS). A general flow diagram and a brief description of the system is presented.

Key words: Colorado; data standardization; Data Base Management System; MARS VI; water; Water Data Bank.

1. Introduction

The Colorado Water Data Bank Project was initiated July 1, 1972. The Project is funded entirely by the State of Colorado, and consists of a developmental and initial data capture phase to be completed in the first three years, followed by a continuation phase where additional data will be added from year to year as new records become available. The Project represents a cooperative endeavor between the Division of Water Resources (Colorado State Engineers Office) and Colorado State University.

---

[1]Authorization for publication granted by Colorado Division of Water Resources.

[2]Project Leader, Colorado Water Data Bank Project and Assistant Professor, Civil Engineering Department, Colorado State University, Ft. Collins, Colorado.

[3]ADP Programming Supervisor, Colorado Water Data Bank Project, Department of Civil Engineering Colorado State University, Ft. Collins, Colorado.

The Division of Water Resources (DWR) has a contract with Colorado State University to provide the computer facilities (CDC 6400) and to use its technical expertise to develop and implement the data bank system including programming, documentation and user procedures for capturing, storing and retrieving the data. The Division of Water Resources is responsible for capturing the historic and current records in a machine-readable format.

Colorado's rapidly increasing population and the corresponding development of competition between the agricultural, municipal, recreational, and industrial water users has magnified the water administration problems. Currently, the Division of Water Resources is required to administer both ground and surface water within the existing laws. Changes in administrative policies are continuously being evaluated to provide complete management of both ground and surface water supplies to minimize water shortages and to provide maximum beneficial use of Colorado's limited water resources. Different types of data are required for each administrative decision.

Recent legislation required that the Division of Water Resources also provide different types of data and administrative decisions to be incorporated into land use planning. A comprehensive land planning bill is being prepared in 1974 by the State Legislature and will require certain water-related data to be incorporated into comprehensive land-water use plans. Extensive use of the Water Data Bank is expected by federal, state, and local water administrators, as well as engineers, lawyers, economists, planners and the general public.

The primary reason for establishing the Colorado Water Data Bank was to provide at a central location all types of water data. The need for rapid administrative and management decisions requires that water data be readily accessible in a form which can be incorporated into simple or complex computer programs. The need to cross-reference different types of data also requires that the records be compatible and available at a central location.

Prior to establishment of the Data Bank Project, most of the data had been processed manually with data storage consisting of handwritten ledger books, keypunched cards, and in some cases, data stored on magnetic tape. For example, gaging station records were available from the U.S. Geological Survey's data bank in Washington D.C. and climatological data were available from U.S. Weather Bureau publications or on magnetic tape from the Weather Bureau Record Center at Asheville, North Carolina. Other examples include Colorado water well data stored on magnetic tape in the State Engineers Office and records for historic diversions, water rights, and descriptive data on dams which exist as typewritten or handwritten records in the State Engineers Office. The incompatibility of the data and the major time required to access and retrieve data from the many sources is quite apparent to those using the data.

2. Data Description

The nine different types of data to be incorporated into the data bank in the initial phase are illustrated in figure 1. These include information on climatology, gaging station records, ditch diversions, reservoirs, dams, water rights, wells, stock ponds, and eventually water quality. The lines connecting the circles in figure 1 indicate that cross-referencing between the connected types is needed. For example: In evaluating the adequacy of a water right to provide water for a proposed new subdivision, it is necessary to evaluate the water right as well as the historic amount of water which has been diverted. Development of cross-referencing identification numbers will be described later in this paper.

The oldest water right in Colorado dates back to 1852 and numeric records on the amount diverted have been kept since 1881. Table 1 indicates the magnitude of the different types of data which are to be entered into the Water Data Bank. The decision was made by the Division of Water Resources to place 30 years (1942-72) of historic diversion, reservoir, and climatological data into the Data Bank. To provide complete and accurate records it was necessary to develop the capability for capturing current data beginning with 1973. Only those types of data which had been recorded in the past were to be included in the data base; however, the system was to be capable of handling new types of data at a later date.

Table 1. Types of data to be initially placed in Colorado Water Data Bank with indication as to whether it is descriptive, numeric, or both.

Type of Data	Descriptive Data	Numeric Data
Water Rights	37,000 Records	--
Reservoirs	2,200	30 years historic monthly + current
Dams	2,500	--
Gaging Station	530	Daily values for entire record
Diversions	12,000	30 years of historic daily + current
Wells	75,000	--
Climatology	248	30 years of historic daily + current
Stock Ponds	12,500	--
Water Quality	Unknown	Unknown

Methods had to be devised for the capture and processing of both historic and current records considering data quality, economics, time requirements, and including the necessary identification system to provide flexibility in access and retrieval. A more detailed description of the overall data bank system, including procedures, follows in a later section.

A review of the material in table 1 indicates that both numeric and descriptive data exists. The format of the descriptive data for a well is considerably different than that required to describe a dam or a water right. The wide variation in descriptive data required special consideration in selecting record formats to be used in the Colorado Water Data Bank.

The MARS VI DBMS will handle only fixed length records and thus several different subrecord types were defined which allow processing of what might be considered a variable length record. In the case of diversion records, the numeric data for some ditches were recorded daily; however, in other instances the amounts were recorded periodically, or in some cases, lumped as monthly values. Due to the legal requirement that the Data Bank must be able to exactly reproduce the observed historic records, it was essential that a record format be devised which would allow retrieval of actual observed amounts. To satisfy this legal requirement, strict control on data accuracy and number of records in the Data Bank is maintained.

The amount, type and format of water data varies from state to state and thus a standardized water data bank for all states is not feasible. Although the specific elements to be included in a record format may vary, it is felt that the logic and philosophy which are the basis for the CWDBS could be applied to other states.

3. Selection of Data Base Management System

From the outset, it was apparent the Colorado Water Data Bank Project would require a Data Base Management System (DBMS). Because of the time frame specified in the contract, it was not feasible for the Data Bank Project to write its own DBMS and a search of private vendors having available software was undertaken.

The selected DBMS had to be available for the Control Data Corporation (CDC) 6400 computer owned by Colorado State University. This computer system had at that time 65,000 decimal words of central memory; five 841 disk drives with public packs, three of which could be used for permanent file storage; and five 7-track tape drives.

Four candidates for use as a DBMS were found. They were: (1) Remote File Management System (RFMS) from the University of Texas at Austin; (2) SYSTEM 2000, marketed by MRI Systems Corporation of Austin, Texas; (3) MARS VI Version 2.1, marketed by Control Data Corporation; and (4) SISTER, marketed by Temple University. Two of the systems, RFMS and SISTER, were judged to be impractical because of the extensive programming effort required to make them operational. An extensive evaluation of SYSTEM 2000 and MARS VI was carried out by personnel at the Colorado State University Computer Center. The evaluation is described in detail in a project technical report by McMillin [1].

The MARS VI DBMS was chosen over the SYSTEM 2000 DBMS. In general, it was felt that the MARS VI DBMS more closely adhered to industry standards. When the Colorado Water Data Bank Project began operation, on July 1, 1972, the Conference on Data Systems Language (CODASYL) Data Base Task Group (DBTG) "April 1971 Report" was barely a year old. Personnel on the project felt that there was a need for a standardized data base management system. The DBTG Report proposed such a system. While MARS VI certainly did not adhere to the specifications of the report, its file structure was somewhat compatible. Control Data had made a corporate committment to develop and implement a DBMS which was compatible with the DBTG recommendations to CODASYL. This product is known as QUERY/UPDATE.

The MARS VI DBMS has a data base structure which allows user programs to access the data base either through the MARS VI DBMS or by using an entirely external program. This was an important factor in the choice of MARS VI.

4. Characteristics of the MARS VI DBMS

There are several characteristics of the MARS VI DBMS which should be discussed in order that the reader might understand the functioning portion of the Colorado Water Data Bank System (CWDBS). These characteristics have a bearing on the internal structure of data in the Colorado Water Data Bank (CWDB).

1. FILE STRUCTURE - MARS VI has an index sequential file structure with multiple key capability. This results in a partially inverted data base. Those data elements declared as keyed items may be used to make a direct access of all index sequential records containing the keyed value.

2. TABLES - MARS VI maintains a set of internal tables. The internal tables contain unique values for all items which have been declared as keyed. Associated with the unique values are pointers to the index sequential records containing these values.

3. FILE RESIDENCE - The MARS VI DBMS may access data through Rotating Mass Storage (RMS) files or from magnetic tape files. The RMS files may be local non-permanent or permanent files.

4. PROGRAM INTERFACE - A MARS VI data base may be accessed by user programs written in COBOL. The MARS VI DBMS does not communicate directly with these user programs; however, interfacing subroutines are available which enables the data base created by the DBMS to be accessed by user programs written in COBOL.

5. VARIABLE LENGTH RECORDS - The MARS VI DBMS has a limited capability for handling variable length records. Each record type which is of a different length must be on a separate index sequential file. MARS VI allows ten of these files which may be managed concurrently and collectively as a data base.

6. DATA DEFINITION LANGUAGE - MARS VI has a Data Definition Language (DDL) which is used to describe the format of the data elements on each record file. The definition is used by the MARS VI DBMS in all subsequent uses of the RETRIEVAL and UPDATE modules.

7. RETRIEVAL CAPABILITY - Data may be retrieved from a MARS VI data base in two ways. The first method allows the user to retrieve data and process the retrieved data using the MARS VI DBMS directly. This makes use of a RETRIEVAL module followed by a REPORTER module, which allows selected data items to be printed in a very readable format with a minimum of report formatting effort. Basic statistics are also available through the use of these two modules. The second method of access allows the user to retrieve data directly from the data base using the MARS VI DBMS, which writes a sequential file. The sequential file of retrieved data may then be processed by user programs.

8. USER PROGRAM DIRECT ACCESS - Should the user not desire to access the data in the MARS VI data base by using the MARS VI/COBOL interface or using the MARS VI RETRIEVAL module, he may access the index sequential file directly. That is, a user program written in a language such as FORTRAN or COBOL may read the sequential file portion of the index sequential file directly. Thus, when it is desireable, user programs may access data stored in the data base without using the MARS VI DBMS.

5. Development of the Colorado Water Data Bank System (CWDBS)

Project personnel were required by the first year contract to incorporate existing computerized water rights data into the data bank within the first six months. Capture of other historical and current records had to be initiated within the first year. These requirements prohibited initial development of the overall CWDBS and an interim procedure was implemented for storing and capturing data while correction, update and verification procedures were not addressed until the complete system design was initiated in the second year. It was imperative that the project demonstrate its capability by implementing a data base using the MARS VI DBMS.

The water rights data existed on magnetic tape and had been pre-edited and verified and it was possible to directly input these data into the MARS VI DBMS without editing. Updating and correction procedures were tried with this data base and it became apparent that development of the overall CWDBS was imperative to success of the project. Because of personnel limitations, an outside consultant, Fritz & Associates, of Ft. Collins, Colorado, was retained to design a system which could be used for capturing, editing, verifying, updating and retrieving data from the CWDB. The consultant was retained for three months and at the end of that period, submitted a report, Fritz & Associates [2], which was to serve as the working document for further development of the CWDBS.

Implementation of the CWDBS began in July, 1973. Software requirements necessitated some minor modifications to Fritz's system design. Implementation of the system has clarified the user/machine interactions and has allowed development of some universal software and procedures which have been used to process several types of data. This has minimized software overlap and has standardized user procedures for coordinating data capture, correction, verification and updating.

5.1 Structure of Record Formats

Each of the data types listed in table 1 and illustrated in figure 1 has a different length of record to be stored. Because of these variable record lengths, it was decided to implement each of the data types as a separate index sequential file within the MARS VI DBMS. Because of user requirements, it was necessary to be able to cross-reference data between the index sequential files. That is, having used some criteria to select a data record on one index sequential file, it may be necessary to retrieve several associated records from one or more other index sequential files. The MARS VI DBMS allows this type of access to a data base provided a common identifier is specified on each index sequential file in order to link the two types of data.

For example, (reference figure 1), it may be necessary to retrieva all diversion data for a given water right. This data access might be compared to a personnel data base where

the financial records are on one file and address information is on another. The social security number would be the common key to link these files together.

The State of Colorado has developed its own identification system. For administrative purposes, the state is divided into seven large geographic areas where each represents a major river drainage basin. These areas are called divisions (DIV) and each of these is further subdivided into smaller drainage basins called water districts (WD). There are 80 WDs in Colorado. Within each WD a unique five-digit number is assigned to each data collection point. The WD number, when combined with the data point number, creates a unique common identifier (ID) for each data collection point. Using the ID, it is possible to access interrelated data elements from different files in the same retrieval.

Several different types of data may be associated with a single data collection point; e.g., water rights, diversion and water quality. The assignment of the unique ID for the collection point allows the desired cross-referencing and also eliminates the need to assign a different identification number to each record for every data type.

a. Choosing the Keyed Items

The MARS VI DBMS allows a partially inverted file structure. For those data elements within an individual record that the user desires to directly access, MARS VI creates data base keys. The data elements which are chosen as MARS VI keys are said to be inverted and unique valued tables are constructed for each of them. Relative pointers to the index sequential file are constructed for each of the unique values within the corresponding table. Retrieval of data elements which have been inverted requires only that the unique value be looked up in the index tables and the relative position in the index sequential file obtained. The MARS VI DBMS may then directly access the record or records containing the desired value.

For each keyed data element within a data record, on-line storage will be needed for the index tables in addition to that required for the sequential file. MARS VI DBMS users must be careful in the selection of keyed items to provide random retrieval and update capability without increasing the storage requirement excessively.

For the CWDB, three basic data elements were chosen to become keyed data elements in nearly every record type. They are Division (DIV), Water District (WD), and the common identifier (ID).

The primary reason for making a data element a keyed item is to facilitate either updating, retrieval or a combination of both. Within a record there may be data elements that lend themselves to being keyed items for that particular data type; however, these elements may not be common to all record types. To reduce storage and simplify the data base definition, it may be desireable to change some keyed elements to non-keyed elements following the correction, verification and updating of specific data bases. Such a condition is described in section 5.2.

b. Mapping Identifier Numbers

The implementation of data from federal data bases requires that at least a Colorado-assigned ID be inserted into each data record. This is necessary for cross-referencing. There is no standardization between the chosen collection points of the federal data network and the state-chosen collection points for the Colorado water data network. The collection points of the federal network that the State chose to use are a small subset of the entire Colorado data network.

The mapping process whereby a federal ID is mapped to a state-assigned ID to facilitate cross-referencing data within the Colorado Water Data Bank is not a complicated one. However it does seem that this step is unnecessary and would not be required if there was a standardized method for assigning IDs to data gathering networks. Currently, water data captured under federal control may be obtained by all state agencies and cross-referenced through the federal identification system. In some cases, state agencies supply data captured under state control to the federal data base system. In these cases, the state agencies have cooperated and used the federal ID system. What is not easy to do is to make use of federal

data in conjunction with state data. Even more unfeasible is to share data between state agencies. For example, sharing of diversion data between the states of Colorado and Wyoming would be most difficult at this time. Both states have a different identification system and it's not clear whether the respective state agencies address the same type of data as being diversion data. For engineers who have computer modeling applications, it would be most desireable to be able to interchange water data at all governmental levels.

5.2 Working Versus Official Data Base

As indicated earlier in this paper, much of the data in the CWDB which is collected by the state is intended to be a legal record. In order to make this data a legal record, there is an extensive verification process. This process is described in detail later in this paper. To facilitate this extensive verification process, the Colorado Water Data Bank Project has developed the concept of a working data base and an official data base. The working data base contains both verified and non-verified information while the official data base contains only verified records. The structure of these two data bases may differ considerably.

It is intended that the working data base be smaller than the official data base. The working data base contains only that data which has not been verified by the agency or individuals responsible for data capture. Once verified and declared to be correct, data will be transferred to the official data base. A primary difference between the two types of data bases is that the structure of the working data base allows it to serve as both a data base which can be "read from" and a data base which can be "written to". The working data base may be updated by adding new data or by correcting existing data within the data base.

The official data base is thought of as a "read only" data base. It is intended that the official data base will be accessed only to retrieve data for a user. Data which has been verified in the working data base may be transferred and added to the official data base. Once data elements become a part of the official data base, it will be most difficult to make changes to these data elements. Provisions have been made for changes to be made to data in the official data base, but the process involves technicalities much as would be expected in changing any type of legal record. This process is expensive, both in terms of computer cost to perform the updating and time required for an individual to process the change.

In structuring the official data base, several changes have been made in the MARS VI data definition. The changes reflect the fact that the official data base is primarily designed to be read from. Therefore, keys which exist in the working data base for updating purposes are removed. Only items which will be specified frequently for retrieval purposes and those data items that are used for cross-referencing data types are kept as keyed values. Therefore, the storage requirement for the official data base structure versus the working data base structure is significantly less.

5.3 Data Collection Network

The CWDBS identifies three main points in its data collection network. They are: (1) Data collection and verification, (2) The Data Base Administrator (DBA), and (3) The computer software. The data flow between these points is shown in figure 2. This figure details only the data processing for current diversion or current reservoir data. Other types of data employ variations of this data processing procedure.

Data enters the CWDBS from two sources. The largest source is from within Colorado. The second source is from other agencies such as the U.S. Geological Survey. The discussion below presents the collection of data from each source. The acronyms correspond to those used in figure 2.

a. Colorado Water Data

Two points are identified in the network for capturing and verifying data. They are the water commissioners (WC) and the office of the Division of Water Resources (DWR). These two points in the network are primarily responsible for the coding of new data, coding of data

corrections, and verifying data which has been entered into the CWDBS. Modes of data capture include the coding of OpScan mark sense forms and load sheets. Both the WC and DWR must transmit the captured data to the data base administrator (DBA).

The water commissioners are involved in a hierarchical structure. Therefore, the network necessitates their submitting the captured data to the Division Engineer's office, (DIV). Under the control of each of the seven Division Engineers' offices are several water commissioners within the different water districts (WD). Each Division Engineers office is responsible for batching all data submitted by water commissioners in his division. The data is transmitted periodically to the data base administrator (DBA). Water commissioners capture only current diversion or reservoir records.

Historical data is captured by the State Engineer's office, DWR, and is batched and transmitted directly to the DBA. This data is also captured utilizing either the OpScan mark sense sheets or load sheets.

The data base administrator (DBA) is responsible for logging and submitting data received from either DIV or DWR. This data is received in either OpScan or load sheet form. The OpScan data is submitted to be captured on the OpScan 100DM to 7-track tape. Load sheets are submitted for keypunching. The DBA is then further responsible for maintenance and updating of the CWDB. This is accomplished by using the CWDBS computer software.

After the data base has been updated, the DBA is responsible for distributing either error lists or the verification reports published by the CWDBS software. This distribution process involves returning the reports and error lists to the respective point in the data network from which the data originated. Therefore, these reports are returned either to DIV or to DWR. If the report and error list are returned to DIV, they are then further distributed to each WC. In the case of DWR, which is an originating source, no further distribution is required.

At each originating DWR or WC, additional manual processing is performed. In the case of edit error lists, each error is resolved. Corrections for the errors are coded and the processing begins a new loop.

In the case of the verification reports (see fig. 4), the originating source must check the data values associated with each data element in the report. The report is verified on a page-by-page basis. On each page is a signature block (no. 9, fig. 4), which is signed by the individual who coded the record for original input. The signature is affixed to the verification report page only if all data on that page is correct. The data on that page is then eligible to be moved to the official data base and the report is forwarded to the DBA. Should there be errors on the page, then corrections must be coded for the incorrect data. These corrections then enter the data processing loop.

It is up to the DBA to determine when a logical batch of data from the working data base has been verified as being correct. At the discretion of the DBA, the data from the working data base is moved to the official data base. At the same time, the signed verification reports are distributed to DWR to be entered into the archives as an official legal record. The CWDBS software is responsible for removing the data from the working data base to the official data base.

The object of the verification process is to move all data from the working data base to the official data base. Since data is constantly being captured, this objective seemingly may never be reached. However, the capturing of water data in Colorado is oriented around an irrigation year which begins on November 1 of the first calendar year and continues through October 31 of the following year. Therefore, it is intended that on or about October 31 the old working data base will be "frozen" and a new working data base will be initiated. It may take a few weeks into the new irrigation year to remove all remaining data from the previous year's working data base.

b. External Water Data

Not all data entered into the CWDB is data which has been captured under state control. Data may come from separate state agencies or a federal data collection agency. When entering this data into the CWDB there may or may not be a verification process. For the most part, this data is accepted at face value. However, general editing for obvious data errors is performed in the data processing system.

In lieu of the working data base concept, which is required for state gathered data, there are intermediate data files generated for external sources of data. Generally, this intermediate data file represents the procedure of extracting only the needed data from the external source and mapping the state assigned identifier to the external data. In some cases, data conversion or modification may take place. The resulting external data file then is loaded directly to the official data base. In keeping with the concept of the official data base, it is not intended that external data appearing in the data base will be modified. Data may be added through an add-on load.

5.4 Computer Software for the CWDBS

The software which the DBA uses to maintain the CWDB is written in two computer languages in conjunction with the MARS VI DBMS. Programs exist in FORTRAN and COBOL as well as input specifications to the MARS VI DBMS. The CWDBS software obtains most of its control information through user-supplied tables. These tables are maintained by the system by entering table information as data. Header information identifies the data as tables and the tables are updated.

The use of tables allows the user more control over the CWDBS software. Old record formats may be changed and new record formats added without software modification.

Figure 3 presents the general flow of data through the CWDBS software. The acronyms presented here correspond with those in the system flow diagram.

a. DBAC--Preprocessor

Program DBAC is responsible for processing the data to be input into CWDBS. This initial processing involves reading of data from external magnetic tape sources, 80-column data cards, or magnetic tape generated by the OpScan 100DM. DBAC reads the data from these sources and adds unique sequencing information to each record from the input source. Header records precede each type of data to be entered into the system. Information on these header records, combined with a sequential numbering system, creates a unique identifier for each data record.

In the case of OpScan input, a further requirement for DBAC is that it unscrambles and decodes the magnetic tape input which is generated by the OpScan processor. The UNSCRAMBLE/DECODE software is table-driven and these tables exist on a permanent file accessible by program DBAC.

Program DBAC then sorts the output by data type and generates a 7-track magnetic tape of this data. A disk file of the data is used as input to a subsequent program DBAD in the CWDBS.

b. DBAD--Edit/Update

Program DBAD is responsible for all data editing. This editing is done within the data record at the data element level and between data elements. As records are edited, they are either accepted or rejected. The rejected records are written to an edit reject file which exists on 7-track magnetic tape. The accepted records are processed.

Processing of accepted records involves direct updating of the CWDB through the MARS/COBOL interface software or indirect updating through the MARS VI DBMS. Direct updating can only be done to those data elements which are non-keyed. If the non-keyed data being

entered into the system is original and no record exists where this data can be added, a record is written to a MARS VI transaction file. The transaction file will be processed by the MARS VI DBMS. If updating is to be done to keyed data elements, then program DBAD generates a MARS VI transaction file which will later be input to the MARS VI DBMS. Program DBAD is also responsible for updating the control tables which are used by all CWDBS software. Updating of these tables makes use of random access files. DBAD does not use the MARS VI DBMS to update the tables.

As part of the updating process, program DBAD compares the update transactions against the edit reject file. The software is capable of performing modification to the edit reject file to correct the errors that occurred in the data records when they were written to the edit reject file. When a data record on the edit reject file is corrected, it is removed from that file and input into the normal data base edit and update procedures. The objective is to eventually remove all records from the edit reject file.

In addition, program DBAD is responsible for writing the edit error report. These error lists are taken by the DBA and distributed to the peoper points in the data collection network.

c. MARS VI--Update/Add-On Load

The MARS VI DBMS is utilized to update the data base for keyed data elements and new data records. Updating of keyed data elements is necessarily more expensive and experience has indicated that updates to keyed data elements should be batched together. This is because it is less expensive to update five keys in one session than to do so in five sessions.

The new data records are processed through the add-on load feature of the MARS VI DBMS. This is the most common type of update transaction.

d. Report Generation

After the updates have been processed, the MARS VI software reads retrieval specifications from a card data file. Retrieval from the CWDBS may be done using either the MARS VI DBMS or the MARS/COBOL interface. MARS VI provides the user the capability of having a quick look at data in the data base. Using the MARS VI RETRIEVAL and REPORTER modules, the user can create reports in a short time. However, because of format and logic limitations of the REPORTER module, most of the project's reports are created using special report generation software.

A sample report generated by the special report software is included as figure 4. This report is a complex report requiring cross-referencing of multiple index sequential files, data computations, and data interpretation. Item 1 indicates this report is for an irrigation year. Items 2 and 4 specify location. They require three accesses of two files. The structure number (03551) in item 2 requires an access to a location file to retrieve the structure name. When retrieving the stream number (001) in item 4, an access is required to the location file to obtain the stream name. The information obtained in item 3 requires yet another access to a file. The names in item 5 are stored in tables within the program. Item 6 indicates observed data as indicated by the asterisk. Observed data is the only data entered into the CWDBS. However, the report requirements state that if data is missing then values are to be interpolated from the last observed value. That is, the last observed value is carried forward until the next observed value. This is indicated by item 7. Item 8 indicates the computations which are performed. If a verification page is correct, the page is signed in the lower right corner as indicated by item 9.

The report generation software is responsible for generating all verification reports for the CWDBS. This software is required to produce quite complex reports. There is often a requirement to merge data from several of the MARS VI index sequential files. The MARS VI DBMS is used to retrieve the desired data and the report software reads the intermediate files to produce the reports.

5.5 Verification Procedure

The verification software of the CWDBS is imbedded almost entirely in program DBAD. Since it is a stringent requirement that the data captured under state control be verified as being absolutely correct and entered as a matter of legal record into the archives, this software is quite important. Basically, the software must keep track of the status of each data element in the working data base. The possible status conditions are: (1) The data element has been entered into the data base but no verification report has been produced for it, (2) The data element has been included in a verification report and is assumed to be correct, or (3) The data element has been corrected through use of a verification report.

The general logic is that a data element enters the CWDB as status 1. When this data is included in a verification report its status is changed to status 2. Data elements which have a status of 2 are assumed to be correct. Should the verification report reveal that a data element is in error it is corrected through the CWDBS software. At this time its status is changed to status 3. Additional verification reports are produced on status 3 data and the status is changed to 2 again. The data element is again assumed to be correct, until reported to be in error. The goal is to have all data with a status 2. The DBA will determine after receiving signed verification reports when to move status 2 data to the official data base.

An aesthetic problem exists in having an individual sign the verification report and its becoming a legal record. What guarantee does this individual have that the data he verified as being correct was the actual data (combination of binary zeros and ones at the most elementary computer level) that was transmitted to the official data base? It has been suggested that in order for the data to become an official record, that an additional report must be produced from the official data base after the data has been moved from the working data base. Currently, these problems are still being resolved between data bank personnel and DWR. Basically, the problem is to what degree can one trust computer software? If the computer software is 100% logically correct, to what degree can computer hardware be trusted?

6. Conclusions

The Colorado Water Data Bank Project has been in operation approximately 1 1/2 years. The major effort during this period has been to develop the logic, procedures, and programs to be incorporated in the overall Colorado Water Data Bank System (CWDBS). The Control Data Corporation MARS VI Data Base Management System has been incorporated as an integral part of the overall system. Several different types of records have been captured and placed in the data bank and more recently requests for access and retrieval of data have been processed. As would be expected, the project has experienced both success and setbacks on meeting certain objectives within the selected time frame.

Although complete implementation of the Colorado Water Data Bank System is not expected prior to June, 1975, considerable progress has been made and it is possible to draw the following conclusions:

1. There is a need to have localized water data banks which will contain many different types of data and which will provide the capability for access and retrieval of all the information required to make administrative or management decisions at one time. Centralizing the data location will minimize retrieval costs, allow cross-referencing, and provide the information within a reasonable time frame.

2. Currently, it is not possible to directly interchange water data with other federal or state agencies. While variations may be subtle, each agency has implemented its own identification system for the data collection points within each agency. Data from other agencies may be entered into the CWDBS. However, the external agency identifier must be mapped to a Colorado identifier.

537-860 O - 74 - 17

3. It is possible to utilize a commerically available data base management system as an integral part of a complex water data bank system. Utilization of a commercial data base management system requires standardization of input and output procedures. Utilization of the MARS VI DBMS permitted this project to begin capturing of data at least six months earlier than would have been possible if all programs had been written by project personnel.

4. The wide variety of data to be placed in the data bank has required the establishment of several record files with different record formats. Organization of these files has permitted the treatment of both fixed and variable length records.

5. Although data was originally captured using interim procedures, development of Colorado Water Data Bank System has provided the software and procedures for capturing different types of data with a minimum of effort by the Data Base Administrator (DBA).

6. Data is being captured with mark sense forms, allowing the water commissioners to prepare the machine readable document, thus minimizing transferral errors and time required to put the data in the data base.

7. The processing of both water rights and diversion records has used the overall CWDBS. During the next year, the necessary tables and edit routines will be defined and incorporated into the system for the other types of water data. Development of detailed user documentation is underway which should allow the Data Base Administrator (DBA) to process all incoming data as well as honor data retrievals.

7. Acknowledgements

The authors wish to acknowledge the funding of the Colorado Water Data Bank Project by the State of Colorado. The Colorado Water Data Bank System described in the paper was developed jointly by personnel from Colorado State University and the Division of Water Resources. Correspondence on this paper may be addressed directly to the authors at Colorado State University; however, programs, data, or report requests should be sent to the Division of Water Resources, 1845 Sherman Street, Denver, Colorado, ATTENTION: Chief, Computer Services Branch.

8. References

[1] McMillin, Norval E. Evaluation of a Data Base Management System for the Colorado Water Data Bank Project. Colorado Water Data Bank Project, Colorado State University, Technical Report Number 72-01, October 6, 1972.

[2] Fritz, Terrence L. Proposed System Design of Manual and Automatic Data Processing Procedures for Colorado Water Resources Data Bank Project. T. L. Fritz and Associates, Fort Collins, Colorado, June 26, 1973.

9. Addendum

Some modification of the preliminary draft was made for clarity and to emphasize discussion points raised during the symposium. Questions forwarded to the authors are presented and answered in this addendum.

Question: Why can you not interchange data with Wyoming or Nebraska:
Answer: Assuming data exists, it can be exchanged; however, the data format would probably not be compatible between states. The record formats for two different states might contain different data elements. Most likely location identifier for the data collection points would reflect individual state location systems and would not allow physical referencing of a point in one state to a similar point across

the state line in another state. In some instances reformatting of data records to a common format will allow compatible usage of the information, but the lack of uniformity in the data elements included in each record can not be easily overcome. See section 5.1 and 6.0 for more discussion.

Question: Could you elaborate on the conflict of federal data and state data?
Answer: The federal versus state data uniformity problem is similar to that discussed above between two states. Also see section 5.1.

Question: What is the function of the Data Base Administrator (DBA)?
Answer: Currently, one individual performs the function of the DBA. All transactions which will update or modify the data base must be processed by the DBA. In addition, requests for verification reports must be submitted to the DBA for processing.

Question: What is the error rate using the OPSCAN mark sense technique? What was the degree of acceptability by the users of the mark sense forms.
Answer? We have found the existing OPSCAN machine to adequately capture the marked forms with a very low machine reject rate, a small fraction of one percent. Existing edit programs and the verification procedure define miss-marked or improper data. Numbers to evaluate the errors due to improper marking versus machine read problems are not availabe, but our success in data capture has encourage us to use the same mechanism for another year. Some attempt will be made in June, 1974 to evaluate the relative merits including cost, of mark sense capture versus key-punching.

Education programs were held to acquaint personnel with the mark sense technique. Assuming data codes and marking procedures are well defined prior to education meetings, most of the personnel have adapted to mark sense capture of data. Some redesign of forms has been undertaken to incorporate suggestions from the users. Careful layout of forms is most important to the success of the technique.

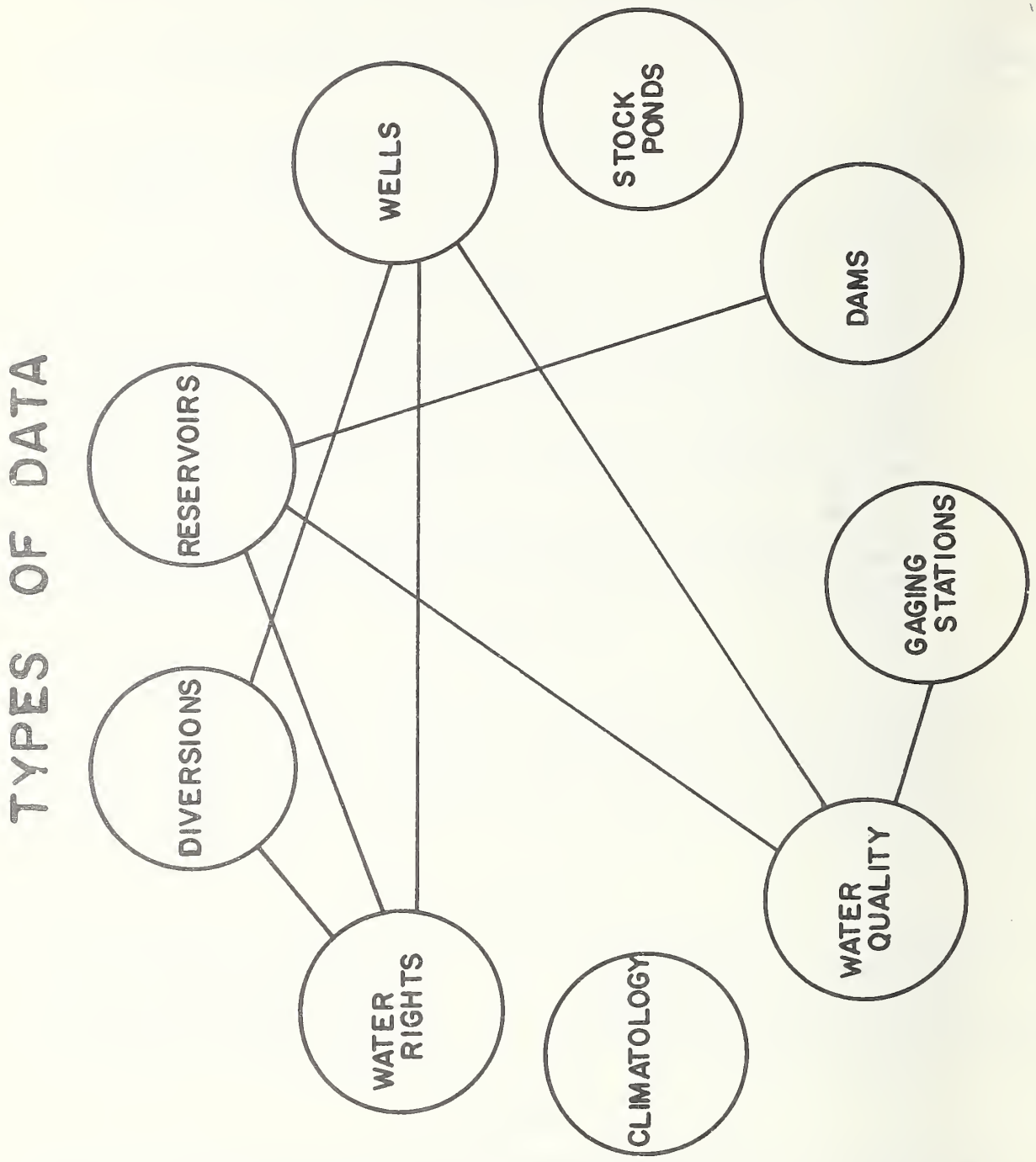


Figure 1. Types of data to be included in the Colorado Data Bank showing those types to be cross-referenced.

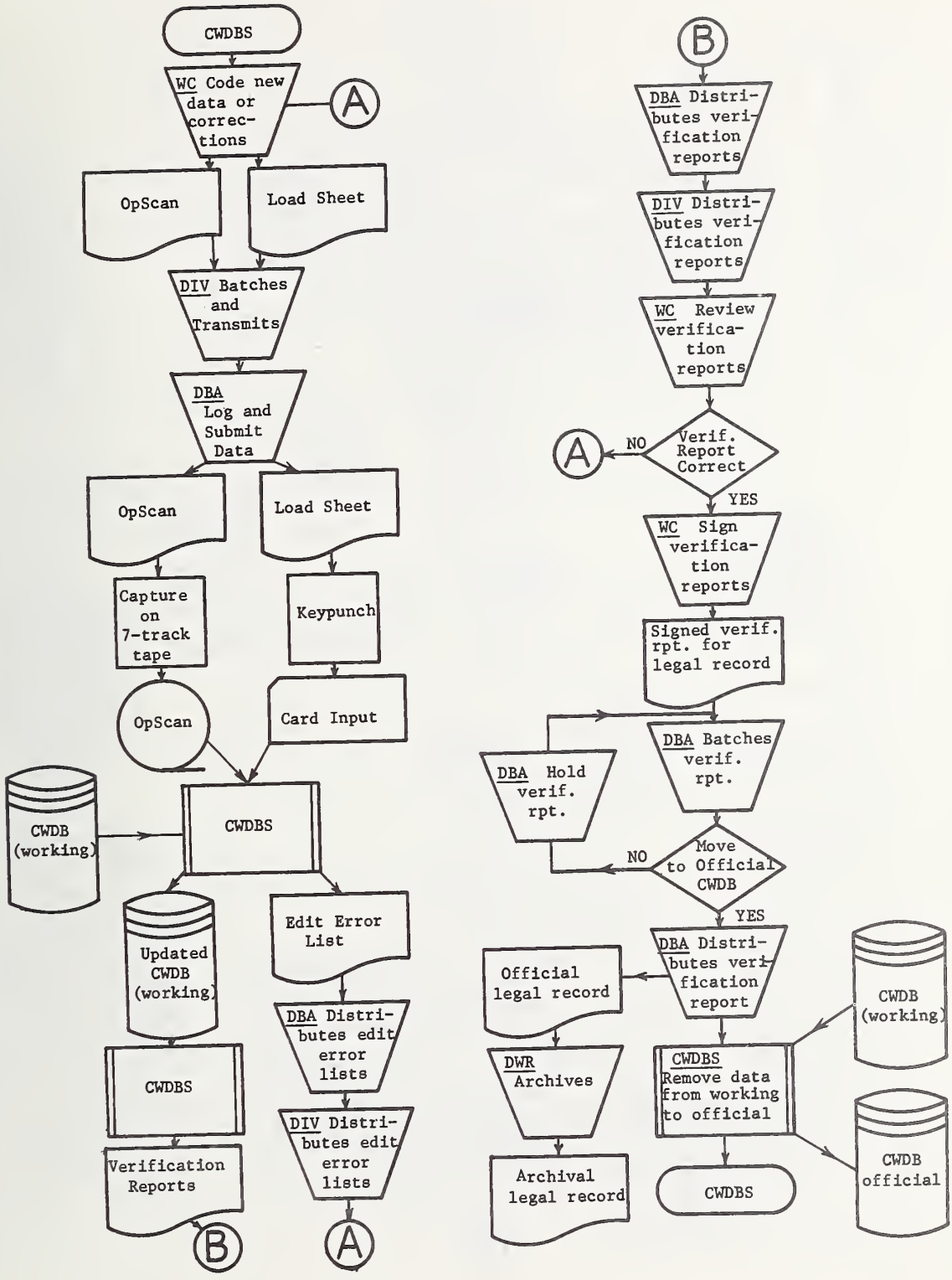


Figure 2. Flow diagram illustrating procedure for capturing current reservoir or diversion data.

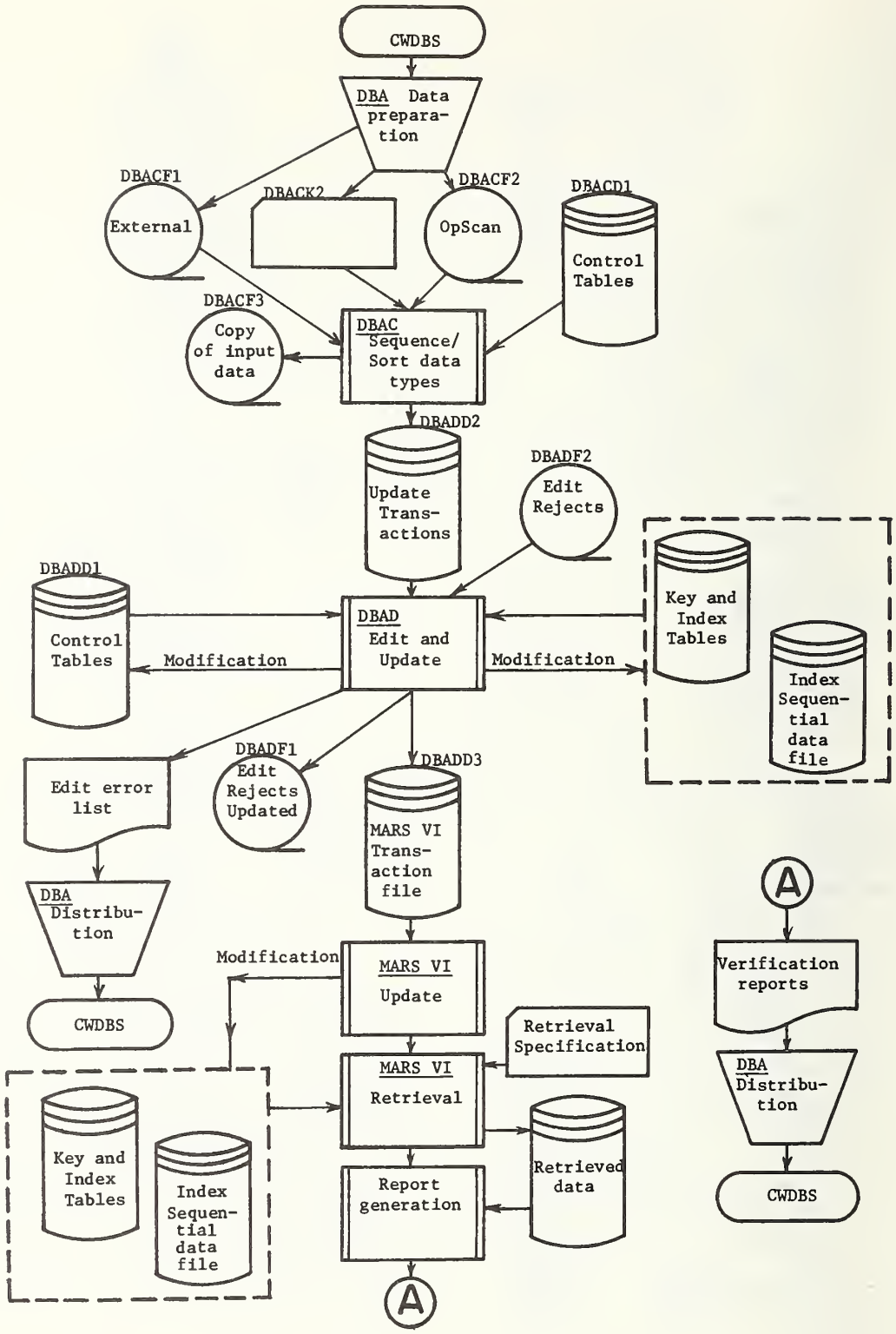
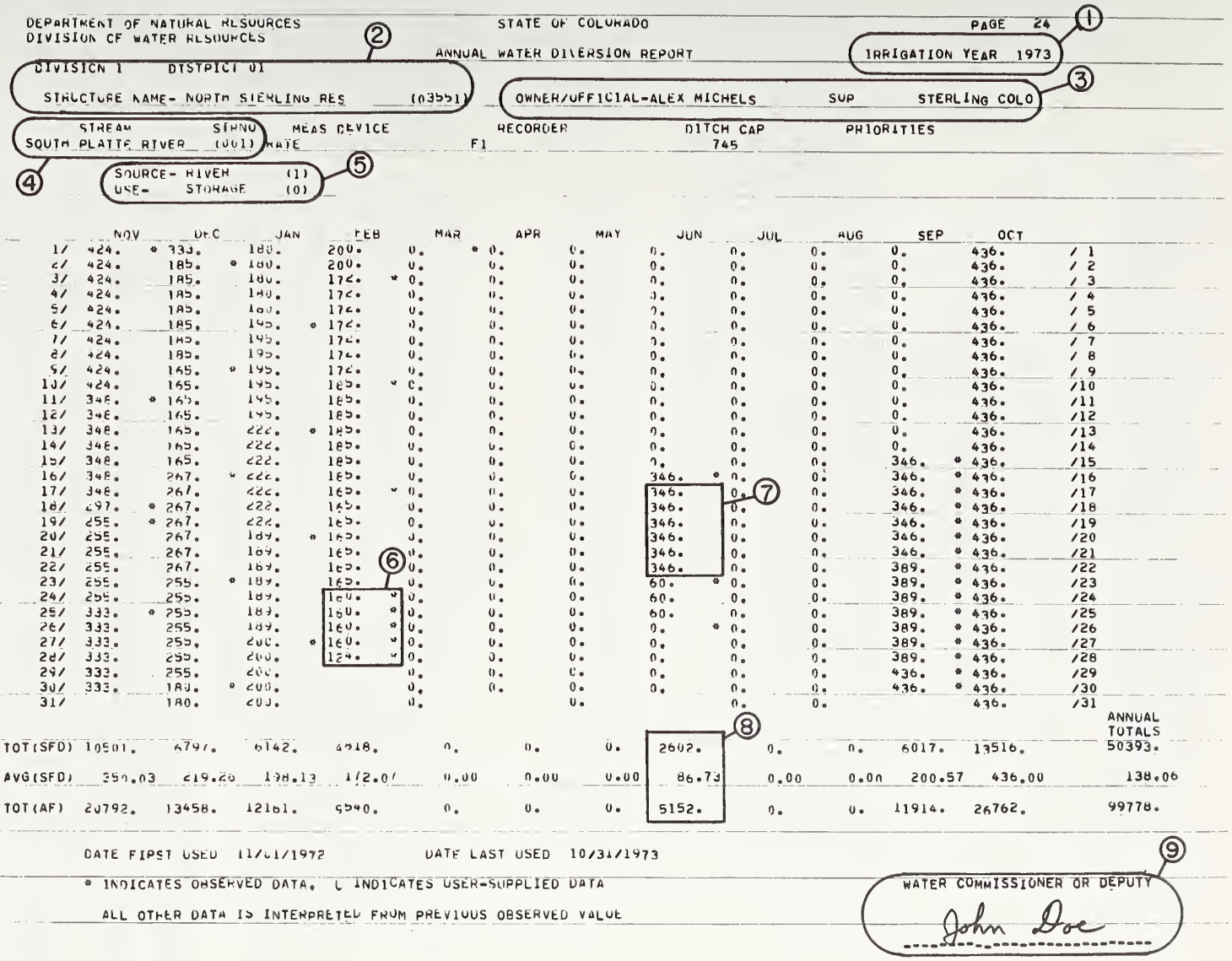


Figure 3. Skematic showing connection between software programs.



DEPARTMENT OF NATURAL RESOURCES
DIVISION OF WATER RESOURCES

STATE OF COLORADO

ANNUAL WATER DIVERSION REPORT

PAGE 24

(1) IRRIGATION YEAR 1973

(2) DIVISION 1 DISTRICT 01

STRUCTURE NAME- NORTH STERLING RES (03551)

(3) OWNER/OFFICIAL-ALEX MICHELS SUB STERLING COLO

(4) STREAM SOUTH PLATTE RIVER STRNO (001) MEAS DEVICE RATE F1 RECORDER DITCH CAP 745 PRIORITIES

(5) SOURCE- RIVER (1)
USE- STORAGE (0)

	NOV	DEC	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	
1/	424.	* 333.	180.	200.	0.	* 0.	0.	0.	0.	0.	0.	436.	/ 1
2/	424.	185.	* 180.	200.	0.	0.	0.	0.	0.	0.	0.	436.	/ 2
3/	424.	185.	180.	176.	* 0.	0.	0.	0.	0.	0.	0.	436.	/ 3
4/	424.	185.	180.	176.	0.	0.	0.	0.	0.	0.	0.	436.	/ 4
5/	424.	185.	180.	176.	0.	0.	0.	0.	0.	0.	0.	436.	/ 5
6/	424.	185.	195.	* 176.	0.	0.	0.	0.	0.	0.	0.	436.	/ 6
7/	424.	185.	195.	176.	0.	0.	0.	0.	0.	0.	0.	436.	/ 7
8/	424.	185.	195.	176.	0.	0.	0.	0.	0.	0.	0.	436.	/ 8
9/	424.	165.	* 195.	176.	0.	0.	0.	0.	0.	0.	0.	436.	/ 9
10/	424.	165.	195.	185.	* 0.	0.	0.	0.	0.	0.	0.	436.	/10
11/	348.	* 165.	195.	185.	0.	0.	0.	0.	0.	0.	0.	436.	/11
12/	348.	165.	195.	185.	0.	0.	0.	0.	0.	0.	0.	436.	/12
13/	348.	165.	222.	* 185.	0.	0.	0.	0.	0.	0.	0.	436.	/13
14/	348.	165.	222.	185.	0.	0.	0.	0.	0.	0.	0.	436.	/14
15/	348.	165.	222.	185.	0.	0.	0.	0.	0.	0.	346.	* 436.	/15
16/	348.	267.	* 222.	185.	0.	0.	0.	346.	* 0.	0.	346.	* 436.	/16
17/	348.	267.	222.	185.	* 0.	0.	0.	346.	0.	0.	346.	* 436.	/17
18/	297.	* 267.	222.	185.	0.	0.	0.	346.	0.	0.	346.	* 436.	/18
19/	255.	* 267.	222.	185.	0.	0.	0.	346.	0.	0.	346.	* 436.	/19
20/	255.	267.	189.	* 185.	0.	0.	0.	346.	0.	0.	346.	* 436.	/20
21/	255.	267.	189.	185.	0.	0.	0.	346.	0.	0.	346.	* 436.	/21
22/	255.	267.	189.	185.	0.	0.	0.	346.	0.	0.	389.	* 436.	/22
23/	255.	255.	* 189.	185.	0.	0.	0.	60.	* 0.	0.	389.	* 436.	/23
24/	255.	255.	189.	160.	* 0.	0.	0.	60.	0.	0.	389.	* 436.	/24
25/	333.	* 255.	189.	160.	* 0.	0.	0.	60.	0.	0.	389.	* 436.	/25
26/	333.	255.	189.	160.	* 0.	0.	0.	0.	* 0.	0.	389.	* 436.	/26
27/	333.	255.	200.	* 160.	* 0.	0.	0.	0.	0.	0.	389.	* 436.	/27
28/	333.	255.	200.	124.	* 0.	0.	0.	0.	0.	0.	389.	* 436.	/28
29/	333.	255.	200.		0.	0.	0.	0.	0.	0.	436.	* 436.	/29
30/	333.	180.	* 200.		0.	0.	0.	0.	0.	0.	436.	* 436.	/30
31/		180.	200.		0.		0.		0.	0.		436.	/31

(6)

(7)

(8)

													ANNUAL TOTALS
TOT(SFD)	10501.	6797.	6142.	4818.	0.	0.	0.	2602.	0.	0.	6017.	13516.	50393.
AVG(SFD)	350.03	219.26	198.13	172.07	0.00	0.00	0.00	86.73	0.00	0.00	200.57	436.00	138.06
TOT(AF)	20792.	13458.	12161.	9540.	0.	0.	0.	5152.	0.	0.	11914.	26762.	99778.

DATE FIRST USED 11/01/1972 DATE LAST USED 10/31/1973

* INDICATES OBSERVED DATA, (INDICATES USER-SUPPLIED DATA

ALL OTHER DATA IS INTERPRETED FROM PREVIOUS OBSERVED VALUE

(9) WATER COMMISSIONER OR DEPUTY

John Doe

Figure 4. Typical annual summary and verification report for diversion records.

Records, Computers, and the Rights of Citizens: The Report of the Secretary's Advisory Committee on Automated Personal Data Systems

David B. H. Martin

Special Assistant to the Secretary of
Health, Education and Welfare

The report of the Secretary's Advisory Committee on Automated Personal Data Systems recommends enactment of a Code of Fair Information Practice applicable to all record-keeping operations that process data about identifiable individuals. The report also recommends constraints on the use of the Social Security number as an identifier of records and people. This paper describes the work of the Committee and the main features of the Code of Fair Information Practice.

Key words: Automated personal data systems; fair information practice; privacy; records; right of privacy; Social Security number; standard universal identifier.

1. Introduction

There are no easily constructed bridges between the topics being discussed at this symposium and the concerns of the Secretary's Advisory Committee on Automated Personal Data Systems. Partly this is due to the fact that the issues the Committee addressed would merit our attention whether or not we manage to improve our current capacity to exchange information among computer-based record-keeping systems. Partly it is due to the fact that the Committee purposefully avoided questions about how systems should be structured, what types of data they should contain, and, save for its examination of uses of the Social Security number, what techniques should be used for transferring information from one system to another.

My remarks, therefore, will focus on the background and rationale for the Committee's posture in the hope that when you come to examine its recommendations in detail, you will have a better basis for evaluating them against your own experience and against the recommendations that have been put forward by others. I do not plan to review the recommendations one-by-one. That would take more time than I have been allotted, and would not be as useful as if you yourselves took the time to read the Committee's report with a view to its relevance to your own work, and, particularly to your own capacity to influence the ways in which we, as a society, make use of this powerful new technology.

2. Background and Membership of the Committee

The Secretary's Advisory Committee on Automated Personal Data Systems was established in February 1972 by then HEW Secretary Elliot L. Richardson. The formation of the Committee rested upon a public interest determination which stated in part:

> the use of automated data systems containing information about individuals is growing in both the public and private sectors.... At the same time, there is a growing concern that automated personal data systems present a serious potential for harmful consequences, including infringement of basic liberties. This [concern] has led to the belief that special safeguards should be developed to protect against potentially harmful consequences for privacy and due process.

The Committee that was established pursuant to Secretary Richardson's determination was asked to analyze the harmful consequences for individuals, for record-keeping organizations, and for the society as a whole, that may result from uncontrolled application of computer and telecommunications technology to the collection, storage, and use of personal data about identifiable individuals. In addition, the Committee was asked to recommend:

(1) Safeguards to protect against any potentially harmful consequences that might be identified;

(2) Measures to afford redress for any harmful consequences that might occur; and

(3) Changes in policy and practice relating to the issuance and use of Social Security numbers.

The formation of the Secretary's Advisory Committee coincided with the publication of an internal HEW task force report on the issuance and use of Social Security numbers. That report recommended the creation of a public advisory body to consider the broad question of linkages and exchanges of information among computer-based personal data systems. Also, the Committee counted among its precursors and friends the Senate Subcommittee on Constitutional Rights, chaired by Senator Sam J. Ervin, Jr. (D.-N.Car.), which, in the spring of 1971, had held hearings on computer data banks, the use of the SSN in them, and other related matters having to do with government record-keeping policy and practice.

The Secretary's Advisory Committee on Automated Personal Data Systems had 25 members drawn from State government, private industry, the social service professions, the academic research community, the legal profession, and from private life. Many had had practical experience in operating or using automated personal data systems in settings ranging from a nationwide credit-bureau network to the program management information systems of a State Department of Finance. Indeed, as some members of the press have pointed out, the Advisory Committee's recommendations are striking (1) for having been developed by a group of individuals well acquainted with the many beneficial uses of computer-based record-keeping systems and (2) for having been developed by a group that started its work by questioning whether the issues it had been asked to address really merited serious attention at that time.

As the Chairman, Dr. Willis Ware of the RAND Corporation, wrote in the preface to the Committee's report:

> Many, indeed probably most, did not initially feel a sense of urgency about the potential ill effects of current practices in the design and operation of automated personal data systems. Some

agreed that computer-based record keeping poses a latent danger to individual citizens, but looked optimistically to technological innovations, particularly access-control devices, to prevent problems from arising. Others painted dramatic portraits of the potential benefits of large-scale data networks to citizens in a densely populated, highly mobile society.... Slowly, however, the attitudes of the members changed. Shared concerns took root as [the Committee] heard testimony from over 100 witnesses representing more than 50 different organizations, and as [it] reviewed a substantial collection of written materials, including reports by similar commissions in this country, Canada, Great Britain, and Sweden.

3. Why This Change? What Did the Committee Find?

As the Committee listened to data system managers describe what their systems do and why they do it, it became clear that computerization is having three principal effects on personal-data record-keeping policy and practice: First, the computer enables record-keeping organizations to enlarge their data processing capacity substantially. Second, the computer greatly facilitates the retrieval of recorded information and its transfer across great distances and traditional organizational boundaries. Third, the computer creates a new class of record keepers whose functions are technical and whose contact with the original suppliers and ultimate users of personal data are often remote.

With respect to the first of these effects, the Committee noted that although the computer can greatly increase the efficiency of an organization, and its capacity to serve its clients readily and fairly, computerization can also have quite the opposite impact. Computerization is expensive, and the expense gives some organizations, particularly small ones, an incentive to spread the cost over an enlarged data-processing volume. A typical result is that clients receive erroneous bills, unjustified dunning letters, duplicate magazine subscriptions, and the like.

A strong incentive to concentrate on efficiency may also foster a tendency to behave as though data management were the primary goal of a computer-based record-keeping operation--with the result that unnecessary contraints may be placed on the gathering, processing, and output of data. "Please check one of the following boxes" becomes a common request which contributes, unjustly in some cases, to the dehumanizing image of the technology.

Scale also becomes significant here. It is one thing for a record-keeping operation to make errors in the checks written for five hundred people; it is quite another for five hundred thousand people to be affected --as the failure last year of the French family allotment payments system for the Paris area demonstrates.

Easier access to recorded information across vast distances and across organizational boundaries is the second effect of computerization that I mentioned. Although we may be inclined to dismiss the Orwellian nightmare of wall-to-wall databanks as too impractical to warrant serious discussion, we should still not let the "hard realities" of the matter blind us to the problems we face today, and for the foreseeable future, as a consequence of the ease with which computer and telecommunications technology permit information to be retrieved and moved about with great speed.

These problems have to do with: a) the different <u>interpretations</u> that can be made of information generated in one context, but used in another; b) the different <u>laws</u> that govern the use of personal information in different political and legal jurisdictions, e.g., the wide variation in State

laws requiring arrest-record checks before issuing licenses for an amazing range of occupational activities; c) the inability to enforce uniform rules with respect to the design, construction, and operation of automated record-keeping systems where the construction and operating costs are shared by different levels of government that jealously guard their respective jurisdictional prerogatives; and d) the imbalances that can be created among theoretically co-equal institutions by unevenly distributed computing capability--imbalances, for example, between the courts and the prosecutor, the legislative and executive branches, and, in general, between the organized and less organized elements in our society.

Finally, there is the new class of record keepers that computerization fosters--a class of data-processing specialists and advocates, if you will, who tend to be more concerned about efficient data management than with whether we, as a society, might be moving in the direction of more labeling, and numbering, and tracking of people, than might otherwise be thought desirable.

I do not say this contentiously. I think that in many respects computer-based record-keeping operations--and those who design and operate them--are being held responsible for problems that are not their fault. But I also think that groups like yourselves should think carefully about the long-term social implications of record-keeping practices that are promoted in the name of heightened system efficiency.

To give you an example of the kind of practices I have in mind--the kind that I think should induce pause in all of us--let me tell you about one case, described in the Committee's report, of a "cradle-to-grave" health data system developed for an Indian reservation in the Southwest. It is a system that records every contact that any member of the resident population has with any segment of the reservation's health-care facility. The system has a "surveillance component" designed to assure that every member of the population adheres to a prescribed schedule of preventive treatments (examinations, inoculations, etc.), and it is acquiring a statistical-reporting component that will be used to identify members of "high risk" sub-groups in the patient population. Furthermore, the developers of the system are now trying to improve its method of identifying patients for the purpose of updating and retrieving the information maintained about them. In this particular situation, the Social Security number happens to be considered a poor identification device because many patients have more than one. But the patients also tend to go by different names at different times, so the system managers are trying to develop their own unique numbering scheme cross-referenced with all known "aliases" for each patient.

Now, you might ask what is wrong with such a system. Clearly it is being used to help a population in need. Clearly its potential for realizing the Orwellian nightmare is limited by its exclusive concern with health matters. Clearly it is a special system to deal with a very special set of problems under very special circumstances; the likelihood that such a system would be replicated in Palo Alto, or White Plains, or Chevy Chase seems too remote to warrant serious concern. Or does it?

Systems of this sort, I would submit, should be of concern to all of us for several reasons. In the first place, the bare fact of their existence, no matter how restricted the circumstances under which they operate, attests to our capability to develop them in other settings if we so wish. Second, we tend to view systems like the Indian health system as if they were mere tools for delivering services, overlooking, or ignoring, the fact that they markedly increase the capacity of organizations to anticipate, and thus to control, the behavior of individuals. Third, at the present time, the individuals who are ultimately affected by these systems, and by most personal data record-keeping systems for that matter--the people about whom notations are made, the people who are being labelled and numbered--have very little

role to play in determining whether these systems should exist, what data they should contain, and how they should be used.

4. The Safeguard Recommendations

It was primarily the Advisory Committee's recognition of how sytems come to be, of who has a say in determining what goes into them, and of who does or does not have the authority to hold them to their declared purposes and operating procedures that led the Committee to recommend enactment by the Congress of a Code of Fair Information Practice for all automated personal data systems. The Committee concluded that what is needed today is a way of assuring close congruence between the uses that people expect to be made of information in records about them and the uses that are actually made. And the Committee felt that a good match between expectation and fact could be achieved if the operations of a personal data record-keeping system could be held to five basic principles:

- o There must be no personal data record-keeping systems whose very existence is secret.
- o There must be a way for an individual to find out what information about him is in a record and how it is used.
- o There must be a way for an individual to prevent information about him obtained for one purpose from being used or made available for other purposes without his consent.
- o There must be a way for an individual to correct or amend a record of identifiable information about himself.
- o Any organization creating, maintaining, using, or disseminating records of identifiable personal data must assure the reliability of the data for their intended use and must take reasonable precautions to prevent misuse of the data.

In the Committee's view, these principles should govern the conduct of all computer-based personal-data record-keeping operations. Departures from them should be permitted "only if it is clear that some significant interest of the individual data subject will be served or if some paramount societal interest can be clearly demonstrated." In no case should any exception be made that is not specifically provided for by law.

These five principles were translated by the Committee into safeguard requirements--or minimum standards of acceptable record-keeping practice--to be incorporated in the recommended Code of Fair Information Practice. Although focused on automated systems,[1] the Committee thought that the Code would be applied wisely to all personal data systems, whether automated or manual. It observed that the distinction between an automated and a non-automated system is not always easy to draw; that uniform application of the Code to all systems would ease conversion from manual to automated processing when it does occur; and that broad application of the Code seems warranted by the growing contrast between the capacity of organizations to make effective use of the information they record about individuals on one hand, and, on the other, the capacity of individuals to protect themselves against negligent, arbitrary, or malicious uses of such information.

---

[1] Defined as "a collection of records containing personal data that can be associated with identifiable individuals, and that are stored, in whole or in part, in computer-accessible files."

The legislation proposed by the Committee would make any violation of a safeguard requirement subject to both civil and criminal penalties. The Code would give individuals the right to bring suits for unfair information practices to recover actual, liquidated, and punitive damages in individual and class actions. It would also provide for recovery of reasonable attorneys' fees and other costs of litigation incurred by individuals who bring successful suits. It is important to note that an individual bringing suit for violation of any safeguard requirement would not have to demonstrate actual injury to himself or to a class of individuals to which he belongs. Evidence of a record-keeping organization's failure to meet the standard set by any one of the safeguard requirements would constitute sufficient ground for a civil suit or for bringing criminal charges.

4.1. Safeguards for Administrative Systems

Recognizing the important functional distinctions between administrative record-keeping systems and systems that are devoted exclusively to statistical reporting and research, the Committee recommended separate sets of safeguard requirements for each. However, I will deal here only with the requirements for administrative systems, they being the more elaborate and also the basis for the modified rules recommended for statistical-reporting and research systems.

The safeguard requirements for administrative automated personal data systems are divided into three categories. The first is focused on the operating rules that an organization establishes for a computer-based personal-data record-keeping system that it maintains; the second seeks to assure public awareness of a system's existence and of the procedures whereby an individual can affect the content and disposition of any record that it maintains about him (or her, as the case may be); and the third establishes the rights of an individual to make effective use of the access and review procedures that the record-keeping organization is required to establish and maintain.

Thus, the first category of safeguards, called "General Requirements," would fix the responsibility for maintaining a system, and for assuring that the system complies with the standards set by the Code of Fair Information Practice, in one clearly identified individual. In addition, a record-keeping organization would be required to do whatever is needed (a) to assure that each of its employees observes the ground rules set by the Code (i.e., through education, establishing incentives and penalties, providing adequate technical security, keeping a record of all non-housekeeping accesses to the system, enforcing appropriate standards of accuracy and completeness for all information entered into the system), and (b) to assure that all other record-keeping organizations to which the system transmits data observe the standards set by the Code, at least in regard to the information that is being transmitted to them. As a practical matter this would, of course, preclude, or at least make very risky, the transmission of personal data to any record-keeping organization not subject to the Code.

As I mentioned earlier, the "Public Notice Requirement," the second category, is a way of creating the conditions necessary for an individual to be able to exercise the rights guaranteed by the third category, the "Rights of Individual Data Subjects." In addition to information on the type of data in it, the annual notice a system would be required to publish would have to specify the procedures whereby an individual can find out if he is the subject of any data in the system, have access to that data, and contest their accuracy, completeness, pertinence, and the necessity for retaining them.

Finally, there is the third category, the "Rights of Individual Data Subjects," which constitutes the indispensable core of the Committee's recommendations. Under the proposed Code, any organization maintaining an administrative automated personal data system would be required to:

(1) Inform an individual asked to supply personal data for the system whether he is legally obligated, or may refuse, to do so;

(2) Inform any individual who asks whether he is the subject of data in the system, and if he is, permit him to have access to those data, in a comprehensible form, if he asks to see them;

(3) Obtain an individual's explicit, informed consent for any use of data about him which is outside the stated purposes of the system as reasonably understood by the individual;

(4) Inform an individual, who inquires, of all uses of data about him, including the identity of all persons and organizations involved;

(5) Notify an individual that data about him have been subpoenaed before responding to the subpoena; and

(6) When there is a disagreement between the record-keeping organization and an individual about whether a correction or amendment should be made in a record maintained about him, assure that the individual's claim will be noted and included in any subsequent disclosure or dissemination of the disputed data.

These safeguard requirements, in the Committee's view, constitute a set of minimum standards. There may be some instances when a persuasive case can be made for exempting a system from one or more of them on the ground that the exemption would serve some overriding, but clearly demonstrable, societal interest. The Committee expected such cases to be rare but it did countenance the possibility of exemptions so long as the organizations seeking them do so openly--through the legislative process or through the process of formal administrative rule making. In no case, however, did the Committee wish to dilute existing protections for data subjects that are stronger than the recommended safeguard requirements. And the Committee hoped that in the long run the demonstrated practicality of the safeguards--their confidence-building impact on people's attitudes toward record-keeping organizations and their effect on the accuracy of information in systematically maintained files--would induce at least some systems voluntarily to provide individuals with stronger protections than those required by law.

5. Recommendations on the Social Security Number

I suspect that you would like me to say a few words about the Committee's recommendations on the Social Security number.

I should begin by saying that the Committee saw the SSN issue as essentially symbolic. It is true that the SSN can facilitate data linkage and that strong arguments can be made against widespread use of the SSN as a personal identifier in the absence of stringent safeguards against possible abuses. Among possible abuses, moreover, I include systematic linkages that are not expected by the individuals whose records are being linked. (Heretofore our thinking about record-keeping abuses has been strongly influenced--and, I would submit, unduly and unhealthily constrained--by the concept of "unauthorized access," i.e., isolated instances of surreptitious entry by persons who are "just curious" or who want to help out their "buddies" in some other record-keeping organization.) By and large, however, the key issue raised by the use of the SSN as a personal identifier is the issue of whether, as a society, we want to have a universal scheme of unique personal identification--be it the SSN or, as the Committee asks in its report, some

other, more reliable identification device. And here, three questions arise: First, do we realize what having a Standard Universal Identifier (SUI) implies in the way of bureaucratic arrangements for assigning and verifying numbers? Second, what would we want to do with an SUI if we had one, i.e., who should be permitted to use it for what purposes? And third, by what process should the decision be made to have or not to have an SUI?

The Committee strongly suspected that if people realized the difficulty of making an SUI work--for example, an effective system would very likely require everyone to carry an identity card--the majority of Americans today would reject the idea out of hand. On the question of permissible uses, I think I have made it clear that today there are few rules governing the exchange of information about individuals among record-keeping organizations, and in the absence of such rules it does not seem wise, as a matter of public policy, to institutionalize a powerful piece of information exchange technology before deciding how that technology may or may not be used.

Finally, there is the question of process--in this case, the process for deciding what ought to be done and how. The Committee's report documents the gradual metamorphosis of the SSN from an account-numbering system for Social Security programs into something approximating a personal identification device for a broad range of activities that have no relationship whatsoever to the administration of the Social Security system. The evidence suggests that the Federal government itself has been in the forefront of expanding the use of the SSN (more often through administrative action than through explicit legislative initiative) but it also suggests that we have now reached a point where decisions being made throughout the society are greatly accelerating the drift toward making the SSN an SUI. Today, even organizations selecting a single-system personal identifier are likely to choose the SSN "because it is available," and, therefore, convenient and efficient to use. What is critical to note, moreover, is that these decisions are being made, in effect, behind closed doors. The public is not being consulted even though there is no reason to think that the long-run consequences for individuals in our society will be benign.

The Committee's estimate of our current situation with respect to the use of the SSN as a personal identifier led it to make a strong recommendation against the adoption now, or in the near future, of any nationwide, standard, personal identification format, with or without the SSN. It called for a halt to the drift toward making the SSN serve as an SUI, coupled with prompt action to establish safeguards providing legal sanctions against present and potential abuses of automated personal data systems. Once such safeguards have been created, and have been shown to be effective, the Committee felt that the question of expanded SSN use might properly be reopened.

So far as Federal policy on the SSN is concerned, the Committee recommended that all decisions be consistent with three general principles:

(1) Uses of the SSN should be limited to those necessary for carrying out requirements imposed by the Federal government.

(2) Federal agencies and departments should not require or promote use of the SSN except to the extent that they have a specific legislative mandate from the Congress to do so.

(3) The Congress should be sparing in mandating use of the SSN, and should do so only after full and careful consideration preceded by well advertised hearings that elicit substantial public participation.

When the SSN is used in instances that do not conform to these principles, no individual should be coerced into providing his SSN, nor should

his SSN be used without his consent. Furthermore, an individual should be fully and fairly informed of his rights and responsibilities relative to uses of the SSN, including the right to disclose his SSN whenever he deems it in his interest to do so.

The Committee recommended specific, pre-emptive Federal legislation giving individuals these protections and providing in addition that no organization or person required by Federal law to obtain or record the SSN of any individual may use or disclose the individual's SSN without his consent, except as may be necessary to the Federal government purposes for which the SSN was required to be obtained and recorded.

With respect to HEW policy on the SSN, the Committee recommended that there be no positive program of issuing SSNs to children below the ninth-grade level; that the Social Security Administration provide "SSN services" only to organizations or persons that are required by Federal law to obtain or record the SSN, and then only as necessary to fulfill the purposes for which the SSN is required to be obtained or recorded; and that the Secretary limit affirmative measures taken to issue SSNs pursuant to Section 137 of Public Law 92-603 (H.R. 1) to applicants or recipients of public assistance benefits supported from Federal funds under the Social Security Act.

6. Concluding Remarks

These, plus the Code of Fair Information Practice, constitute the principal recommendations of the Secretary's Advisory Committee. The Committee considered other approaches, notably the creation of a centralized, independent, government agency to regulate the use of all automated personal data systems. Such an agency, if authorized to register or license the operation of data systems, could make conformance to specific safeguard requirements a condition of registration or licensure. The Committee felt that a regulatory approach might be appropriate for certain types of record-keeping operations. For example, it noted the provisions of the Fair Credit Reporting Act which assign specific enforcement responsibilities to a few existing regulatory agencies, and it was also aware of experiments with quasi-regulatory mechanisms for applying safeguard requirements to so-called "integrated municipal information systems." In general, however, the Committee doubted that the need exists, or that the necessary public support could be marshalled at the present time for an agency of the scale and pervasiveness required to regulate all automated personal data systems.

Enactment of a Code of Fair Information Practice, coupled with limitations on use of the Social Security number, seemed to the Committee the best way to begin. It will require the creation of no new institutions and may well give record-keeping organizations all the incentive they need to bring their operations into conformity with the desired standards of fairness.

537-860 O - 74 - 18

Enforcing Naming Standards Through Use of a Data Dictionary[1]

Patricia A. McNamee[2]

Celanese Corporation
Box 1414
Charlotte, N. C. 28201

With the advent of third generation computer technology a trend toward centralization of data processing equipment and services was seen in many businesses. When development activities were also centralized, the problems of data classification, recognition, and definition were compounded by the sheer mass of data required. A diversified manufacturing corporation with many divisions or companies needs new methods to aid in identifying similarities or differences in data definitions. A data dictionary used with new disciplines in the definition of data can go a long way towards solving this problem.

Key Words: Acceptance; accessibility: data dictionary; data element definition; data element names; enforceability of data element standards; glossary.

1. Introduction

Naming standards for data elements should interest anyone who must deal with the corporate data base philosophy. This subject is too often overlooked in developing corporate data bases. I would first like to give you a brief background of the organization with which I was associated at the time of first publishing this paper.

Revised by author for this Symposium.

[2]Standards and Education Manager

2. Background

The company is a diversified manufacturing corporation. Like many businesses which have numerous divisions or companies within the corporation, it had an almost like number of computer centers. The availability of third generation computer technology made centralization desirable.

If you ask, "Why consolidation?" the following points may answer your question.

The first point is management consolidation. The second point stresses reducing development redundancy by sharing new systems, knowledge, and data. It is this point, in particular, that will be addressed today. The third point involves computer systems development. Consolidated technical support, computer design, programming, and data base design allow better scheduling of manpower. The fourth point, pooled hardware resources, provides economy of scale, reduces hardware limitations, allows simulation. The fifth point is increased purchasing power. Consolidation carries a bigger stick. Better service and more negotiating strength result. The sixth point better satisfies the requirement for training. Consolidation permits in-house training, shared knowledge, and specialization.

3. Corporate Data Base Concept

When a company embraces the corporate data base concept, new disciplines are required. In addition to the new technology and expertise required to support the mechanics of a corporate data base, new standards are required.

4. New Tools Required

4.1. Need for a Data Dictionary

To support the goals for a corporate data base, new tools must be introduced. IMS provides some of the tools needed to eliminate or reduce redundancy. Redundancy at the data element level cannot be reduced without recognition of the sameness of data elements. We hope to achieve this recognition through use of a data dictionary and the establishment of new guidelines and standards.

4.2. Data Dictionary

The data dictionary concept should support and relate many elements. These include:

- Data Element Definitions
- Segments
- Data Set Groups
- Programs
- Transaction Codes
- Physical Terminals
- Logical Terminals
- Lines

Today's discussion is centered on data element definitions, in particular, standardized names for data elements.

5. Data Elements

5.1. Definition of a Data Element

The definition of a data element which I am about to offer is a quotation from the draft of the proposed technical report, "Guide for the Development, Implementation, and Maintenance of Standards for the Representation of Computer Processed Data Elements", which was prepared by the American National Standards Institute.

The guide says, "In information processing and exchange the data element is used to identify the intended field in a record. The data element thereby forms the fundamental building block out of which all information structures (records, files, and data bases) are made".

"Therefore, the basic unit of information, the data element, has a name which serves to identify it and to distinguish it from other data elements".

It is this identifying and distinguishing name that I will address today.

5.2. What We Hope to Achieve by Standardization of Data Element Names

Suppose by magic you could have all the data elements which are defined in all of your existing data files or data bases dumped into a dictionary tomorrow! Do you realize what you would have? Could you identify redundancies by your current definitions? Could you identify some of your elements at all? The point is, mechanical control of data elements alone is not going to solve your problems.

This is the problem confronting us today. Differences in nomenclature between systems is a carry-over from the decentralized period. Differences in nomenclature within a system can result from lack of guidelines to follow or indifference on the part of development personnel.

Standardization becomes absolutely necessary if there is to be a high level of recognition of data element names.

There are distinct objectives which must be met in standardizing data element names. Our goal is to eliminate or reduce ambiguity by providing one common accepted name for each data element. To reduce redundancies, again recognition of a common accepted name is paramount. We should be able to accelerate and facilitate the development process by providing commonality of terms throughout development on through to operation and maintenance.

We should also be able to facilitate training by standardizing data element names. Every manufacturing organization has a language all its own, from the raw materials which make up the finished product to the machinery used to produce it. Glossaries of standard data element definition can provide textbooks of knowledge on the industry they cover. This information, made easily accessible, can provide the background to make a new analyst effective months sooner.

5.2.1. Control

Controlling data elements in a data dictionary provides a vehicle for standardization. A dictionary is not an end unto itself. It is a user responsibility to put standardization in the driver's seat.

5.3. Criteria for Achieving Standardization

Let us now look at the criteria for standardizing data names.

5.3.1. Acceptance

The method used to standardize data element names must be psychologically acceptable. Standardizing a name implies restriction of that name to current acceptable words.

This must be done with care so as not to cause resistance by those who will deal with the standard name. Individual preference and tradition will almost always make this task difficult initially. It is important that those affected by standardization understand the long range goals and benefits.

5.3.2. Universal Identification

The method used to standardize data element names must provide universal identification. That is to say, the name used in system specifications should also be identifiable in a COBOL program by a close approximation of that same name.

5.3.3. Accessibility

If a dictionary is to be effective for maintaining data element definitions, each element definition in the dictionary must be accessible individually. Additionally, the elements should be accessible collectively be a category specified by the user.

5.3.4. Enforceability

No standard is any good unless it can be enforced. Webster states that "Standard applies to any definite rule, principle, or measure established by authority."

The group which sets standards for data element names must have the authority to do so. But even given authority and support by upper management, the standard-maker must have some means of enforcement.

6. Methodology for Standardized Data Element Names

6.1. One Approach

After looking at several methods of formulating data names, we decided to explore further an adaptation of the method used in IBM's IUP Data Dictionary/Directory. This method was chosen because it most closely paralleled those names we found, in our analysis of existing names, to be highly recognizable.

6.2. Parts of the Data Element Name

A data element name is an English language series of words. The name must contain enough identifying, qualifying, and modifying information to distinguish this data element from another. The words which make up the name we shall call "designator" words.

The designator words are of three types: prime words, modifying words, and class words.

The prime word is the thing being described. Modifying words provide further definition and qualification of the prime word. The class word is used to indicate type or class of data contained in a data item. Typical class words are quantity, code, amount, and number.

In this scheme, each data element name must contain one or more modifying words, one prime word, and one class word, in that order. Applying the above rules, we can look at the representative list of words in Table 1 and formulate many data element names. The example shows that "Style Number" occurs in two data element names, "Finished Style Number" and "Greige Style Number." Similar, yet entirely unique as a result of the modifying words, "Greige" and "Finished". In this example "Style" is the prime word and "Number" is the class word.

Table 1. Modifying, prime and class words combined to form a standard data element name.

Modifying	Prime	Class
Finished	Order	Number
Greige	Department	Amount
Company	Color	Date
Customer	Style	Code
	Data Element Names	
	Finished Style Number	
	Greige Style Number	
	Company Order Number	
	Customer Order Number	

When analysis was made of data element names in use which had no standard applied to them, I found 702 differently spelled words in 3700 data names in a COBOL COPYLIB. These data names represent IMS data bases developed since 1970. In this analysis a word was that group of characters between hyphens or the beginning or end of the data name. Of these 702 words 58 were completely unrecognizable even in the context of the names in which they were used. Of the remaining 644 words, analysis proved that 97 standard words could replace them. All of this reduction could be accounted for merely by condensing variations of spelling and abbreviations. No attempt was made to isolate synonyms for which a standard word would suffice.

7. Universal Identification - Applying These Names

If we are to achieve a goal of accelerating the development process then we must apply the standard data element name across all stages of the development process.

The documentation of a system, starting even at the feasibility study level, should use the standard names if it is necessary to refer to data elements. In the early stages of designing a system the analyst starts to define the data which the system will manage. In the process the data elements will be listed and grouped by function. The data base designer will review this list and identify those elements which already exist in the dictionary. He will make new entries for those that do not.

As data base design continues, the necessary data elements are linked to the segment being defined. Any documentation concerning this segment and its elements will be produced by the dictionary.

As computer programs are designed, references to data elements should be made only by the names produced by dictionary documentation.

COBOL programs will use the element names in the COBOL COPYLIB members which, in turn, are generated by the dictionary. To a great degree, the program documentation may be obtained from the dictionary itself. Those portions which the programmer himself must write should reflect the standard data element names. Systems documentation will also use standard terminology.

If a standardized approach is to succeed, the user should also refer to the data element by standardized terminology. This system will do nothing overt to enforce this particular item. It will provide the information necessary to point out such inconsistencies in the design phase. It will then be the responsibility of the standard-maker to sell the standard name to the user.

Documentation for the user will provide the standard name. User documentation works hand-in-hand with user inquiry responses and reports.

Universal identification does not mean the same exact spelling of the names for all of the above uses. Instead, it means two basic types of representation. The first type as shown in Table 2 for the data element name "Manufacturing Order Number" is the English language spelling of the name. At this point the element need not be associated with a particular data base. It will be used in system specifications. It will also be used in data base documentation, system documentation, user reports, and user documentation.

Table 2. Two basic representations for the same data element name.

Data Element Name-Manufacturing Order Number	
Representation	Used In
Manufacturing Order Number	System Specifications Data Base Documentation System Documentation User Reports User Documentation
KBØ1-MFG-ORD-NBR	Data Base Documentation Program Specifications COBOL Programs Program Documentation System Documentation

The second type of representation occurs after the element has been associated with a data base. In the prefix, the "KB" identifies the data base, and the "01" identifies the segment. This type will be used in data base documentation, program specifications, COBOL programs, program documentation, and system documentation where required.

It is worth mentioning that the data element with its prefix should not be stored in the dictionary as a separate entity. Instead, the identifier prefix becomes a factor of the relationship of the element to the segment. The element may be related to any number of segments.

8. Need for Accessibility

Maintaining data element definitions in a dictionary or thesaurus is of little use if the definitions are not easily accessible. If a dictionary is to be used as a development tool in addition to its better known role as a control facility, then accessibility must be provided in terms of what the user seeks.

8.1. Requirements for Accessibility

The requirements for accessing data element definitions vary depending on the user and the task at hand. In the earliest stages of development of a system, the analyst would be interested in seeing the definitions which relate to his subject.

9. Glossary

A glossary could provide the information he is seeking. A glossary, by standard definition, lists all terms applied to the subject it covers, with corresponding definitions. A glossary should prove useful as reference material for standardization and comprehension of terms. This glossary must be provided in the subset of terms of interest to the user. Only in rare cases would a glossary of all the data element definitions maintained in the dictionary be required.

Since a glossary does imply a subset of terms on a specific subject, a way must be provided to place a subject category on each data element defined to the dictionary. One method of achieving this is by a published list of categories. A data element definition will always belong to one of these categories more than to another.

A representative list of categories might include customer orders, product description, inventory, financial accounting, manufacturing, and process control. Affixing a category to each data element makes it possible to obtain a glossary for that category. Any one data base could contain elements from several categories.

Multiple users may be associated with a data element. When addressing standardization and conformity, the concept of multiple users of a single data element is very important. When this actually occurs it can be construed as a measure of success for the standardization effort.

As a result of associating category and user with each data element, it is possible to produce glossaries by user or category, or category within user. Such glossaries provide lists of existing data elements to use as development aids.

10. Keyword Search

When access is required at an individual element level, a keyword search technique provides a useful tool. In this case our designator words now become keywords.

In the case of on-line entry and response a minimum of one modifying or prime word would be required for entry.

In our example, the user wishes to search for an element whose prime word is "Order". Since no modifying word or class word was supplied, the response will include all data element names whose prime word is "Order".

This response may be rather voluminous, so it is likely the user will supply more than one designator word.

11. Synonyms

The user should be allowed to inquire in terms that are familiar to him. These terms may be synonyms for the approved standard terms. Where it is known such synonyms exist, they should be linked to the proper standard entry.

12. Enforceability

Let us now look at a method to enforce these standards. As new data element definitions are added to the dictionary they must be edited for conformity to standards. This means that each word in the data element name must appear in the list of approved words. The name must also have one prime word, one class word, and one or more modifying words.

As new standard words are approved, they must be added to an edit data base. When a new data element definition is added to the dictionary, it must be edited positionally. The word in the class position must appear on the edit data base as a class word. The same rule applies for the prime word and the modifying word.

If the new data element definition passes all edits, a linkage is established from the edit data base to the data element definition. The edit data base can now function as a keyword data base.

13. Acceptance

I would like to conclude with a word about acceptance. There seems to be no one best way to introduce standardization into a data processing environment. But do remember that the goal is to make standardization welcome.

The Need for Standardization of Data Elements and Data Codes -- from Origin of the Effort to Partial Fruition

James W. Pontius

Finance and Service Operation
General Electric Company
Schenectady NY 12345

The need to investigate at the national level the feasibility of an effort to standardize data elements and data codes for use in information interchange first surfaced late in 1964. The first meeting of people interested at the national level took place on 1965 April 7.

The paper then relates the steps taken in 1966 and 1967 by a large industrial manufacturer, General Electric Company, leading to formulation in 1968 of a Corporate program for development and adoption of standard data elements and data codes. The organization of the Corporate effort and the continuing practical implementation of the program are described.

Key words: Automatic interchange; calendar date; company standard; corporate program; data code; data element; Federal standard; industrial manufacturing; international standard; national standard; standard; standards register; work scope.

1. Origins of the National Effort

1.1 Identification of Needs

The need to investigate at the national level the feasibility of an effort to standardize data elements and data codes for use in information interchange first surfaced late in 1964. While it turned out that the need appeared to have been recognized by various organizations, it was the desire of a specific supplier and a specific customer to exchange business transactions automatically that led to the formal effort. The supplier was General Electric Company and the customer was a large electric utility in the eastern United States.

The two organizations wanted to launch a demonstrative project to identify and automatically execute without human intervention the exchanges of data required to handle one complete business transaction between the two parties -- from the original request by the customer for a quotation through quotation, placement of the order, engineering, manufacturing, shipment, billing, and payment for the product. The last step would have involved banks as third parties of course.

1.2 Lack of Common Data Elements

Adequate information processing and data transmission facilities and software appeared to be available and presented no real problems. The project faultered when it became evident that no basis existed for common identification, definition, and specification either of data elements, data items, and data codes or of data interchange procedures.

1.3 Organization of Effort

a. Sponsorship

Once the needs were identified, a way of pulling together an appropriate effort under a common umbrella was attempted. One of the individuals whose counsel was sought was Charles L. Phillips, then Director, Data Processing Group, Business Equipment Manufacturers Association (BEMA) and, as many of you know, formerly of the Office of the Assistant Secretary of Defense (Comptroller) in Washington. While potential sponsorship by at least five different organizations was considered, there seemed to be a natural gravitation to support of the effort by BEMA as a part of its sponsorship of Sectional Committee X3 Computers and Information Processing under procedures of the American National Standards Institute, Inc. -- provided that substantial participation by responsible individuals and organizations could be demonstrated.

b. First Meeting

The first meeting of fifteen to twenty people representing both non-government and government organizations interested at the national level took place in New York on 1965 April 7, and the organizational meeting followed one month later. It is interesting to note that seven of fifty individuals on the membership list as published following the second meeting in 1965 July are still on the current list of more than eighty members, alternates, and observers for Subcommittee X3L8 Representations of Data Elements.

c. Formulation of Work Program

The development of objectives and work scope soon followed, and the organization, planning and division of work into task groups was underway. The formulation of definitions, criteria, and methodology consumed a great deal of effort, but gradually some order was brought to the unchartered ground and the Standards Planning and Requirements Committee (SPARC) of Sectional Committee X3 gave approval to establishment of Subcommittee X3.8 Data Elements and Coded Representations, as the Subcommittee cosponsoring this Symposium was named at the time. It should be recognized that the Chairmanship of the Subcommittee passed through several capable hands, including John F. McCarthy, the first Chairman and his successor David V. Savidge who became Chairman in the spring of 1967. Present Chairman Harry S. White, Jr., then followed and took over the leadership in the fall of 1968.

d. Manpower and Implementation

My notes show that during the early discussions of the potential scope of the work program back in 1965, I ventured an estimate that the intermittent participation of perhaps 500 individuals, with about half of them working an average of three days each month over a period of ten years, would probably be required to produce a reasonable number -- probably over 100 -- of the more important standards. At the time, those estimates seemed very bold, and the problem of obtaining the participation of so many capable people over such a long period of time was very real. It is clear now, of course, that the estimated period of elapsed time was too short, and, in retrospect, the relatively long period required for gestation of a true data standard was not understood. There is recognition now that the seemingly very slow process has the authoritative advantages of helping to assure simplicity, inherent rightness, and longer life for a standard.

1.4 Issuance of First Standard

The first American National Standard relating to data elements and data codes was approved and issued on 1971 July 1 as ANSI X3.30-1971 Representation for Calendar Date and Ordinal Date for Information Interchange. Several others have since been issued. An increasing number are nearing approval, and each year should bring the goal closer to reasonable fruition.

2. Origins of the International Effort

2.1 Need for Extension of National Effort

Individuals attracted to the effort soon came to realize that the common scope of the work extended beyond national to international requirements.

2.2 Organization

At a Plenary Session of Technical Committee 97 Computers and Information Processing of the International Organization for Standardization (ISO) held in Tokyo, Japan in 1965 October, the formation of a Working Group for standardization of data elements and data codes was authorized. The United States accepted the Secretariat and six countries indicated their intention to participate: France, Italy, Japan, Switzerland, the United Kingdom, and the United States. Accordingly, in 1966 January, Working Group K Data Elements and Their Coded Representations was organized at a meeting held in Geneva, Switzerland and William E. Andrus, Jr., of the United States was elected the first Chairman.

2.3 Issuance of First Standard

It is interesting to note that the same individuals attended a TC 97 Advisory Committee meeting also held in Geneva in 1966 January and made the first TC 97 recommendations relating to data elements and data codes -- the writing of dates and the numbering of weeks. The TC 97 recommendations were submitted to the ISO Coordinating Committee on the standardization of the writing of dates (DATCO). Five years later, the first ISO Recommendations relating to data elements and data codes were approved and published in 1971 January as R 2014 Writing of Calendar Dates in All-numeric Form and R 2015 Numbering of Weeks.

2.4 Suggestions for Future Participants

In this connection, and based on limited experience in international standards work, there are two suggestions I would like to pass on to future participants as aids in accomplishing their goals. First, make important presentations in two languages, such as English followed by French. Contrary to some impressions, not everyone understands English -- even though it may be designated as the official language of the meeting. Presentation in a second language will help to bring about understanding on the part of multilingual participants whose comprehension of English is non-existent or poor. In addition, presentation in a second language by someone from the United States will help to make friends and win votes since it is seldom done. Secondly, volunteer for assignments such as chairman of the editing committee, secretary, or drafter of resolutions. The power of the holder of the pencil in a small group of people attempting to reduce verbal deliberations to writing in the presence of language barriers is substantial.

3. Corporate Program of a Large Industrial Manufacturer

3.1 Responsibility

a. Assignment by Chairman of the Board

As one example of corporate implementation by a large industrial manufacturer, the effort to standardize data elements and data codes within the General Electric Company was formalized in 1966 June when Mr. G.L. Phillippe, Chairman of the Board, recognized the need to correct a "code explosion" which had taken place over the years by assigning responsibility for "a continuing panel on data systems codes for the benefit of all operating components" to what is now the Corporate Computer Planning Operation and is directly responsible to the Vice President and Comptroller.

b. Establishment of Corporate Committee

Responsibility for formulation of objectives and for development and approval of standards is in the hands of a Corporate Information Standards and Codes Committee. The eight members of the Committee are drawn principally from the various functions represented in the Corporate Administrative Staff and serve on the Committee in addition to carrying out their regular responsibilities. The Chairman of the Committee is Walter F. Schlenker, Consultant-Application Development, Corporate Computer Planning Operation who also has been serving as Chairman of Task Group X3L84 Geographic Units for the past few years.

3.2 Organization of Work

The corporate program is presently divided into seven areas of work,each carried out by a separate subcommittee of five to ten members and chaired by a member of the parent committee. In general, the members of the subcommittees are individuals with fairly long functional or information systems experience and are drawn from widely varying operations throughout the Company for particular authoritative contributions they can make. As in the case of the parent committee, the members of the subcommittees carry out standards work in addition to their regular responsibilities. Each subcommittee meets seven or eight times each year.

With the benefit of early participation in the ANSI X3L8 Subcommittee's efforts, a plan for implementing a corporate data standards program was developed and implementation was begun in 1968 October. While the plan calls for the principal efforts to be directed toward the development and adoption of standard data elements and data codes, the way is left open for specification of other standards. For example, one major corporate standards project undertaken outside the area of data elements and data codes resulted in the standardization during 1973 of invoices issued to customers.

3.3 Documentation of Standards

Patterned in part after Federal procedures, approved standards are published as General Electric Information Standards in the General Electric Information Standards Register which is distributed throughout the Company. The terminology developed in the work of Subcommittee X3L8 is used in the standards.

3.4 Status of Work

A total of eighteen standards has been adopted to date, each with a specified implementation date which is usually one to three years after issuance of the standard. In general, mandatory use of specified data elements and data codes is required where data is exchanged between major components of the Company.

About thirty standards are currently in the development stage, and five to ten requests for standardization of data elements and data codes have been rejected as not warranted.

The standards work of both ANSI and the National Bureau of Standards is monitored very closely and, to the extent applicable and feasible, the American National Standards and Federal Information Processing Standards are adopted. So far there are no variances. Naturally, some of the data elements standardized are not used outside General Electric.

4. Implementation in One Component of the Company

4.1 Responsibility

Responsibility for maintaining the Information Standards Register and for implementing the standards in each operating component of the Company is assigned to the Manager-Finance.

4.2 Benefits

The component which I represent serves several large pooled sales organizations which account for about one-third of the Company's annual sales volume. The interest in standardization of data elements and data codes has been very high for at least three important reasons:

> Adoption of standard data elements and data codes eliminates duplication, removes ambiguities, clarifies communications, and results in operating economies.
>
> Large volumes of related records are interchanged between headquarters, 15 large and 45 small offices throughout the United States, and about 65 departments within the Company or about one-third of the total number of departments.

The specification of national standards both for data elements and data codes and for information interchange procedures must take place prior to any appreciable automatic interchange of records between customers and the Company as a supplier, or by either with other parties such as banks.

4.3 Implementation of Information Standards

a. Integration with National and Federal Standards

With the momentum of the Company-wide program and fewer complexities to deal with, progress within the Company is beginning to move at a faster pace than the national efforts. This has the disadvantage that a standard adopted by the Company and implemented by its components may turn out to be different from an American National Standard issued at a later date. Care is being taken to keep abreast of both the National and Federal efforts to hold such cases to a minimum. In any event, any subsequent conversion from a particular Company standard to an American National Standard should be simplified in that the various names and definitions previously used to identify the same data element and the different data items and data codes used in the past will already have been abandoned in favor of one Company standard.

b. Timing

In general, the least effort is required in implementing a standard data element and data code if this can be done coincident with redesign of an information system. However, this is not always possible. In certain difficult cases, translation techniques may be used to achieve implementation by the effective date.

4.4 Media Used

The media presently used for interchange of data presently run the gamut from the U.S. mail for transportation of hard copy, paper tape, punched cards, and magnetic tape to various low and high speed communications facilities for remote batch and on-line data transmission. Generally speaking, managers use the data in serving customers, filling orders, sales direction, management of assets, control of expense, budgeting, and business planning.

4.5 Data Interchanged

a. With Customers

Among the principal records interchanged with customers are quotations, orders, acknowledgements of orders, shipping promises, invoices, and payments.

b. Within the Company

Data exchanged within the Company is included in records of customers, suppliers, orders received, pending business, shipments, billing to customers, cash received from customers, open orders, accounts receivable, financial transactions and accounts, expenses, employees, mailing lists, property, and other transactions.

THE NECESSITY AND MEANS OF DISCIPLINING DATA ELEMENTS IN A COMPUTER SYSTEMS ENVIRONMENT

MERLE G. ROCKE

Caterpillar Tractor Co.
General Offices
100 N.E. Adams St.
Peoria, Illinois 61602

The underlying premise of this presentation is that "all data elements - and data codes in particular - must be rigidly disciplined and controlled before shared data bases and common information processing systems can fully achieve their objectives." The application of automated data processing systems to business management functions usually points out the need for new or revised methods of classifying, coding and representing the data involved. Manual systems benefit when data is well-defined, organized and consistent; computer systems, however, require rigidly disciplined data to fully utilize logical decision rules. The "stand-alone" computer system is rapidly becoming a thing of the past. Large shared data bases and common application systems are now technically feasible. It is becoming increasingly critical that data elements be rigidly disciplined if we are to realize the full impact of the many potential benefits afforded by current technology. Accordingly, this paper describes 1) the need for, 2) methods of, and 3) benefits of controlling the development, usage, and modification of data elements and coding structures. Included are discussions of 1) problems which result from lack of control, 2) methods of resolving and avoiding those problems, and 3) actual examples and experiences to illustrate various points.

Key words: Computer; control; data base; data code; data codification principles; data element; data processing system; information interchange; data representation, standardization.

537-860 O - 74 - 19

1. Introduction

1.1 Importance of Data Codes

The necessity for managing and controlling data elements in automated data systems is widely recognized. In fact, this entire Seminar is devoted to the broad subject of "The Management of Data Elements in Data Processing." This particular paper, however, is primarily concerned with a specific type of data element - the data code. Data coding is fundamental to all types of data systems since it is codes which serve as "keys" which identify the information stored on mechanized media (e.g., magnetic tape, disk, etc.). Codes are the fundamental units of which information structures are built and are necessary to facilitate the logical recording, accumulation and presentation of facts for management. Accordingly, this writing directs attention to a new information systems function - that of controlling the development and usage of data coding structures.

1.2 Definition

The term "data code" is important enough to merit specific definition before proceeding further. A data code may be defined as a brief label, composed generally of one or more letters and/or numbers, which identifies an item of data and has the capability of expressing the relationship of that data item to other items of the same or similar nature. The complexity of the relationship governs the complexity of the coding structure.

2. The Problems

Because of previous limitations in both hardware and software, each system designer designed his own system of data representations to fit the problem at hand. Inevitably, many found unique solutions useful for specific applications in limited areas. Collectively, however, these unique solutions set the stage for a computer age Tower of Babel. The various functional areas of a business organization (e.g., the Accounting, Manufacturing, Engineering and Sales departments) frequently develop unique, incompatible data codes for representing a common entity set. Each segment of the company, for example, often has its own code structure for classifying and identifying the various products produced by the firm. Today these circumstances result in tremendous duplication of stored data within an organization, extensive cross-reference lists, and extreme difficulty in sharing data among the indivdual units of that organization - primarily because a multiplicity of code structures exists where a single structure would suffice.

To further elaborate, many or all of the following undesirable circumstances could be avoided if data code structures were better disciplined.

1. Redundant code structures - multiple structures exist where one would suffice; they may differ in format, size, or meanings.

2. Incompatible code sets - code values of one set cannot be converted to equivalent values of a "synonymous" code set because of incompatible definitions.

3. Inadequate flexibility - the structure, due to its design, has limited application.

4. Insufficient expansion capacity - little capability exists for coding additional entities.

5. Cumbersome code structures - because of inadequate design, the code structures are unwieldy.

6. Complex computer programs - programs are more complex as a result of 1, 2, 3, and 5.

7. Instability of individual code values - frequent updates or "replacements" are necessary.

8. Confusion among code users - this results when several code structures are similar and have the same name but still have essential dissimilarities.

9. Unnecessarily high response time - desired changes take longer to implement because of inadequate code design.

10. Duplication of effort - this occurs when many similar but slightly different code sets must be maintained.

11. Undetected errors; invalid data - these result when the format or definition of a data code is changed without properly notifying affected areas.

12. Erroneous decisions - these result from invalid data, use of improper code structures, etc.

13. Operating costs higher than desired - as a result of the above circumstances.

3. A Solution

These undesirable circumstances may be avoided by the establishment of an effective, central data code control function, usually within the information processing organization. The two basic reasons for this function are to ensure (1) the development and use of standard data codes (i.e., standard formats, definitions and code values) and (2) that new data codes conform to established principles.

The responsibilities of this central data code control function should include:

1. Publication of standards and procedures necessary to support data code standardization and discipline.

2. Publication and maintenance of a directory of "standard" data code structures for use by application systems. This publication should document the size, format, definition, source of specific values, controlling user, etc. for each code.

3. Provision of necessary education for systems analysts in the areas of data codification principles and methods (discussed in greater detail later in this paper).

4. Provision of assistance and guidance to operating areas which have ultimate responsibility for the design of new coding systems or modification of existing ones.

5. Review and approval of proposals for new or changed data coding structures to ensure conformance to established data codification principles.

6. Liaison between the organization's physical locations, functional units, and data processing application systems groups to review the impact of proposed changes, establish effective dates for code changes, etc.

It is important to note that the ultimate responsibility for a given code structure should lie with the organizational unit most directly affected by or concerned with the code. For example, Engineering controls part numbers; Purchasing controls supplier codes, Accounting controls the chart of accounts, etc. These areas, therefore, should do the actual design and development of their own data codes because they are most familiar with their specific informational needs. The central control authority, however, provides consultation services to each operating area during the development of a code structure and is charged with the responsibility for final approval before the new structure can be implemented.

The operating area (e.g., Engineering) responsible for a specific data code controls the day to day assignment of specific code values (e.g., new product serial numbers, new part numbers, etc.) to ensure adherence to the assignment conventions established for the code. This area is also responsible for appropriate publication and maintenance of all current values of the code.

4. Data Codification Principles and Methods

An indepth understanding of the principles and various methods of data coding is essential to support effective reviews of proposed data coding structures. This information must also be given to initial designers of data codes.

4.1 Ten Traits

The following are ten characteristics of a sound data coding system. These traits must be considered if the information processing systems supported by the coding system are to be viable, effective and stable. It should be noted, however, that all these traits are not completely compatible with one another. Trade-offs must be considered when several traits are in conflict. The data coding system eventually selected for implementation should reflect these characteristics to the greatest possible degree.

1. Uniqueness
   The code structure must ensure that only one value of the code, with a single meaning, may be correctly applied to a given entity.

2. Expandability
   The code structure must provide reasonably sufficient space for entries of new items within each classification.

3. Conciseness
   The code should require the least possible number of positions to adequately describe each item. Brevity is advantageous for human recording, communication line transmission, and computer storage efficiencies.

4. Uniform Size and Format
   Uniform size and format is highly desirable in mechanized data processing systems.

5. Simplicity
   The code must be simple to apply and easily understood by each user, particularly workers with the least experience.

6. Versatility
   The code should be easily modified to reflect changing conditions, characteristics, and relationships of the coded entities.

7. Sortability
Obtaining reports in a predetermined format or order is desirable. Reports are most valuable when sorted for optimal human efficiency.

8. Stability
Code users need a code which does not require frequent updating. Individual code assignments for a given entity should be made with a minimal likelihood of change. Uncontrolled and unlimited changes are laborious, costly, and likely to breed error and reduced confidence.

9. Meaningfulness
For greater meaning, code values should indicate some of the characteristics of the coded entities, such as mnemonic features, unless this causes the code to become inconsistent, inflexible or unwieldy.

10. Operability
The code should be adequate for present and anticipated data processing mechanization as well as human reference. Care must be exercised to insure that clerical effort or computer update and maintenance time necessary to preserve required relationships does not grow to unmanageable proportions.

4.2 Specific Principles

In addition to these general characteristics, a number of more specific data coding principles have evolved. Among the more important of these are:

1. Character Content. Characters other than letters or numbers (such as the hyphen, period, space, asterisk, etc.) are to be avoided in code structures (except for separating code segments, where a hyphen may be used). Upper case letters only, i.e., ABC...Z (not abc...z) are to be used in data codes.

2. Visual Similarities. When it is necessary to use an alphanumeric random code structure, characters that are easily perceived as, or confused with, other characters should be avoided. Some examples are: letter I vs. number 1; letter O vs. number zero; letter Z vs. number 2; slash, or virgule, / vs. number 1; and letters O and Q.

3. Acoustical Similarities. Nonsignificant codes should avoid characters that can be confused when pronounced (acoustically homogeneous); for example, the letters B, C, D, G, P, and T or the letters M and N.

4. Vowels. Avoid the use of vowels (A, E, I, O and U) in alpha codes or portions of codes having three or more consecutive alpha characters to preclude inadvertent formation of recognizable English words.

5. Multiple Code Set Compatibility. More than one code or representation is necessary in some instances to meet most systems requirements. A single code is the ideal objective, but is not always the most practicable solution. Multiple codes, if needed, should be translatable from one code to another, i.e., the data items remain unchanged, only the codes are variable.

6. Mnemonic Codes. Mnemonic codes may be used to aid association and memorization, thus increasing human processing efficiency, provided they are not used for identification of long, growing lists of items. Mnemonic structures must be carefully chosen, however, to ensure that flexibility is not sacrificed. Mnemonics should generally not be used if the potential code set exceeds 50 entries, because the effectiveness of the mnemonic feature decreases as the number of items to be coded increases.

7. Code Naming. All independent data code segments must be individually named with standard, unique, consistently applied labels.

8. Calculation of Code Capacity. When calculating the capacity of a given code for covering all possible situations while maintaining code uniqueness, the following formula applies (assuming 24 alpha characters and 10 numeric digits are used because the letters I and O should be avoided whenever possible):

$$C = (24^A) \cdot (10^N)$$

where
C = total available code combinations possible
A = number of alpha positions in the code
N = number of numeric positions in the code
(A + N, when combined, equal the total positions of the code)

Note: This formula assumes a given position is either alpha or numeric - never both. If both alpha and numeric characters can appear in all code positions, the formula becomes

$$C = (34)^{A+N}$$

9. Segmentation. Codes longer than four alphabetic or five numeric characters should be divided into smaller segments by a hyphen for purposes of display and reliable recording, e.g., XXX-XX-XXXX is more reliable than XXXXXXXXX.

10. Alphabetic versus Numeric. The recording of numeric codes is more reliable than that of alphabetic (all letters) or alphanumeric codes (letters and numbers). Controlled alphanumeric codes (i.e., where certain positions are always alphabetic or numeric) are more reliable than random alphanumeric codes. For example, AA001 (where the first two characters are always letters and the last three are numbers) is a more reliable code than when letters or numbers can appear in any position.

11. Character Grouping. In cases where the code is structured with both alpha and numeric characters, similar character types should be grouped and not dispersed throughout the code. For example, fewer errors occur in a three character code where the structure is alpha-alpha-numeric (i.e., HW5) than in the sequence alpha-numeric-alpha (i.e., H5W).

12. Code Position Sequence. If a code divides an entire entity set into smaller groupings, the high-order positions should be broad, general categories; and low-order positions should be the most selective and discriminating (including any prefixes and suffixes). An example is the date (YYMMDD).

13. Check Characters. When the number of characters of a proposed code exceeds four characters and when this code will be for purposes of identification of major subjects (e.g., organizations, projects, materials, individuals, etc.) consideration should be given to the addition of a self-checking character to avoid errors in recording.

14. Codes for Numeric Categories. Quantities or numbers should not be coded since this introduces additional translation and a loss of preciseness. For example, the numbers 1 to 99 could be coded A, 100-99 coded B, etc. This may be desirable for purposes of categorization, but statistical value is lost since the actual numbers cannot be derived once they are coded. Categorizations can be performed during later phases of data processing rather than in precoding of the input data.

15. Use of "Natural" Data. A code structure should not be developed if the specific data in its natural form (such as specific percentage amounts) is appropriate and adequate.

4.3 Coding Methods

The variety of data coding methods available is fairly extensive. Knowledge of the primary uses, the advantages and potential disadvantages of each of these methods is essential to enable selection of the best method for a given application. Codes may be significant (provide meaning in addition to simple identification of an entity) or nonsignificant. They may be assigned sequentially, randomly, or mnemonically. Codes which collate entities (place them in a predetermined sequence) may be alphabetic, hierarchical, chronological or classificatory. In short, the selection of codification structures and structural combinations is quite broad. Matching the coding method to its particular use is necessary for optimal effectiveness of the data code and the information processing systems it supports.

The developments resulting in expanding emphasis on and importance of data codification have been rather recent. Accordingly, not much literature is available on the emerging data code control function. Detailed information on data coding techniques such as those listed above is currently difficult to find. Additional information may be obtained from a technical report entitled GUIDE FOR THE DEVELOPMENT, IMPLEMENTATION AND MAINTENANCE OF STANDARDS FOR THE REPRESENTATION OF DATA ELEMENTS being developed by Committee X3L8 of the American National Standards Institute (ANSI), 1430 Broadway, New York, NY 10018. This handbook should be available from ANSI sometime this year.

5. Principles in Practice

As stated previously, the two primary objectives of central data code control are to ensure (1) the development and use of standard (commonly used) data codes and (2) that new data codes conform to established coding principles. Several examples of actual experiences may provide insight concerning practical application of coding principles. The following are examples of coding difficulties which could have been avoided had the code control function at my company been implemented sooner.

The first of the two code control objectives concerns standard data codes. When we developed our directory of standard data codes, we discovered about twenty different company facility identification coding structures being used in mechanized systems. Since much information is keyed by facility code, this information is quite difficult to compare or share - especially since code value "B" in one system may be the equivalent of both "23" and "47" in another system or "D30" and "M30" in still another system. By requiring that all new systems now use the corporate standard facility code, the others are now being phased out through attrition.

The importance of sound data codification principles cannot be overemphasized. We have experienced extensive problems because our dealer and customer identification codes are tied to geographic regions. The first code position designates the geographic region of the dealer or customer. This region identifier is necessary for code value uniqueness. (Without it, there are duplicate codes). Hence, when the Marketing Department redefines the geographic regions, many of the codes must be changed to avoid duplicates. Dealer and customer codes on all historical records and files must also be updated to reflect the changes. Problems 7, 9, 11, 12 and 13 as previously described all result because these codes were designed improperly. These codes were also initially designed with insufficient expansion capability, so field sizes must now be expanded in many computer programs and on forms - a costly process.

We have found that through deliberate efforts to correct existing data code deficiencies and by avoiding known pitfalls as new structures are developed, we have been quite successful in reducing the thirteen undesirable circumstances previously described. And reduced costs are the ultimate result.

6. In Defense of Codes

From time to time, a disgruntled user or manager may be heard to grumble about the tendancy of systems designers to dehumanize systems - forgetting the people who must make them work and the people they are designed to serve. The concern is usually over the use of codes in place of uncoded (English words and phrases) data. It is desirable then to review the five basic reasons for coding data in an information processing environment. They are:

1) To translate from a difficult-to-use source language (words/phrases) to one which is more oriented toward the needs of data translation and analytical activities.
2) To decrease required data element size per unit of information.
3) To supplement the information available in the source language (i.e., to provide more than simple identification of the item).
4) To distinguish between alternative ideas or words which are not easily distinguished (are ambiguous) when described in text.
5) To enhance accuracy of data translation processes.

Note that people benefit every bit as much as computers from usage of codes. In fact, people were using codes before the computer was ever invented! And both benefit from use of codes which are well disciplined. Advent of the computer only made us aware of disciplines we should have been applying all along.

Computers are wonderful. But they do impose some constraints (or inconveniences) on people - whether we like it or not. People prefer recognizable words and phrases. The computer requires disciplined codes if we desire it to make mechanical decisions based on the data. To the extent we can provide software to perform the necessary encoding and decoding function for us - great! But data disciplines are still necessary, simply because we deal with a machine. The point remains that if we are to successfully interface with the computer, we must adjust to its requirements. We must recognize, accept and cope with its idiosyncrasies - but must minimize and control them in order to reduce inconveniences for its users. In short, both machine and people needs and efficiencies must be considered when we design data coding structures and computer systems.

7. Conclusion

Discipline of computer processed data elements - and data codes in particular - is becoming increasingly necessary. Without it, integrated information processing systems will flounder. One should not, however, expect immediate results on a Profit and Loss Statement after implementing a data element control function as suggested in this paper. Be patient. Desirable results will be forthcoming.

Data Element Dictionaries for the Information Systems Interface

E. H. Sibley and H. H. Sayani

National Bureau of Standards and
University of Maryland, College Park, Md.

This paper deals with the design of data element dictionaries to structure the process of design, implementation, maintenance, and effective use of modern information systems. It identifies types of users (from data base administrators and systems designers to ad hoc querist) and classifies the necessary entries with respect to their various needs. Present data base management systems and information systems design techniques are examined relative to these needs, and a set of necessary research objectives identified to aid in fulfilling these requirements. Finally, the potential of future systems methodology is examined in the light of present and future technology trends.

Key words: Data base; data base administrator; data base management; data definition; data directory; data element; data element dictionary; data structure; file management; system development.

1. Introduction

Many of the papers in this symposium have dealt with the use and implementation details of Data Element Dictionaries. This paper is intended rather to discuss the need for, and scope of Data Element Dictionaries within the entire information system design, implementation, and usage cycle.

If we presume that each person already has his own concept of a Data Element Dictionary at this point, and we ask for a set of definitions of the term "Data Element Dictionary", in conjunction with a short paragraph describing its potential use by each reader, we should expect a disparity. This is presumably because the Data Element Dictionary has different potentials for different people within the information system cycle.

As an example, the ultimate user of the system may like to know the exact data item name: he may need to use this to access some interesting data; he probably also needs information about the data item, such as the fact that it has a numeric value which represents a weight in tons, and so forth. To the designer or administrator of a system, the usage of a data item (in the sense of: "Which programs access it?" or "What are the statistically averaged accesses per day?" in a working system) is probably much more important than the specific format of the name. One part of this paper therefore deals with the discussion of the category of users of information systems in order both to contrast and distinguish their potential needs.

Because these different classes of users of Data Element Dictionaries are interested in different attributes of Data Elements at different phases of the systems development process, a brief discussion is also given of the authors' views of these phases, and the significant users during each phase. These are illustrated in figure 1.

In essence, each facet of the use of an information processing system brings its own need to control the flow of data. The need to control requires a focal point which sensibly resides within the computer system. The use of data base management systems has provided one facet of control, but many implementors of systems using data base technology have found it in itself is not enough: they realize that it must be supplemented by other means such as those we normally call a Data Element Dictionary.

Researchers in better implementation techniques and better methods of stating the requirements of large scale information systems have also perceived a need for a coordinated and coherent directory, which in itself bears great resemblance to the modern Data Element Dictionary. It is therefore our contention that the Data Element Dictionary can and will become one of the most important components of the information processing systems life cycle.

In order to substantiate this statement, the major part of this paper is divided into three parts. The first of these defines both the information processing system life cycle and the types of potential users; the second discusses an analogy between information systems and the manufacturing process, whereas the third part is a discussion of potential attributes of a Data Element Dictionary in the light of this analogy and our general experience in information systems technology.

2. Definitions

Although it is probable that everybody has intuitive knowledge and understanding of these terms, it is felt necessary to include them for purposes of clarity and precision.

2.1. The Stages of the System Life Cycle

The phases of the information processing systems life cycle start with a "perception of need." This entails the understanding on some part of the enterprise that there is a need for some new aspect of data processing. This may be as trivial as the knowledge that a particular subroutine in the mathematical library of functions is inefficient and needs to be improved, or the sudden understanding that a new corporate marketing system will require a totally redesigned automated information system. Generally speaking, some sort of preliminary study evolves from this perception of need, but no detail of design would be expected at that time.

The next phase is one of "stating the requirements." It involves a determination of the way that the users expect the proposed system to operate (or even the way that they have operated in the existing system, with anticipated improvements due to increased capabilities of the new system). For our two examples, the result of a study of the mathematical routine may be merely the feeling that:

> "Inputs and outputs should be the same, but the results should be obtained either more rapidly, or more economically, or both."

Of course, this also involves a careful understanding of what <u>are</u> the present inputs and outputs, and how the effectiveness of the new implementation is to be evaluated.

In the case of the marketing information system, the study is more complex, and though it involves much the same elements as the simpler example, it requires a greater volume of data and probably more complex measures. The inputs and outputs are now a mixture of the previous system's parameters plus those new elements which motivated the new implementation. But apart from the additional volume and complexity, the techniques are essentially the same.

The "design" phase now follows. This may, or probably should, be split into two parts: the logical and the physical. The distinction between these two parts will sometimes be quite fuzzy, and in other cases quite distinct. As an example, in dealing with a subroutine of some mathematical nature, the logical design normally deals with the provision of an effective algorithm. The physical design, which may be almost subliminal to the implementor (who probably does his own design in this case), probably involves decision on a higher level language for implementation, and, for certain rather complex mathematical functions dealing with large volumes of data, a method of storing the information on secondary devices during processing. Of course, in this case the physical environment (which is naturally a considerable part of the physical design process) has normally been pre-defined by the task.

In more complex systems, the design phases are much more strongly identifiable. The development of the new marketing system probably involves some knowledge of the way in which the company is about to implement its new marketing organization, which has been generated in the requirements phase, followed by a study of the computer methods that may be chosen for implementing the system. At the end of this logical design phase, some alternative methods have been examined, and some preliminary trade-offs for these methods identified. The second part of the design phase therefore examines the needs for processing relative to the existing environment: which presumably includes any computing facilities already available. The physical design therefore determines effectiveness of implementation of the various logical designs within the existing or several different proposed hardware and software environments. The outcome of the physical design will therefore be a single recommended logical and physical design including documentation of detailed program and data organization specifications and hardware, in conjunction with plans for the other phases, which will be discussed later.

The "construction" phase is the detailed programming and coding of the algorithms, with any support functions that are necessary, including overall documentation of the methodology. This is followed by, and to some extent paralleled by, the "testing" phase, which shows that the programs indeed reflect the correct logical design specifications. Methodologies somewhat differ at this point, going all the way from debugging in the commonly accepted sense to the use of postulates in a formal logic. It might even be said that a certain amount of testing can be performed on the original logical system's design, because certain continuity tests (and even complex logical tests, if necessary) can be performed at that time. Obviously, the more complex the system, the more difficult the construction and testing phases.

The next phase of the systems life cycle is that of "conversion and education." In this phase, the people who will use the system are re-educated (as necessary) in the way that it will operate. In the more trivial example, this presumably requires no effort, whereas in a complex marketing information system, it may involve substantial efforts in teaching either new technology or new tools. The conversion effort involves both the replacement of the physical equipment and software and the loading or conversion of any new or old data into the new system; it also involves a careful description of the way in which the computer room operations staff must now view the system.

The next phase is the outcome of most of the planning: the "operation" phase. This involves the original start-up of the entire system after reasonable testing and conversion, followed by the fine tuning or honing of the system so that it works efficiently and truly provides the service required by the users. It is at this time that the operating room staff are potentially able to gather statistics on the effectivenss of the implementation, and to determine changes in operational policy to provide continuous yet efficient service.

The final phase is one of "modification and maintenance." This is, in fact, an arguable phase: it might merely be a total or partial iteration through the entire systems life process. Thus a modification can be predicated on a new perception of need, with consequent iteration throughout the whole series of previous dates, or maintenance might be brought about by better understanding or the operational phase imposing a slight modification of the physical design. On the whole, however, the changes during this phase are small compared to those during the original system implementation.

2.2. The People Associated with the System Life Cycle

In many of the phases of the system, it is relatively easy to categorize the type or class of person involved. There are, however, difficulties in degree of interest of a particular type of person at any particular time. In consequence, only those people who are particularly concerned will be considered.

The process of perception of need generally involves a management function in conjunction with the particular technologists associated with a user. The technologist are not necessarily heavily computer-oriented, although they presumably must have some feeling for the potential.

At the time of statement of requirements, the ultimate system users are presumably communicating with some preliminary systems designers, who are performing simple analysis on the initial design.

In the actual design phases, technologist in the logical design and physical implementation details are required.

During construction, programmers who know both the logical design and computing system environment, in conjunction with some low-level coders and documenters are implementing and testing pieces of the system.

The ultimate testing probably includes efforts in conversion, and requires much the same talents as construction, with some additional use of operations staff.

With operation, the computing room staff and ultimate users are the principal proponents.

For modification and maintenance, one expects to see a mixture of the all "types", depending on the degree of modification.

There are probably at least eight important categories of people who are involved with the system; these categories will now be briefly defined:

The Application Programmer is generally expert in writing higher level language programs which produce, store, or utilize the data. They are generally writing these programs to perform some well-defined task for a parametric user (i.e., an end user who wants to know nothing of computer technology, but just to invoke, possibly with data input, some pre-defined program).

The Application System Administrator is an operation administrator who makes portions of the data base available to programs. He therefore deals with the security and integrity of data in the database by limiting what and who can access and update it. Generally, he also has certain other scheduling types of functions which may not require database functions.

The Database Administrator is the guardian of the physical entity which is the database. His functions therefore are to care for the data: limit access to it (though he may delegate some of this responsibility to the Application Systems Administrator), make sure that it is kept intact or can be easily replaced (its integrity), provide efficient service by using certain storage devices and storage structuring methods.

The Enterprise Administrator is the prime administrator of the corporation. His function is to set policies on the use and applicability of data: its real meaning and sensible functional operational definition. He then is in charge of the data element dictionaries and controls how and on what the Database and Application Systems Administrators shall operate.

The Inquiry, Update and Report Specifiers are three similar types of users. Generally they must be provided with a very high level language interface: for though they are not programmers, they none-the-less expect to be able to access and modify data without the intercession of an application programmer. Naturally, the inquiry specifier normally deals with relatively small volumes of output data, whereas the report specifier expects a relatively large but highly formatted output. Also the update specifier is generally carefully limited in his ability as he holds the possibility of destroying the integrity of the database.

The Operations Clerk is a parametric user of the application programmers' work; i.e., he may invoke processes pre-defined by the inquiry, update or report specifier.

The Site Operator is generally an operator of the computer room staff whose particular function is to tend to the normal operation of the database system. Consequently he must receive reports (often automatically generated) on the minute by minute functioning of the system.

The System Programmer is either making planned modifications to the entire information system (possibly including the database system itself) or else helping the site operator and database administrator staff;e.g., in re-starting or recovering the system after some abnormal behavior.

The above categorization of users of information systems is hardly unusual except in its addition of a particular trio of administrators. The term data base administrator has been used in the past to distinguish a functional unit or authority associated with the definition of data structures, and to some extent on their administration. The essential difference here is that this function has been split by allowing those systems or run-time aspects to be handed over to the application systems administrator.

The essential difference between the conventional data base administrator concept and the trio of administrators is therefore the addition of an enterprise administrator. The main task of this administration function is to look to the long term plans and methods of the enterprise. This function therefore might deal with many aspects which are not presently (or even in the near future) being implemented on a computer. However, the elements associated with an enterprise administrator would still be identifiable, and well understood as elements of interest to the organization. The intent of this function could be termed long-range planning, but might better be a reflection of the way that the enterprise sees its total

business picture. In some sense, the way that current data is structured and stored (the purview of the data base administrator) and the efficiencies of the program use of such data (the bailiwick of the application systems administrator) are of no consequence to the view of the enterprise administrator. Of course, the converse cannot be said to be true, because presumably the application systems and data base are implemented as some subset of the way that the enterprise sees its overall function.

Having determined some terms, it is now possible therefore to discuss the roles of Data Element Dictionaries, first by analogy, and then in the various systems design phases. For further discussion see Benjamin [2].

3. Perspectives of Data Element Dictionaries

As was previously suggested, perspectives on Data Element Dictionaries will be motivated by an analogy. It might very well be asked at this point why this is necessary. The answer, we feel, is that we are still dealing with a new technology. In fact, we highly espouse the statement (which unfortunately we cannot reference, but feel it is extremely apt):

> "If computers continue to make the gigantic strides forward in future as they have in the past, then the industry will soon reach its infancy."

In a highly technological area such as computing, one tends to feel that all problems are unique: but we should ask whether they are. Furthermore, we should try to learn from other technologies, and see whether they have sensible comparisions which allows to project some new ideas. For this reason, we shall spend a fairly substantial amount of time on discussion of a manufacturing operation.

3.1. Electronics Manufacturing as an Analogy

Because most people have an intimate knowledge of do-it-yourself kits, we have chosen the analogy of an electronics company which is manufacturing a set of kits for making radios, amplifiers and similar electronic components. For purposes of reference we shall coin the name Easykit Corporation, and assume that they will provide either kits or fully assembled products to the public.

If we consider the manufacturing process, we see that the kit or fully assembled product is built from a series of atomic elements which are the parts. These may consist of such elements as: a resistor, a condenser, a transistor, an integrated circuit board, a chassis, and a cabinet. Presumably some of these elements occur in different sizes and qunatities, and the manufacturer has the choice of either building or purchasing. The building process involves manufacturing from raw materials, which themselves may consist of some of the elements previously discussed (as an example, a circuit board may contain some resistor and condenser elements which have been pre-assembled because of some special precision requirements). Other parts may be purchased from separate vendors.

Let us now consider what information is required about the atomic elements or parts. First of all, we find that the type of information varies depending on the type of part, and secondly the information requirements vary depending on the role of the person within the manufacturers organization.

As an example of the former, information about the element named a resistor will be a series of specific attributes such as its resistance (in its relevant units, in this case ohms), and probably its power campacity

or wattage. If, however, we were dealing with a named element termed a condenser, we should be introducing its value in some unit such as a microfarads, and probably, be interested in somewhat different attributes such as its peak voltage. As far as differences in usage are concerned, the designer may be interested in available values and tolerances from different vendors, whereas the manufacturing or production engineers are interested in the quantity in stock.

Let us then follow the entire process of kit manufacturing from original concept to final building:

The first phase of the process is essentially parallel to the computing systems life cycle: the realization that a new product is marketable. This perception of need is presumably a decision on the part of marketing staff that a new kit could be sold to a willing public. The knowledge available at that time is probably somewhat sparse, and therefore details are not specifiable. However, this gives rise to a request for a preliminary design of a potential product.

The first part of the design phase will be one of deciding what are the overall expected characteristics of the new kit. This is parallel to the specification of user requirements. At this time the new product which will carry the name "amplifier" may be required to deliver as an attribute some power, e.g., sixty watts rms, with attributes of certain levels of distortion at certain inputs of voltage and frequency. It will be seen, once again, that these are the principal specifications of the relationships between the inputs and outputs expected from the product. It is at this time that certain constraints may be put upon the ultimate implementation. Once again these are by the users, or their mentors (in this case the marketing staff may suggest a maximum cost based on their assumption of reasonable market price for the ultimate product). It may, be possible to identify some of the requirements as being constraints and vice versa, but this is possibly an unnecessary differentiation.

The designer is now faced with the first part of his design. This involves his use of technology in determining potential circuitry which will provide the necessary product characteristics within any constraints such as cost and weight. In this, the designer is performing his logical design. During this time he will be discovering what parts should make up the product. He may also be defining the characteristics of those parts (such as the acceptable tolerances), as well as considering some side problems such as mechanical size, electrical interference between various components, and the overall energy which must be dissipated within a relatively small volume without producing excessive temperatures. These are all associated with the environment in which the component must function, and the way that it may affect or be affected by other components.

The next part of the operation, but still a part of the design process, is to physically build the component, and use this "breadboard" layout to determine methods of assembly of the unit. This, in fact, parallels the physical design portion of an information systems life cycle. At this time, quite different attributes are required to be understood by the designer. He may still be interested in tolerances and power consumption, but he is now also vitally concerned with physical dimensions and difficulty of access for manual operations such as soldering. The results of this physical design phase will be a series of specifications to manufacturing, testing, etc as well as information that may be used by the marketing, costing, or procurement departments for their operations.

The next operation within the manufacturer, assuming that there have been no stumbling blocks requiring changes of specification (in the event of impracticality of relative requirements verses costs to build) is that of manufacturing. This, of course, parallels system implementation. The

manufacturing process generally involves the normal controls of industry such as needs for knowledge of parts in inventory, the ability to retrieve these parts based on some part number (which presumably is unique and will enable all parts of the same type such as 47K resistors of the same tolerance and power) to be located in one portion of the warehouse. This also assumes that there is no redundance such that the same component can have different part numbers when used in different operation. Similarly, there are probably restrictions on authorization of which station can call for which part number. This provides a mechanism for self-checking, as well as protection against theft by employees. The control of inventory generally requires reordering or general availability of the parts, as well as some knowledge of the production schedules so that reasonable planning is possible on how to restock the warehouse etc.

The quality control function in a manufacturer normally involves both testing and sampling techniques. These are carried out according to a series of specifications drawn up in the initial design, and will involve some sampled-testing of components, probably total testing of a set of sub-assemblies, and checking for completeness of the product. Of course, in the case of kits certain packaging and assembly instructions must be provided. These constitute a parallel to the debugging of an information system.

As far as the manufacturer of electronics components is concerned, he has now completed the greater part of his task. He is not interested in the operation of the assembled kit unless some unexpected troubles arise. To the purchase of the kit, the operational phase is now about to occur. His instructions manuals will help in day-to-day operations, and provide him simple testing in the case of component failure. Thus some simple maintenance functions are available to him.

In the event of major problems in the design, which come to the attention of the manufacturer, or in the case that the engineering staff determine that there is better way to implement the circuitry, a maintenance and modification phase has begun. This may, in fact, be the result of bad design or bad components, in which case it will present a portion of debugging service. In any case, the designer or maintainer is interested in many of the attributes of the system that previously existed, and potentially some others which only appear due to the operation (such as an unexpectedly high localized heating causing burning out of some adjacent components).

The above analogy has dealt with a manufacturers view of components, but not in all of the senses. There is still the managements viewpoint, which is generally associated with the value of the product as reflected in its sales price, and the cost of all parts of the operation including raw materials. This therefore involves cost associating with each element, each sub-assembly, and the final product (even though it maybe a kit of parts it is still, to the manufacturing corporation, a single component).

Thus we see substantial similarities between the manufacturing operation and an information system. The analogy will now be further examined by endeavoring to relate some of the components with the data in an information system.

3.2. Relating the Analogy to Data Elements

First let us consider the relation between the atomic parts of an electronic component and the data elements within an information system. To some extent the question of what is the atomic part parallels the question of what is a data element. As an example, a resistor may be looked upon as either an atomic part or a molecule made up of a carbon body and two end-caps. In the same sense, a data element called <u>date</u> could

alternatively be considered to be three data elements termed year, month, and day. Thus the atomic element is very much in the eye of the specific utilizer of the system. However, if we allow the parts to be equated to data elements, then we may say that the sub-assemblies and assemblies of these parts become the assembly of data elements into larger groups or data structures. Then the total assembly or product becomes the equivalent of an output document, whereas the raw material and parts received are the equivalents of input documents.

We may, if we wish, look on these circuit boards as being parts which allow assemblies to be made, but we are not stretching the analogy very far when we relate circuit boards to storage structures. In other words, a circuit board represents the place where specific instances of the data elements are gathered together as an instance of the data structure. Of course, the parts inventory, in conjunction with all the necessary mechanisms for obtaining parts, represents the data base and its retrieval mechanism.

If we extend the concept of circuit boards to being the equivalent of storage structures, we might consider the kit to be the equivalent of queries to the data base. In other words, the kits are a set of component parts which represent a specific product, and these are brought about by a request on the part of a specific buyer for the kit. In the analogy, a user poses a query, which causes retrieval of instances of the data element from the storage structure to produce an output document. When we consider the analogy in this light, the instructions for assembling the kit then bear close analogy to a query language for the data processing system. Furthermore, the relative cost of a kit against that of a pre-assembled component can be looked on as the difference between a query and a known formatted report in an information system.

There are, of course, some substantial areas of difference in the analogy. The most obvious of these is in the fact that each part tends to be homogenous (that is all of 47K one watt resistors of 1% tolerance may be considered the same), whereas instances of data elements are normally distinguishable from one another in some specific fashion, such as their key.

In the same way restart and recovery of manufacturing process normally implies some loss of in-process inventory, with some idle time due to the problem of resupplying the farthest station that has been affected. As an example, if there is a breakdown in a conveyor belt delivering an ingot of hot steel to a seamless pipe mill, then the ingots which have left the furnace will cool down, with potentially some spoilage, representing a loss of in-work inventory. The loss of time occurs in repair followed by starting a new ingot from the furnace to all portions of the system which have been idle due to lack of material. This is partly ameliorated in a manufacturing process because the equivalent of information systems check-point occur at each of the assembly stations. It is interesting to note that the normal manufacturing operation has quite a large overhead due to the cost of this in-work inventory. In some sense, the information system designer concerns himself about the cost of quality and integrity without noticing that he is probably extremely low in this cost relative to the average manufacturing system.

The event of deliveries in a manufacturing process parallel input documents to an information system. It should be noticed that these are quite dissimiliar in several ways: e.g., vendor deliveries are seldom broken down and reassembled into different groupings in the same way that data in input documents often are.

Now that we have developed the analogy, we should look at the different views of data as seen through different systems users, carrying with it any technology from our analogy.

537-860 O - 74 - 20

4. Usage Relationships in a Data Element Dictionary

Because the perception of need is a somewhat tenuous process, it has been unusual to consider a Data Element Dictionary to be an effective tool at this phase of the systems life cycle. Exceptions to this are only recently being discussed; a dictionary could be used in providing a uniform terminology so that the requirements specifier, and various users are communicating effectively using a common definition of terms: this is presumably one of the functions of the enterprise administrator. The other use (also an enterprise administrator function) is in providing the enterprise's narrative view of the data relationships, and consequently of the way of doing business.

The specification of requirements is the first place where it becomes obvious that a Data Element Dictionary is not only required but mandatory. The first of the needs is to provide communication between the ultimate user and the requirements (systems) analyst. If they are not talking the same language (as shown by their commonality of definition) then the communication is bound to fail. Too often, it is assumed that our fallible definitions in English are perfectly understood by both parties. This is seldom the case and the literature is filled with horror stories of implementation difficulties which prove this point. Indeed, the ultimate user is often laboring under some delusions as to the way in which he receives his data, and is very inexact in his definition of precision and formats of his data. It is only by having an exact mechanism for communications that this irritating and potentially disastrous condition can be resolved. Of course the fact that a Data Element Dictionary exists will not in itself solve compatibility problems nor does it ensure a correct design of the target system. It is just a basis for communication and control.

This control can be enforced by only allowing approved data element descriptions in the statement of requirements and would be feasible if the Data Element Dictionary was available during the requirements statement process. Furthermore, it is our thesis that no effort at automating the system building process can succeed without an automated data element dictionary management system.

The elements most interesting in this phase are:

- Name. This is the normally accepted way of describing the element. It may need some qualification in order to make it unique amongst others of similar nature (as an example the word date may need to be qualified by a suffix like received or a prefix like sent). This will usually be associated also with any synonyms or abbreviations. Of course, the abbreviations may ultimately become the implementation name when dealing with a language for systems implementation or query, but this probably is not a part of the original systems specification or requirements statement analyses.

- Usage. This describes how and where the named element should be referenced. In general, because named elements are often used in many different locations and for many different purposes, the usage may either have to be summarized, or else all instances may have to be recorded.

- Structure (or association). This is the way in which the element bears relationships to other elements. It describes what reports, both as input and output documents, reference this data element. It will also at some later time during the logical design or physical design process relate the structure of this element to others: however, this probably is not of interest to the specifications phase.

- Picture. This is the representation of the way that the user views the data. In the senses of a picture, this represents the format either during input or output due to some mechanical device at the interface, or it may represent the accuracy of some numerical item. Once again, the information at this time is related only to the inputs and outputs. As a result, the picture may have different instances for different documents, and this may not reflect in anyway the form in which the ultimate element is stored in instances in the data base.

- Units and Dimensions. The fact that an element called SHIPPING-WEIGHT has the dimensions of WEIGHT and is recorded in TONS is of obvious interest to the developers and users of the information system. This is obviously of importance to the coder or program designer, but it may be possible in future to have the system (e.g., a data base management system driven by the data element dictionary) itself check units and invoke correct conversion routines (e.g., change the units from lb. to ton., as necessary, in much the same way as we now change from floating point to fixed point representation).

- Quality control. Although this may have some substantial relationship to the previous attribute, insofar as it may predicate the size and characteristics, it may also involve some more complex validation criteria.

- Timing information. This deals with information about the time at which inputs and outputs are available or required. Information on how often the documents enter or leave the system, statistical relationships, or knowledge of the triggering of events by either external phenomenon (such as a timing clock or date change), or by some condition within the data base, such as the reduction of inventory below some reorder entry value. It should be noted that the timing information is therefore in two parts: the first of them deals with some probabilistic or exact statement of occurrence, whereas the other states how this condition may be known to have occurred. As an example, the fact that a report is required daily at 4:00 P.M. contains both the information on the frequency (i.e. daily) and also on the mechanism for knowing (the fact that the clock has now arrived at 4:00 P.M. on a particular day). As another example, in inventory control, we need to both know that reordering occurs whenever the stock point drops below a certain level, which is the criterion for the report, but also that this is expected to occur approximately once per month for each named element in the inventory (that is, the frequency).

- A textual description. This relates the information, in some English phrases, about the exact meaning of the named element. Obviously, the degree of success of this depends very much on the capabilities of the systems specifier, and should be tested carefully by the user-specifiers of the system.

- Responsibility. This is the question of which group or organizational entity is responsible for changes to either the data definition or for access authorization.

- Security. This involves the access and update requirements and rights as delegated by the responsible authority, and the means whereby these are monitored and enforced.

- Integrity. This involves the question of who has the ultimate responsibility for providing continuity of service for this element, as well as details of who is responsible and how this should be maintained.

- Worth. This involves an attempt on the part of the requirements specifier to determine the particular value of each element of the data either as a whole or else in its relationship to various reports in which it is a member.

During the requirement statement phase several methods, some of them automated, can be applied (see for example [2]). These include analyses that can be utilized to check the consistency of some parts of the process of requirements and partial implementation. In these systems the data element dictionary is produced as a result of the analysis rather than used as a standard of reference.

In the design phases, both logical and physical, most of the previous attributes of the data element dictionary are relevant. In fact, without good definition of timing, volume, speeds of response, security, and integrity the design is based merely on a particular designer's prejudices. Moreover, although very difficult to determine, a set of worth figures becomes mandatory for effective and efficient design.

The principal efforts during the design phase which transcend those of the previous phases are in the grouping of elements and their association with the various processes which use or produce them. This is then followed in the physical part of the design phase in definitions of potential storage groupings and accessing criteria for the entire data base as it is now being defined. This therefore requires at least the following additions to the data element dictionary:

- Logical design relationships. These are an extension of the structural relationships of interest during systems specification. They represent the data administrator or logical system designer's view of the best way to group information, and the relationships which might be provided for these processes. As an example, the element called age may not be physically stored within the system. Presumably some other element such as date of birth is however stored, and consequently it is possible to invoke a procedure which, knowing the present date as well as the date of birth, can compute the age. This is being referred to as virtual data in some recent documents. In the same way, if we have a data base involving people with data on their sex and parent names, it is possible to determine other relationships like cousin and aunt, however this involves invoking some procedural computation rather than an immediate access to the data base by an explicit relationship.

- Association with processing. This is the association of elements with the processes in which they play a part. This enables the user to determine how and where his data is obtained or modified. It also provides a mechanism whereby the data base administrator and application systems administrator may determine better groupings during operation. Finally, this enables some other problems such as the security and integrity relationships to be better examined through a knowledge of the potential accessing paths and processes.

- Storage groupings. This is the purview of the data base administrator and his ability to define how the data structure is mapped into the particular storage. This, generally, also involves the ability to define any accessing criteria and the physical devices under which the element may be supported.

The additional information described above is sometimes generated as requirements analysis reports or design specifications, and used for checking the completeness and correctness of the design. It is our contention that this information, which is the coalescence of the various statements of requirements and design decisions, must become part of the data element dictionary; further, we believe that any structured design process cannot really succeed without the availability of such a data element dictionary.

During the construction phase, the data element dictionary may both require additional information, and provide substantial aid to the programmers. The additional requirements are:

- Information about the "sub-schemas" in which the element resides. This information can be of use in partitioning the data base for various users needs. The sub-schema concept involves the statement of which part or parts of of the data base are visible to any particular application program or user type. It also allows different filtration of names for different languages, and potentially the transformation of logical schemas to suit the way that a language processes them. As an example, a FORTRAN sub-schema may utilize the matrix concept, whereas some other language may look on this as a vector made up by the transformation of rows or columns of the matrix.

- Test data generation. The testing and debugging of large scale systems normally requires quite a large volume of test data. In dealing with large scale systems and complex data bases, it is not always possible to provide raw data without danger to integrity during testing. The data element dictionary may very well be the repository of information on how to generate data which tests programs based on a generation of random but theoretically possible values.

By marrying the data element dictionary to the data base management systems, a concept which is gaining more wide-spread support in industry, it is possible to allow the data element dictionary to provide much of the support functions otherwise required by the data base management system. Thus the dictionary, in having the data structure and information about the sub-schemas may be used to generate the data definition for the data base management system. It also is the repository of information about potential users security and integrity. It could therefore provide

the necessary parameters to the data base management system so that these are automatically generated rather than relying on systems support activities.

In the day-to-day operation of data base management systems, the data element dictionary can provide some extremely important functional interfaces. These are generally mechanisms whereby the dictionary becomes the repository of important information about the way in which the system is, in fact, working. Examples of this are:

- Date and time stamps. These can be associated with each data element in order to show when the element was either last accessed, last updated, or originally input. It is, in fact, sometimes necessary to keep a total history of the element.

- Description for query. Here the data element dictionary can form the basis of the menu which the querist examines to determine the information that may be obtained from the system. It is an invaluable aid for systems designed to guide the user through the system. It is also important here to note the difference between the standard data element definition and the instance values. When a set of correct values is given (e.g., for validation purposes) these may also be required by the user, or be of interest to him as a list of available parameters. As an example, if we have alist of correct military ranks, the query user may like to "pick" from these rather than have to remember them. This means that the ranges of values must be stored, and also available.

- Usage statistics. If the data element dictionary is used to retain statistics on the use of the data element, it will be possible for the data administrator to utilize this information during reorganization. The data element directory seems to be the obvious place to retain information about the usage of the element, not only in a logical but practical way. Thus the statistics collection mechanism can provide information to the data element dictionary on the frequency of usage and potentially even what processes are using elements. Such statistics are also necessary for the design of good user prompting systems (for the querists) described above.

- Security and audit. It is possible to use information quite similiar to the statistics gathering for determing both the frequency of access to potentially secret or sensitive information, and also for the ability to determine accuracy of data either from a financial, or quality control aspect.

- Modification. When some portion of the information system is to be modified, it usually involves changes to the data within the system. These may range from a change in the values that the data may now obtain, to the introduction of new data items. Hence the presence of a good data element dictionary is imperative here.

5. Conclusion

The preceding discussion has emphasized the need for a data element dictionary; which, of course, implies that there will be a system which will allow convenient access to the dictionary. The manufacturing analogy, while not necessarily exact, helped to show that the data element dictionary system is akin to the system devised to keep track of parts in the manufacturing system; i.e. inspite of the special terminology we may use to talk of the data element dictionary, it is just information about our (working) information! However, current technology of system building does suggest several areas of research which would be relevant to furthering the effective use of data element dictionaries, which in turn would contribute to system building and usage.

It has been presumed, based on some experience, that data element dictionaries are desirable; however, there is need for some methodology by which the effectiveness of the dictionaries may be measured (the dictionaries do not have to be automated to do this). A similiar question addresses the effectiveness of the dictionary over the whole life cylce: when is it highly cost effective? Can its use be delayed (e.g. not used in the requirements statement phase)?

An implementor of a dictionary is faced with the question of the type of interface that a dictionary system must provide to the user. This type of user ranges from the occassional querist to a data base administrator. Should there be a common interface? This question leads to another, namely, should the whole dictionary (i.e. with all its attributes) be available as one monolithic piece, or should it be partitioned (e.g. as a sub-schema)? A suggested starting point for a schema is shown in figure 2.

One of the current trends in providing an interface to a user for multiple systems, is to give him a "HELP" package that allows him to work across multiple systems with varying capabilities. How can the dictionary aid this process (and "learn" from its usage by keeping statistics)? An allied area of interest is the use of the dictionary by the data base administrator: he can use it to collect usage statistics of an operational system and hence aid in data reorganization.

There are, currently, many approaches to the system building process - using data base management systems, using requirements statement languages, other structured methods, or the traditional approach. Each of these approaches implies a different attack to the task of setting up a data element dictionary. Is the ultimate dictionary any different in each of the cases? Is there a way to complement the system building approach? At least two criteria may be used to describe the dictionary: the type of user and the building process.

Finally, there are several issues such as:

- To what degree should the data element dictionary contents be allowed to be dynamic?
- What security problems arise due to the availability of a central repository of a company's data description (no different than the problems encountered in any information processing system!)?
- How can it be used to aid networking? May there be differences between elements throughout the system?

In summary, we can say that the data element dictionary (and the system to use it) is really just an information system to support the management of our data. Building it involves all the problems of building a normal information system, along with the special challenges of having to be introspective - performing a systems analysis on our needs, worrying about integrating our requirements, providing an interface to us, handling security problems, as they concern us, etc. Using a data element dictionary approach, we may even learn to build systems correctly!

6. References

[1] Benjamin, R.I., Control of the Information System Development Cycle. Wiley, New York (1971).

[2] Couger, J.D., Evolution of business system analysis techniques, ACM Computing Surveys, Vol. 5 No. 3, (Sept. 1973) pp. 167-198.

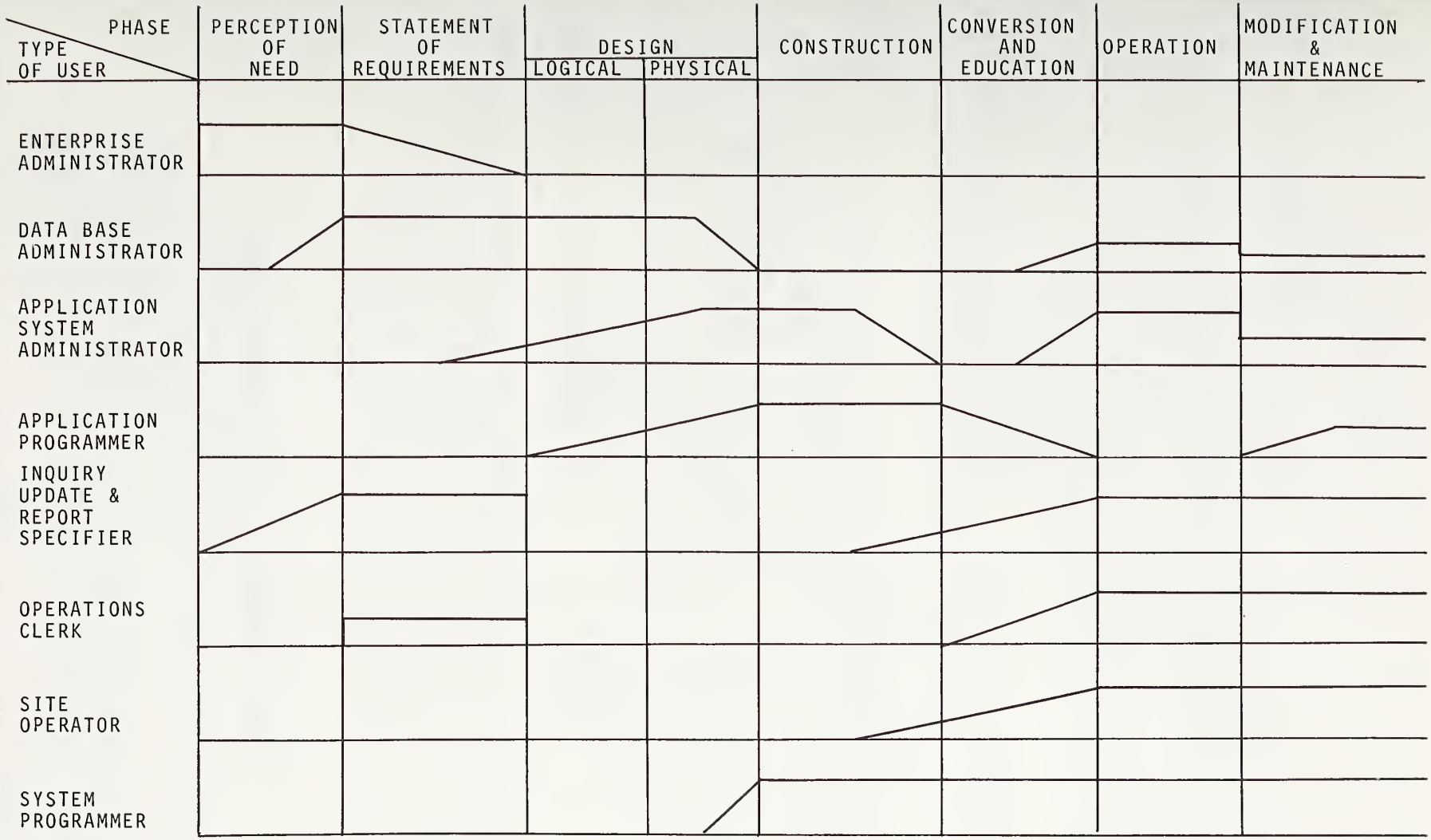


Figure 1. The Various Users and Their Relative Need for a Data Element Dictionary.

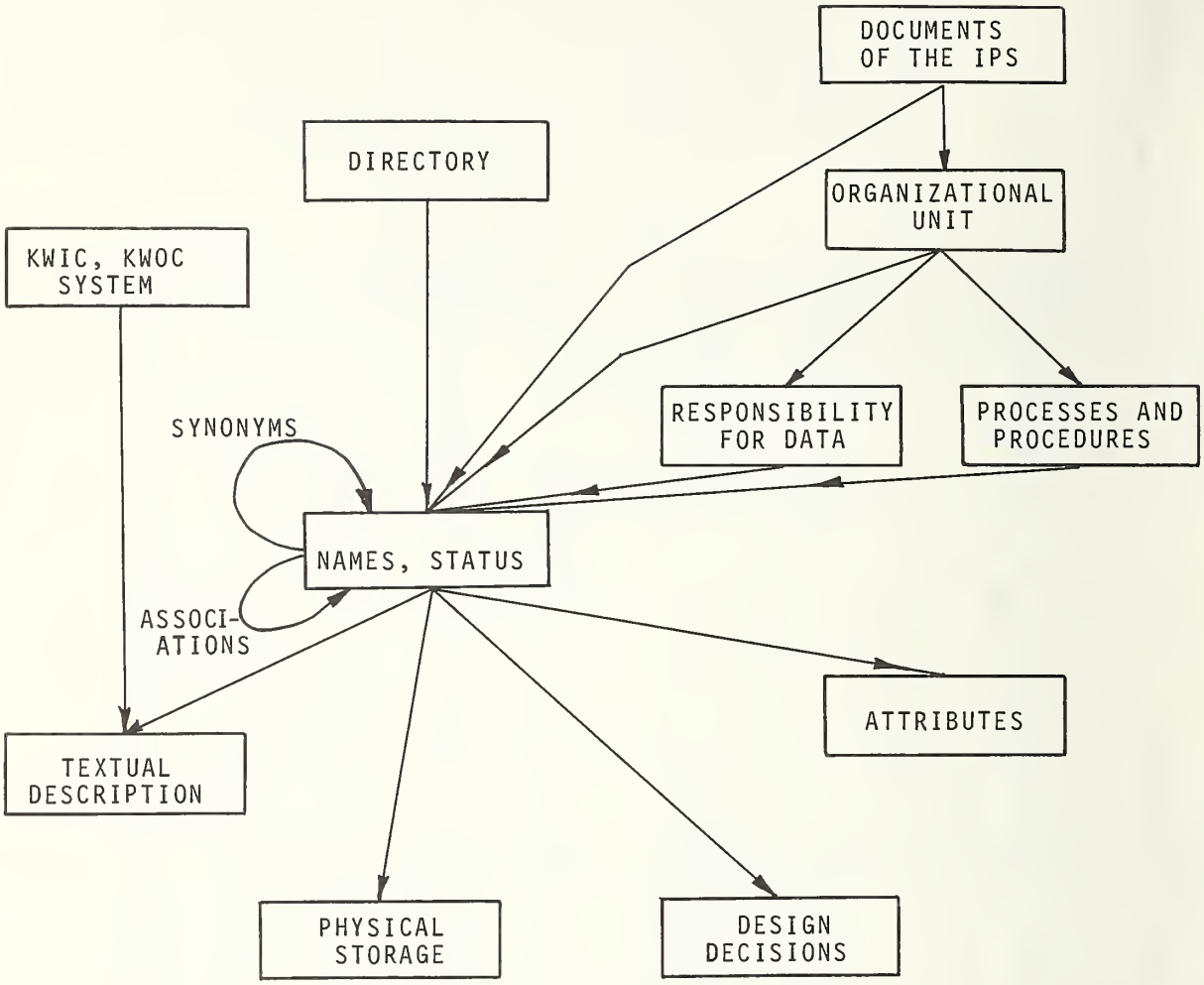


Figure 2. A Schema for a Data Element Dictionary

QUESTION: Isn't the Data Element Dictionary the same as the Data Definition Language of the proposed CODASYL DBMS?

ANSWER: The data element dictionary (ded) is a structured collection of information about the data elements of interest.
The Data Definition Language (DDL) is a language; statements made in the DDL represent the data schema; processed by a DDL processor (such as the one available in UNIVAC's DMS 1100) it could provide, as one of its products, a data dictionary. This dictionary, however, is devoid of many of the attributes of the data elements that one would expect to have in a ded - eg textual description, usage volumes.
It is our contention, that given a comprehensive data element dictionary system - similar to the one proposed in our schema - one could write a processor to derive DDL statements necessary to describe the target schema: in fact, some commercial systems presently available do just this (e.g. UCC-10 provides a DDL for IBM's IMS version II)

QUESTION: Is it more important for a dictionary to describe the data element names that are being used (though these may be redundant) or the data element names that should be used (standards)?

ANSWER: Surely, we must be honest with ourselves. Although we may conclude that there is every reason to go to standard names, there is almost bound to be resistance. If we force users into unwanted (by them) standards, they may develop a sort of paranoia - where they either deliberately or subconsciously try to sabotage the system.
What we propose is a reasonably forgiving system with a synonym capability. Such a system as that developed by BB and N with Mass General Hospital allowed the user (a nurse) to ask a question using a non-standard drug name: the order was fulfilled, but the system replied using the standard name. Maybe (it may still take 20 years) we could thus educate our users into a standard.
Of course, we, the information technologists, should never become so conceited that we are sure that we know best in other fields of technology (e.g. marketing).

QUESTION: Which user "owns" the data?

ANSWER: We presume that the "user" is, say, an employee of the corporation which maintains the data base; and, further, that the user is interested in this data as a means for discharging his duties. If our presumption is correct, then we do not see why data should not be considered as just another resource of the organization. The jealous guarding of the data as a personal possession usually arises out of either of the following situations:

a) the user is responsible for maintaining the data, its accuracy, timeliness etc.

or

b) the user had to go through a (sometimes) painful phase of collecting, and massaging the data.

It is natural to feel possessive about this data - but it is no more a user's possession than is a scientist's discovery strictly his own when done as part of his work for the company.
If, however, our presumption about the question was wrong, and the ownership pertained to an individual's data about himself, then the owner has the right to see that his data does not get misused; he should even be given rights to access it, verify it, and have a say as to who receives copies of it, and

what the receivers does to it (e.g. not further disseminate it, or else maintain it on his own without the above provision).

QUESTION: Is ownership a viable concept when flowing data between agencies?

ANSWER: The answer to the question may be deduced from the answer to the previous question on ownership of data by substituting "agency" for "user". There is, however, one point of difference; should there be some need for special "ownership rights" on parts of the data, the distributed data base concept allows this; i.e. have a local data base from which an agency could contribute to the central (pooled) data which other agencies may access. An example of such a situation might be a working budget being developed for an agency: it may not wish to have this released till it was in a finished form to be released to other agencies. In such a case the local data would be "owned" by the local agency.

Health Care and The Delivery Mechanism

Sheila M. Smythe[1]

Blue Cross-Blue Shield
622 Third Avenue
New York, N.Y. 10017

Increased demand for health services as well as increased costs for the provision of such services within the past decade, makes it imperative that a comprehensive health care delivery system be developed in this country. Essential to the design and implementation of such a health care program is the development of an appropriate, computerized information system which would provide necessary data elements on the needs of the population, utilization of services and resources, and measures of effectiveness.

You and I have spent from sixteen to twenty-four years of our lives in the classroom being educated, we are affected for up to forty years of our lives by the particular phase of life in which we have chosen to earn our living. We are affected for brief intervals as we go through life by telephones and transportation.

There is nothing in this material world however, that affects us more than health - from the point of entry - birth -- and thereafter - our own health and that of others impacts on every other aspect of living. It is therefore especially meaningful that as part of this information processing symposium we devote ten minutes to this subject.

In considering how to deal with this subject in so short a period of time, I came upon a quotation in Man, Memory and Machines (An Introduction to Cybernetics) from Samuel Butler. He said -

> "It must always be remembered that man's body is what it is through having been molded into its present shape by the chances and changes of millions of years, but that his organization never advanced with anything like the rapidity with which that of the machine is advancing."

With the awareness that this statement was made in 1872, over 100 years ago, by a man of letters and that we have had the opportunity of vast knowledge and participation in both the generic and specific aspects of computer operations and data gathering (its use and misuse) for more than a century since Samual Butler wrote that statement - from the viewpoint of my own professional experience and commitment, I would like to make a few observations about data elements and their interchange as they relate to the national and local health care system.

[1]Vice President

The nation's public and private costs for health and medical care were an estimated 94.1 billion dollars in 1973 - or 7.7% of our gross national product.

The magnitude of this expenditure is more evident when one considers that the health service industry is now considered to be the second largest in America. We have 345,000 doctors, 116,000 dentists, 141,000 pharmacists and 815,000 nurses; 20,000 nursing homes, and more than 7,000 hospitals with a combined capacity of over two and a half million beds. The health transactions, many of them computerized, number in the billions each year.

Today, it is estimated that one out of every twenty of the nation's employed are involved in the health industry - over a recent fifteen year period the number of persons involved in the health professions increased ninety percent. So today well over four million persons are employed in the health industry, and this number is increasing every day.

The twenty year rise in health care expenditures can be attributed to costs for health services such as hospital care, physicians' services, as well as hospital construction, research, disease control and detection programs. Nearly 50% can be attributed to the increase in prices; in excess of one-third is due to increased use of services such as: seeing the doctor and dentist more often; going to the hospital more often or staying there longer; and using many miracle drugs and life-saving (but expensive) new techniques not available in 1950; and the remainder is the result of population growth.

Cost is only a single facet of health care, and the answer to improvement is neither more money nor necessarily cost control. However, I do believe that some of the corrective measures rest with computer expertise and its appropriate, immediate application. This is not simply a personal stance since many here today had the opportunity to participate in the two-day conference in January 1972 sponsored by HEW where technology and the health system in the 1980's was discussed in great detail by knowledgeable and acknowledged experts. But that was two years ago, and I suggest that today it is a necessity to think of the now and not 1972 nor the 1980's.

The health status of human beings is my primary, professional concern, as it is with my most immediate, day-to-day colleagues. Health status is admittedly an abstract term, an end in-and-of itself -- a goal. Meanwhile, while we are seeking to attain that goal what will we settle for? May I suggest: The development of an effective health care delivery system responsive to human needs. An equally abstract notion? I do not think so, and furthermore, one toward which this audience could provide needed incentive and leadership.

In order to show how data elements, computers and technologists can relate to such a system, its organization and structure, some explanation is needed as well as some common agreement. An effective health care delivery system can be divided into two interrelated components:

1. the service elements

2. and, the desired characteristics

The service elements of a responsive health care delivery system include at least the following: prevention, primary care, emergency care, acute general and specialty care, long-term care, home care, and rehabilitation.

The desired characteristics of a system would include at least (individually or collectively) the following; availability, accessibility, quality, financing or freedom from financial barriers, flexibility, coordination, continuity, acceptability, accountability, capacity for dealing with populations at special risk, economy, efficiency, and research. Quite a laundry list!

To illustrate the interrelationship of these components, let us visualize a matrix with the service elements as the vertical array, and the desired characteristics as the horizontal array. They interrelate through what could be called indicators; that is, quantitative measurement. To make the process useful, there would be a series of cells or boxes to be filled in. For example, population ratios, travel time, free structure, insurance coverage, patient flow, utilization review, formal/informal relationships with other services, patient attitudes - in other words - a series of indicators.

The components of these indicators are all data elements. These indicators are measures of both the present status and the desired level of health within the optimal system. With this knowledge, appropriate resource allocation and health programming could be predicated. (I am not however, saying that such comprehensive information would necessarily result in either)

I also do not wish to imply that all the direct measurements (the indicators) exist or that they necessarily involve all the elements or involve all the desired characteristics. What does exist in this country is the possibility of a measurement where needed, whether it be on a national, state or regional basis. What does exist is the capacity to define the data elements of a health care system, and this ability is most certainly within the purview of technologists. And, such expertise is and cannot be limited to the health professional computer personnel. This activity rightfully falls within the responsibilities of all data professionals who are engaged in any function which would impact on social well-being.

How all this is brought into a unified, interrelated activity is the unanswered question. And, that is begging the question.

However, I do know the capacity of Blue Cross and Blue Shield national systems, the hospitals and health industry to fulfill both regulatory and nonregulatory functions with the use of computer technology.

In the Blue Cross/Blue Shield system, we must relate our customers vertically, horizontally and diagonally from their point of entry (when they are enrolled as our customers); we must monitor their changes in status (births, deaths, marriages), as they change from one employer to another, or as they become directly responsible for their own coverage. We must relate them from the point they enter the hospital through emergency, inpatient and outpatient situations, or visit a doctor's office for covered services. We must be concerned with their hospital, doctor, dental, drug, and home care bills; we must relate them to the federal government (Medicare) to state and local government (Medicaid) etc., and we must keep track of them as they vacation, travel or retire in California, Florida and Arizona, and keep track

of California's and Florida's and Arizona's subscribers as they vacation in New York and elsewhere.

The mass of data necessary to control the above situation in an orthodox data processing environment is almost impossible to comprehend, let alone process.

In the past decade there has existed, to name a few: Hill-Burton, Hill-Harris, Medicare, Medicaid, OEO, Model Cities, Comprehensive Health Planning, and the Regional Medical Programs. All of these made and make different health and health-related data demands both on the public and private sector.

Now the nation is faced with the implementation of HR 1 (Public Law 92-603) -- it too requiring multiple, additional, supplemental, new and different data. In addition, demographic projections, as well as health needs of a given community will certainly be a requirement of the new HMO legislation and in a more technical sense also the local implementation of Professional Standards Review Organization activity.

Then there are the current Congressional proposals calling for the establishment of regional "decentralized" health authorities across the country. And lastly, on the 30th of this month, the President of the United States is "expected to outline major initiatives designed to improve the nation's system of health." (New York Times - 1/20/74)

During the past decade, several presidential health messages have indicated that the nation's health policy is the attainment of adequate health for all citizens and the assurance of equal access to quality medical care for all. The crux is converting such a policy into a specific system and operating programs. This is the subject of public, political debate at the present time.

Whatever the outcome, the development and implementation of any process of planning, monitoring and evaluating health services, demands the development of a comprehensive health information system. This system would provide data on the health needs of the population, utilization of services and resources, and measures of effectiveness. Such a health information system would eliminate the multiple and incompatible data collection and reporting requirements currently in operation from the national to the zip code level.

In these few minutes, I have tried to outline in a primitive manner what I think would be the core of such a system, as applied to the development of a health care system responsive to the nation's need and to indicate the necessary and continuous role which computer technology and its professionals can, and must play.

A Syntax for Naming Data Entities

Helmut Earl Thiess

Naval Command Systems
Support Activity[1]
Washington D. C. 20374

The names of data elements are used for recording, storing, indexing, manipulating, and retrieving most of the data in automatic data processing systems. Hence, these names should be unambiguous to people and to machines. The syntax described in this paper provides rules for such unambiguous naming of data entities in normal English.

The syntax for names of data entities (the collective term used for what usually are called data elements, data chains, data use identifiers, and fields) defines the use of the operators OF, AND, OR. The extension of the syntax to include the operators AT, BY, FOR, NOT is indicated.

The use of the syntax permits the construction of names for easy use by keyword indices, classification of entities, and retrieval on contiguous or noncontiguous keywords of the names. The syntax permits incorporation of its features into or recognition of them by existing computer programming languages.

Key words: Data element; data element dictionary; data element name; data element syntax; data management; data processing; information processing; information storage and retrieval; keyword indexing; programming language.

1. Introduction

1.1. Terms Used

A datum appearing in a data processing system usually is a numeral in digital notation representing a number, a coded representation, or a word or phrase in a natural language. This datum is a *data item* of some *data element*. Data elements can be combined into *data chains*. *Data use identifier* is the name, title, or description given to the use of a data element. For instance, the data element "country" may be identified as a country "of birth" or country "of residence", depending on its intended use. The terms *data element*, *data chain*, *data item*, and *data use identifier* are used as defined by the Department of Defense and published by the Department of the Navy [1][2].

Data entity is used in this paper to mean data element, data chain, or data use identifier when the distinction among these terms is not needed.

[1] The views expressed in this paper are the author's and do not necessarily represent those of the U. S. Department of Defense.

[2] Numerals in brackets indicate the references at the end of this paper.

537-860 O - 74 - 21

1.2. The Problem

Data are resources. The frugal use of data requires that data be managed. This management of data is done by some technique [2, 3, 4, 5]. However, the catalog, directory, or dictionary of data entities in use by data managers usually has no rules for naming these data entities, except for the necessary requirement that the name of a data entity be unique. As a consequence, exchange and use of data to provide information is frequently hampered by storage and retrieval problems that were caused by the unsystematic naming of the data entities.

2. Acknowledgements

The syntax developed in this paper or the notion of having such a syntax was inspired by COBOL-derived query languages which have been used in military command and control applications since the advent of second-generation computer hardware. The following example illustrates this background:

IF SHIPTYPE (CRUISER) COLUMN SHIPNAME, COMMANDER-RANK, COMMANDER-NAME.

This query statement is adressed to a formatted file containing records on ships. The data element SHIPTYPE (also a field in this file) has, among others, a data item CRUISER. For any ship in the file that is identified as a cruiser, the name of the ship and the rank and name of its skipper will be on the output in columnar form.

After the syntax was tested in an application area and was discussed with an associate, the associate called a specification of the "Of-language" [6] to the author's attention. Since then, publication [7] has also been reviewed by the author.

When the syntax was originally developed, the author wished to arrive at names that can fit into a query language for processing by a computer. No specific query language is intended to accommodate the names. The degree of processing by computer also is left undecided. Indeed, processing by computer need not be done at all.

3. Specifications for Designing Names of Data Entities

If one were to specify the characteristics the name of a data entity should have, one would begin by prescribing unique and unambiguous names expressed in natural language. Further, these names should be amenable to categorizing the entities in one or more manners. As these names are used by or for automatic data processing systems, they should also lend themselves to automatic processing, including recognition of such names by automata.

4. The Syntax

4.1. Reserved Words

Strictly speaking, a syntax deals only with the arrangement of words. The ultimate goal of the syntax described here is to lead to a "language" for naming data entities. Hence, reserved words are needed at this stage.

The following words have special meanings as connectors, delimiters, or operators and should be used for these purposes only. A reserved word hyphenated with any word is not a reserved word unless this hyphenated word itself is a reserved word.

AND -- Operator requiring that the condition specified on either side of the AND be met:

AT -- Operator requiring that the following phrase is either a representation of a geographic location or of a specific "moment" (minute, hour, day, etc.) in time.

BY -- Operator requiring that the following data element (field) be answered or is applicable for all data items (values) of the data element in the "domain" of the question; e.g., NUMBER OF ENLISTED MEN BY RANK FOR "SHIP". SHIP is the domain (a specific vessel); BY RANK specifies E-1 through E-9, NUMBER would (in a query) respond with the number of enlisted men in every pay grade on the specified vessel. See the operator FOR.

FOR -- Operator requiring that the following data element (field) be answered or is applicable for one specific data item (value), if one exists, which is represented by the name of the data element. See the operator BY.

NOT -- The logical negation.

OF -- Operator establishing the specific-to-general name of a data entity.

OR -- Operator requiring that one and only one of the conditions specified on either side of the OR be met. The inclusive OR is ruled invalid, because we are concerned with the representation of the contents of fields in records. Such a field can contain only one of two or more possible alternatives.

Because data entities are potential input for a query, the following usual query-language words should be reserved and not be used for naming data entities:

ADD, ADDITION	EQUAL	PLUS
DIVIDE, DIVISION	GREATER	MINUS
MULTIPLY, MULTIPLICATION	LESS	SUM
SUBTRACT, SUBTRACTION	THAN	PRODUCT
DISPLAY, PRINT, READ, WRITE	EITHER	QUOTIENT

If the query-language chosen or to be chosen should resemble an existing language, such as COBOL, appropriate additional words should be avoided.

It is recognized that these usual query-language words may constitute a problem. Armies have "divisions"; businesses manufacture "products"; etc. Reserving these words is a design goal, in case the names of data entities should fit into a query language without the need for separate symbols delimiting the names.

4.2. Use of Predicates

All the names of data entities are assumed to be phrases that do not contain a predicate (an action, state, or condition which is stated, ordered, or exclaimed by the use of a finite verb). Infinitives or participles cannot be used as predicates, hence they can be used in the name of a data entity. Loosely stated, a name of a data entity should not contain a verb that makes a statement about a subject of a sentence. Briefly, a name must not be a sentence.

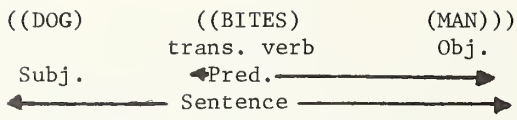
4.3. Punctuation

As part of the name of a data entity, only a hyphen may be used as a punctuation symbol. This hyphen is used to connect words in normal English writing or to connect a reserve word with another word.

4.4. Operations and Parsing Rules

a. General Remarks

Whether one reads a sentence or a mathematical formula, one analyzes the string of symbols representing the sentence or formula and parses the string from left to right for determining the meaning. The sentence DOG BITES MAN consists of the subject DOG and the predicate BITES MAN. The predicate, in turn, consists of the transitive verb BITES and the object MAN. The implied components are:



The mathematical notation 5 x 3 + 8 can mean 23. One arrives at that meaning by knowing that one first multiplies five by three and then adds eight. The implied components are:

((5 x 3) + (8)).

A person reading the sentence or the mathematical notation is aware of the implied components. A computer program used to assign meanings to sentences or to calculate results must understand the sentence structure (subject, transitive verb, object) or the order in which mathematical operations are performed (multiply before adding). However, people and machines can encounter ambiguity. In order to avoid ambiguity in naming data entities, an order of operation and parentheses are assumed to be present for knowing where one syntactic element stops and the next starts and for knowing which words or phrases are linked more closely than others. One can check for ambiguity by inserting the implied parentheses based on the order in which the operations are "executed".

The outermost opening and closing parentheses are used as surrogates for explicit or implicit delimiters understood by people or machines. For instance, a person or a computer program could understand READ DATE OF BIRTH by knowing that READ is a command and DATE OF BIRTH the name of the data entity. The end of the command READ signals the beginning of the name. The end of text on a line indicates the end of the name. The order of operation, i.e., the sequence in which the implied parentheses could be inserted or are assumed to be present, is described in the following. In all instances, the number of implied opening parentheses must equal the number of implied closing parentheses.

b. Operators and Reserved Words Other Than OF, AND, OR

All the operators given in paragraph 4.1, other than OF, AND, and OR, perform unspecified operations. They are currently not used in naming data entities.

c. Operator OF

The operator OF has the highest order. If one OF appears in a name, the implied opening parenthesis appears to the left of the first word of the name, and the implied closing parenthesis appears to the right of the last word of the name. In other words, the whole name is assumed to be in parentheses.

If two OFs appear in one name, the two implied opening parentheses appear to the left of the first word of the name; the first implied closing parenthesis appears immediately to the left of the second OF; the second implied closing parenthesis appears to the right of the last word of the name. By extension, this applies to three or more OFs. For examples see paragraph 4.5.a.

This rule makes it impossible to construct names such as ABBREVIATION OF NAME AND DATE OF ADMISSION OF STATE OF US. To make the OF order of operation work, this data chain would have to be renamed ABBREVIATED NAME AND ADMISSION DATE OF STATE OF US. This name is good English. The retrieval by keywords would be the same in either case.

The commutative law, unlike in algebra, does not apply, because the word (or phrase) to the left of OF modifies the word (or phrase) to the right of OF and not vice versa. For example, COUNTRY OF BIRTH and BIRTH OF COUNTRY have different meanings.

d. Operator AND

If one OF is present, the parentheses implied by it need not be inserted, as one knows that they are at the beginning and at the end of the name. If two or more OFs are present, the parentheses implied by the OFs should be inserted before the parentheses implied by AND are inserted. The parentheses implied by AND are inserted before those implied by OR. In all cases, it is recalled, one proceeds by inserting parentheses for each operator from left to right.

The commutative law does not apply because the components of a data element are read in a specific sequence. While in algebra a times b equals b times a, A AND B in our data element language is not the same as B AND A.

e. Operator OR

After the parentheses implied by OF and by AND are inserted, one inserts the parentheses implied by OR. If one or more OFs are present these parentheses must be within a pair of parentheses implied by an OF. If one or more ANDs are present and parentheses have been placed around the phrases for the ANDs the phrases in parentheses on either side of an OR are put in parentheses proceeding from left to right as for OF.

f. The Space Character

The space character between words of a phrase delineates the individual words which are used for keyword indexing. Obviously, normal English-language usage should not be violated. Authoritative general-purpose dictionaries and technical vocabularies should provide guidance in spelling. Whether one or more space characters differ in meaning is not decided here. Probably, an implementing query language would have appropriate specifications.

4.5. Using Operators and Parsing Rules

a. Using OF

Using SEX as an example of a name of a data element, the names of its DUIs (data use identifiers) should be formed by using SEX OF ... as the initial words for DUIs, such as SEX OF CIVILIAN EMPLOYEE. Other data elements describing the civilian employee could be HOME ADDRESS OF CIVILIAN EMPLOYEE, FULL NAME OF CIVILIAN EMPLOYEE, etc. In each instance, the keyword index would permit the retrieval of all the data elements pertaining to a CIVILIAN EMPLOYEE. No specific, separate classification into "civilian personnel data elements" is needed.

Similarly, we may have the data elements NAME OF STATE, ABBREVIATED NAME OF STATE (or ABBREVIATION OF NAME OF STATE), and CODE OF STATE. Again, simple look-up in the keyword index would lead to the retrieval of all the data elements for STATE.

If more than one OF is used in the name of a data element, such as NAME OF DATA ITEM OF DATA ELEMENT, or NAME OF COUNTY OF STATE OF THE UNITED STATES, the leftmost word or phrase delimited by the leftmost OF is to be joined to the second word or phrase delimited by the second OF. Then the first phrase, OF, and second phrase together are joined by the OF to the third phrase, etc. The assumed parentheses would be:

(1) phrase 1 OF phrase 2 OF phrase 3 OF phrase 4 ...
(2) (phrase 1 OF phrase 2) OF phrase 3 OF phrase 4 ...
(3) ((phrase 1 OF phrase 2) OF phrase 3) OF phrase 4 ...
(4) (((phrase 1 OF phrase 2) OF phrase 3) OF phrase 4) ...

If NAME OF COUNTY ... were used in a DUI, such as NAME OF COUNTY OF BIRTH OF CIVILIAN EMPLOYEE, retrieval by keywords or key phrases would lead to the data element NAME OF COUNTY OF STATE OF THE UNITED STATES, as well as to the sets of data elements on CIVILIAN EMPLOYEE; and would also permit the retrieval of related or associated data elements or DUIs on STATE and UNITED STATES. For purposes of classification or categorization, the data use identifier NAME OF COUNTY OF BIRTH OF CIVILIAN EMPLOYEE could fall into classes such as geopolitical data elements, civilian personnel data, names, etc., depending on the ad hoc need of a user. Keyword and key phrase indexing permits such dynamic classification. For static classification the same keywords could be assigned to the respective classes of data elements just as can be done for ad hoc classification.

b. Using AND

Assume a data chain consisting of NAME OF COUNTY OF STATE OF THE UNITED STATES and NAME OF STATE OF UNITED STATES. Let us use the respective letters N, C, S, and U to represent the phrases separated by OF or by AND. We can then form the name of this data chain by writing

N OF C AND S OF U

by recognizing the rule that AND has a lower order of operation than OF. Reading from left to right (by people or computers) we can insert the implied parentheses in the following order:

(1) (N OF C AND S) OF U
(2) ((N OF C AND S) OF U)
(3) ((N OF (C AND S)) OF U)

These parentheses serve people as a check on the uniqueness of meaning and serve a computer as the markers it assumes for parsing or concatenating words or phrases.

As already noted in paragraph 4.4.d., the name

N OF C OF S OF U AND N OF S OF U

is invalid. Inserting the implied parentheses would result in

(((((N OF C) OF S) OF (U AND N)) OF S) OF U).

The underlined phrase UNITED STATES AND NAME, modified by NAME OF COUNTY OF STATE, does not make sense. Expressed differently, UNITED STATES AND NAME, concatenated by the rules of our grammar, is a meaningless phrase.

Again, as already indicated in paragraph 4.4.d. paraphrasing is necessary when a data chain consists of two (or more) data elements with differing characteristics. ABBREVIATED NAME and ADMISSION DATE were given in paragraph 4.4.d. as an example. Using the date of birth and the county of residence of an employee as a data chain, we could express it as BIRTH DATE AND RESIDENCE COUNTY OF EMPLOYEE.

c. Using OR

Assume a field that contains either the name of a county and the name of a state of the United States or the name of a province and the name of a country other than the United States. For our purposes, it is irrelevant how one would know which of the two alternatives is represented in the field.

Because a field cannot contain two values simultaneously, we can assume that the OR is the exclusive OR, otherwise explicitly expressed by the use of "either ... or ...".

Let us use the letters N, C, S, P, and K for name, country, state, province, and "Kountry" to abbreviate the explanation. Applying what we know about the use of OF and AND, we can express the contents of the field as:

N OF C AND S OR P AND K.

Inserting the implied parentheses step by step, based on the order of operation, we may proceed as follows:

(1) (N OF C AND S OR P AND K)
(2) (N OF (C AND S) OR (P AND K))
(3) (N OF ((C AND S) OR (P AND K))).

If this field were given a full name, such as N OF C AND S OR P AND K OF BIRTH OF EMPLOYEE, we could confirm that the short name was chosen correctly for adding the use phrases. We use B for birth and E for employee as the use phrases in constructing the full name of the field from this assumed DUI name.

(1) N OF C AND S OR P AND K OF B OF E
(2) (N OF C AND S OR P AND K) OF B OF E
(3) ((N OF C AND S OR P AND K) OF B) OF E
(4) (((N OF (C AND S) OR (P AND K)) OF B) OF E)
(5) (((N OF ((C AND S) OR (P AND K))) OF B) OF E).

5. Other Nomenclatures

5.1. Inverted Nomenclatures

In supply management, the inverted nomenclature, also known as storekeeper English, is used. It is claimed to have the advantage of bringing like terms together in an alphabetical listing of items [8]. This claim no longer holds. Modern indexing, with the advent of key-word indexing [9] produces alphabetical listings that bring together like items by more than one category.

Examples of crass abuse of storekeeper English are the following data use identifiers of a data chain called YEAR AND MONTH:

(a) Year-month, of source document, projected assignment area
(b) Year-month, port arrival
(c) Year-month, expiration current service agreement.

Indeed, like "items", namely "year-month", are brought together. However, like categories for these three DUIs may be as follows, if one knows the subject area:

document control for name (a)
ship movement; personnel; logistics for name (b)
personnel for name (c)

5.2. Other "Normal" Nomenclatures

It is not claimed that the syntax proposed in this paper is the best method for naming data entities. Experts in linguistics are urged to comment on the correctness of the proposed syntax. Does it lend itself to creating names in natural language which can be understood by computers? Is the proposed order of operations (OF higher than AND, etc.) appropriate?

6. References

[1] U. S. Department of the Navy; ADP Glossary; U. S. Government Printing Office, Washington, D. C., 1970.

[2] Charles J. Bontempo; Data resource management; Data Management, vol. 11, no. 2, February 1973, 31-37.

[3] Jerome E. Dyba; Benefits of data base and doing one better; Data Management, vol. 11, no. 3, March 1973, 31-32.

[4] Michael L. Wearing; Upgrade documentation with a data dictionary; Computer Decisions, vol. 5, no. 8, August 1973, 29-31.

[5] Richard A. Nerad; Data Administration as the nerve center of a company's computer activity; Data Management, vol. 11, no. 10 October 1973, 26-31.

[6] IBM Data Dictionary/Directory, System Specification, September 1969; document X3.8/162 of March 1970 of National Standards Subcommittee X3.8, received in May 1972.

[7] International Business Machines Corp; Data Dictionary/Directory System, Data Specification Methodology Details, Version 2; International Business Machines Corp., White Plains, NY, 1971.

[8] Defense Supply Agency; Item Identification, Chapter 2, Federal Manual for Supply Cataloging; Defense logistics Service Center, Battle Creek, Mich., 1967, page 220-1.

[9] H. P. Luhn; Keyword-in-context index for technical literature; paper presented at the 136th meeting of the American Chemical Society, Division of Chemical Literature, Atlantic City, NJ, September 1959.

ADDENDUM

A. Operator Precedence

Dr. J. A. N. Lee, from the Department of Computer and Information Science, University of Massachusetts reviewed the paper at the request of the author. In a letter, Dr. Lee stated:

"I believe that the most striking thing is that one should be able to establish an operator precedence table for the connectives in the same way as precedence tables exist for arithmetic and logical operators. From a brief study I would say that the rules for precedence are as follows (from highest to lowest):

NOT
AND
OR
OF
AT, BY, FOR

with the usual rules of left-to-right precedence governing the cases of equal hierarchy. This means for example that the statement Name of State at port of entry would be parsed into (Name of State) at (port of entry), and so forth. I am not totally certain that AT, BY, and FOR are on the same level and have not developed examples to either prove or disprove my hierarchy above.

The notion is great--I would only like to see more relationships with existing methods of parsing such as precedence techniques."

B. Why not Esperanto?

The oral presentation and the paper stress that the names of data entities be expressed in natural language (even though in conformance with a prescribed grammar). One of the questions raised during the discussion was: "Why not use Esperanto as a worldwide programming language?"

English fills quite adequately a need for which Esperanto was designed. English is the most widely used lingua franca. English is the native language of more people than any other language, except Mandarin. Either English or Spanish is the native language of the largest number of countries. English appears to be the second language of more people than any other language. Given these circumstances, one need not wave the flag, but merely be practical and use English as a language of international exchange.

The problem and its solution can be looked at in another manner. The smallest number of people engaged in international exchange needs to learn another (natural) language in order to understand a computer program written in English. As a written language. English is learned relatively easily. Native English speakers are very tolerant of English spoken with a foreign accent.

C. "Data Are" or "Data Is"?

The question arose whether "data" is the plural of "datum" or "data" is a collective

noun used in the singular. Another speaker expressed the view that "data" is a collective and that "data are" is archaic.

Webster's Third New International Dictionary (1961) states that "data" is the plural of "datum".

The American Heritage Dictionary of the English Language (1969) states:

> "Usage: Data is now used both as a plural and as a singular collective: These data are inconclusive. This data is inconclusive. The plural construction is the more appropriate in formal usage. The singular is acceptable to 50 per cent of the usage panel."

There are several practical reasons for retaining the distinction between singular "datum" and plural "data".

(1) If "data" were singular, one datum, e.g., the amount of money written on a paycheck, would have to be expressed by some awkwared linguistic contortion, such as "item of data". How does "item of data" differ from "data item"? If the singular and plural of "datum" are used, one can express oneself with clarity and precision with relative ease.

(2) Sooner or later some people, innocently or ignorantly, will construct a plural "datas are", if "data" were to be a singular noun. One symposium speaker, during his oral presentation, helped himself to "medias are...", a usage condemned very severely by The American Heritage Dictionary.

(3) People learning English as a foreign language usually learn formal English. Their teachers explain the English -um and -a endings for the singular and plural of certain nouns of Latin origin. Many foreign speakers of the English language question the lack of knowledge of English and Latin when they hear or read "data is".

A Technical Information Network Serving A Decentralized Manufacturing Company

C. L. Tierney[1]

Whirlpool Corporation
Benton Harbor, Michigan 49022

The Whirlpool Information Network (WIN) is a system by which technical information generated or applied anywhere in the corporation is condensed, organized and communicated.

Users (the technical community itself) assumed the key role in developing the system. This was a major factor in modifying attitudes regarding user responsibility in *managing* the technical resources.

Whirlpool's Information Network (WIN): (1) Provides an internal communication medium for documenting and orienting information regardless of source or original format into a company-wide information pool, allowing input and output from the entire technical community. Input is by means of the *WIN Summary*. (2) Either a hard copy file or microfiche file of *WIN Summaries* is duplicated and located at each division or subsidiary. The numerical sequencing of this file gives a chronological organization with broad subject categories. This permits informal browsing of file data. (3) A computer generated alert is issued monthly which cites all information entering the system during the month and arranges the citations in a scannable format by subject category. (4) A detailed computer generated subject and author index to *WIN Summaries* is developed on an annual basis with quarterly cummulated supplements.

Key words: Communications; current awareness; information network; information retrieval; technical information transfer; user attitudes.

1. Introduction

In an industrial environment, the effectiveness of technical communication lies not so much in the systems used as it does in the *attitude* of those using the system. Whirlpool's Information Center, in developing a technical information network, utilized those people whose attitudes were critical to effective transfer -- the technical community itself -- in the key development role. Utilization of users to develop a network shifted the emphasis from the conventional correlation of system capabilities with user needs, to communication factors.

---

[1]Manager, Information Center

2. Communication Factors

Three basic communication factors were examined by the technical community in developing this network:

2.1. Motivational Factors Involved in Communication

The users of technical information examined such things as: what attitudes are involved in accepting a new concept; what technical information warrants transmission beyond one immediate department and why; what incentives or rewards are there in transferring information horizontally to company peers beyond one's department. A probe into user attitudes revealed that one's concept of his own job function is the primary factor in determining either choice of information source or choice of recipient, and also in determining the selection of the communication channel employed. Receiving *technical* information is to a large extent dependent upon the sources or channels that an individual selects. This individual selection is unlike the communication flow pattern of *operational* information, which tends to follow the organization chart and one's position or assignment largely determines the operational data one needs and gets.

Users generally felt that the broader the range of relevant data one could bring to bear on a particular problem, the greater the quality of one's decision. And that the range of relevant technical information one brings to bear is largely influenced by the way one sees his technical job assignment.

Not too surprisingly, this probe of motivational factors also indicated that there is greater acceptance of new concepts which were interpreted and evaluated by Whirlpool employees than concepts submitted directly from outside sources.

2.2. Environmental Factors Involved in Communication

A look at the role that organizational structure and managerial style play in technical communications revealed that technical information, like operational data, tends to flow vertically within divisional organizational structures with little horizontal transfer among company peers at other divisions. Media used to communicate technical information was essentially sender oriented in that the sender determined who would get a report or attend a meeting based upon the sender's knowledge of individual interest. This further accentuated the parochial characteristics of technical communications as one's knowledge of subject interest was limited primarily to local division personnel.

2.3. System Factors Involved in Communications

A probe of the system factors involving both oral and written communication processes revealed a strong orientation toward verbal channels in communicating technical information. Verbal channels (meetings, informal discussions, telephone conversations) were felt to be faster and more responsive by permitting dialogue. This preference for verbal channels however, also further enhanced the parochial characteristics of technical communication.

The major barrier to the effective technical communication of *documented* information was felt to lie in the fragmented char-

acteristics of documented information structured in terms of media or source rather than subject content.

Both the technical community and information processing departments (EDP, Records, Information Center, Telecommunications) participated in this probe of Whirlpool communication characteristics and in the subsequent development of a communication network. Although the development of a technical network focused on documented data, care was taken to encourage the extension of verbal communications beyond division structures. In addition to, and separate from, the development of the documentation network (WIN) described below, several verbal communication channels evolved. An internal technical referral service, consisting of a directory of technical expertise (who knows what) was established, extended telephone service (WATS) permitted easier personal contact with peers throughout the company, and company-wide technical conferences fostered contact with one's counterpart at other divisions.

3. Whirlpool Information Network (WIN)

WIN is basically a system that transmits and stores documented technical information which has been analyzed or interpreted by a Whirlpool employee in terms of specific job assignment. The system is essentially user oriented rather than sender oriented. Although the sender can designate specific recipients, the user searches and selects applicable information from a company pool of technical information.

Any technical information that an employee finds relevant to his job assignment which he feels has interest beyond his particular assignment is transmitted. This may be information developed internally or applicable information from outside sources.

3.1. WIN Summary: A Means of Documenting Condensed, Company-Oriented Information

The format of the *WIN Summary* form permits about 900 words of documented information. The written summary usually includes results, observations or conclusions, and, if applicable, recommendations. It aims to be highly informative in itself. It may include sketches, drawings, be typed or even handwritten.

Fourteen data elements comprise the *WIN Summary* sheet. One includes a title, or citation. (A citation is used when the information submitted did not originate with the sender.) The citation, when employed, indicates title, author, and source. It is, of course, essential that the title or citation be descriptive of the summary as this is a key factor in selecting information within subject areas. Other identifying elements include sender's name, division and department location of the sender, and date.

If the *WIN Summary* is part of a full report, or summarizes other documented local data, two data elements are used. One indicates the local file number or location of the full report or supplemental data and the second indicates the size (number of pages) or format (lab notebook -- raw data).

Another data element fulfills a commentary function. It permits the sender to indicate: restrictions, if any, to the distribution of the summary, or sources for additional information, or any qualifying factors the sender may wish to designate.

Two final sender-generated data elements include both the distribution the sender designates for the *WIN Summary*, and the retrieval terms the sender feels describe the contents of the summary. These terms may include code names, company names as well as subject terms, and are for this particular data element, uncontrolled terms.

Three elements comprising this program are submitted by the processor. One indicates a broad subject classification number. At the present time 18 broad subject categories are used such as ecology, computers, electrical engineering, food technology, mechanical engineering, manufacturing and shipping, economics and marketing, materials, testing, safety and consumer protection, etc. This subject classification number (two digit) is the prefix to the identifying *WIN Summary* number and serves to arrange the summaries in broad discipline categories for browsing.

A second processor generated element is the *WIN Summary* number. This consists of a seven digit number. The first two digits consist of the last two numbers of the current year and the remaining five digits are a unique accession number. This number serves to arrange summaries in a hard copy file chronologically within the broad discipline category. This data element also permits a suffix letter code to indicate information source e.g. internal data, published literature, vendor data.

The third processor generated data element consists of subject terms. As this network uses a controlled vocabulary the author suggested subject terms may be modified to conform with Whirlpool's list of subject terms. Cross references are used in all indexes referring an author generated term to a controlled term. Various subject terms "authorities" are used in developing the Whirlpool list of controlled subject terms. Engineers Joint Council's, *Thesaurus of Engineering and Scientific Terms* is used for technical vocabulary, *Thomas Register* for company names; *Business Periodicals Index* for business and management terms.

Of the fourteen data elements comprising the *WIN Summary*, seven are currently processed by the computer. These are: the title (or citation), sender's name, division, supplementary notes if a restriction is indicated, the class category, *WIN Summary* number and the controlled index terms.

3.2. WIN File

A hard copy file of *WIN Summaries* is duplicated and maintained at each of eight separate divisions or subsidiaries of the corporation. The numerical arrangement of this file places *WIN Summaries* chronologically within eighteen broad subject categories, thus permitting local browsing of summaries. This browsing capability is utilized primarily by new employees with new assignments as it offers a comprehensive look at what has been going on in major areas of company interest. Retrieval from this file is, of course, primarily by *WIN Summary* number as cited either in the computer generated index or alert. This file is also available to divisions in microfiche format.

3.3. WIN Index

A computer generated subject and author index is developed on an annual basis with quarterly cummulated issues. The December issue, which covers the full year, becomes the permanent annual index. The file is thus searched manually by using annual indexes plus the current quarterly supplement. The computer stored data is purged once the annual index is developed.

3.4. WIN Alert

A computer generated monthly alert to all summaries entering the network during the month is widely disseminated. This *WIN Alert* indicates the descriptive elements of the *WIN Summary:* Title/citation, sender, division, location, file number, form code, and an indication whether the summary is restricted or not. These elements are listed on the alert in the same order in which summaries are actually filed -- by broad subject category. Whirlpool personnel preferred this scannable alert format of all data entering the system each month, to a selective dissemination of information (SDI) format.

4. Evaluation

It is difficult to provide a precise account of developmental costs of the *WIN* program due to the broad level of company-wide participation. However an estimate of total number of participatory man hours involved runs about 1,500 hours or slightly less than one man year.

The primary tangible benefits of the *WIN* network over the past four years has been a sharp reduction in the cost of printing and storing multiple copies of technical

reports. The *WIN Summary* is used as a report surrogate and full copies of technical reports are reproduced xerographically only on request. However, the intangible benefits, in the opinion of users, are of primary corporate value. The users feel that the network provides a means of upgrading technical decision making by establishing a broader company-wide information base. And the network also provides a means of extending technical productivity by providing a company-wide framework for transmitting an individual's analysis beyond his particular problem.

These intangible benefits have been effective, however, only to the degree that individual attitudes have been touched. Probably of greater importance than the *WIN* program itself in enhancing technical communication has been the growing awareness within the technical community that information is indeed a company asset and must be managed by the community to the same extent that other company resources are managed. This awareness and shift in attitude resulted in part from the technical community's active participation in the *process* of arriving at a network rather than from the network itself.

A Computerized Model For Determining Brand Position By Small Geographic Areas

Dik Twedt[1]

University of Missouri-
St. Louis 63121

Manufacturers of food products and other packaged goods have long sought a practical, economical way to determine brand position by relatively small geographic areas. Demographic data are increasingly available by ZIP Code, and some companies have converted sales administrative territories to modules based on ZMA's (Zip Marketing Areas), which may be aggregated by routes, districts, and regions. Marketers can determine market shares and share changes over time, for as many as 137 ZMA's. Approximately 600 ZIP Sectional Centers (first three digits) serve as modules. They are stable, may be aggregated according to individual marketer's needs, and permit use of a wealth of data from government sources. In addition to data on distribution, display conditions, price competition, and levels of point-of-sale promotion, the model provides data on actual brand share of volume for a given manufacturer's brand. Information generated is used in conducting and interpreting marketing experimentation, such as determining effectiveness of alternate promotional strategies. Most important application is "management by exception" capability, which permits marketing management to make prompt, effective response to out-of-limit conditions.

Key words: Advertising experimentation; brand share; competitive status; display share; distribution; market share; marketing experimentation; marketing measurement; pricing; retail audit; ZIP Marketing Areas; zone analysis.

1. Background and Purpose

"Market share" is a key concept in marketing, since many far-reaching management decisions are based upon knowledge of total potential, and that fraction of the total held by a given brand. For example, the nature and magnitude of promotional support given to Campbell Soups or Morton's Salt, both of which enjoy relatively high market shares of their respective product categories, may be quite different from that given a newly introduced brand. Decisions about optimum length of line, capital expenditures required to expand production facilities, size of sales force required, probable payback, and the like are profoundly influenced by knowledge of where a brand stands relative to its competitors.

[1] Professor of Marketing

537-860 O - 74 - 22

Although some companies are able to share information on total production and sales (usually through their trade associations), the precise detail required for decision making is usually missing --- partially because of the natural reluctance of the marketer to divulge information which may prove helpful to his competitors, and partially because of fear of governmental constraints associated with anti-trust laws.

Need for market share data is not new, and over the past half-century, many organizations have attempted to provide such information. Two leading examples are the A. C. Nielsen Company, which initiated its Food & Drug Index in the mid-1930's, and Market Research Corporation of America, which began its syndicated reporting service a few years later. Nielsen share data are based upon a retail audit of food stores, and MRCA data are based upon weekly purchase data obtained from a panel of households. A more recent syndicated service is SAMI (for Selling Areas-Marketing, Inc.), a subsidiary of Time Inc., which bases its reports on food chain warehouse withdrawals.

When I joined the staff of a major marketer of refrigerated food products in 1963, my first assignment was to review existing data sources for total product category movement, and competitive brand shares and prices. Objections to existing sources were found to center around such factors as: 1) high annual cost (up to $30,000 per category), 2) unavailability of data by small geographic area, 3) unavailability of data by major retail food chain, 4) reporting lag, 5) some services had problems with refrigerated foods, which are often delivered store-door rather than to warehouses, 6) some major chains refused to cooperate in providing data, and 7) syndicated information sources were equally available to competitive marketers. The purpose of the model to be described was to provide timely, valid data on brand position, by small geographic areas, and at reasonable cost.

2. Description of Sample

The United States (including Alaska and Hawaii) were divided into 137 ZMA's (ZIP Marketing Areas). Criteria for setting ZMA boundaries included: 1) the entire U. S. must be accounted for, 2) no overlap permitted between ZMA's, 3) no County lines broken (every County is wholly within a ZMA), 4) areas of dominant media coverage will be taken into account, and 5) consideration will be given to natural transportation patterns. The sampling frame included about 3,000 retail store locations at 132 sampling points, representing 105 of the 137 ZMA's, accounting for more than 85% of total U. S. consumption of the five product categories measured.

3. Method of Data Collection

Observers entered each of the 3,000 retail food stores, and recorded pertinent information about distribution, display, location, price, special promotions, etc. about each product category and brand. More than 1,200 different brands were identified in the study, conducted twice yearly.

4. Data Processing

Observers recorded their findings on printed forms which were carefully checked, transferred to tape via punched cards, and all data were again subjected to elaborate error-detection routines, with a machine-edit program designed to print out data exceeding predetermined limits. Time lapse between data collection in the field, and final printed report to management, was about 30 days. A major efficiency of the model in its present state is the production of data in a printout format that is ready for the camera, with a preprinted frisket providing typographically superior headings.

5. Characteristics of Information Collected

Detailed data are provided for total product category, and for major brands within each category, on distribution, display share, retail pricing, and special promotions, by the following informational hierarchy:

Total U. S.
- -Major Corporate Sales Regions, Districts, Routes
- -ZMA's within Districts and Regions

-By Major Chains within the above groupings

As an example of the level of detail available, the printed report cites the share of display space enjoyed by a given brand (and its competitors) in Safeway stores in San Diego. No syndicated service offers such detailed information on a regular basis.

A variation of the described model, based upon estimates of total industry volume and internal sales records, provides answers to such questions as:

a) What is the total consumption of (product category) in the Washington, D. C. ZMA?

b) What proportion of total consumption is accounted for by each major chain within this ZMA?

c) What is the marketer's brand share, by ZMA and by chain within ZMA?

6. Uses of the Data

The model provides information twice yearly on key marketing factors, by small geographic areas (ZMA's), which can be aggregated according to the individual company's sales territories. Because data are recorded by approximately 600 ZIP Sectional Center areas (first three digits), the model is flexible, and can be changed as sales territories change. Information generated is essential to conducting and interpreting market tests, including experiments in productivity of alternative promotional strategies. The most important application of the model is its "management by exception" capability. Computer-generated graphic reports call attention to out-of-limit conditions, which permits prompt and effective response by marketing management.

7. Controls

The greatest control problem associated with this model was the construction and maintenance of a zero-defect Brand Index. It was found to be impractical to program the computer to recognize "near misses" in spelling, and not until a valid and complete Brand Index was constructed, together with a carefully supervised edit routine, did the system become efficient. Other major controls were relatively straightforward, involving checking validity of recognition codes and visual inspection of printout of out-of-limit data. Obviously, in a time of rapidly rising food prices, the limits change correspondingly.

8. Improvement of Data

Potential efficiencies could result from elimination of the manual edit program, and manual keypunching from source documents. These operations could be improved by alphanumeric hand printed entry data, by more sophisticated machine editing, and by optical scanning transference of validated data to machinable form.

Standardization of Data Elements in Library Bibliographic Systems

R. E. Utman

Princeton University Library
Princeton, New Jersey 08540

National and international standardization of bibliographic data elements is progressing under the aegus of the Library of Congress, the American National Standards Organization, and the International Standards Organization. This is having important effects on developments in automated information systems, information retrieval, libraries and library networks, documentation, publishing, and in the general realm of human communications. The brief history of this bibliographic standardization activity is accounted for, with emphasis on the important interrelationships that characterize data element standardization in an information systems world. The necessity to assure compatability amongst computer-communications hardware systems and the wide range of human communications media -- written, visual and oral, current and archival -- requires constant awareness, coordination and control of the several data element and systems-related standardization activities of the past, present and future.

Key words: American National Standards Institute; bibliographic standard; data element standardization; Library of Congress; MARK II; Ohio College Library Center; standard.

Presentation

The library automation field has as its goal nothing less than the capture in machine readable form, and the storage, organization and retrieval of all recorded knowledge, regardless of the form of the recording, i.e., whether it is in print or non-print form, on film, audio records, tapes or otherwise. It has been a flourishing business the last ten years. The use of computers in libraries got a late start as far as data processing in general is concerned, but through the middle 1960's quite a bit of development activity really gave library automation a boost. Then, with the realization of a few basic standards around the period 1969-72, library automation has really begun to take off.

This again impresses one with the importance of standardization to our ability to realize the full potential of a field of computerized endeavor. My talk tells the story of a particular standard in library automation. Hopefully this story will encourage everybody here to begin to participate actively in the national standardization procedure, and particularly to represent the user and the user opinion therein. It is essential in healthy standardization that user opinion and needs be brought directly to the attention of the manufacturers, who have to date in ANSI X3, X4 and Z39 (Information Systems, Office Machines, and Libraries, respectively) carried a predominant role. This is not necessarily the most representative manner in which standards should develop. Therefore, again, if any of you represent users, by all means participate actively in presenting the user's requirements for standards.

The story I want to tell is that of the American Standard Z39.2-1971, the standard for bibliographic interchange on magnetic tape. It has al-single-handedly unlocked the real future for the use of computers in controlling, accessing and retrieving recorded knowledge of all types. The basic requirement for accessing recorded knowledge is to have a bibliographic description of each piece or recorded unit. What is a bibliographic description? It is a record of the information that uniquely describes and identifies a piece of recorded information. Only with such a unique description can one get to that particular recording of knowledge. Therefore, a standard for bibliographic description is basic and essential to the use of computers in this most demanding and flexible of all non-numeric areas of computer endeavor.

The subject standard got started in the early 1960's at the Yale Medical Library under Fred Kilgour, in a joint project with Harvard and Columbia to develop medical literature cataloging in computerized form. Kilgour went on at Yale to develop a general literature catalog in machine readable form using a bibliographic description based on the record originally developed for the medical literature.

In 1965 the Library of Congress (LC) began its major computerization effort, by investigating how to put its catalog information in machine readable form. LC borrowed heavily from the Yale experience in developing the current LC MARC catalog record format, which is really derived from this Yale work that preceded it. In about 1968, when the program for MARC distribution of bibliographic information on magnetic tape began, subscribers received a weekly tape of LC cataloging data in machine readable form. Also in 1968 a Z39 Subcommittee was formed, under Henriette Avram of LC as Chairperson, to develop a standard format for bibliographic data and communication. In 1971 their resultant proposal was approved as an American Standard (Z39.2-1971). Now this MARC standard is of such basic import as to be comparable in the library world to the American Standard Code for Information Interchange (ASCII) in the computer industry, or to a standard bill of lading in the shipping business and all of industry. There is nothing else in information systems standardization that is as completely flexible in its capability of describing a bibliographic entity (a book, serial, map, etc.), and I wish to emphasize the work flexible. As you users, or ex-users of library catalogs can well appreciate, there could hardly be anything more diverse in data element content and structure than the unit record (catalog card) that describes a title or volume in a catalog; especially when you consider that for all the information forms, from print to non-print, sound recordings, films, maps, etc., the standard recorded description of a piece or title would have to be extremely flexible in order to provide quality and unique identification, subject analysis, classification and location, etc.

Such bibliographic quality is an essential requirement in any research library, such as the one I am fortunate to be associated with (Princeton has a leading research library, particularly in the areas of western history and literature, the humanities, etc.). In order for the serious scholar to be able to differentiate between all the writings under a particular title (all the editions in all its language versions, and so on), and also to be able to use this catalog for subject analysis and research in depth, the bibliographic description must not be restrictive. It has to be as flexible as is conceivable, and yet for interchange purposes it has to be in a standard format. This was the problem, and the excellent group of people who worked to develop this standard format for bibliographic interchange on magnetic tape came up with a structure (or format) within which upwards of a hundred different data elements can be employed to uniquely identify and describe virtually any form and/or item of recorded knowledge.

How is the standard being used? The LC through its MARC tape distribution program has been sending out its English language monographic cataloging since 1968, to over 100 subscribers throughout the USA and the world. One such subscriber, the Ohio College Library Center (OCLC), an organization in Columbus, Ohio, under the directorship of Fred Kilgour, that provides shared cataloging in machine readable form on-line to upwards of 180 terminals throughout the Eastern USA at the present time, promises within the next couple of years to be providing on-line interactive access to over one million bibliographic records or titles (monographic and serial) to upwards of 600 CRT terminals throughout most of the USA. OCLC thus has the potential of becoming a model or de facto national network for on-line cataloging and access for the entire USA library need. Also, OCLC is the only such network and data base resource at this stage of library automation development, of real consequence and experience that is. Their success is based on the fact that they get LC cataloging in standard form, to or from which the OCLC computers can then provide information on-line to/through their extensive telephone-terminal network.

A new experiment is about to be established by OCLC and the Federal Library Committee (FLC), which will put the current 770,000 titles of OCLC's catalog data base (in this standard format) on to the commercial TYMSHARE net. This network currently serves over sixty metropolitan areas in the USA, and it serves the Canadian provinces, several European metropolitan areas, and also South America. This approach will make the extensive OCLC bibliographic data base available almost world-wide, as well as throughout the USA on a commercial basis. Selected Federal Libraries will experiment with access through the TYMSHARE net from local CRT terminals, with their catalogers having instantaneous access to this considerable data base. The LC is also now cataloging in machine readable MARC format the new French and German monographic imprints, as well as all new serials and periodicals, films and maps. You can begin to see from this how just one standard of machine readable format for bibliographic description is going to effect the future of libraries in their efforts to present in an organized and systematic manner the recorded knowledge of mankind.

The standard itself consists of a specification of a data structure and data element content. It also observes or adheres to several other American Standards: e.g., it requires recording in ASCII code; it employs standard 800 bpi magnetic tape; it uses American Standard magnetic tape labelling. However, it contains over 100 different data elements, only three of which are themselves standardized. And here is the relevant problem to this conference -- e.g., there is a standard book identification number element, there is a standard serial number, and there is a standard form of the date, but there is no standard country code to identify the country of publication, and there is no standard code for the language used. This lack of data element standardization goes on and on throughout this standard. It leads me to a realization that without active participation in and support to the ANSI standardization process (for instance, in Z39 and its subcommittees), the great potential of this bibliographic format standard will continue to be retarded until the data elements within its structure are given appropriate standardization attention.

Addendum

1. We have in the audience, Mr. Sundblad, who represents the Secretariat of the International Standards Committee ISO/TC46, the counterpart of the ANSI Z39 standards committee on libraries, documentation and publishing. He wishes to report (and I should have, but was restricted by time) that the ANSI bibliographic interchange standard described above has an international counterpart in the ISO-2709 standard. Therefore, the international library scene is benefiting from this attempt at orderliness as well. He also reports that ISO/TC46 standardization activity is underway in the areas of character sets for bibliographic interchange, content designators, and filing rules for cataloging.

2. Question--You said there was no American standard country code in the MARC record of 100 data elements. Why not FIPS 10? (Neil Wallace)

Answer--Right, there is no American Standard Country Code. FIPS 10 is a Federal standard, and is practically a de facto USA standard, and every day it becomes a little more so. But FIPS 10 has serious short comings in international applicability, and it has yet to achieve widespread international acceptance, although offered (or profered) as an American position.

3. Question--Is there any work being done to adopt a universal language to eliminate translation problems and costs? This seems to be one of the most essential standards needed, especially in information fields.

Answer--No such work is going on within the auspices of ISO or ANSI to my knowledge, unless (jokingly) one wishes to consider the universal appeal (especially for computer interchange use) of binary forms.

Management of a Business Information System in a Multinational Envirnoment through Standardization of Data Elements

Carroll P. Weber

Corporate Banking Systems
Bank of America
San Francisco, Ca. 94137

Banking is changing to meet the changing needs of businesses. This includes systems that operate in a multinational environment. To meet new challenges, Bank of America is developing computer-based business information systems to help in serving its business customers on a world-wide scale. Data standardization is a prerequisite to such a system. The management of data elements within the business information file gives account officers the information they need to serve both the financial needs of the business customer and the profit requirements of the bank.

Key words: Address standard; banking; business information system; country code; currency code; data standardization; D-U-N-S number; industry classification; name standard.

1. Introduction

Banking, in today's world, is involved in much more than just taking deposits and making loans. Businesses have changed, and banking has changed with them.

Businesses needed flexibility in the managing of their cash reserves. Banking responded with variable term and rate time deposits, with acceptance financing, with negotiable certificates of deposit, with money market trading desks. Businesses needed to expand their customer bases and to smooth the flow of their sales. Banking responded with credit card plans, overdraft privileges, courtesy-check guarantee cards. Businesses needed innovative help in the daily management of income and expenses, in the monthly management of benefit packages, in the annual planning for capital expenditures. Banking responded with billing and payroll services, with pension fund plans, with lease-purchase arrangements.

The past decade has seen the growth and development of the multinational corporation. Logically, banking has also grown and developed multinationally. A business customer may need a stand-by credit facility of $1,000,000 over a twelve month period. And he may specify that the funds will be drawn down

as needed anywhere in the world. The loan might eventually be allocated for raw materials purchased by pesos in Argentina, for wages and production costs paid by yen in Japan, for shipping fees paid by drachma in Greece, and for accounts receivable financing by deutschemark in Germany. Banking must be able to meet this customer's needs. The bank that does it best, prospers.

2. Business Information System

2.1 The Problem

Bank of America would like to prosper. To prosper we must serve our customer. To serve our customer we must know him. Not as easy as it sounds. Our customer, or one of his branches or subsidiares, may have any number of separate accounts in over 60 different service lines, at any number of our 1023 offices in California or our 102 overseas branches.

2.2 The Solution

We are in the process of developing a computer-based central information file containing all the account relationships that a business, including its branches and subsidiaries, has with the bank. We call this the Business Information System - BIS for short. Data is passed directly from each computer accounting system to BIS. We are approaching this modularly with demand deposits (checking), time deposits (savings), loan commitments, and commercial loan drawdowns in the first module. The data in the system will include both balances and profitability.

2.3 The Benefits

By being able to pull together the total worldwide relationship that a business has with the bank, including the profitability of each service as well as the overall profitability of the relationship, BIS provides three broad benefits.

a. Financial Decisions

More complete knowledge of the customer's total banking relationship supports the information needs of the bank account officer. He can price our services fairly for the customer as well as the bank. If he underprices our services, we lose money. If he overprices them, we lose our customer.

b. Marketing Research

Complete knowledge of accounts and their profitability permits better marketing research. The use of management sciences techniques (regression, modeling, linear programming) can profile the profitable customer.

c. Selective Marketing

Improved knowledge of customer profiles enables our marketing staff to select specific prospects for specific service lines to maximize profitability for the bank and service benefit to the customer.

3. Data Standardization

3.1 Essential to BIS

Data Standardization is an essential prerequisite to the implementation of a central information file drawing from multiple sources. We established a Data Standardization project in support of BIS. Its purpose has been to research, develop, publish and implement Bank of America standards for those common data elements which identify, locate, and classify business customers.

3.2 Key Data Elements

The key data elements requiring standardization are account number, name, address, location codes, and industry classification. Standards establish computer programming conventions and, as a result, simplify new account procedures throughout the bank.

4. Standards Adopted

4.1 Common Account Number

The key to BIS is the assignment of a unique cross-reference number to each of the accounts which a business establishes with any office, branch, or department of Bank of America - worldwide. We evaluated potential numbering systems - ranging from Employers Identification Numbers (EIN), to telephone numbers, to primary checking account numbers, to CUSIP securities numbers, to Department of Defense DOES numbers, and finally to D-U-N-S numbers.

a. D-U-N-S Number

A D-U-N-S number is a special number assigned to a business by DUN & Bradstreet. The term "D-U-N-S" is actually an acronym for the "Data Universal Numbering System." The D-U-N-S number has gained wide acceptance as an identifier. It is nine digits long. It is randomly assigned. It has no intelligence built into it. It is unique to a specific business establishment at a specific address. There are over 3,000,000 D-U-N-S numbers assigned in the U.S. and Canada. Numbers have also been assigned in the United Kingdom. Number assignments are beginning in Europe, South America, and other parts of the world.

b. Why the D-U-N-S Number?

We came to the conclusion that D-U-N-S is far superior to any other existing system. These are some of its stronger points:

(1) D-U-N-S is the only feasible nationwide system for non-personal entities. As such it is the only non-government identification number recommended in the "Proposed American National Standard Structure for Identification of Organizations for Information Interchange" developed by a subcommitee of X3L8. (NOTE: the only other recommended number was the employer identification number - a government identification number).

(2) D-U-N-S is the only system that is hierarchically structured to individually identify a corporation, subsidiaries within the corporation, and branch offices within the subsidiaries.

(3) D & B advises business establishments of their D-U-N-S number.

(4) D & B will assign a D-U-N-S number to an establishment at no cost - if the request is made by the establishment itself (third party requests are chargeable, however).

(5) D & B offers marketing data on establishments. The data includes D-U-N-S number, name, complete address (location and mailing), standard industrial classification codes, number of employees, annual sales, net worth, credit rating, etc. By using D-U-N-S as a common account number, this external data can be incorporated into an internal information system.

(6) D & B has expanded its coverage to firms outside of the U.S. particularly in the United Kingdom. (NOTE: D & B has recently announced the forthcoming publication of a Marketing Directory of some 45,000 international businesses in 135 countries).

4.2 Name

We investigated name formatting techniques. We are in general agreement with the "syntax rules" contained in the "Proposed American National Standard Structure for Identification of Organizations for Information Interchange" developed by a subcommittee of X3L8.

Full implementation of the name standard will require modification of our computer systems to permit variable length name records.

4.3 Address

Our initial emphasis has been on United States addresses. We have established field lengths for four data elements comprising the address: Street address - 32 alphanumeric characters; City - 24 alphabetic characters; State - the two alphabetic character Postal Service abbreviation; ZIP code - 5 numeric characters.

Implementation is proceeding in all new computer applications.

4.4 Location Data Codes

a. County Data Codes

We have adopted the coding structure published by the National Bureau of Standards (FIPS PUB 6-2) and by the American National Standards Institute (Standard X3.31-1973).

b. State/Province Data Codes

For the United States we have adopted the coding structure published by the National Bureau of Standards (FIPS PUB 5-1) and the American National Standards Institute (Standard X3.38-1972).

For other countries we have adopted the coding structure published by the Defense Intelligence Agency (DIAM 65-18 - "Geopolitical Data Elements and Related Features").

c. Country Data Code

We have developed an internal Bank of America standard containing codes for countries and territories. The geographic code consists of three digits. The first digit identifies the continent and follows a basic pattern of North-South moving East. The last two digits denote a country's or territory's position on a continent. Numbering begins in the North and proceeds in West-East sequence with numbers available for additional political units. Examples are United States 102, Argentina 241, West Germany 340, Egypt 438, Uganda 518, Bangladesh 762, Japan 880, Fiji 950.

Incidentally we have also adopted the same code as a currency code: U.S. dollar 102, Argentina peso 241, West German deutschemark 340, Egyptian pound 438, Uganda shilling 581, Bangladesh taka 762, Japanese yen 880, and Fiji dollar 950. We also retain the flexibility to handle multi-tier currencies, e.g., the Argentine commercial peso 241C and the Argentine financial peso 241F in addition to the basic currency designations.

4.5 Industry Classification

a. Legal Status Data Codes

We have developed an internal Bank of America standard containing codes for legal status. Examples include bank 07, central bank 36, cooperative 45, corporation 03, foreign bank 09, government-owned corporation 47, non-profit organization 02, partnership 41, sole proprietorship 42, state of California 22.

We have also adopted the 4 digit standard industrial classification code contained in the "Standard Industrial Classification Manual - 1972", published by the Office of Management and Budget.

5. Management of Data Elements

5.1 People

Businesses deal directly with account officers located in our offices, branches, and departments. In establishing a new banking relationship, the account officer is the primary data element source for the basic information necessary for proper coding of the standard data elements. However, we evaluate our account officers on how well they serve our customers business requirements, not on how proficient they are on assigning codes.

We have established a centralized staff to manage coding of data elements. Their first responsibility when a new account is opened is to insure that the business in contained in or added to the Business Information System with a D-U-N-S number and that the same D-U-N-S number is entered on the new account record. Their second responsibility is to properly code and maintain the data elements describing the customer on BIS.

5.2 Computer

Once people have made a one-time decision on the data elements in the Business Information System that describe our customer, information interchange takes place. The BIS computer system transfers these data element values to each of our customer service line accounts. Conversely the service line computer systems transfer balance and profitability data back to BIS, using D-U-N-S as the linkage.

6. BIS - Success or Failure

As with any complex data system relying on many people in many places processing data on many computer systems, BIS is not a total success nor is it a total failure. Our successes have been that each module of BIS does do what we wanted it to do. It brings us one step closer to meeting our goal which is also our customer's goal - intelligent managing by our account officer of the customers total financial relationship with the bank - worldwide. Our failures have been that a module may take longer and cost more than we anticipated. The farther we get from our home base in San Francisco, the more time and costs tend to escalate. Data element decisions that appear solid as a rock in San Francisco may begin to crumble in London and turn to sand in Tokyo. We reevaluate, correct, and go on. BIS is a success today. Tomorrow it will be a greater success.

Addenda

The following questions were submitted following presentation of this paper at the "Symposium On The Management of Data Elements in Information Processing".

1. How will BIS affect other banking organizations?

It won't. BIS is strictly internal to Bank of America.

2. What percentage of your customers have D-U-N-S numbers?

Percentage depends on type of business. Virtually all manufacturers have numbers. Most retail establishments have them. About half of the service industries are covered. Very few professionals (doctors, accountants, attorneys, etc.) have numbers. Overall, perhaps 60% of our business customers in California are contained within the California D & B file of D-U-N-S numbers.

3. How long does it take to get a D-U-N-S number? How do you handle interim numbers?

It is possible to get a new number by phone. Its not very practical, though. We don't make any special effort to obtain a new number. We check a new business customer against the D & B file of numbers (which we obtain on a bimonthly basis). If we find a number, we use it. If we don't, we assign a user or interim number from the block of numbers reserved for that purpose. Each new D-U-N-S record in the bimonthly D & B file is checked against our interim file. When we find a match, we replace our interim number with the D-U-N-S number. BIS is programmed to pass the D-U-N-S number to each account that the business customer maintains in our service line applications. One entry corrects BIS and each interfaced application.

4. How do you link subsidiaries to their parent company?

In our basic customer record, a business establishment's D-U-N-S number is its account number on BIS. However, we also carry up to three additional "pointer" D-U-N-S number where appropriate. These are the D-U-N-S numbers of its parent company, and in the case of a multi-level corporation, the D-U-N-S number of the top company in the hierarchy. The latter number is sometimes referred to as the "ultimate" D-U-N-S number.

5. How do you identify foreign businesses corresponding with your international department in San Francisco? Does lack of standard translation/transliteration require more manual intervention than in most other operations?

When a foreign business establishes an account relationship with our international department, we add it to BIS with either a D-U-N-S number or an interim number, as we would for any other new business customer. Translation problems are normally handled by specialists in each foreign language. More manual intervention may be required, but this is normal in international transactions.

6. Will the D-U-N-S number enable you to research the history of a customer - mergers, name changes, etc? Can you trace a customer backwards in time?

We can't do this. This ability would require retaining prior names, prior hierarchy numbers, prior statistical data. It would be quite costly and have limited value at Bank of America. There may be some necessity for an account officer to have some back-up data of this nature. However, a 3 x 5 card or a notation on loan documentation would satisfy his needs.

7. Do you have an established, regularized method for keeping your name and address file up to date? If so, what is the frequency?

We do not have a formal review of names and addresses. We pick up some changes from a bimonthly D & B file of California business establishments. Our primary source for changes, though, is the account officer and teller at our banking offices. They are in direct contact with customers. As name and address changes occur, the information is routed to our centralized data base management staff for entry into BIS.

8. What was your reason for inventing your own country code instead of using the standard published by the National Bureau of Standards (FIPS PUB 10)?

We needed a code in 1968 - badly! ANSI had not yet approved one. FIPS PUB 10 had not yet been published.

9. By deviating from the ANSI/FIPS country code aren't you committing the basic procedural error that this symposium is addressing itself to?

In this case, I think we have made the correct choice. We needed to standardize within our own organization before an external standard existed. Ours works very well. A final standard will have to come from the International Standards Organization (ISO). It will probably be a numeric code with an alphabetic abbreviation. When a code structure is agreed upon at the ISO level, we can make a one-for-one conversion from our code to a universally accepted code.

10. Are there any current efforts underway to establish standard currency codes? Is your code used outside of Bank of America?

I am not aware of any current efforts to establish standard currency codes. A currency code study may be undertaken this year either by ANSI X3L8 (Representation of Data Elements) or X92 (Bank Operations). Our code is used only by Bank of America.

11. Are the standard data elements you are working on directly related to those that are being developed for the transportation industry by the Transportation Data Coordinating Committee and the Department of Transportation (CARDIS system)?

We are not working directly with either TDCC or DOT on codes. However, we are all represented on ANSI committees - which permits a interchange of views and works towards adoption of standards that we can all accept.

Standardization of Data Elements and Representations

Harry S. White, Jr.

Associate Director for ADP Standards
Institute for Computer Sciences and Technology
National Bureau of Standards
Washington, D.C. 20234

The full advantages of recent advances in computer and communications technology cannot be realized until standards are developed and implemented to provide for the uniform identification, definition, and representation of the data being interchanged. There is an ever increasing need to interchange data and programs with state, local and other governments, and with industry and the public, all of which adds further emphasis and dimension to the need for responsive standards that will facilitate meaningful interchange.

This paper addresses three areas of formal standardization activities, those being carried out by the International Organization for Standardization (ISO), those by the American National Standards Institute (ANSI), and those presently underway within the Federal Government. Copies of referenced standards documents cited in this paper are contained in Appendices A through E of these Proceedings.

Key words: American National Standards; computers; data elements and representations; data interchange; data processing systems; Federal Information Processing Standards; information processing; international standards; U.S. Government.

1. Introduction

Standardization is a tool in the management tool chest which if utilized appropriately can achieve increased efficiencies in the operations and the functions of an organization. The principal goal for any organization in standardizing data is to make maximum utilization of its data resources and to afford a more effective means of collecting and exchanging data with others. Data standards are consensus agreements between the sender and the receiver. Before meaningful interchange can take place, there must be understanding and agreement on the identifications, meanings, and representations of data.

This paper will address three areas of formal standardization activities, those at the International, National, and Federal Government levels.

2. International Standards

The first of these is the data standardization effort under the sponsorship of the International Organization for Standardization commonly referred to as ISO. ISO was established in 1947 to promote the development of standards for international commerce and trade and to further areas of intellectual, scientific, technological, and economic

537-860 O - 74 - 23

activities. Currently there are over 3,000 standards that have been adopted within the member bodies of the International Organization for Standardization. There are several activities within ISO involved in the development of data standards. The first of these is concerned with the development of common representations used in data processing applications. This work is being undertaken by Subcommittee 14 of Technical Committee 97 (Computers and Information Processing). Standards in the field of banking are being developed under the sponsorship of ISO Technical Committee 68 (Banking Procedures). Data standards used in administration, industry, and commerce are being developed in ISO Technical Committee 154. Another effort is the development of standard codes for country names by ISO Technical Committee 46 (Documentation). Appendix A of these Proceedings contains a list of data standards and proposals that have been approved or are under active consideration within ISO, the American National Standards Institute, and the Federal Government.

3. American National Standards

Nationally, voluntary industry standards are developed under the auspices of the American National Standards Institute (ANSI). ANSI is the national clearinghouse for standards and provides the mechanism for obtaining national consensus. ANSI is also the official member body that represents the United States at meetings of the International Organization for Standardization. Within ANSI, there are several committees involved in the development of data standards. The X3L8 Committee, Representations of Data Elements, has responsibilities for the development of standards for representing the common elements used in automated data systems and interchange. X3L8 is also responsible for the development of guidelines to be used by other standards groups either within the formal standards bodies or by other organizations in the development of standards to serve their particular purposes. A copy of these guidelines is included as Appendix B of these Proceedings. These draft guidelines will be published formally when approved by the American National Standards Institute. Also within ANSI, the X9 Committee is involved in the development of data standards for banking operations. Data elements to be used in vehicle and motorist data bases are being developed by the ANSI D20 Committee.

4. Federal Standards

Federally, data standards are developed and implemented under the provisions of Part 6 of Subtitle A, Title 15 of the Code of Federal Regulations. A copy of this regulation is included as Appendix C. The National Bureau of Standards provides for the day-to-day monitorship of the Federal data standards program. Standards are approved by the Secretary of Commerce on behalf of the President for Federal-wide implementation. There are currently six approved data standards which have been published by the National Bureau of Standards as Federal Information Processing Standards Publications, commonly referenced as FIPS PUBS.

5. Approved and Proposed Standards

Some of the standards that have been developed and approved and also some of those that are pending in a proposal status are:

Dates. The first data standard to be approved is a standard representation for calendar dates. This has been approved as an International, National, and a Federal Standard (FIPS). The elements are arranged in a logical, high to low order; that is, year, month, day. Hyphens may be used in this representation to facilitate human readability. Also included in the standard is a representation of ordinal date. (Reference Appendix A for complete identification and availability of standards mentioned in this paper.)

Standards for the Representations of the States of the United States. There are two representations--a numeric code that sequences the States in alphabetical order and the abbreviations established by the Postal Service. These representations are two characters in length. These have been adopted as both American National and Federal standards.

Standard Codes for the Counties of the States of the United States. This standard provides three character numeric codes to be used in conjunction with the State representations and will order the counties alphabetically within each State. There are special provisions for Alaska and those States having independent cities that are treated as county equivalents. These codes have been adopted as both American National and Federal standards.

Country Codes. Country codes have been approved as a Federal standard. Work is still continuing on developing country code standards at the National and International level. One of the problems in the development of country codes at the international level is the differences in languages for the names of the countries to be represented and the derivation of a mnemonic codes.

Standard Metropolitan Statistical Areas. For statistical purposes in the United States a four character numeric code for representing metropolitan statistical areas has been adopted as a Federal standard.

Congressional Districts of the United States. A two character numeric code used in conjunction with the State representations to represent each district including those districts that are identified at large has been approved as the Federal standard.

Codes for Organizations. A proposed standard for the identification of organizations has been developed by the ANSI X3L8 Committee. A copy of this proposed standard is provided in Appendix D to these Proceedings. The proposed standard consists of a three part representation for identifying organizations, and identification code designator, a code part, and a name part. The Employer Identification Number (EIN) used by the Internal Revenue Service and the D-U-N-S number of Dun and Bradstreet are the major codes used. Also user agreement (local) codes can be accomodated with the proposed standard. It is expected that this draft standard will be approved by ANSI in July 1974.

Time and Time Zone Representations. This proposed American National Standard provides for time representation in both the 12 and 24 hour time keeping systems. The International standard for time as currently proposed provides for representation of time only in the 24-hour system. As a matter of interest, the moment of midnight is represented in the standard by four zeros (0000). However, there are provisions for special agreements between interchange parties and those who want to treat the moment of midnight as the end of the day (2400).

Points and Places. Internationally, the basic elements in point locations are longitudes, latitudes, and altitudes. In the United States other grid reference systems are also being considered for inclusion in the standard. A proposed place code structure has been developed for representing all named places in the United States (approximately 130,000). The code is a five (5) character numeric code that is assigned alphabetically based upon the name of the place within each State. It is used in conjunction with the standard two character State representation for a total of 7 character positions. It is planned to make the code tables available in both printed and magnetic tape form. (A copy of this proposed standard is included as Appendix E.)

Identifiers for Individuals. A proposed standard for identifying individuals developed by the X3L8 Committee of ANSI is currently in a holding status. This proposed standard uses the Social Security Account Number and the name of the individual as a standard identifier. It was deemed essential by the standards activities involved that clarification be obtained on the use of the Social Security Account Number before proceeding further with the processing of this standard. It is expected that the clarification on this matter will take the form of either a Federal Regulation, an Executive Order, or a Statute.

Other standards under various phases of development include mailing and shipping addresses, commodity codes, occupation codes, industry codes, curriculum coes, codes for representing sexes, and merchandise identification codes.

6. Future of Data Standards

In the future as far as data standards activities are concerned, I see the trend being more and more away from general subjects and more toward standardization for particular applications. Federally, these are identified as Federal Program Standards.

In summary, data standardization is an important aspect in the role of data management. Hopefully, the experiences that we have gained thus far in our standardization efforts will prove useful in the field of data management.

The ADP user community generally has been very complacent when it comes to the matter of standards. If standards are to be responsive to user needs, ADP users need to be leaders and active players in setting standards, and not observers. The National Bureau of Standards nor any other Government agency or professional society alone can represent the interest of users. Standards should be just as important to the user community and the computer profession as are its societies and its trade journals. NBS and ANSI need and welcome your support in the development and implementation of standards. Please contact me concerning your particular needs and interests in standards and standards activities.

(Detailed information concerning Federal, National, and International ADP standards activities is provided in FIPS PUB 12-1 entitled Federal Information Processing Standards Index. Available from the Superintendent of Documents, U.S. Government Printing Office, Washington, D.C. 20402. Price $1.25 a copy.)

Fundamental Tools and Techniques Toward Evolving A Data Element Management System

Arthur J. Wright

Bell Laboratories, Inc.
Common Language Department
Piscataway, N. J. 08854

This paper treats the fundamental tools (disciplines) and techniques which are necessary in carrying out the Bell System Common Language Standardization program. Some examples of the fundamental tools are the Common Language Coding Guidelines, Common Language Training Courses, Common Language Reference Library and the Common Language Code Development Procedure. Equally important to the standardization program are the methods of processing and distributing standards throughout the Bell System. These tools and methods, and others, are necessary in the design of the Bell System's evolutionary data element management system.

Key words: Abbreviations; codes; Common Language; communications; data element; language; management system; representations; standards; User Labels.

1. Introduction

In November 1966, the Bell System established the Business Information Systems Programs (BISP) Area with the objective to centrally design, implement, and maintain major business systems required by the twenty-three Bell System Operating Telephone Companies. Such business systems include Customer Service Order Operations, Circuit Engineering, Plant Installation, Equipment Ordering, Traffic Message Forecasting, Directory Operations, Inventory of Central Office Equipment, Trunks Integrated Record Keeping and several others.

A key requirement, throughout all these systems, is standardization of the information that will appear on new system documents, which will be used by people in their day-to-day job activity. Information such as user labels (field headings), codes and abbreviations (representations used to populate the fields), and descriptions of the information being presented must be carefully developed to take into consideration the human aspects, and to recognize peoples' natural reluctance to accept change.

We call our work in this area "Common Language". It is a language that people use to communicate with a machine, that a machine uses to communicate with people, and that people use to communicate with each other.

The points I will address include the Bell System's concepts about data element standardization, the techniques and disciplines needed to develop good data element products, and the evolution of data element standardization into a manageable system.

2. Background and Evolving Technology

In 1961, AT&T initiated a project aimed at simplification of engineering procedures throughout the Bell System and with the ultimate objective of mechanizing the equipment provision function. One of the most significant problems identified was the diversity in engineering nomenclature and sources of reference used throughout the Operating Telephone Companies. It became evident that there was a growing need for the establishment of "Common Language" - a system of codes or abbreviated language which could be used throughout the Bell System, with minimal misinterpretation, by people in their interactions with computers or with other humans.

In 1963, as a first step, a Common Language Task Force, whose members were drawn from the Operating Telephone Companies, the Long Lines Department of AT&T, the Western Electric Company and the Bell Laboratories, began work in the trunk facility field (trunk facilities connect circuit switching offices). The job of the Task Force was to standardize facility assignment and usage information entered on a key document called a Circuit Layout Record Card (CLRC).

Today, there are approximately ten million CLRCs in use throughout the Long Lines Department and the Operating Telephone Companies. These records serve the maintenance of local, long distance, and special service circuits that form the nation-wide telephone network. Information included on the CLRCs relates to circuit terminations, central office equipment assignments, and cable and radio relay facilities provided to establish the communications paths. Other information covering signalling, transmission levels and the like is also included. For the most part, the information is represented by either codes or abbreviations, in order for all the information to be accommodated by the CLRC card and mechanized.

In 1965, the Common Language Task Force completed its assignment and published the first Bell System Common Language Practices (Standards). As a result of the Task Force accomplishments, a small group was formed by AT&T to continue development and ensure maintenance of the Practices.

Creation of the Business Information Systems Programs (BISP) project, in November 1966, stimulated the on-going need for Common Language. The BISP philosophy is to centrally design systems for the Operating Telephone Companies and develop programs that will permit interfacing between systems without conversion of information. Such a plan requires not only standardizing the software (e.g., programming languages, data base structures, etc.), but also demands standard representation of the data elements.

3. Data Element Management Concepts

One traditional approach to developing codes is the creation of a task force or committee to focus on specific code problem areas. These groups search out and investigate relevant documents, make concessions as to data element structures (sizes) and their code sets, present recommendations, and push to get acceptance. However, the realization comes quickly that in order to standardize the thousands of fields that will appear on the input and output forms, and Cathode Display Tube (CDT) masks of our new systems, a much faster and more technical approach than traditional must emerge.

First, we had to get the basic procedures in order; one of these was the set of rules to guide the technical people in establishing data elements that would best serve the users' needs. It was found that information relating to the code design process and a listing of coding principles were needed as

guidelines. A member of the Common Language Department undertook the task of creating "Coding Guidelines"[1] and the Western Electric Company Information Systems people are continuing studies in this area. (My colleague, Mr. Martin Gilligan of Western Electric, will present a paper at this symposium on this subject.)

Next, we established the methods and procedures to relate with the system designers of the various projects. Since our BISP approach to system development has many steps in itself - from Formation of Objectives, through to Design, Testing and Implementation - we determined that Common Language should be included as a requirement early in design. The Common Language concept is, therefore, included in BISP's documentation.

Recognition of Common Language alone won't necessarily get system designers to request central development of data element standards. It seems everybody knows how to code and abbreviate, so why not use what's familiar? I submit that this attitude caused the need for Common Language to surface. To offset this possibility of each system designing its own standards, a member of the Common Language Department was assigned to directly support one or more BISP systems. This person works with the system designers to understand their input and output language needs, and ensure development and implementation of standard Bell System data elements.

One other important area of concern was the documentation of the data element development effort in a technical manner. This serves two main purposes. First, it allows those who review the development study and recommendation to assure themselves that a thorough job was done. Second, it provides a formal reference for the data elements in the project's documentation and in the Common Language User Label Manual. (I'll address the User Label Manual later).

Equally important as any of the above items, is the need for maintenance of the data elements. No matter how well a code set is developed initially, changes may be required as time goes on, and the coded representation must be updated. When such changes occur, they must be reflected, in a timely manner, in order to maintain the credability of the standard.

I hope I haven't left the impression that the evolution of Common Language all took place in the exact order in which I have presented it, or that many differing opinions were not considered; however, we did succeed in evolving the basic approach.

4. Common Language Development Tools

In order to establish and maintain any management system, I believe that what we are attempting to manage must be of value and make a contribution to the system. If data element standards are developed considering only the needs of the project, and other needs within the Bell System are ignored, then we would not truly have Common Language Standards. We, therefore, must be aware of and consider in our development, all present uses and similar needs for language representation.

A Common Language Reference Library was established to help accomplish this end. It is a specific technical library containing only documents which include language representations related to the human, in his role of interacting with people and machines. It is devoid of any documents that deal with programming needs (language or codes) or data processing center

---

[1] L. Sonntag, "Designing Human-Oriented Codes", Bell Laboratories Record, Vol. 49, No. 2, (February, 1971).

operations, since such documents primarily serve the needs of machines. Documents are obtained by reviewing every practice in the AT&T General Departments (such as, Comptrollers, Plant, Engineering, Treasury, Traffic, Marketing, etc.), all Operating Telephone Company standard practices, and appropriate documents from Bell Laboratories and Western Electric. In addition, other references include selected U. S. Government Federal Information Processing Standards (FIPS), American National Standards Institute (ANSI) and International Standards Organization (ISO) standards, and documents containing data element references from various industries and trade organizations. All reference documents are indexed according to document source and data element subject. In this manner, we are in a position to research data element usage throughout the Bell System and other important areas where we might exchange information.

Another of our tools is the code development package. This is the BISP documentation that presents all the data and considerations for a Common Language data element recommendation. The code development package includes a bibliography of referenced information, pertinent exhibits from referenced documents, a pictorial "evidence sheet" (which is actually a matrix of the data element usages), expository remarks, and a recommended standard. It is in the area of expository remarks (which are patterned after the ANSI X3L8 Committee work) that coding guidelines and reference library tools are applied in relation to users' requirements, and direct us toward making sound recommendations.

A significant language standardization tool, that is simple yet powerful, is the Code/Abbreviation Table (CAT). We often have difficulty attempting to explain when a meaningful alphabetical representation is considered a code, or an abbreviation. For example, "EDT" (Eastern Daylight Time) can be viewed as being an abbreviation if used in text, but can also be construed as a code if it's part of a structured set which might include other time zones (such as CDT, MDT or PDT). The argument is also presented that field names, represented in less than full English, are not codes, but are abbreviations, acronyms, or initialisms. Whatever might be the argument, and there are many others, there is accord in that a word should not be abbreviated (or coded) differently for the same size structure - no matter how it is used. Herein lies the utility of the "CAT". It allows us to control the assignment of codes or abbreviations by establishing all possible representations of the word. The Table, below, of a family of words is an example of the concept.

Code Abbreviation Table (CAT)

	Characters					
Word	2	3	4	5	6	7
Locate	LT	LOC*	LOCT	LOCAT		
Located	LD	LCD	LOCD*	LOCTD	LOCATD	
Locating	LG	LCG	LOCG*	LOTNG	LOCTNG	LOCATNG
Location	LN	LCN	LOCN*	LOCTN	LOCATN	LOCATIN
Locator	LR	LCR	LOCR	LOCTR*	LOCATR	

It lists two, three, four, etc., character representations for almost all words used in the Bell System. The asterisk designates the standard to be used unless other constraints exist.

If the designated standard does not fit the requirements of the user - perhaps because of a space constraint - an alternate can be immediately selected. This choice in no way suggests that the standard should not be utilized in as many applications as possible. However, the concept of the "CAT" does recognize that we can exercise judgement in applying standards to meet user requirements.

Supporting the data element standardization effort is a Common Language Training Course, which has two parts. One part is an overview aimed at stimulating system designer's awareness of the need for Common Language early in the application design. As hard as one may try, there always will be people or groups that may not understand the role of a support organization working towards Common Language. The formal training (1 day) creates an awareness in the minds of system designers that there is more to coding than meets the eye, and teaches them how to prepare a data element standardization request. The second part is a two day training course for those analysts doing the actual data element design. It highlights the application of coding principles, as they relate to the users' needs, and teaches the code analyst how to write and present the code study documentation.

5. Data Element Naming and Usage

When data elements are developed in the Bell System BISP programs, or in ANSI Committees, the scope and purpose of the data elements indicate that they are to serve as universal a group of applications as possible. The names of the data elements indicate their universal application. For example, the official name for ANSI X3.30-1971 is American National Standard Representation for Calendar Date and Ordinal Date for Information Interchange. In short it's called "Calendar Date". This shortened universal data element name probably will not be used for particular field headings. More familiar field heading names, such as, Employee Birthdate, Plant Test Date, Service Date, Due Date, and literally tens and perhaps hundreds of other specific date labels will be used. Whatever the name, or label, it should relate to the single ANSI "Calendar Date" standard.

These "user labels" (field names) must be stored in the projects' data base along with the coded representations that populate the field. Entering this information into the data base, and following specific programming language labeling conventions, the data base names might be changed to: Date - Employee Birth, Date - Plant Test, Date - Service, etc.. The user labels and the corresponding data base labels become difficult to associate.

The difficulty in association is due to structuring of the names. The data base software people structure their names in a logical machine serving manner, whereas users of the system need names in "people language" (one might call it illogical). The conflict in data element naming is a language communications problem.

Another problem in naming is caused by differing human interpretations in describing the user labels; thus, different user labels might represent the same information.

Interchanging information between systems requires that the interchange parties must be in complete agreement as to the data element name and its description and code structure. When this information requires interpretation, the system designers must get together and work out their differences. This method circumvents our Common Language Objectives.

In Business Informations System Programs, we have designed and developed a User Label Reference Manual to alleviate these communication problems. The Manual contains a listing of each user label used on any form or mask in the BISP Area. For each user label the following information is entered:

- User Label Name (in full English)
- Description (as contained in the project documentation)
- Code Structure (such as alphanumeric, length, etc.)
- Code Reference (Authority for code)
- Use in Project (e.g., Form A123)
- User Codes (e.g., a partial listing of Place Codes)

We view the User Label Reference Manual as a key tool in managing the data elements. By designing programs to aggregate the user label information by key word, commonality of usage on various forms and masks, or just listing the information in different formats (which permits visual detection of conflicts), we are now in a position to bring about Common Language in an organized way. By organized, I mean working with the concerned project or application system people, resolving any differences, and making appropriate changes to the documentation. This technique might not appear much different than that which we stated earlier, regarding system designers working out their interchange problems. However, the main difference is that once we keep track of the user label information in the User Label Reference Manual, the resolution of the conflict is not lost and we are able to build our base of standard information. System designers, from that point on, would be ill-advised to choose user labels which are in conflict with those in the Manual.

I hasten to add that our program in management and control of user labels is a continuing and demanding task seeking agreement on names, descriptions and representations. However, we feel that specific application users must have a say in the final resolution of the language representations, to ensure that we don't come up with unilateral decisions which will not fit users' requirements.

> "A man convinced against his will, is of the same opinion still."
>
> - Unknown

6. Bell System-Wide Acceptance Techniques

It is our view, in BISP, that a good data element management system is primarily based on user acceptance. User acceptance will come about if Common Language Data Elements meet users' needs. The needs of the user will be included in the code design, if the user can contribute to the code development.

Business Information Systems Programs are only a part of the total computer systems under development by the Bell System companies. However, BISP is the only area in the Bell System whose scope is large enough to warrant a Common Language Department. The AT&T Company, in recognizing the need to further implement Common Language throughout all Operating Telephone Company operations, established a Bell System Common Language Bureau in June, 1971. The working arm of the Bureau is the Bell Labs Common Language Department, that works together with the various AT&T General Departments. All data element recommendations are processed for review by AT&T, representing the Operating Companies. Simultaneously, the Western Electric Company, and other areas of the Bell Laboratories, are forwarded copies of Common Language recommendations for approval by their respective organizations. This latter procedure is a two-way street whereby data element standards, utilized as interchange information between the companies in non-BISP systems, are processed for review by BISP as Common Language Standards.

Although the Bell System is diverse in its complex operations, which serve the manufacturing, research, and operating areas of our business - each having authority to function as best fits corporate objectives - we do find unity in Common Language in the areas of interchange and commonality of data element usage.

7. Implementing Common Language Products

Actually, there are two major classes of data elements in the Bell System - one uniquely identifies equipment, locations, facilities or circuits - which is called the "identification" series; the second is the "general" series.

The identification series of data elements has been around for several years and is accepted in new designs with minimal difficulty. However, the general series of data elements, those whose names are less known (such as Manufacturing Company, Corporate Structure, Independent Telephone Company, Restoration Priority and a hundred others), must also be implemented in as many areas of Bell System company operations as possible. Publishing them in practices and writing letters telling about their availability, won't necessarily get them implemented in new or existing systems.

We are presently investigating the feasibility of having Common Language products (data elements and their coded representations, user labels, etc.) placed on line for Bell System-wide, real-time access. Such a computer system could share existing communication networks already established for other projects, and would greatly encourage the acceptance and use of Common Language throughout the Bell System.

8. Summary

The importance of standard data elements in the Bell System has consistently gained significance in the past decade. The Bell System, driven by the need to manage the massive information flow of the mechanized system, has pledged itself to language standardization. The approach we have developed - Common Language - incorporates human aspects into the otherwise sterile environment of machines. We are in a dynamic environment, continually growing and learning from our own experiences and from our association with other members of American and International Standards organizations, with various branches of the U. S. Government, and through professional exchange of ideas through symposia, such as this one. As Henry Ford once said, "If you think of standardization as the best that you know today, but which is to be improved tomorrow - you get somewhere."

In managing Common Language, our emphasis is on human communications. We believe that an effective data element management system will evolve through the central development and maintenance of data element standards, and the continued recognition of users' requirements.

1974 January 24

AVAILABILITY OF DATA STANDARDS

FEDERAL INFORMATION PROCESSING STANDARDS

FIPS 4 Calendar Date (SD Catalog No. C13.52:4) - 20 cents
5-1 States of the United States (SD Catalog No. C13.52:5-1) - 20 cents
6-2 Counties and County Equivalents of the States of the United States (SD Catalog No. C13.52:6-2) - 65 cents
8-3 Standard Metropolitan Statistical Areas (SD Catalog No. C13.52:8-3) - 55 cents
9 Congressional Districts of the United States (SD Catalog No. C13.52:9) - 10 cents
10 Countries, Dependencies and Areas of Special Sovereignty (SD Catalog No. C13.52:10) - 35 cents

Note: Copies should be ordered from the Superintendent of Documents, U.S. Government Printing Office, Washington, D.C. 20402. Include FIPS PUB number, title, SD Catalog No., and price when ordering.

AMERICAN NATIONAL STANDARDS

Published X3.38-1972 Identification of States of the United States (Including the District of Columbia) for Information Interchange - $1.50
X3.30-1971 Representation for Calendar Date and Ordinal Date for Information Interchange - $2.50
X3.31-1973 Structure for the Identification of the Counties of the United States for Information Interchange - $1.50

(The above American National Standards should be ordered from the American National Standards Institute, 1430 Broadway, New York, N.Y. 10018 at the prices indicated above.)

Proposed BSR X3.35 (X3L8/190) (73-06-18) Structure for Identification of Organizations for Information Interchange
BSR X3.43 (X3L8/177) (73-10-24) Representations of Local Time of the Day for Information Interchange
BSR X3.47 (X3L8/104) (72-02-09) Structure for the Identification of Named Populated Places and Related Entities of the States of the United States
X3L8/183 (73-03-14) Representations for U.S. Customary, SI and Other Units to be Used in Systems with Limited Character Sets
X3L8/186 (73-08-01) Draft Technical Report, Guide for the Development, Implementation and Maintenance of Standards for the Representation of Computer Processed Data Elements
X3L8/188 (73-10-23) Representations of Universal Time, Local Time Differentials, and United States Time Zone References for Information Interchange

(A copy of the above proposed American National Standards may be obtained without cost from the Computers and Business Equipment Manufacturers Association (CBEMA), 1828 L Street, Washington, D.C. 20036.)

INTERNATIONAL STANDARDS

R2014 Writing of Calendar Dates in All Numeric Form - $2.80
R2015 Numbering of Weeks - $3.15
ISO 2711 Representation of Ordinal Dates - $2.80
ISO 2955 Representations for SI Units and other Units to be Used in Systems with Limited Character Sets

The above may be obtained from the American National Standards Institute, 1430 Broadway, New York, N.Y. 10018.

X3L8/186
73-08-01

AMERICAN NATIONAL STANDARDS INSTITUTE

PROPOSED TECHNICAL REPORT

GUIDE FOR THE DEVELOPMENT, IMPLEMENTATION AND MAINTENANCE OF STANDARDS FOR THE REPRESENTATION OF COMPUTER PROCESSED DATA ELEMENTS

This GUIDE was prepared by the X3L8l Task Group (Data Standardization Criteria) of the X3 Sectional Committee (Computers and Information Processing) of the American National Standards Institute (ANSI) to assist ANSI groups and others in developing, using and maintaining standard representations of computer processed data elements. This GUIDE contains considerations intended to aid in the design and development of voluntarily adopted uniform practices and standards. The GUIDE is not itself a standard nor is any part of it to be considered mandatory or binding on any individual or organization. A definition of "data element" may be found in Appendix A.

The GUIDE is aimed at both administrative and technical levels of decision-makers. Both groups will require answers at some stage in their involvement with information processing to such questions about coding, codes and forms of data representation as, What are the current standards and where can I find out about them? Who has standardized common data related to my field of interest? How does one engage in data standardization? How can one develop optimum codes and other representations of data? This GUIDE offers some hints and special recommendations along these lines.

It should be pointed out that this report addresses alpha-numeric data only. It does not address, for example, geometric entity data. The material is organized into three main topical areas covering the background and concept of data standardization, codes and coding, and the current organization and activities of data standardization. This GUIDE is intended to be comprehensive while being "modular" in design to permit independent reference to individual sections as required.

None of the material in this document should be considered final. Much of the content is opinion. Some is controversial. Not even all the members of the X3L8 Subcommittee agree completely on all points. Nevertheless, this document does represent the current state of the art according to current authorities on the subject. Since the content is evolutionary, details of the readers' experiences and recommendations for improvement to this work will be appreciated.

Credit must be given to various people who have made the GUIDE possible. The major tasks of writing the primary text and supporting its completion have been borne by Harry S. White, Jr. of the National Bureau of Standards. Thanks are also due to Thornton J. Parker III of the Office of Management and Budget of the Office of the President for the valuable input provided the Task Group in the form of a number of documents and informally expressed views and insights. Additional credit is given to the Bell Telephone Laboratories, Inc., particularly to Arthur J. Wright and Lou Sonntag who furnished a major contribution in providing the Task Group with material contained in their document "Common Language Coding Guide." Merle G. Rocke of Caterpillar Tractor Co. has provided invaluable assistance to the X3L8l Chairman both by making available substantial portions of his document "Data Codification Principles and Methods" and by volunteering much time, effort, and interest in preparing the GUIDE. Thanks are also forthcoming to the American Institute of Physics for the time made available to the X3L8l Chairman, Arthur R. Blum, which has made possible the completion of the writing and editing chores for the final version of this document.

CONTENTS

537-860 O - 74 - 24

Section

Page

SECTION 1 - BACKGROUND

In recent years there has been an enormous expansion in the collection, processing and exchange of data required for governmental, industrial, commercial, scientific and computer processed technical information. Such information is essential to the life and operation of modern society.

To serve the vital need for improved communication of information within our society, further technological advances in computers, communications and allied fields have continued to make possible an increasingly broader integration of data systems and ever greater aggregation and exchange of data among them. These advances have achieved both substantial cost reductions and important improvements throughout the spectrum of data systems and services.

However, the full effect of these advances will not be realized until the data processing and management communities reach a uniform understanding about the common information units and their expression or representation in data systems. This can only be done by developing and applying appropriate standards.

The need for data standards is not new, but it is ever more pressing. The expansion of data needs within small and intermediate as well as large-scale computer systems--and the prospects of even more sophisticated electronic tools--re-emphasizes the need for data standards. Future applications dictate that action be taken to hasten their development and use. The GUIDE recognizes that standardization must never be undertaken for its own sake, but to promote greater efficiencies and economy, including those cases where the benefits derived are not always self-evident. The GUIDE also recognizes that the community of data users has already grown too large to expect a resolution of all problems. This GUIDE is therefore offered as a means by which those concerned with the development and implementation of data systems can gain an appreciation of the need for more uniform practices and standards and can concentrate on the areas of greatest importance and potential benefit.

SECTION 2 - CONCEPT

Data and information are fundamental to human communication.(*) No communication can occur without data or information having been transferred and recognized (or at least conveyed and accepted).

For the purposes of this GUIDE we will call the specific unit of information a data element. In information processing and exchange the data element is used to identify the intended field in a record. The data element thereby forms the fundamental building block out of which all information structures (records, files, and data bases) are made.

The increasing use of sophisticated and rapid methods of handling data has intensified the problem of dealing with meanings. Computerized society is not compatible with ambiguities of language or erroneous numbers. Man can no longer afford to apply ambiguous words or symbols to describe or to fill in the records used in daily life. Woeful is the life of the person -- whether a customer, employee, employer, or taxpayer -- who tolerates ambiguous meanings or entertains erroneous data values in such computerized records as credit cards, personnel files, purchase orders, tax forms, airline tickets, or utility bills. Considerable effort has been made in recent years to bring together the large-scale users of information in government and industry to achieve greater uniformity through clearer understanding (definition) and to facilitate processing of common data through standard data representations.

Standardization of the basic units of data requires that variations in the data being interchanged be eliminated or at least minimized wherever possible. It is generally possible at more or less expense to translate identical or very similar data of one system to the format or arrangement of a second system, despite differences between the names, codes or other representations of the data elements used in the two systems. This is often the case in the trivial instance where two or more data element names refer to the same element, e.g. "Purchaser" or "Name of Customer". But where the same name is assigned to different elements or informational units translation may not be possible (for example, two fields may be called "Status", the first requiring marital condition in a personnel record and the other field querying the condition of a body at a hospital emergency admissions office).

Therefore, the basic unit of information, the data element, has a name which serves to identify it and to distinguish it from other data elements. Typical examples of the names of data elements, which serve to identify the meanings attached to the data fields in records, are "Applicant's Name," "Sex," "Date of Birth," "Place of Birth," "Number of Dependents," and "Social Security Number."

---

*To keep this document as informal as possible such formal distinctions as that between data and information are taken relatively laxly throughout the text. Where a strong contrast is needed, it is assumed that information is the holistic meaning, possibly derived from the assembly, analysis or synthesis of the data into a previously unknown, unpredicted and meaningful form. Data by contrast provide the atomic or molecular fragments to be connected.

The data element, i.e. the meaning of the data field, can usually be identified by the name of the field. But it remains an open or empty or "unsatisfied" meaning until a specific value is applied to the field. For example, an "Applicant's Name" (data element or data element name) could be "Jones, John Adam" (data item). The meaning of the field is unassigned until the specific value (called data item) is given to it. The data items may be names as in the case above, or in other forms such as abbreviations (of variable length), codes (fixed length), or quantities. His "Sex" would be "Male" which could be coded "M". His "Date of Birth" could be "February 21, 1969" which could be represented "690221." His "Place of Birth" would be "Springfield, Illinois" which could be coded as "1782202". "Number of Dependents" would be "3". Nevertheless, the data standardizer is more concerned with the meaning of a particular field than with the particular names which are applied to it (although uniformity here is very important). For it is the meaning which must be unique and unambiguous and which requires a specific and precise representation.[1]

The gist of the problem of data standardization is that before meaningful data interchange can occur, the sender and receiver concerned must understand the identification and definition of the data elements and data items involved. The codes used in the interchange must be identified and defined. The position or location of the data elements in the record or form must be described.

Mutual understanding and agreement by the parties who interchange data form the basis of data standardization. But success of such standardization depends upon the comprehensiveness of the agreement. The greater the agreement on the national and international levels and the more inclusive the forms of representation, i.e. the names of elements, the codes, the coding methods, and the record forms that are standardized, the more effective will be the efforts in data standardization.

---

[1] It should be noted that the terms "data element" and "data item" are understood as identical with the COBOL data description terms "data item" and "data value", respectively, although the area of application of data standardization concepts is much wider.

SECTION 3

DATA CHARACTERISTICS

SECTION 3 - DATA CHARACTERISTICS

3.1 Introduction. This section treats the relationship between the data processed in an information system and the entities, events and properties which the data represent.

Data standardization is essentially concerned with the representation of data elements. It is not things and their attributes which are of primary concern, nor are the data contents, syntactic structures and applications or machine operations necessarily important in themselves. Although the data standardizer deals with the objects of the everyday world and with many formatting and organizational problems of systems, he sees these two realms from the viewpoint of the representational function of the data.

An essential task of the data standardizer is to obtain agreement on a method of representing data elements (e.g. natural language names, abbreviated names, codes, or even such special use indicators as the name or surrogates of the name of the data field on a record). However, there must be control of the relation between the world and the machine-sensible records and files. The data standardizer seeks to control this relation by working with the formatted data in question, by probing the data characteristics, by testing the suitability of the designation: names, codes, numeric data; perhaps also by structuring and otherwise organizing the designations. He works mostly with data already in records, data expressed in terms of controlled and uncontrolled vocabularies, code and term sets.

Our task can be summarized in two questions: How can we manage and understand things of our world in terms of data characteristics? How can the data characteristics be defined, represented, and then formatted in data transmissions?

Practical answers to both questions can only be found in the everyday work and long-range accomplishments of the people engaged in data standardization. However, to facilitate their work, the present section on Data Characteristics will provide a tentative reply to the first question, while the remainder of the document will give some hints as to how to answer the second question.

3.2 Viewpoints, Things and Classes. The world around us is made up of natural and manmade, physical and conceptual, as well as hypothetical or imagined entities. These entities have their own properties and can be related to one another. All of these individual things and notions can be known and designated, and therefore provide potential data to be recorded. For example, the political subdivisions we call countries, or states, or cities, or the physical objects that are arranged and labeled in a warehouse can be considered as data. The characteristics of these data correspond to the characteristics of the original things or notions or attributes.

Another example of such a relation might be found in the personnel record of a person where the data element (the meaning of the data field) is "State of Birth"; here the data item or value "Massachusetts", as one unique and unambiguous choice among the other allowable items for States, will satisfy the requirement.

Several processes must take place before a thing or notion can be meaningfully represented in a data processing system, and particularly before data records can be interchanged between systems.

3.2.1 Acceptance. There must be a common acknowledgement of the existence of thing, notion, or characteristics within the given context.

For example, we must agree that Maryland and Virginia exist and are "States of the United States" before they can be coded within the code set for states. Before attempting to communicate about a thing, it is essential to know that a thing is, to describe what it is, and to standardize the communication. Consequently, the existence of the object must be accepted by us before we can, for example begin to disagree about whether the boundary line between the two states mentioned is the high or the low water mark of the Potomac River.

3.2.2 Common Viewpoint. People must perceive an object and relate it to existing schemes or to familiar subject fields. However, to share the object with others there must be a mutual agreement about what it is, so that a common viewpoint concerning it may be established.

It is at this point that the boundary line between the two states becomes important, because uniform specification demands that the things be perceived and described in the same manner. Where the common viewpoint is lacking, specification and standardization may be impossible. For example, a common viewpoint is required to decide whether the Canal Zone, Puerto Rico, Guam, etc. are assigned to either the class "Countries of the World" or the class "States of the United States" or neither.

The standardization process cannot proceed until one has achieved a common viewpoint on whether the object in question is specified as an individual (a thing) or a class. For instance "The United States of America" can be the real individual that belongs to "Countries of the World". From another viewpoint it can be the class which contains "States of the United States: whose members are Alabama, Alaska, etc.

The issue of individual versus class overlaps the problem of uniqueness in the case of common names. For instance, "John Jones" is the name of twelve different individuals in a local telephone directory. Not knowing additional attributes, such as address, the problem of deciding upon uniqueness might be resolved by elimination. This could mean telephoning until the correct individual is located among the other members of the class "John Jones." But even when the individual is identified, true uniqueness is not established until a common viewpoint is determined as to which "John Jones" is meant -- John Jones, employee? John Jones, father? Data standardization must resolve all questions concerning class membership. This requirement extends to classes, as in an industrial classification, and to individuals, as in a warehouse inventory.

3.2.3 Terms. A person cannot know and control things or notions unless he can designate them and use the facilities of language to convey his designation to others. Furthermore, the terms or symbols in the designation must also be understood. When we have the name, indicate, identify, describe and quantify (that is, to tell how much, how many, how large, or how long in time), we find ourselves involved in the complex field of semantics.

Data characteristics can be based on physical characteristics. Meaningful data may be derived directly from physical signs. For instance, analog instrument readings convert physical variations into a variety of representations and measurements. There are many other familiar representations of physical states which may be recorded, stored or displayed in a variety of states and dynamic forms, such as motion pictures, photographs, or drawings.

Data characteristics can also be based on the data derived from such primary readings. One form of machine sensing can be translated into another, e.g. analog readings can be converted to a binary form, or on-line data can be subjected to various forms of interpretation. Such data are meaningful and may be denoted by terms. For example, signs of certain types can be interpreted, named and quantified by word symbols and numbers. The data characteristics of all such signs exhibit clear cut digitally codable patterns.

Typically, however, data standardization is concerned with conventional symbols, particularly those which express word meanings and discrete quantities. These symbols are commonly applied to data structures, i.e. to the data items and elements, and to the logical records and files organized into data bases and systems.

In an area as complex as data standardization, problems of meaning that present communication stumbling blocks arise quite often. An example of a difficulty that could occur when assembling data items is the "language barrier." Different language designations for items of the element "Countries of the World" cannot cross specific language lines without causing confusion or total unintelligibility:

English Language Term	Native Language Term	
France	France	
Germany	Deutschland	
China	Chung Kuo Zhongguo	Depending upon transcription scheme
India	Bhārata	

Influences stemming from general record keeping and data processing contexts, and especially from specialized "closed" systems tend to make the individual and associated terms more rigid, structured and controlled than would be the case for natural language vocabularies. The clearest example of structured vocabularies may be found in general classification systems, where all terms in the system are ordered. The ordering of terms can be discovered whether in its full or conventional name form or in its condensed representation. Terms can be ordered intrinsically or extrinsically.

For example, data terms can have an intrinsic order given by an ordinal numbering system such as a catalog number used as the unique identifier for stocking and ordering purposes, or a license plate, or a street address number. Extrinsic order can be given to terms in a classified arrangement where subject terms are ordered alphabetically according to their ranking within the scheme.

Quite often, unless code numbers are applied to the terms, there can be no intuitive way of knowing the interrelations of terms just by examination of the terms themselves. Most terms encountered in experience, such as words, names of persons, places and things are unordered.

Ordering interrelates individuals. But regardless of whether the individual member (or members) of a family of terms is taken singularly or is related to the others, it is usually essential to know whether the term itself relates to:

(1) a single unique thing;

(2) a class or group of things that are accepted as a unity, a composite whole, or a manifold;

(3) many things.

Therefore, by understanding what the term relates to, the data characteristics of the term become controllable. Control of the term makes it possible in turn to control the objects and motions referred to, as well as messages that contain the term. Such control extends to the term in its original mode of presentation, say in its full name form, or in alternate modes or representation, as in coded or abbreviated forms.

The criteria related to how the data standardizer copes with terms can be crucial. For example, to control the language performance of the terminology used, attention must be given to the denotative precision or expressiveness (as specified in the term definitions), uniqueness (or seen from the other side, zero ambiguity), compactness, and cost in development and implementation of the whole set of terms.

3.2.4 Condensed Representation. Efficiency and economic considerations in data processing require that data elements be represented in a condensed and accurate symbolic form.

The data must be controlled in such a way that the objects and notions are effectively designated and identified, and their meaning is faithfully conveyed throughout the process of representation. The terms or names of the data contents must be abbreviated or coded according to specific rules, but cannot lose any of their precision or uniqueness ... for human or machine processors ... in any of their condensed forms. Thus, we expect that the "State of the United States" named "California" may be abbreviated as "CA" and coded "06" with increased efficiency and economy without detriment to the performance criteria mentioned above under "Terms."

Only by working with and through the four basic processes of definitions--acceptance, common viewpoint, terms and condensed representation can data description be formalized. For instance, it is possible to standardize a code for a unique item or a unique class only if:

(1) The uniqueness of the thing or class has been established;

(2) There is acceptance of the specification, description, limits or properties of the thing or class;

(3) An accepted and unambiguous term is established for the thing or class;

(4) There is acceptance of the code as standing for the term.

3.3 Data and Data Representations

3.3.1 Fundamental Approaches to Data Standardization. The distinction between things, classes of things, and pure classes is so fundamental that it characterizes the individual approaches to data standardization.

In principle and in actual systems design, one of three methods derived from this distinction is often emphasized. Accordingly, the data tend to be treated as:

(1) a unit usually associated with its physical or at least nominal occurrence in data fields of records

(2) a class based on intrinsic or assigned relations between units which belong to the class; and

(3) classes of information which form part of a defined classification scheme.

The definition of data elements, data items and activities of data standardization can depend on which approach is adopted:

(1) The unit approach - where the fundamental meaning of the data element is identified with the unit of meaning that occurs in the particular data field of a document or data record. The association is considered so close that the same name is given to the data field, the identifier of the data field and to the contents as well as the meaning of the contents of the field. The basic unit of meaning, the data element, is duplex. It consists of a general part and a specific component (the data item). For instance, this approach maintains that there is a data element "Date of Birth" that is different from the data element "Beginning of Employment", although both elements will have common data items or values that follow the same formatting specification, e.g. the values of both may appear as "September 8, 1950" or "500908."

Standardization activities are based on the data items, so that uniform representations may be used for the most significant data elements. Consequently, although there are many data elements that apply, for example, to time or to countries, one does not attempt to standardize the data elements but concentrates instead on the methods of formatting and representing the codes for sets of data items e.g. (list of geo-political entities).

(2) The class approach - where the data element is considered as a class or category, independent of its appearance or use in any particular record context. The class is considered to be denoted by the intrinsic or assigned relations or attributes of the data items, the members of that class.

Consequently, the class is abstracted from the concrete instances of its occurrence and use (although types of use may be documented and controlled). Considered as the fundamental unit of data, the class itself is standardized: it is treated as a semantic entity and may be analyzed, defined, put into thesauri or dictionaries, and controlled. The class "Date" will, for example, be considered the data element, to be defined and allotted certain allowable names, abbreviations or code structures (perhaps formatted as 720101, or as January 1, 1972). Its uses may be documented as "Date of Purchase" or "Date of Birth", which may or may not share the same name as the data field identifier.

(3) Classification approach - where an entire subject field, perhaps the totality of human knowledge (as in a bibliographic scheme such as the Universal Decimal Classification), a library book location scheme (such as Dewey Decimal Classification) or an industrial classification (e.g. the Standard Industrial Classification) is considered as the main information unit. All particular instances of the component classes or entities then form subdivisions which are hierarchically ranked. Each class or entity may have its own code value, but one which is representative of its relative position within the total scheme. Data can them be either subordinated to a class (under this "subject heading") or comprise the class itself (subject display).

Standardization for this approach requires the specification and structuring of the main scheme and the precoordination of terms for the body of knowledge. It also requires rules for expanding subordinate segments of the scheme, and a method for coding classes and perhaps special attributes of classes (as in facetted schemes).

Common to all three approaches is the basic reliance upon the value of the unit of meaning (found primarily in a data processing context such as a data field), the value which we have called the data item or data variable above. The data item is always the expression of what is selected as the unit of meaning (or that which is considered fundamental) which is listed as one of the items in a code or other representational structure.

For the sake of simplicity and to preserve the elementary open relation which binds entities and attributes to a data field only by a non-committal linguistic tie, we will adhere to the unit approach through these Guidelines.

3.3.2 Data Elements. The data element is the meaning of the data field, and the data record which contains this field will be only as accurate as the data which it contains.

To ensure optimum accuracy, data handling systems are carefully designed to preserve the precision of the data characteristics throughout all operations. Trained specialists are employed to give particular attention to the design and organization of forms, reports, files and data formats. Words and symbols used in the procedures and systems descriptions are carefully chosen so that

effective communications are facilitated. In most instances, forms are so designed and partitioned that each element on the form can be completely described in detail. Instructions for completing or filling out the form are devised to permit the recorder to provide the needed information both accurately and without ambiguity. The individual units of information which are found in these boxes or fields on forms, records, and files have open meanings which require certain data for the meanings to be satisfied. These units and meanings are the data elements.

A data element is a unit of meaning made up of two parts, a general component which designates the information required (something previously unknown and meaningful to the recipient), and a specific part which supplies the data required, i.e. that which when recorded indicates a particular fact, condition, qualification, or measurement. The specific part stated as a value or the representation of a term is called the data item. Data items can therefore be expressed as names, abbreviations (including name truncation), codes, or numeric values. For example, the specific values associated with the general component "Color of Dress" can be expressed as a name "Blue," abbreviation "BL," or code "12." Alternatively, one could also use an instrument to measure the color temperature and express the result quantitatively, such "5500° K" (Kelvin).

3.3.2.1 Complex Data Elements (Data Chains). The meaning of a data field is usually simple. I.e., the data element or the data element name connotes a singular object or notion, such as "Color of Dress" connotes "Blue" or "State of Birth" connotes "California." The basic meaning requires only one thing or notion to satisfy its unique intent.

On the other hand, some data elements are complex. Their total meaning requires a chain of secondary meanings and, as a result, a composite group of data items to be entered into the data field to fulfill their primary meaning.

For example, the data element named "Mailing Address" may require data items which express the notions of name, street number, street name, apartment number, room number, building number, organization, city, state, county, and ZIP Code. Similarly, the specific representations associated with the data element named "Birth Date" convey the notions of year of birth, month of birth and day of month of birth.

3.3.2.2 Data Elements used for Matrices. Related to the complex data element, although more highly structured by extrinsic ordering (see 3.2.3), are the data elements used in matrices or tables or arrays of data elements (see especially 4.4.1.1). The name of the matrix may be considered a complex data element which refers to or intends the subordinate data elements that form the headings of the rows and columns.

The subordinate elements are organized in arrays that are peculiar to th type of matrix at hand, for example:

Educational Level of ADP Management and Supervisory Personnel

Educational Level (V_1)	Management Category (V_2)		
	Data Processing	Systems Analysis	Programming
College Degree	49.1%	50.7%	34.8%
Some College	38.2%	38.6%	47.6%
High School Graduate	11.5%	9.6%	16.3%
None of the Above	.4%	.1%	.2%
No Response	.8%	1.0%	1.1%

In this case, the name of the data element and its specific value can be identified in the following way:

> The percentage of ADP management and supervisory personnel with educational level of (V_1) in management category of (V_2).

All the subordinate data elements in the matrix are jointly identifiable by that single complex name, and when values are supplied in the columns for the variables V_1 and V_2, each specific value of the matrix can be explicitly identified. For example:

> The percentage of ADP management and supervisory personnel with an educational level of (V_1 = College degree) in management category of (V_2 = system analysis) is 50.7%.

3.3.2.3 Primary Data Elements and Attribute Data Elements. The data elements within a data system collectively make up larger units of data called records. Within the records (whether they be forms, reports, or logical computer records) we may find that at least one data element, which we will call the primary data element, stands out, and has a certain primacy and logical privilege over the others.

A primary data element is the element which serves as a unique meaning "key" to distinguish a particular entity from others.

The element is therefore used as an identifier for the entity or entities and is qualified by the other data elements in the record. In many such cases, the primary data element is at the same time a record key or provides a sort key in machine sensible records. For example, in a personnel record which contains information concerning a particular individual within the organization, the following data elements may be used: Social Security Account Number, Name, Date of Birth, Personnel Grade, Salary, Job Title, Organization Assignment, and Home Mailing Address.

If the organization is small, the name of the individual usually provides the unique identifier for him and serves as the primary data element which is qualified by the other data elements. In a larger organization where several persons may have identical names, the Social Security Number plus the name may furnish the primary data element or key to identify the individual. If necessary, two or more data elements may be used collectively to provide uniqueness, and each may be regarded as primary. The remaining data elements in the record then simply qualify or further describe the entity (whether it be a person, place, thing, or notion) which has been identified by the primary data element(s). Qualifying data elements are called attribute data elements.

In the example above, Date of Birth, Personnel Grade, Salary, Job Title, Organization Assignment, and Home Mailing Address are attribute data elements.

Attribute data elements can be chained together or "nested." For, in some cases, attributed data elements may have qualities which also are identified in the record. In these instances an attribute data element may be qualified by another attribute data element. In a personnel record like the one mentioned above, we could find attribute data elements named "Spouse's Name" and "Spouse's Birth Date;" the data element "Spouse's Birth Date" is an attribute data element of the attribute data element named "Spouse's Name."

Depending upon the structure of a particular record or file, what might be a primary data element in one record may be an attribute data element in another record. Likewise, an attribute data element in a given record could in another record be a primary data element.

3.3.3 Data Representations Other Than Codes. It was mentioned earlier that both the general and specific parts of the meaning of data, although especially the latter, can be represented in such various forms as names, abbreviations, codes, and quantitative (numeric) expressions. Blue as a specific value associated with the data element named "Color of Dress" can be represented as a name "Blue," as an abbreviation "BL," as a code "12" or can be measured and expressed quantitatively as "5500 K." Similarly, the general portion of the data element can be represented as a name, "Color of Dress," as an abbreviation "CLR-OF-DRESS" or as a code "COD." Each of these forms of representation, names, abbreviations, and general quantitative expressions have characteristics of their own which need and are given further explanation in this Section. Data codes are treated as fixed length representations and are discussed independently in greater detail in Section 4.

3.3.3.1 Names. Natural language terms are the most common designators of data structures. As it was pointed out in Section 3.2.3 -- Terms, the terms used for logical data structures, i.e. data items and elements, records, files, forms, and whole data bases, are basically built up of meanings. These meanings are indicated by a variety of representational expressions, such as names, abbreviations, special symbols, and codes. But names are generally the most suitable universal and familiar forms for representing the meanings of the data elements and, where non-quantitative, their data items. The principal function of names is for the identification of objects, qualities, quantities, and notions, for the purpose of aiding human recognition and manipulation of the things and ideas encountered in experience.

But it must always be remembered that any specialized use of natural language, such as for identifying the meaning of a data field or its content, is governed by the same laws and constraints as any other use of natural language. And natural language is notorious for its imprecision in conveying meanings uniquely and without ambiguity. For example, quite often the only clue to the exact meaning of a name is provided by the format, context or overall situation within which the natural language name appears. For instance, the data element named "Grade" is used in the following records:

School Personnel Record:

Name	Grade
John Smith	4

Employment Personnel Record:

Name	Grade
John Smith	GS-12

College Transcript Record:

Name	Course	Grade
John Smith	Biology 251	B

However, the meaning of "Grade" is different in each case and requires a distinctive full name which in some way reflects the context within which the element is used. "Grade" implies "School or Class Grade" in the first example, "Civil Service or Personnel Grade" in the second, and in the last "Course Grade". In the interchange of data among various data systems or even among the components of the same system, it is necessary that the context in which the names are used be known and specified (explicitly or by default) before communications can be accomplished unambiguously.

The proper context can be established and defined in several different ways: (1) The data elements can be related to the larger context in which they appear, i.e. to the particular records or reports in which they are used or to the primary data elements which they qualify; (2) The data element name can be expanded or otherwise modified to include essential words which establish the proper context; or (3) The definition or explanation of the data element name can mention in which context the data element is used. System descriptions and documentation often employ combinations of these techniques to describe data elements.

In effect, a data element may have more than one name by which it is identified. The same is true for the names of specific data items associated with a data element. The names may be the actual names used internally on the reports and forms in a particular system, or even the field labels or name tags by which the same element is identified. Alternatively, they may be more universally understandable names, perhaps full explicit names with appropriate descriptions, which should be used to communicate the data element outside the particular system. Both the local (or internal) names and the interchange name (the explicit form which indicates the full context) must be identified in any complete data system description.

Both natural language and the terms which name and describe data characteristics reflect the real and conceptual world which contains all of the things and notions of human experience. But it would be a mistake to assume that natural language is or necessarily should be identical with data terms. The language of data terms is in one sense not natural. It is called into being in order to identify only those objects and ideas which find their way into records and other data structures. However, it also borrows heavily from natural language for designators or descriptors to name, identify, classify and otherwise describe much of its contents. As such, it is a subject of natural language. But the data terms also draw upon a variety of formalized representations such as highly structured code sets for rigorously unambivalent connotation of their meanings.

However, even if the contents are different, the processes used in the language of data terms coincide completely with those of natural language when selecting and assigning names for data structures. Therefore, the naming of such data structures as data items and elements involves all the linguistic description and prescription that would apply to naming anything else. Names are called nouns in grammar. Hence, the grammatical conventions that apply to nouns and related word formations also apply to the data item, element, record designator or descriptors.

The assignment of names for data structures must be based upon a variety of considerations. These include grammatical specifications. These grammatical specifications for nouns as parts of speech, and the various criteria of language performance mentioned at the end of Section 3.2.3: 1) denotative precision or expressiveness of the noun or noun embedded syntactic formations; 2) uniqueness of reference, that is, does the data name or noun refer to: (a) a simple unique thing, (b) a class or group of things accepted as a unit, a composite whole, or a manifold, or many things; 3) compactness of expression; and 4) cost in development, implementation and maintenance of the whole vocabulary or set of names. For additional guidance in developing acceptable names, refer to ISO document R704-1968, "Naming Principles."

Grammar

The subject of the grammar of nouns and related formations is far too extensive for exhaustive treatment here. But a brief review of its cogency in the naming of data structures is apt and, it is hoped, will serve to stimulate further inquiry on the part of the reader.

The number of things and notions that people see, even from a common viewpoint, is greater than the number of nouns that people use to designate them. One result of this was shown above in the use of the noun "Grade." A noun can have more than one meaning: each such noun may be a name for two or more things. A second result of the lack of available nouns can be that some names are not single words. A name that is not a single word is not strictly speaking a noun, but rather a syntactic formation where the noun is embedded as a nucleus in a name "cluster." The noun is the essential element in the cluster. For example, the noun nucleus in the data element "State of Birth" is the noun "State", where the other words are modifiers of the nuclear or principal noun.

537-860 O - 74 - 25

Therefore a noun cluster is a grammatical construction which contains a noun as its nucleus, preceded and/or followed by modifiers of the nuclear noun. Nouns may appear singly or as one of several words in such nuclei, and are characterized by their singular or plural forms (usually ending in -s or -es, although some nouns have irregular plurals).

Nouns can be modified by various modifiers or adjectival units that consist of a single word or one or more groups of words. A variety of modifiers occur, such as determiners (of uniqueness or possession), as the articles the, a, your, numerals, such as "First Position Held" or "Choice One"; adjectives such as "Principal Function;" noun adjuncts, as "Data Name;" or phrases as in "State of Birth," or "Status at Time of Resignation."

A noun can be the name of two or more things in two ways. First, there is the way discussed above for the element "Grade," and then there is another, still more far-reaching manner: Two things can have the same name because people recognize that both things are in some sense the same.

If different things are not the same, each is unique, and if it has to be named, generally deserves a proper noun to identify it. On the other hand, if we recognize sameness, and find that the same name can be applied to two more things, we are dealing with a class. The name applied to the class or to each member of this group is a class or common noun. The things or notions that are named by class nouns can be counted: if the class is void, the number of members is zero; if it is a singleton, the number is one. If there are more members than one, a variety of grammatical number words can generally be attached to the member nouns, including ordinal numbers, indefinite articles, etc. Nouns can also denote a multiplicity of members in the case of mass or collective nouns such as the word "Carbon" when describing the composition of diamond, coal and graphite. The same word can be a class noun in the item "Carbon" for data element "Office Supplies." The mass noun never requires an article.

The composite noun is a syntactic formation which includes the nuclear noun cluster. For example, a data processing operation might be called "Personnel Data Throughput" or "A Sort by Name;" The file may be "Salesforce by Major City."

The attributive elements include single words (Personnel Data) and prepositional phrases (by Major City). The formations which cluster about the nuclear or principal noun can become quite complex. Various grammatical forms can adhere to these clusters, e.g. "Current Awareness Alerting Service", where there is a composite adjectival modifier which contains a verbal noun (alerting), all of which are attributive to the nuclear noun "Service." Possessive nouns can be considered under this heading, as in the element "Vendor's Name."

The composition of name parts presupposes a conciseness and compactness of expression. Precision, clarity and familiarity of the words used in names cannot be compromised by the need for compression, and some degree of optimization may be required.

The nouns used for naming entities at various levels in the data structure are not absolute, and can often be used at other levels. For example, "Virginia" may be the name of the data item for the data element "State of Residence." It becomes an important noun modifier in the data element name "Population of the State of Virginia." The question of level assumes great importance in the hierarchical ranking of names, as appears in a classification system. The effort needed to organize the name structures according to the levels required by class distinctions then becomes a significant cost parameter.

The section on names would not be complete without mentioning a few other cost factors involved in the overall process of naming data structures. Development costs as well as operating costs can apply to:

data collection - entity search, naming, definition, and preparation for encoding;

name control - the compilation and implementation of vocabularies in the form of dictionary entries, lists, thesauri, classification schemes;

maintenance - updating procedures, organizational assignments of term and coding control where the centralization versus local file trade-offs are vital; providing access to file contents possibly through publication, display terminals, etc.

Name Definition

A central issue in data standardization is the meaning of the data terms rather than the word forms and word syntax. As a result, the definition of names is of major importance. Improper definition can seriously impede data interchange.

The cost of definition can be very high. But if definition is not performed from the most general yet most common point of view, data interchange may still not be possible. Certain data systems which have developed highly standardized defined vocabularies in unique controlled environments may not be able to converse with systems in different environments. Although term definition may be present, a universal viewpoint related to the names and their definitions may be lacking. For example, both systems in two hypothetical different environments may use the same code set and format for "Date," perhaps expressed as "730325." Yet "Shipping Date" from a military point of embarcation will not have the same sense as "Shipping Date" for the local delivery of a small commercial parcel. A contractor who deals with both environments may find that there cannot be a universal definition which accommodates both meanings. Two definitions may be required.

3.3.3.2 Abbreviations. An abbreviation is a shortened form of a word, term, or phrase. Abbreviations improve the communication process by presenting information to be read by humans quickly, accurately and with ease. The abbreviation saves space and time, and it provides a convenient, compact way of reducing long and complicated words or phrases that may often be repeated.

The names of data structures, particularly the data elements and data items, frequently lend themselves well to abbreviations. Nevertheless, there is no widespread standard method of abbreviation. Among the styles and forms of abbreviations, there are two tendencies toward commonality. First, individual disciplines and organizations produce lists of abbreviations that become authoritative for their industry or special field of interest, such as the list used in the publications of the American Chemical Society. The second movement is to establish rules and algorithms for the generation of uniform abbreviations. An example of this technique may be found in the American National Standard for the Abbreviation of Titles of Periodicals (ANSI Z39.5-1969).

The difference between abbreviation and other forms of coding is not self-evident. Abbreviations are generally developed for human handling, since codes are more suited for such machine applications as on computers, card and paper punches and similar keyboarding devices, as well as on communication machines. Nevertheless, many of the same basic criteria are applicable to both these forms of representation and to the methods of deriving them.[1]

(1) Each word in the name should be compressed to require as little keyboarding time and storage space as possible.

(2) There should be no loss of discrimination and uniqueness between the original name and the compressed representation.

(3) The compressed forms should be at least as readily recognizable, learned and recalled by humans, and as easily transmitted without error as the original names.

(4) To retain optimum discrimination the compressed form should be mnemonically similar to the original name.

(5) Whenever possible, the abbreviated form should be capable of being systematically transformed back into the original name when desired.

(6) Whenever possible the abbreviated words should sort in the same alphabetic order as the original name.

These requirements are basic in the sense that at least two must be used in any efficient abbreviation scheme or code structure, but are ideal in the sense that all can rarely be applied at the same time.

At the risk of arbitrariness, the abbreviation may be generalized from its common appearance in text, and defined as a mnemonic code with a variable length. When existing or constructed abbreviations have a minimal number of characters, will alphabetize in a desired sequence, and are easily manipulated as well as mnemonic, then the abbreviation set is identical to the code.

Thus defined, several techniques for deriving abbreviations are commonly used:

a. contraction - the shortening of a word, syllable or word group by systematic omission of an internal letter or letters. For example, "abbrvtn" for "abbreviation".

---

[1] cf. Charles P. Bourne, Methods of Information Handling, New York, 1963, p. 46.

b. truncation - the shortening of words by the omission of letters at either end. For example, right end truncation retains the proper number of characters at the left end and deletes all the remainder up to the end of the word, e.g. "Wash" for "Washington." Left truncation drops letters from the left end, e.g. in the list:

"h, A.R.	for	"Smith, A.R.
h, Dick		Smith, Dick
h, Paul		Smith, Paul
m, Thomas"		Smith, Thomas"

c. the formation of acronyms - forming words from the initial letter or letters of each of the successive parts or major parts of a compound name. For instance, RADAR for Radio Detection and Ranging.

In the absence of any universal authority for abbreviations one can recommend the use of the abbreviation list and individual entries in Webster's New Unabridged International Dictionary. However, care must be exercised. Only unique and unambiguous abbreviated word forms should be assigned to the data terms in question. The elimination of letters within words or phrases tends to produce undistinguishable character clusters, e.g., "DA" may be an abbreviation for "Day," "District Attorney" or "Department of the Army." "No" may be an abbreviation for the chemical "Nobelium," for the direction "North," or for the word "Number."

In addition to the criteria basic to both abbreviations and other codes listed above, the following suggestions may be useful in the development of uniform abbreviations for data terms:

Abbreviate significant words in the name, allotting a consistent maximum number of words to be used and word types (e.g. articles, conjunctions, prepo-- sitions) to be dropped.

Words with a small number of characters (say, four or five) when used alone should generally not be abbreviated.

For mnemonic purposes the first letter of a name word should be present in the abbreviation.

Initial capitals or all capitals should be used.

Consistency is of major importance: either use periods at the end of all abbreviated words or omit final periods (preferred, since people often inadvertently omit them).

The same abbreviation is used for singular and plural forms of the same words.

Given a choice of deletion, consonants are more important than vowels in the abbreviation, initial letters than final.

If a conventional abbreviation already exists, it is preferable to a newly developed one, provided that it conforms to the other criteria mentioned above.

The abbreviation should be as universally understandable and recognizable to human beings as possible and not merely provide a jargon "shorthand" version of the name, for example, one should avoid giving the initial letters of a data term such as "INOC" for "Identification Number of Consignee."

In a compound name, the order of abbreviated words should follow the same sequence as the original name.

Abbreviations must be developed with consideration given to existing software constraints. For example, in a COBOL environment it is essential that non-connected words should not begin or end with a hyphen, must have at least one alphabetic character, and names used as tags must be restricted to a word-length no greater than 30 characters.

3.3.3.3 Quantitative Data. Quantitative data provide a numeric answer to such questions as "How much?" "How many?" "How large?" "How long in time?" or "How frequently?" The numerals in quantitative data represent numbers which express the limits of quantities and magnitudes. The meaning of the quantity or magnitude is a data element and is connoted by the name of the element, e.g. "Length of Runway." This meaning is satisfied by furnishing the appropriate numeral, which is the proper data item for the specific element, e.g. To satisfy "Length of Runway" one could specify "800 feet." Numerals often include a wide range of expressions, such as whole numbers, ratios, exponents, fractions, and constants.

The degree of preciseness needed within a given system determines the form of the particular quantitative representation. For example, the unit cost of certain supply items may be expressed in dollars, cents, and mills as $1.035, the sales price of these items is expressed in dollars and cents as $1.30. The inventory of a business may be expressed in dollars as $82,520 or in thousands of dollars as $82.5. On the extreme end of the economic scale are the expressions of the gross national product or national debt which are expressed in billions of dollars. Similarly, the precision of irrational numbers (numbers which cannot be exactly expressed as a ratio of two integers) will depend upon the specificity required. Pi, which is used as the symbol to denote the ratio between the diameter and circumference of a circle, may be expressed as 3.14, 3.1416, or 3.14159265... depending upon the requirements of the system.

The same value can also be expressed numerically in several different ways. For example, 3-1/2 hours can be expressed as 3.5 hours, 3 hours 30 minutes, or 210 minutes. Likewise, 100 metres can be expressed as 0.1 kilometres or 10,000 centimetres or 100,000 millimetres.

However, upon closer examination of quantitative data it is found that all have certain fundamental characteristics which can be described. These are:

(1) All quantitative data expressions have some form of numeric expression. The most common form used in human-to-human communications is that of the decimal

(base 10) system. However, numbers represented in computers are generally converted to binary form (base 2) or binary coded decimal form. Some computers have the capability of representing two or more binary coded numerals in a single computer word. This type of expression is commonly called packed numeric representation.

(2) All quantitative expressions have an expressed or implied radix point (called decimal point in decimal representations). Generally, when a quantity is expressed without a radix point, it is interpreted to be an integer (a whole number).

Some computers have a floating point capability. This capability allows a wide range of magnitudes to be represented to a given precision by means of a limited number of digits. For example, in a decimal system which uses only three digits to represent significant digits, the number 134,000,000 ($= 1.34 \times 10^8$) would appear as 1.34, 8 (where 1.34 are the significant digits and 8 is the exponent of the base 10). Likewise, 0.0134 ($= 1.34 \times 10^{-2}$) would appear as 1.34, -2, and 1.34 ($= 1.34 \times 10^0$) would appear as 1.34, 0.

(3) Normally, quantitative expressions have an expressed or implied sign (+ or -). Usually, unsigned quantities are considered to be positive. When the quantity is negative, the sign is usually expressed (explicit).

(4) Quantitative data representations which indicate measurement usually require an expressed or implied unit of measurement (e.g. dollars, meters, degrees, percent, etc. Some measurements, however, do not have a unit of measure expression (e.g. dress, hat, and shoe sizes.)

Quantitative data reflect the degree of preciseness, approximation, range, or tolerance either as part of the representation or in the definition of the data element (e.g. a ship's position may be defined to be accurate within plus (+) or minus (-) one nautical mile, or the result of a computation may be expressed as being accurate within certain maximum (+) or minimum (-) given limits.)

Quantitative representations are frequently rounded in systems applications. Rounding is a systematic way of shortening an expression (e.g. Pi expressed as 3.14159265... when rounded to four decimal places would be represented as 3.1416).

Another method of shortening is that of truncating a representation (this applies to indicative as well as quantitative expressions). Truncating simply is the act of dropping a certain number of characters (or digits) from an expression (e.g. Pi expressed as 3.14159265+ when truncated to four decimal places would appear as 3.1415. Both rounding and truncating degrade the preciseness of expression.

In the interchange of information among or between systems, it is essential that these fundamental characteristics of quantitative data be thoroughly described and understood by both the sender and receiver.

3.4 Summary (Section 3). The following is a summarization of Section 3, DATA CHARACTERISTICS. Presented is a short statement that attempts to summarize the major concepts presented in Section 3. Paragraph references correspond to those in Section 3 above.

3. Data Characteristics

3.1 Introduction. Data standardization is concerned with:

> analysis and control of the relation between the data processed within an information system and certain entities, events and properties in the world of human experience
>
> representation of these things and notions by names, codes and numeric description.

3.2 Viewpoints, Things and Classes. Characteristics of data correspond with appropriate degrees of precision to the original things, notions or attributes.

3.2.1 Acceptance - common acknowledgement of the existence of things, notions or characteristics is essential to begin data collection.

3.2.2 Common Viewpoint - to achieve standardization the objects concerned must be perceived from a common viewpoint, related to familiar subject knowledge, specified and subjected to mutual agreement concerning what it is. Definition must be made according to the uniqueness, individuality and class membership of the objects of the data.

3.2.3 Terms - objects to be standardized, that are seen from a common viewpoint, must be named, described and quantified. Typically, data standardization is concerned with conventional symbols, particularly with terms which express word meanings and discrete quantities. Data terms relate to data structures, i.e. data items, elements and logical records and files that are organized into data bases and systems.

Terms may be ordered intrinsically or extrinsically, or be unordered. To be standardized they must relate to 1) a single unique thing, or 2) a class or group of things accepted as a unity, a composite whole, or a manifold, or 3) many things.

To control the language performance of the terminology used, attention must be given to the denotative precision (accuracy) or expressiveness (definition), uniqueness, compactness, and to the cost of developing and implementing the set of terms.

3.2.4 Condensed Representation - A condensed and accurate symbolic form is needed to represent data terms. Objects and notions are effectively designated and identified and their meaning is effectively conveyed by abbreviating and coding their names. Four minimum coding requirements are given.

3.3 Data and Data Representations

3.3.1 Fundamental Approaches to Data Standardization - There are three methods of approaching data standardization based on the distinction between particulars, classes of individuals and pure classes: 1) the unit approach; 2) the class approach; 3) the classification approach.

3.3.2 Data Elements - The data element is the meaning of a data field, which may also be found to be represented in records, forms, reports, and other formatted data in files. It is composed of two parts, a general component and a specific part (the value or data item).

3.3.2.1 Complex data elements - a complex data element entails a chain of secondary meanings and, therefore, requires representation by a composite group of data items, as "Mailing Address" requires name, street number, street name.... city, state...etc.

3.3.2.2 Data elements used for matrices - The name of the matrix is used as a complex data element which refers to or intends subordinate data elements that form the headings of the rows and columns.

3.3.2.3 Primary data elements and attribute data elements - The data element used as an identifier for the given entity or entities and which is qualified by the other elements in the record is the primary data element. The element or elements which qualify it are attribute data elements.

3.3.3 Data Representations Other Than Codes - The general and specific parts of the meaning of data can be identified and represented by names, abbreviations and quantitative expressions.

3.3.3.1 Names - Names are the most universal and familiar forms for representing the meaning of data elements and items. Specifically, they provide natural language identifiers for the data counterparts of objects, qualities and notions within reports, forms and record data fields. Language ambiguities may be reduced by proper use of grammar and possibly eliminated by reference to context or with definition. Differentiation is needed where the same data structure has more than one name, just as when one name applies to more than one data structure. This problem may be resolved by definition, although there are situations where more than one definition is required.

3.3.3.2 Abbreviations (Variable length representations) - The abbreviation is a shortened form of a word, composite term or phrase considered as a variable length mnemonic code. Basic criteria are given for word compression. Several techniques are described for the derivation of abbreviations, particularly by contraction, truncation and the formation of acronyms. A number of further suggestions for the derivation, formatting and style of abbreviations are offered.

3.3.3.3 Quantitative data - Quantitative data provide a numeric answer to such questions about quantities and magnitudes as "How many?" "How large?" "How long it time?" or "How frequently?" The degree of precision and various quantitative expressions are treated as significant data characteristics.

SECTION 4

BASIC CODING METHODS

4. Basic Coding Methods

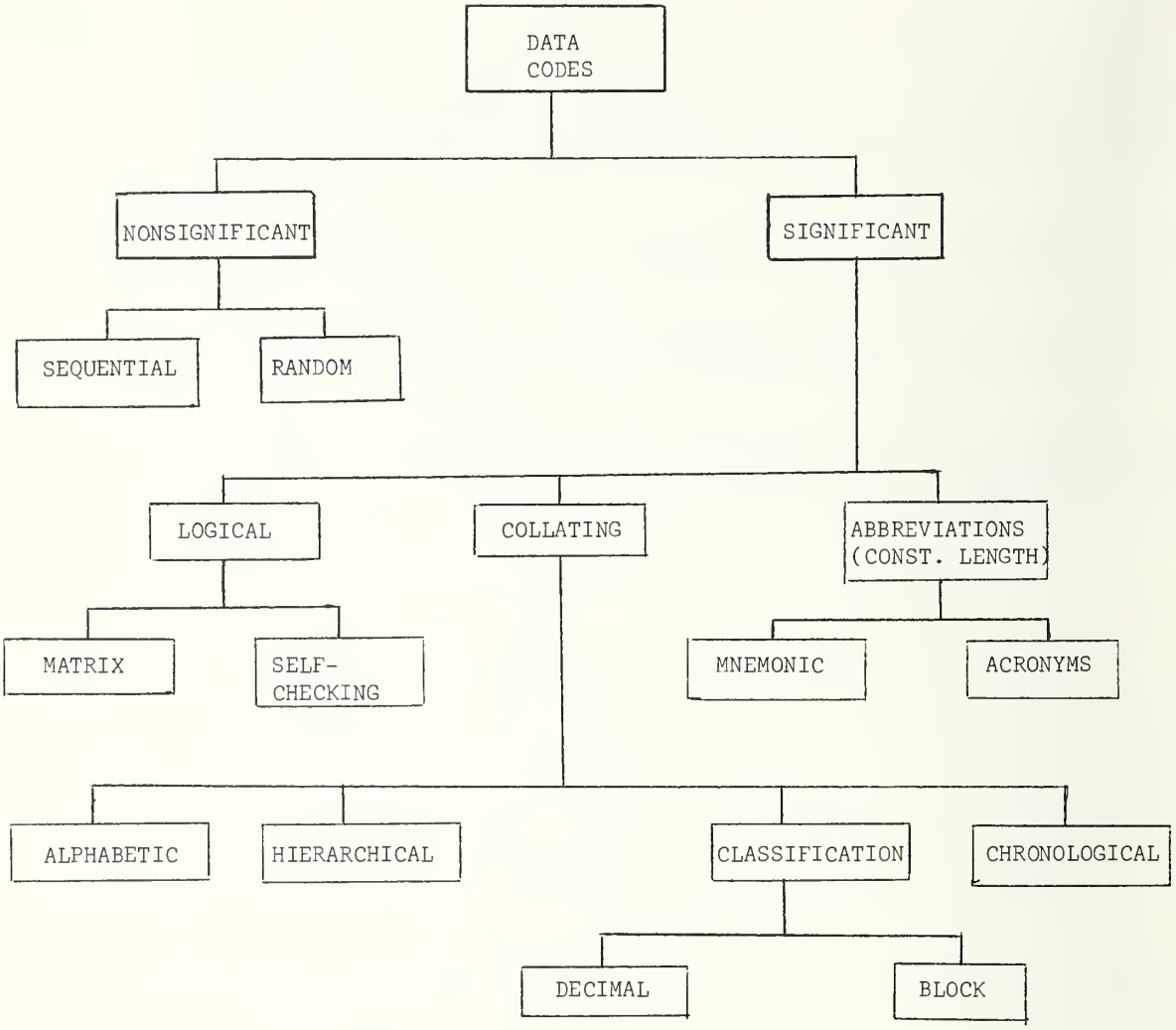
4.1 Introduction. This section provides a description of basic coding methods, including advantages and disadvantages of each method. It is intended to assist data standardization task groups in selecting the most appropriate code structure for each particular application.

A code is an ordered, shortened, fixed-length data representation. Codes are designed to provide unique identification of the data to be coded. To accomplish this, there must be only one place where an identified word or phrase can be entered in the code structure and, conversely, there must be a place in the code for everything identified. It is imperative that this "mutually exclusive" feature is built into any code structure.

The choice of code structures is fairly extensive. The following information, however, should help lead toward selection of the best method.

Section 4.2 is a chart outline of the coding methods discussed in Section 4. This set of code structures is not entirely comprehensive, but does include all the significant types. Further, these are "pure" codes -- and many data codes are actually combinations of these basic types.

For additional information on coding methods as well as indepth reports on psychological studies, etc., from which much of the content of this GUIDE was taken, refer to Appendix C, "BIBLIOGRAPHY."


DATA CODES
NONSIGNIFICANT
SIGNIFICANT
SEQUENTIAL
RANDOM
LOGICAL
COLLATING
ABBREVIATIONS (CONST. LENGTH)
MATRIX
SELF-CHECKING
MNEMONIC
ACRONYMS
ALPHABETIC
HIERARCHICAL
CLASSIFICATION
CHRONOLOGICAL
DECIMAL
BLOCK

4.3 Nonsignificant Codes. Individual values of nonsignificant codes are meaningless without some defined relationship to another entity set or sets and are assigned only to provide unique identification to the entities coded. The sequence number and the random number are the two most commonly used nonsignificant codes.

4.3.1 Sequential Code

Sequential (Serial or Tag) Number. The simplest to use and apply, the sequential method of coding is merely the arbitrary assignment of consecutive numbers (beginning with, say, "101") to a list of items as they occur, just as employee numbers might be assigned to employees as they are hired. The code value has no significance in itself but does uniquely identify the entity.

This method makes no provision for classifying groups of like items according to specific characteristics and cannot be used where such requirements exist. It is practical only for coding entity sets where the only requirement is a short, convenient, easily applied representation.

The advantage of the sequence code is its ability to code an unlimited number of items by using the fewest possible code digits. As new items occur they are simply assigned the next-higher unused number in sequence.

This number is frequently used to give a unique reference number to entities (e.g., countries) which are composed of several elements identifiable in their own right (e.g., states, cities). With proper controls it is extremely useful in many applications and usually exists as a part of other more specialized coding schemes.

4.3.2 Random Code. The term random number is frequently applied erroneously to the sequential code just described. The difference between a sequential and a random code is the number list from which the code values are assigned. The random code is drawn from a number list which is not in any detectable order or sequence. There are computer programs available to produce these random number lists. Each additional item to be coded is given the next number in the random list. This method forces the coder to look up the next number on the list because there is no logical way to predict what that next number will be when the last used number is known.

In a sequential list, if 200 were the last number assigned, the next one will be 201. The next number on a random list might be 163.

This forced look-up is supposed to reduce errors in coding, but in actual use it tends to introduce problems of control. Properly controlled sequential lists have proved less error-prone than random lists.

4.4 Significant Codes. Codes are designed to provide unique identification of the words or phrases being coded. In other words, in a coded set of entities, no two entities should be assigned the same code. If in addition to providing unique identification of entities a code is so designed to

furnish additional meaning, this type of code is called a significant code. The additional meaning supplied by the significant code can yield logical significance, collating significance, or mnemonic significance.

4.4.1 Logical Codes. Individual values of logical codes are derived in conjunction with a consistent, well defined logical rule or procedure (algorithm). Two examples are the matrix code and self-checking code.

4.4.1.1 Matrix Code. This code is based on x-y coordinate locations or longitude-latitude coordinates. It is useful in coding two component relationships. Code values can be formed by assigning the "XY" coordinate numbers or by assigning sequence numbers. (The squares in the example are numbered both ways for illustration.) A code value is merely read from the appropriate square in the table when assigning code values to an entity. When decoding, the code value is located in the matrix and appropriate XY attributes are obtained. For example:

X \ Y	1=Round		2=Square		3=Rect.		4=Oval		5=Irreg.	
1 = Round	11	(01)	12	(05)	13	(09)	14	(13)	15	(17)
2 = Square	21	(02)	22	(06)	23	(10)	24	(14)	25	(18)
3 = Hex.	31	(03)	32	(07)	33	(11)	34	(15)	35	(19)
4 = Oct.	41	(04)	42	(08)	43	(12)	44	(16)	45	(20)

(Note: Numbers in parentheses are merely the matrix location sequence numbers; the other numbers are the resulting code values.)

4.4.1.2 Self-Checking Codes. It is possible to append to a code an additional character which serves the purpose of checking the consistency or validity of the code when it is recorded and transferred from one point to another. This character, which is commonly called a check character, is derived by using some mathematical technique (algorithm) involving the characters in the base code. The check character feature when utilized provides the capability of detecting most clerical or recording errors. These errors are categorized in four types, i.e. transposition errors (1234 recorded as 1243), double transposition errors (1234 recorded as 1432), transcription errors (1234 recorded as 1235) and random errors (1234 recorded 2243) which are multiple combinations of transposition and transition errors.

Several different techniques are employed to generate the check character. Each method has its advantages and disadvantages based upon the complexity or capability of the equipment involved in the data system and the degree of

reliability essential to the particular application. For purposes of demonstrating the technique, one typical system which is prevalently used in credit card applications is described below:

Given the base code 457843, the check character is derived in the following way.

Each position of a character in the base code is given a weight (the amount by which it is multiplied to derive a product). In this example, the least significant position (rightmost position) is given a weight of 2, the next one, and so forth (alternating positions 2, 1, 2, 1, 2, 1....) until all positions are assigned weights.

4	5	7	8	4	3	(base number)
1	2	1	2	1	2	(weight)
4	10	7	16	4	6	

Each character in the base number is multiplied by its weight producing the above products.

The individual digits of these products are then added to produce a sum of the digits:

$$4 + 1 + 0 + 7 + 1 + 6 + 4 + 6 = 29 \text{ (sum)}$$

This sum is then divided by 10 which produces a quotient of 2 and a remainder of 9. (10 is referred to as the modulus, i.e. the number which is used to divide the sum of the digits to arrive at a remainder):

$$29 \div 10 = 2 \text{ plus } 9 \text{ remainder.}$$

The remainder is then subtracted from the modulus (10 in this case) to produce the check character

$$10 - 9 = 1 \text{ (check character)}$$

Thus the base number plus the check character would be

4578431

In application, the full number including the check character is recorded. The check character is then used in the following way to determine the validity or consistency of the recorded number.

Weights are assigned to the positions as before except the check character is given a weight of 1 and other positions are alternately assigned weights of 2, 1, 2....)

4	5	7	8	4	3	1	(base number plus check character)
1	2	1	2	1	2	1	(weights)
4	10	7	16	4	6	1	(products)

Products are generated as before.

Digits are added as before.

4 + 1 + 0 + 7 + 1 + 6 + 4 + 6 + 1 = 30

This sum is then divided by the modulus (10), producing a quotient of 3 and a remainder of 0.

30 ÷ 10 = 3 plus 0 remainder

Now examine the remainder. If it is zero, then the number checks. If other than zero, an error has been detected.

This particular self-checking system will detect 100% of all transcription errors, 97.8% of single transposition errors, and 90% of random errors. It will not detect double transposition errors. For additional methods, refer to the texts on error detecting and error correcting codes in Appendix C.

4.4.2 Collating Codes. Collating codes are by far the most directly useful and the most frequently used. The collating code structure is designed so that when sorted by the code number, the items represented by the codes are placed in a predetermined sequence. This sequence is frequently the sequence of the output required from the computer for optimum use by people.

4.4.2.1 Alphabetic Codes. For maximum effectiveness, alphabetic coding requires placement of all items in alphabetic sequence, then assignment of a code of ever-increasing value. Future sorts on the code put the items in the original alphabetic sequence. For example:

01 - Apples
02 - Bananas
03 - Cherries
04 - Dates

Normally, space is left between each item for future expansion. This code has some very strong points in its favor:

* Ease of sorting into desirable output format.
* Ease of maintenance.
* Accessibility to the code list without initial encoding.

Unfortunately, this code has some disadvantages that can result in problems that are extremely expensive to correct. This is especially true in large, scattered data systems where high rates of corrections or additions are necessary to maintain the list.

These disadvantages include:

* The necessity of coding the entire item list at one time to get reasonable spacing for new entries.

* Crowding that requires renumbering to maintain sequence of new entries.

* Relatively short life.

* The necessity of central control of number issues.

This code does, however, have a very useful place. Proper system design can utilize its good points and eliminate many of its shortcomings for certain applications.

4.4.2.2 Hierarchical Codes. The hierarchical code is a collating code which ranks entities or attributes by relative levels. It is very useful for many diverse applications. In its simplest expression, the hierarchical code arranges items in a predetermined sequence. The sequence may be increasing weight, length, diameter or other single attribute of the items.

As code requirements become more complex, pure hierarchical coding is seldom sufficient for large systems. New ways to create hierarchies have been developed using the basic technique in combinations with other codes. Hierarchical codes are still of great value in specialized applications or supplementary to a larger code system for indicating increasing values, organization structures or levels of data summary control.

4.4.2.3 Chronological Codes. As the name implies, a chronological code is assigned in the order of events so that each code has a higher value than the last code assigned. This is essentially the same approach as nonsignificant sequential. The difference is the attachment of time significance to the code number assignment.

4.4.2.4 Classification Codes. Classification is best described as the establishment of categories of entities, types and attributes in a way that brings like or similar items together according to predetermined relationships. A classification is by nature an ordered systematic structure.

The design of a classificatory structure must satisfy two basic requirements: (1) comprehensiveness and (2) mutual exclusiveness of its categories. Its scope must be broad enough to encompass all the items that need to be included in the various classes, and the definition of the classes must be exact enough to assure the existence of only one place for every item. Further, that place must be the same for every user of the classification. The underlying logic is simple; every question must have a unique, unambiguous binary answer: 'yes' or 'no'; 'true' or 'false'; 'present' or 'absent'; 'included' or 'excluded'; and so on.

Entities, types and attributes change continuously in a dynamic world. A viable classification system which contains them must be flexible enough to accommodate such changes. Its classes must be expandable. To be comprehensive, new and mutually exclusive classes may have to be added to the structure. Old classes may in addition have to be modified or deleted.

Classification schemes are based on the viewpoint of particular people, called upon to do certain tasks at a specific point in time. As experience grows and circumstances change, the systems too must grow and change.

537-860 O - 74 - 26

Decimal Codes. One of the most widely known classification codes is the Dewey Decimal System used primarily for indexing libraries or classifying written correspondence by subject matter. The following is a representative example:

300.	Sociology
400.	Philology
500.	Natural Science
510.	Mathematics
520.	Astronomy
530.	Physics
531.	Mechanics
531.1	Machines
531.11	Level and Balance
531.12	Wheel and Axle
531.13	Cord and Catenary
531.14	Pulley
531.141	Pulley, Compound

The decimal method of coding is designed to be used for identifying data in situations where the quantity of items to be coded cannot be limited to any specific anticipated volume. It is particularly well suited for classifying and filing abstracts of written material because it is able to handle an infinite number of items as they are added to any given classification.

Pure decimal code construction does not lend itself readily to mechanized data processing methods because fixed-code field definition is inconsistent with the decimal code expandability. A number of devices may be used for machine processing of the decimal code, such as tagging variable length fields, special indentation and spacing, and blocked construction as in the following example.

Code	Subject
531000	Mechanics
531100	Machines
531110	Level and Balance
531120	Wheel and Axle
531130	Cord and Catenary
531140	Pulley
531141	Pulley, Compound

In this example, the decimal code has been converted to a six-digit fixed-field block classification code.

The organization of the decimal code is retained, but the degree of expandability has been limited to ten subdivisions for each machine class. The next section describes block codes in greater detail.

Block Codes. The block code dedicates each code position or groups of digits to some characteristic of the items to be coded. There are several variations of block coding. One of the simplest forms is the high

order block. This form uses only the first digit in a blocking mode, the rest of the code is some other type. If several company locations are involved, for instance, employee identification numbers may be blocked like this:

First Digit	Location
1	New York
2	Chicago
3	Denver
4	San Francisco

Hence, "200001" might be the number of the first man hired at the Chicago location. This use of block coding is common when duplicate employee numbers which existed at several previously autonomous locations are incorporated in a central information processing system. The blocking first digit eliminates the duplicates. This technique also allows each location to continue issuing new numbers without the necessity of establishing a central number control point.

Dependent Codes. In most classification systems, classes are divided into subclasses, and subclasses are divided further into sub-subclasses. When coding these classes and subclasses, usually the code assigned to subclasses is unique only within the subclass since the same codes are used to code members of another subclass. By example, the following illustration demonstrates the dependency of the identification of the class for unique identification of the subclass.

Class: States of the United States

Members: Alabama - Coded 01
Arizona - Coded 04

Subclass: Counties of the States of the United States:

Alabama

Autauga County - Coded 001
Baldwin County - Coded 003
Barbour County - Coded 005

Arizona

Apache County - Coded 001
Cochise County - Coded 003
Coconino County - Coded 005

In this example, the code 001 as a county code represents two different entities (Autauga County, Alabama and Apache County, Arizona). In order to be unambiguous, the county code must be used with the state code as 01001 for Autauga County, Alabama and 04001 for Apache County, Arizona. In this example, the county code is dependent upon the state code in order to yield unique identification. The three character county code is also unique within a given state and can be used when the application is restricted or limited to counties of only one state.

When classified and coded in this way, the county code is a dependent code. When the county code is used with the state code, this collective code is also a significant code, because the code structure not only identifies the county, but also the state to which it belongs.

This concept of dependency is not limited solely to classes and subclasses. For example, in certain applications different transactions are identified by a code consisting of parts which represent the organization, the data of the transaction, and a serial number assigned to each transaction on that date. In this example, all three code segments must be employed to produce a unique transaction number derived from all other transaction numbers. This too, is a dependent significant code of the composite data element named "Transaction Number."

4.4.3 Mnemonic Codes (Constant Length Abbreviations). Mnemonic code construction is characterized by the use of either letters or numbers or letter-and-number combinations which describe the items coded, the combinations having been derived from descriptions of the items themselves.

The combinations are designed to be an aid to memorizing the codes and associating them with the items which they represent.

Unit of Measure codes are frequently mnemonic codes. For example:

FT - Foot or feet
BD - Board
BF - Board foot or feet

It should be noted that not all codes used by humans are truly fixed length. To facilitate computer processing, high- or low-order blanks or zeros must frequently be added to make the code values constant length.

There are some problems connected with the use of mnemonic codes to identify long, unstable lists of items. Wherever item names beginning with the same letters are encountered, there may be a conflict of mnemonic use. To overcome this, the number of code characters is necessarily increased, thus increasing the likelihood that the combinations will be less memory-aiding for code users. Also, since descriptions may vary widely, it is difficult to maintain a code organization which conforms with a plan of classification.

Mnemonic codes are used to best advantage for identifying relatively short lists of items (generally 50 or fewer unless the list is quite stable), coded for manual processing where it is necessary that the items be recognized by their code. A common problem, however, is that the code is likely to be misapplied when specific code values are subject to change and users rely too heavily on memory. Thus, to be effectively coded with mnemonics, entity sets must be relatively small and stable.

Acronyms

The acronym is a particular type of mnemonic representation formed from the first letter or letters of several words. An acronym often becomes a word in itself. For example:

RADAR = RAdio Detecting And Ranging
HEW = Department of Health, Education & Welfare

Only when they are of fixed length are acronyms considered data codes.

SECTION 5

PRINCIPLES OF DATA CODE DEVELOPMENT

5. Principles of Data Code Development

5.1 Introduction. The need to communicate with and by means of computers has made increasing demands on data systems designers and users to work out, work with and understand computer codes and printouts. The difficulties of natural language, and particularly the English language, which were examined above must be overcome in any efficient data code. But it must always be remembered that a data code will be used by human beings, including people who do not have much familiarity with data processing. Data codes should therefore be designed with two features in mind: optimum human-oriented use and machine efficiency.

This section provides guidelines to assist in the design and development of data codes which support both features.

5.2 Ten Characteristics of a Sound Coding System. The most viable and useful coding system is one which contains the greatest number of the following ten features:

(1) Uniqueness. The code structure must ensure that only one value of the code with a single meaning may be correctly applied to a given entity, although that entity may be described or named in various ways.

(2) Expandability. The code structure must allow for growth of its set of entities, thus providing sufficient space for the entry of new items within each classification. The structure must also allow existing classifications to be expanded and others added as required. Generally considered, at least a doubling of the original set must be accommodatable, with normal expansion between presently assigned positions; an anticipated life span, depending upon the collection and the dynamics of the environment, should be scheduled.

(3) Conciseness. The code should require the fewest possible number of positions to adequately describe each item. Brevity is advantageous for human recording, communication line transmission, and computer storage efficiencies.

(4) Uniform Size and Format. Uniform size and format is highly desirable in mechanized data processing systems. The unauthorized addition of prefixes and suffixes to the root code is a common problem and is incompatible with the first trait -- uniqueness. Because such prefixes and suffixes are often of variable length and do not always appear, inconsistencies and confusion result.

(5) Simplicity. The code must be simple to apply and easily understood by each user, particularly workers with the least experience.

(6) Versatility. The code should be easily modified to reflect necessary changes in conditions, characteristics, and relationships of the encoded entities. However, every change in the nature of the defined entities must be accompanied by a corresponding change in the code or coding structure.

(7) Sortability. It is desirable to obtain reports in a predetermined format or order. Reports are most valuable when sorted for optimum human efficiency. Although data must be collatable and sortable, the representative code for the data does not have to be sortable, if it can be correlated with another code which is sortable.

(8) Stability. Code users need codes which require infrequent updating. Individual code assignments for a given entity should be made with a minimal likelihood of change, either in the specific code or in the entire coding structure. Changes are costly, laborious and cause errors, and can damage the system when uncontrolled.

(9) Meaningfulness. Meaningfulness should accompany the codes to the greatest extent possible. To instil greater meaning, the code values should reflect characteristics of the encoded entities, such as mnemonic features, unless such a procedure results in inconsistency or inflexibility.

(10) Operability. The code should be adequate for present and anticipated data processing both geared to machine and human use. Care must be exercised to minimize the clerical effort or computer update and maintenance time required to continue operations.

5.3 Code Design Principles. This summary of data codification principles is intended to serve as a checklist for system designers. Its use may help them to avoid the potentially expensive results of inadequately conceived and developed data codes.

It should be noted that, in many instances, these traits may be conflicting. For example, if a coding structure is to have sufficient expandability for future needs, it may have to sacrifice conciseness to some degree. Hence, all trade-offs must be appropriately considered to enable optimum efficiency within a given structure.

5.3.1 General

Planning a Coding System. Sufficient effort and, if need be, time must be spent in preliminary study, definition and planning when designing a new coding scheme. Potential problems must be anticipated and all design alternatives thoroughly evaluated prior to implementation of the new system.

(1) Code Significance. When properly used, significant codes provide a basis for additional information and tend to be easier and more reliable for human use than non-significant codes. However, caution must be exercised in the development of significant codes to assure that significant parts are connected to stable entities. For example, a significant code for an organization should not be associated with the location of the organization when a change in location would result in a change in the code. Excessively significant codes can become unmanageable and lack expandability, and should thus be avoided. For extremely simple tasks, numeric characters are preferable. However, alpha characters are more meaningful and thus better suited to complex tasks.

(2) Use of Standard Codes. Existing codes should be used wherever possible. New codes should not be designed unless absolutely necessary. In all cases, the preference of the code users should be taken into consideration. It is advantageous to consider all code systems employed by the intended users of a new coding system.

(3) Multiple Code Set Compatibility. More than one code or representation is necessary in some instances to meet most systems requirements. A single code is the ideal objective, but is not always the most practicable solution. Multiple codes, if needed should be translatable from one code to another, i.e., the data items remain unchanged, only the codes are variable.

(4) Mnemonic Codes. Mnemonic codes may be used to aid association and memorization, thus increasing human processing efficiency, provided they are not used for identification of very long, unstable lists of items. Mnemonic structures must be carefully chosen, however, to insure that flexibility is not sacrificed. Mnemonics should generally be avoided if the potential code set exceeds 50 entries, because the effectiveness of the mnemonic feature decreases as the number of items to be coded increases. Where mnemonic or otherwise meaningful codes cannot be provided for all codes in the system, preference should be given to codes having the highest use frequency.

(5) Code Naming. All independent data code segments must be individually named with standard, unique, consistently applied labels.

(6) Calculation of Code Capacity. When calculating the capacity of a given code for covering all situations while maintaining code uniqueness, the following formula applies (assuming 24 alpha characters and 10 numeric digits are used because the letters I and O should be avoided whenever possible):

$$C = (24^{A})\ (10^{N})$$

where

C = total available code combinations possible

A = number of alpha positions in the code

N = number of numeric positions in the code

(A + N, when combined, equal the total positions of the code.)

NOTE: The above formula assumes that a given code position is either alpha or numeric -- never both. If a given position can have both alpha and numeric characters, the formula becomes $C = (36)^{A + N}$ or $(34)^{A + N}$ when the letters "I" and "O" are not used.

5.3.2 Code Length

(1) Conciseness. Codes should be of minimum length to conserve space and reduce data communication time, but at the same time optimized in terms of the code users capabilities.

(2) Fixed Length. A code of a fixed length (e.g., always three characters, not one, two, or three) is more reliable and easier to use than a variable length code.

(3) Segmentation. Codes longer than four alphabetic or five numeric characters should be divided into smaller segments for purposes of reliable recording, e.g. XXX-XX-XXXX is more reliable than XXXXXXXXX. The code designer should take advantage of common English usage to divide or link long code phrases.

(4) Potential Expansion. The code structure should provide for adding new items without having to recode existing items or extending the code length.

5.3.3 Code Format

(1) User Considerations. Code components and phrases should be formatted according to user needs for information, considering greatest ease of scanning for accuracy and completeness, and compactness of the message. Message formatting should be coordinated among system users.

(2) Alphabetic versus Numeric. (*) Human recording of numeric codes is generally more reliable than that of alphabetic (all letters) or alphanumeric codes (letters and numbers) where no mnemonic characteristics exist. Controlled alphanumeric codes (i.e., where certain positions are always alphabetic or numeric) are more reliable than random alphanumeric codes. For example, AA001 (where the first two characters are always letters and the last three are numbers) is a more reliable code than when letters or numbers can appear in any position.

(3) Character Grouping. In cases where the code is structured with both alpha and numeric characters, similar character types should be grouped and not dispersed throughout the code. For example, fewer errors occur in a three character code where the structure is alpha-alpha-numeric (i.e., HW5) than in the sequence alpha-numeric-alpha (i.e., H5W).

(4) Code Position Sequence. If a code divides an entire entity set into smaller groupings, the high-order positions should be broad, general categories; and low-order positions should be the most selective and discriminating (including any prefixes and suffixes). An example is the date (YYMMDD). If a descriptive code is formulated consisting of two or more existing independent codes, the individual code segment occupying the higher-order position will be based on usage requirements and processing efficiency considerations.

(5) Separation of Code Segments. Code segments should be separated by a hyphen (when displayed) or exist in complete separation (when stored and displayed) if the positions or segments are completely independent and can stand alone (i.e. no other code is required for complete meaning).

(*) Cardozo, B.L. and Leopold, F.F., "Human Code Transmission," Ergonomics, 1963, 133-141.

(6) Check Characters. When the number of characters of a proposed code exceeds four characters and when this code will be for purposes of identification of major subjects (e.g., organizations, projects, materials, individuals, etc.) consideration should be given to the addition of an error-detecting character to avoid errors in recording. Employment of a self checking code avoids many unnecessary problems of posting data to the wrong record and providing misinformation.

5.3.4 Character Content

(1) Special Characters. Familiar characters should be used, and characters other than letters or numbers (such as the hyphen, period, space, asterisk, etc.) are to be avoided in code structures (except for separating code segments, where a hyphen may be used). Upper case letters only, i.e., ABC...Z (not abc...z), are to be used in data codes. Names and abbreviations may use both upper and lower case letters and other characters. The vocabulary for a given code system should contain the fewest possible character classes. Wherever possible, the character set used for data standards should conform to the American National Standard Code for Information Interchange (ASCII).(*)

(2) Visual Similarities. When it is necessary to use an alphanumeric random code structure, characters that are easily perceived as, or confused with, other characters should be avoided. Some examples are: letter I vs. number 1; letter O vs. number zero; letter Z vs. number 2; slash, or virgule, / vs. number 1; and letters O and Q.

(3) Acoustical Similarities. Nonsignificant codes should avoid characters that can be confused when pronounced (acoustically homogeneous); for example, the letters B, C, D, G, P, and T or the letters M and N.

(4) Vowels. Avoid the use of vowels (A, E, I, O, and U) in alpha codes or portions of codes having three or more consecutive alpha characters to preclude inadvertent formation of recognizable English words.

(5) Collating Considerations. Any specific character position should be either letters or decimal digits in order to avoid collating sequence incompatibility.

5.3.5 Assignment Conventions

(1) Meaningfulness Reduces Errors. Significant or meaningful data codes are preferred over nonsignificant or random codes. This facilitates use by the human coder and reduces errors. For example, in coding the counties of the States of the United States, fewer errors may be expected when the code structure is SSCCC--where the first two characters are the code for a State and the last three characters are the code for a county within that State--than in a code such as XXXX that is randomly assigned to each county.

(*) ANS X3.4-1968.

In this connection, mnemonic data codes produce fewer errors than other types of codes where the number of items to be coded is relatively small and stable. For example, M and F are more reliable codes for male and female than 1 and 2. Y and N are preferred for Yes and No over 1 and 2.

(2) The rules of the data code structure and its derivation should be clearly stated and consistently applied. For example, a mnemonic abbreviation may be formed by deleting all vowels from the names of the coded items as DT for date or GRN for green, or the first letters of the words of the coded items may be used as EOF for End of File or DO for Due Out.

(3) Codes for Numeric Categories. Quantities or numbers should not be coded since this introduces additional translation and a loss of preciseness. For example, the numbers 1 to 99 could be coded A, 100-199 coded B, etc. This may be desirable for purposes of categorization, but statistical value is lost since the actual numbers can not be derived once they are coded. Categorizations can be performed during later phases of data processing rather than in precoding of the input data.

(4) Use of "Natural" Data. A code structure should not be developed if the specific data in its natural form (such as specific percentage amounts) is appropriate and adequate.

(5) Sequence Code Numbering. To maintain fixed code length and avoid confusing leading zeros, codes assigned in sequence may be assigned beginning with "101, "102" or "1001", "1002", etc. rather than with "1". Another advantage of this practice is keeping unauthorized persons from determining the quantity of data in the total entity set from knowledge of a single code (e.g., Product Serial Numbers). Code numbers with lower values may be used to identify miscellaneous or special situations, if so desired, or may be left unassigned. (This procedure does reduce code set capacity, however.)

(6) Use of "0000" and "9999" as code values. One should not use all "0's" (implies nothing) or all "9's" (implies the end) as assigned code values. These values should be reserved for special situations or for use as processing indicators.

(7) "Miscellaneous" Codes. A code category for "Miscellaneous" or "Other" varieties must be used with great discretion. One should not allow the placement of entities in this category which actually belong in a more specific class.

SECTION 6

GUIDELINES FOR DEVELOPMENT OF DATA STANDARDS

6.1 Introduction

6.2 Project Definition

6.3 Formation of Task Groups

6.4 Information Collection

6.5 Criteria for Development of Standard Representations

6.6 Technical Specifications

6. Guidelines for Development of Data Standards

6.1 Introduction. A data standardization project may be initiated at the international or national level, within a trade or professional association, or within an industrial organization. The task may begin when the people responsible for information or data within the organization find difficulty in obtaining and interchanging the data needed to conduct their necessary functions, and recognize the need for standards. At the national level, a data standardization project may be established by the American National Standards Institute whenever it has been determined that a specific standard should be developed.

There are several steps that must be taken to complete the task of standardization, beginning with the precise definition of the project and a thorough inquiry into the background and available resources to undertake this effort.

6.2 Project Definition. The first step to be taken is to define the purpose and scope of the project. The objectives need to be identified and a program of work developed. After these are prepared, a project chairman should be appointed and a task group formed. If a new ANSI project is to be established, the scope statement and program of work must be coordinated with the X3 Standards Planning and Requirements Committee (SPARC) and approved by the X3 Committee on Computers and Information Processing. The planned project should be documented in accordance with X3 procedures.(*)

6.3 Formation of Task Group. It is important that the proper interests and talents be represented in the standards development. Identifying persons with the interest, the resources and the expertise to assist in the work is often difficult. A letter can be sent to individuals and organizations requesting participation. This letter should request the type of person or expertise needed and provide an estimate of the time involved and duration of the project.

The size of the group will depend on the particular project. Generally, a task group should have at least four members.

When the task group members are known, the first meeting should be planned. At the initial meeting, the objectives and planned work should be reviewed, administrative details should be discussed and meeting schedules planned.

6.4 Information Collection. The development of coded representations for a particular class of subjects should begin with the following questions:

a. What are the requirements of the code and what uses of it are anticipated?

b. Are codes really needed, and if so why?

c. What and how many items are to be included in the class of subjects to be coded?

(*) Document X3/SD-2, "Outline for Recommending the Initiation of a Proposed Standards Project to American National Standards Committee X3-Computers and Information Processing (SPARC #209)".

d. What is the most effective code structure?

e. What rules or procedures are necessary for making code assignments?

Certain basic information needs to be collected to answer these questions. This includes seeking answers to further questions:

a. Will the users of the information produced by the systems accept data codes on the output document?

b. How critical is the coded data to the system? What are tolerable error rates? Should a check character be employed to reduce errors?

c. How will the data codes be maintained?

d. Are there codes currently in wide use that are acceptable?

e. What are the machine factors to be considered? (e.g., computer processing and storage capabilities, input media and method of recording--i.e., punched cards, punched paper tape, magnetic tape, on-line terminals, optically read forms, and transmission time.)

f. How and by whom are the data collected or obtained?

g. What human factors (limitations and capabilities) need to be considered?

h. Have the code design criteria in 5.2 and 5.3 been consulted?

These factors are not listed in any particular order of signficance. Trade-offs usually are necessary before final decisions are made because not all factors can be satisfied.

6.5 Criteria for Development of Standard Representations. The discussion of basic coding methods in Section 4 and the data coding principles in Section 5 are provided to assist in the development of specific standard data representations. It must be recognized, however, that some of the criteria in Section 5 conflict. The development task group must analyze the use of the particular representation and decide which criteria are more important to its particular situation.

The relative ease or difficulty users of a data code can be expected to experience can be estimated by the "Information Load Method". This method takes into account the length of the code and the structure of each character in the code. The "information load" of a given code is defined as the sum of the "character load" of each character of the code. The character load is a value equal to $\log_2$ of the total number of different characters that could appear in that character position. For example, the character load for a numeric character code position that could have values of 0 through 9 is the $\log_2$ of 10, or 3.32, and for an alpha character position where the values could range from A through Z, the character load is the $\log_2$ of 26, or 4.70. The information load of a three-character numeric code would thus be:

3.32 + 3.32 + 3.32, or 9.96. For a three character alpha code, the information load would be: 4.70 + 4.70 + 4.70, or 14.10. A code having two numeric characters and one alpha character would have an information load of: 3.32 + 3.32 + 4.70, or 11.34.

This technique is most usefully applied to nonsignificant codes where no secondary meaning can be derived from the code. Nonsignificant codes are used only to uniquely identify the coded subjects in the class. For example, the number 80 would be a nonsignificant code for the month of December, whereas 12 would be a significant code since December is the twelfth month of the year.

When longer codes are broken into smaller units, the information load applies to the smaller units. Whenever the information load exceeds 20, the error rate of data recording can be expected to increase. This rule is stated simply in principle number 3, Section 5.3.2.

6.6 Technical Specifications. The task group should develop the technical specifications of the proposed standard to include:

--- a list of the data items by name (or as appropriate, the characteristics of the data items if these are not names, e.g., Social Security Account Number)

--- definitions of those data items where explanation is necessary

--- abbreviations (as needed), and

--- a unique data code (or codes) for each item.

There shall not be any duplicate codes on the list (or duplicate abbreviations). Names and definitions should be reviewed to insure that each data item is sufficiently different in name and meaning from any other item so that ambiguities are avoided. A concise name for the proposed standard should be determined, e.g., "Calendar Date", "States of the United States", etc.

When a proposed American National Standard is being prepared, applicable procedures and formats should be followed. Applicable ISO (International Organization for Standardization) procedures and guides should be followed if an ISO Recommendation is to be the end product. The task group chairman should obtain the most current procedures and guides either from the appropriate Standards Sectional Committee Chairman or from the American National Standards Institute.

SECTION 7

GUIDELINES FOR IMPLEMENTATION OF DATA STANDARDS

7.1 Interchange

7.2 Internal Files and Records

7. Guidelines for Implementation of Data Standards

7.1 Interchange. Data standards are developed and approved in order to facilitate the interchange of information between and among independent data systems. Data standards should be employed in these interchanges.

It is recommended that use of the standard be specified when data is requested from another organization. The transmitter is urged to consider converting the data to the standard form, especially if the receiving organization so requests.

7.2 Internal Files and Records. The determination of whether to incorporate data standards into internal files and records is a decision which should be left to the installation manager. When a conversion cost can be offset by the continuing cost of translation of data, the use of the standard in internal files and records can be justified on the basis of cost effectiveness. In other instances, the large investment in current systems and files is such that translation of data (especially, if there is an infrequent or limited amount of interchange) is justified. However, in the redesign of existing systems and in the design of new data systems, the use of data standards should be considered and employed to the maximum extent possible.

537-860 O - 74 - 27

SECTION 8

GUIDELINES FOR MAINTENANCE OF DATA STANDARDS

8.1 General

8.2 Maintenance and Information Relevant to Current Data Standards

8.3 Updating and Improvement of Current Data Standards

8.4 Criteria for the Maintenance of Standards

8. Guidelines for Maintenance of Data Standards

8.1 General. Maintenance of the information that makes up a data standard may be viewed from two distinct viewpoints. The first considers the unique administration of the specialized vocabulary or code set which requires peculiar updating and dissemination techniques. The second view sees the problem from the perspective of maintaining and managing a distinctively designed data base, perhaps one responsive to the accounting, inventory, report and control needs of present large-scale management information systems.

Insofar as the second consideration has recently come under the scrutiny of ANSI Committee X3 in its deliberations concerning standard data base management methods and systems, only the first viewpoint will concern us here.

8.2 Maintenance and Information Relevant to Current Data Standards. There are currently at least five kinds of formal data standards in use:

International Standards - which have broad acceptance and the approval of such international groups as the International Organization for Standardization (ISO) and regional groups such as the European Computer Manufacturers Association (ECMA). These are intended for voluntary use and adoption within the national standards of the community of nations. (See Appendix B.)

American National Standards - which include a variety of standards on computer software, data representations such as code sets and structures, and formatting procedures which have been approved and published by the American National Standards Institute. These are intended for the voluntary acceptance and use of industry and government on a nation-wide scale. (See Appendix A.)

U.S. Federal Government Standard Data Elements and Codes for General Use - include Federal general standards for use in the executive branch of government. They embrace such standards as those for countries, states, counties, places, organizations, individuals and elements of time. They are intended for general use by agencies.

U.S. Federal Standard Data Elements and Codes for Program Use - are intended for use in particular related programs concerning more than one agency of the Federal Government. These standards apply to data elements and codes usually limited to applications in weather, personnel, supply, and other unique systems. The same source data are generally used by several agencies, while the information contained in numerous data bases are aggregated and exchanged on a program basis.

Local Standards for data elements and codes - which are maintained for the use of individual disciplines, industries or limited program applications and are either not applicable to international, national or governmental implementation or not yet incorporated into standards with such broad-scale validity.

Existing data standards which have been approved at the international and highest national levels are announced, published and distributed by the national standards organization in each country. In the United States of America, information about such standards may be obtained from the

American National Standards Institute, Inc.
1430 Broadway
New York, New York 10018

The responsibility for announcement, storage and dissemination of information relevant to international, national and Federal data standards may also be carried by certain national information centers, or such announcement media as the Federal Information Processing Standards Publication (FIPS PUB) Series, published by the U.S. National Bureau of Standards.

Local standards are generally maintained by the special group which designed the data base for its proper discipline or purpose-oriented applications. Information concerning the standards and maintenance operations is ordinarily available from the specific organization, trade association or agency involved. An example of an international special purpose standard is the International List of Post Offices, obtainable from the Universal Postal Union.

8.3 Updating and Improvement of Current Data Standards. The maintenance of a data standard must be assigned to an appropriate organization. For the data standard may require any one of a great variety of data bases for its upkeep, control and dissemination.

The data standard can apply to:

Literals - self-identifiable constants such as exact numbers (e.g., dates), serial entites, etc.

Small semi-permanent lists - such as states, counties, countries.

Mission-oriented codes - dynamic lists such as industrials or commodities.

Program or Discipline-oriented codes - as in technical data lists or transaction codes, e.g. for census districts.

Classified Structures - large, semi-constant hierarchically ordered lists such as the Federal Industrial Classification, or the Universal Decimal Classification.

Dynamic lists - such as the Social Security Number files.

Management Information or Command and Control System data elements - file headings required for intelligence or management analysis and report generation.

The files which contain such data must be seen as structures with more or less dynamic features. Depending upon their applications and many internal as well as environmental conditions, these files may often change in content and occasionally in structure, sequence, or storage medium.

Appropriate organizations must be entrusted with the data collection, selection and posting of new entries to the existing files. The efficiency and effectiveness of these maintenance transactions can determine not only the cost but the feasibility of the entire standard data system.

The updating and improvement of current data standards must be channeled through the proper standards body. National Standards must be updated and reviewed by the appropriately appointed groups within the American National Standards Institute. This national organization will forward suggestions for modification and improvement to the proper groups within ISO for updating and revising international data standards. Similar proper governmental, industrial and professional organizational channels should be used to improve existing data standards at these particular levels.

The American National Standards are periodically reviewed and updated when necessary and at least once every five years.

8.4 Criteria for the Maintenance of Standards. To initiate a data standard it is necessary to question whether maintenance of the code can be justified from the viewpoints of:

cost effectiveness
comprehensive coverage
organizational mandate and competence
user needs

It must be determined in advance who should maintain the standard, and by what means of control: centralized, decentralized, or by a carefully designed balance of the two modes.

User needs must be established and a feedback mechanism must be built into the maintenance system. This may require continuous liaison between the maintaining organization and representatives of concerned user groups. This may involve other representatives of industry, commerce, professional organizations as well as Federal, State and local governments.

Periodic review procedures must be established in advance and scrupulously implemented.

Simple file updating procedures must be instituted with special attention given to:

- timeliness of updating
- periodic publication
- efficient and effective promotion and distribution of the basic data base, periodic updates and relevant services, using appropriate media.

Periodic review of the administration and financing of the code or vocabulary data base maintenance is essential.

8. Guidelines for Maintenance of Data Standards (Summary)

8.1 General

8.2 Maintenance and Information Relevant to Current Data Standards. Existing data standards approved at the national and international levels are announced, published, stored and distributed by the national standards organization or by national information centers, or announcement media such as the Federal Information Processing Standards Publication (FIPS PUB) Series.

8.3 Updating and Improvement of Current Data Standards. Channeled through ANSI, the U.S. national standards organization, national data standards are reviewed and updated at least once every five years. Suggestions for improvement may be sent to ANSI for distribution to the proper technical committee for action.

8.4 Criteria for the Maintenance of Standards. Suggestions are given on the maintenance of representational forms . . . vocabularies, abbreviation sets and code structures. Housekeeping and control measures are required to accommodate the changes required in large dynamic lists.

APPENDIX A

SCOPE AND PROGRAM OF WORK
OF AMERICAN NATIONAL STANDARDS INSTITUTE
SUBCOMMITTEE X3L8, REPRESENTATIONS OF DATA ELEMENTS
(As approved by X3 Sectional Committee, January 23, 1970)

Background

The need for a program of data standardization arose with difficulties in interchanging data among the data systems of business and governments. The difficulties stemmed from different organizations using a great variety of representations for the same subject matter, such as places, dates, individuals, organizations and commodities, and using the same representations with completely different meanings, as well as from the lack of a common method for describing the data that was to be interchanged.

The need for standard representations and ways of describing interchanged data had been recognized earlier by particular industries, such as air transportation in the area of passenger reservations. To satisfy this need, programs to establish and maintain data interchange capabilities were initiated. In addition, agencies of the Federal Government initiated standardization programs to facilitate data interchange between agencies. Standardized representations, formats and format descriptions are required among the several needs that must be satisfied for different organizations to interchange data. Early in the 1960's, a standardization program was initiated by the Business Equipment Manufacturers Association (BEMA) and the American Standards Association, now the American National Standards Institute (ANSI), to establish standards related to systems, computers, equipments, devices and media for information processing. This resulted in the formation of the ANSI Committee for Computers and Information Processing, designated X3, with representatives drawn from producer, consumer and general interest groups. In 1966, Subcommittee X3L8 was established as part of the X3 organization and was given the responsibility for standardization of representations of data elements commonly used in interchange. The X3L8 Subcommittee has concentrated on development of standard representations for subject matter of common interest, including standards for times, individuals, organizations, places and numeric values. Interest has expanded to cover other data elements involved in data interchange and to enlist in this program organizations with interest and experience in each of the areas involved.

Definitions

Data Element - A basic unit of identifiable and definable information. In information processing systems, a data element occupies the space provided by fields in a record or blocks on a form. It has an identifying name and a value or values for expressing a specific fact. Examples: Employee number, Employee name, Date of birth, Mailing address, Color of eyes, Height, and Weight.

Representations - Names, Abbreviations, Codes, and Numeric Values used to express a data element.

Scope

1. To develop standards for (1) describing the representations of data elements involved in data interchange; and (2) representing data elements of common interest, such as the elements concerned with the representations of times, locations, individuals, organizations and materials.

2. To develop recommended procedures, criteria, and guidelines in order to provide an organized approach to the standardization of the representations of data elements.

Program of Work

1. To develop recommended procedures, and criteria for the development, maintenance, issuance, and use of American National Standards for representations of data elements.

2. To develop proposed standards for the following items:

   a. Representation of time elements to include dates, times, and time zones.
   b. For identifying organizations, individuals, and accounts to include standards for name formatting.
   c. Representations for States, Counties, Places, and Congressional Districts of the United States, Countries of the World and their Subdivisions, Shipping and Mailing Addresses, and Point Locations.
   d. Representing quantitative numeric expressions.

3. To represent the interests of the United States through the X3 International Advisory Committee and the American National Standards Institute in the development of international recommendations for representations of data elements by the International Organization for Standardization (ISO) or other standardization bodies.

4. To act as the focal point within the American Standards Institute for reviewing proposed representations of data element standards that have been developed by other organizations and which are submitted for adoption as American National Standards and forwarding these with appropriate recommendations through established channels for subsequent standardization actions.

5. To assist, as necessary and resources allow, industry, government, and other groups in the development of proposed standards for representations of data elements.

Other Factors Bearing on the Work of X3L8

1. It is not feasible for one organization to develop representation standards for all the data elements involved in interchange. Accordingly, the most practicable approach is to have a single group develop and establish common procedures and criteria to guide other organizations in developing standards for their particular subject matter or application area. When the results of such developments by other organizations are submitted to ANSI for consideration as American National Standards, X3L8 would review these and prepare recommendations concerning their acceptability or conflict with other established standards and forward these for appropriate standardization actions.

2. Many of the potential standards for representations of data elements are of such a magnitude that their maintenance is beyond the capabilities of X3L8 or the American National Standards Institute. Examples of such standards are those for representing those data elements concerned with identification of organizations and places (i.e., cities, towns, townships, boroughs, etc.) Accordingly, it is essential to depend upon some other organization outside the ANSI structure for this necessary maintenance. This situation does not necessarily forbid the development and establishment of American National Standards. These can be accomplished through agreements with the outside organization as to the procedures and criteria to be used in maintaining the standard. These procedures, criteria, and other considerations then form the basis for the proposed American National Standard.

APPENDIX B

SCOPE AND PROGRAM OF WORK
(As adopted by ISO/TC 97 on June 20, 1972)

Title ISO/TC 97/SC 14, Representations of Data Elements

Scope

Standardization of the representations of commonly interchanged data elements to facilitate information interchange and information processing.

Program of Work

1. To develop international recommendations for describing data elements and their representations involved in data interchange.

2. To develop international recommendations for representing data elements of common interest to include representations for:

a. Dates and time
b. Countries
c. Languages
d. Identification of Individuals
e. Identification of Organizations
f. Identification of Accounts
g. Mailing and shipping address
h. Point locations such as longitude and latitude
i. Units of measure
j. Numeric expressions

3. To develop recommended guidelines and criteria to provide for an orderly approach to the standardization and description of data elements involved in international information interchange.

4. To provide liaison with other organizations and ISO Committees for the coordination of data standards intended for information interchange.

APPENDIX C

BIBLIOGRAPHY

Aume, N. M. and Topmiller, D. A., "An Evaluation of Experimental How-Malfunctioned Codes," Human Factors, 1970, 261-269.

Bell, J. R., "The Quadratic Quotient Method: A Hash Code Eliminating Secondary Clustering," Comm. ACM, February 1970, 107-109.

Bell, J. R. and Kaman, C. H., "The Linear Quotient Hash Code," Comm. ACM November 1970, 675-677.

Blankenship, A. B., "Memory Span: A Review of the Literature" Psychological Bulletin, 1938, 1-25.

Bonn, T. H., "A Standard for Computer Networks," IEEE Computer magazine, May-June 1971.

Brown, J., "Some Facts of the Decay Theory of Immediate Memory" Quarterly Journal of Experimental Psychology, 1958, 12-21.

Cardozo, B. L. and Leopold, F. F., "Human Code Transmission," Ergonomics, 1963, 133-141.

Carson, J. G. H., "Item Identification and Classification in Management Operating Systems," Production and Inventory Management, June 1971.

Cherry, C., On Human Communication, 1966, M.I.T. Press.

Conrad, R., "Experimental Psychology in the Field of Telecommunications" Ergonomics, 1960, 289-295.

Conrad, R., "The Location of Figures in Alpha-numeric Codes," Ergonomics, 1962, 403-406.

Conrad, R. and Hille, B. A., "Memory for Long Telephone Numbers" Post Office Telecommunications Journal, 1957, 37-39.

Conrad, R. and Hull, A. J., "Copying Alpha and Numeric Codes by Hand: An Experimental Study" Journal of Applied Psychology, 1967, 444-448.

Crossman, E. R. F. W., "Information and Serial Order in Human Immediate Memory," in Proceedings of 4th London Symposuum on Information Theory, 1961.

Crowley, E. T. and Crowley, R. C., Acronyms and Initialisms Dictionary, Third Edition, 1970, Gale Research Company.

Crowley, E. T., New Acronyms and Initialisms, 1971, Gale Research Company.

Field, M. M., et al, Guidelines for Constructing Human Performance - Based Codes, Bell Telephone Laboratories Technical Report, 1971.

Frink, W. J., "In Coding It's Structure That Counts," Control Engineering, October 1962.

Gilbert, E. N., "A Comparison of Signaling Alphabets," Bell System Technical Journal, May 1952, 504-522.

Gilbert, E. N., "Information Theory After Eighteen Years," Science, April 15, 1966, 320-326.

Goldman, S., Information Theory, 1953, Prentice-Hall.

Gombinski, J. and Hyde, W. F., "Classification and Coding," Graphic Science, March 1968.

Gombinski, J., "Industrial Classification and Coding," Engineering Materials and Design, September 1964.

Hall, R. A., Linguistics and Your Language, Doubleday, 1960.

Hamming, R. W., "Error Detecting and Error Correcting Codes," Bell System Technical Journal, April 1950, 147-160.

Harris, D. H. et al, "Wire Sorting Performance with Color and Number Coded Wires," Human Factors, April 1964, 127-131.

Hauck, E. J., "Be Kind to Your Data Codes," Journal of Systems Management, December 1972.

Hare, V. C., Systems Analysis: A Diagnostic Approach, 1969, Harcourt, Brace and World.

Heron, A. "Immediate Memory in Dialing Performance With and Without Simple Rehersal," Quarterly Journal of Experimental Psychology, 1962, 94-103.

Hitt, W. D., "An Evaluation of Five Different Abstract Coding Methods - Experiment IV," Human Factors, July 1961, 120-130.

Hodge, M. H. and Field, M. M., Human Coding Processes, University of Georgia, 1970.

Hull, T. E. and Dobell, A. R. "Random Number Generators," SIAM Review, July 1962.

Jackson, R. L., "'Dial 911' Setup Qualified Success," Gannett News Service, July 1972.

Jones, B. W., Modular Arithmetic, 1964, Blaisdell Publishing Company.

Klemmer, E.T., "Grouping of Printed Digits For Manual Entry," Human Factors, 1969, 397-400.

Klemmer, E. T., "Grouping of Printed Digits for Telephone Entry," in Proceedings of Fourth International Conference on Human Factors in Telephony, Munich, 1968.

Klemmer, E. T., "Keyboard Entry," Applied Ergonomics, 1971, 2-6.

Klemmer, E. T., "Numerical Error Checking," Journal of Applied Psychology, 1959, 316-320.

Klemmer, E. T. and Stocker, L. P., "Optimum Grouping of Printed Digits," American Psychological Association Proceedings, 1972, 689-690.

Konz, S. et al, "Human Transmission of Numbers and Letters," The Journal of Industrial Engineering, May 1968, 219-224.

Laden, H. N. and Gildersleeve, T. R., System Design for Computer Applications, 1963, Wiley.

L'Insalata, B. B., "COBOL Module for the Generation and Verification of a Check-bit Using the Modulus 10 Method," Western Electric Co. Technical Report, June 9, 1971.

Little, J. L., "Some Evolving Conventions and Standards for Character Information Coded in Six, Seven and Eight Bits," U.S. Dept. of Commerce, National Bureau of Standards, 1969.

Little, J. L. and Mooers, C. N., "Standards for user procedures and data formats in automated information systems and networks," 1968, Spring Joint Computer Conference.

Mackworth, J. F., "The Effect of Display Time Upon the Recall of Digits," Canadian Journal of Psychology, 1962, 48-55.

Maurer, W. D., "An Improved Hash Code for Scatter Storage," Comm. ACM, January 1968, 35-38.

Mayzner, M. S. and Gabriel, R. F., "Information "Chunking" and Short-Term Retention," Journal of Psychology, 1963, 161-164.

Meltzer, H. S. and Ickes, H. F., "Information Interchange Between Dissimilar Systems," Modern Data, April 1971.

Miller, G. A., "The Magical Number Seven, plus or Minus Two: Some Limits on Our Capacity for Processing Information," The Psychological Review, March 1956.

Miller, G. A., The Psychology of Communication, 1967, Basic Books, 1969, Pelican Books.

Miller, G. A. and Nicely, P. E., "An analysis of perceptual confusions among some English consonants," Journal of the Acoustical Society of America, 1955, 338-352.

Morris, R., "Scatter Storage Techniques," Comm. ACM, January 1968, 38-44.

Oberly, H. S., "A Comparison of the Spans of Attention and Memory," American Journal of Psychology, 1928, 295-302.

O'Reagan, R. T., "Computer-Assigned Codes from Verbal Responses," Communications of the ACM, June 1972, 455-459.

Owsowitz, S. and Sweetland, A., Factors Affecting Coding Errors, 1965, The Rand Corporation.

Peterson, W. W., Error-Correcting Codes, 2nd Ed., 1972, M.I.T. Press

Pollack, I., "Assimilation of Sequentially Coded Information," American Journal of Psychology, 1953, 421-435.

Radke, C. E., "The Use of Quadratic Residue Research," Comm. ACM, February 1970, 103-105.

RAND Corporation, A Million Random Digits with 100,000 Normal Deviates, The Free Press, 1955.

Rocke, M. G., "Data Codification Principles and Methods," Caterpillar Tractor Company, 1971.

Severin, F. T. and Rigby, M. K., "Influence of Digit Grouping on Memory for Long Telephone Numbers," Journal of Applied Psychology, 1953, 117-119.

Shannon, C. E., "Prediction and Entropy of Printed English," Bell System Technical Journal, January 1951, 50-64.

Shannon, C. E., and Weaver, W., The Mathematical Theory of Communication, 1948, Bell System Technical Journal; 1949, University of Illinois Press.

Sonntag, L., "Designing Human-Oriented Codes," Bell Laboratories Record, February 1971.

Stallcup, K. L., "New Growth in Linguistics Produces Clarity, Confusion and Controversey," The New York Times, January 8, 1973, page 90.

Talbot, J. E., "The Human Side of Data Input," Data Processing Magazine, April 1971.

Thorpe, C. E. and Rowland, G. E., "The Effect of "Natural" Grouping of Numerals on Short-Term Memory," Human Factors, 1965, 38-44.

Wicklegrew, W. A., "Size of Rehersal Group and Short-term Memory," Journal of Applied Psychology, 1964, 413-419.

Woodbury, M. A. and Lipkin, M., "Coding of Medical Histories for Computer Analysis," Comm. ACM, October 1972.

Wooldridge, D. E., The Machinery of the Brain, 1963, McGraw-Hill

Woznick, A. M., "Item Identification Standards," Systems Engineering CASO Development Report 56904010, Western Electric Company, February 9, 1971.

Subtitle A of Title 15 of the Code of Federal Regulations is amended by adding a new Part 6, reading as follows:

PART 6.

STANDARDIZATION OF DATA ELEMENTS AND REPRESENTATIONS

Sec.

AUTHORITY: The provisions of this Part 6 issued under 79 Stat. 1127; Executive Order 11717, dated May 9, 1973 (38 FR 12315, dated May 11, 1973).

6.1. Purpose

The purpose of this Part is to implement the provisions of Section 111 (f) (2) of the Federal Property and Administrative Services Act of 1949, as amended (79 Stat. 1127) and Executive Order 11717 of May 9, 1973 (38 FR 12315, dated May 11, 1973). It supersedes and replaces in its entirety Office of Management and Budget Circular A-86 entitled, "Standardization of data elements and codes in data systems", dated September 30, 1967. Office of Management and Budget Circular No. A-86 was rescinded by the Director of Office of Management and Budget on August 29, 1973.

This Part identifies responsibilities and provides policies and guidelines for the management of activities in the Executive Branch relating to the development, implementation and maintenance of standards for data elements and representations used in automated Federal data systems. Its provisions complement the standards and recommendations that have been or may be issued under the statistical procedures prescribed by Office of Management and Budget Circular A-46.

6.2. Background

Recent advances in computer and communications technologies have made possible the wider use of data and programs that are developed or generated to meet mission requirements of Federal departments, agencies, and activities. While the extended use of these data and programs can contribute to reduced costs in Government operations and improved services, the full advantages of these new technical capabilities cannot be realized until standards are developed and implemented which will provide for the uniform identification, definition and representation of data. These standards for data must also be accompanied by supporting standards for representing graphic characters (alphabets, numbers, and other symbols), communications and device controls. In addition, it is essential to have standards that provide for interchangeable media (e.g., tapes, cassettes and disks) covering both physical and logical specifications.

There is an ever increasing need to interchange data and programs with state, local and other governments, and with industry and the public. This adds further emphasis and dimension to the need for responsive standards that will facilitate interchange.

This Part defines a Federal-wide program for standardizing data elements and representations which are used and interchanged in Government data systems. Other approved standards and guidelines issued by the National Bureau of Standards in the Federal Information Processing Standards series of publications address related ADP subjects and areas.

6.3. Objectives

The principal goal in standardizing data elements and representations is to make maximum utilization of the data resources of the Federal Government and to avoid unnecessary

duplications and incompatibilities in the collection, processing, and dissemination of data.

6.4. Glossary

Appendix A of this Part provides a glossary of terms as used in this Part and in descriptions of data.

6.5. Types of Standards

For the purposes of this Part, the following types of practices and standards are identified for data elements and representations:

(a) **De facto Practices.** Those data elements and representations in current use that have not been subjected to official or formal standardization.

(b) **Unit Standards.** Those data elements and representations that have been approved by an authorized official for use within that unit. (A unit for purposes of this Part is any Federal organization within the executive branch of the Government, which is at a lower organizational level than an executive department or independent agency).

(c) **Agency Standards.** Those data elements and representations that have been approved by an authorized official for use within an executive department or independent agency.

(d) **Federal Program Standards.** Those data elements and representations that have been approved by the Secretary of Commerce for use in a particular program or mission where more than one executive branch department or independent agency is involved with their use. For example, those standards that could be approved and prescribed for use are those which include, but are not limited to, Federal-wide personnel, communications and transportation data systems.

(e) **Federal General Standards.** Those representations that have been approved by the Secretary of Commerce for Federal-wide use by executive departments and independent agencies in all Federal-wide programs and for use in all Federal data systems. For example, this includes such representations as calendar dates, state abbreviations and codes, and codes for standard metropolitan statistical areas.

(f) **American National Standards.** Those data elements and representations that have been approved for voluntary national use by the American National Standards Institute.

(g) **International Standards.** Those data elements and representations that have been approved by the International Organization for Standardization (ISO), for voluntary use by member nations and international organizations.

6.6. Policies

The following policies apply to the development, implementation, and maintenance of data element and representation standards:

(a) Data Elements and representations that are prescribed for interchange among more than one executive department or agency or with the private sector including industry, state, local, or other Governments, or with the public at large will be considered for standardization as either Federal general or Federal program standards.

(b) Federal general standards are the highest level standards followed by Federal program standards, agency standards and unit standards in that order. This order establishes a precedence for standards use. For example, a Federal general standard will be used and will supplant a Federal program, agency or unit standard. Likewise a Federal program standard takes percedence over an agency or unit standard.

(c) Approved standards will be implemented by all Federal agencies in all circumstances where technical, operating and economic benefits can be expected to result. These standards will be considered on the basis of their long-term benefits and advantages to the Government at large. Local inconveniences or short-term conversion costs need to be recognized, but such factors will not be considered overriding deterrents to the development, implementation, and maintenance of standards that are capable of reducing overall government operating costs or providing improved Government services.

(d) Existing standards will be considered for adoption as Federal general or program standards when these are determined to meet Federal requirements or can readily be adapted to do so.

(e) Approved standards and revisions thereto will be implemented on a time phased basis in order to minimize disruption and conversion costs. Conversion costs will be identified and considered in the submissions of annual budget estimates.

(f) Although data element and representation standards are developed and implemented to provide for the effective interchange and processing of data, Federal departments and agencies must comply with applicable statutes, regulations and executive orders to assure that sensitive or classified data are adequately protected and that only authorized disclosure or release of such data is allowed.

(g) In the formulation of standards for data elements and representations which will have implementation impact on state and local governments, industry or other segments of the private sector, arrangements will be made to establish necessary liaisons and coordinations with these interests to consider their needs and potential problems in responding to Federally imposed reporting requirements.

6.7. Responsibilities

Responsibilities for the standardization of data elements and representations are outlined below:

(a) **Department of Commerce.** The Department of Commerce will provide leadership of an executive branch program for standardizing data elements and representations. Within the Department the following specific responsibilities are assigned:

(1) **Secretary of Commerce.** The Secretary of Commerce, on behalf of the President, approves all Federal Information Processing Standards. For data elements and representations, this approval will include both Federal general and Federal program standards.

(2) **National Bureau of Standards.** The National Bureau of Standards will:

(i) Arrange with appropriate executive branch departments and independent agencies to assume leadership and undertake responsibilities for the development and maintenance of specific Federal program and Federal general standards.

(ii) Arrange for the publication and promulgation of approved Federal general and Federal program standards. These will be promulgated by the National Bureau of Standards as Federal Information Processing Standards. The responsibility under this subparagraph includes the authority to modify or supersede these standards whether issued under this regulation or prior to the effective date of this regulation.

(iii) Maintain and promulgate selected registers of data element and representation standards and practices that are under development or are in current use.

(iv) Provide procedures, guidelines and criteria to assist Federal departments and agencies in the development, implementation, and maintenance of standards.

(v) Provide technical assistance, as requested and within the limits of available resources to Federal departments and agencies on matters concerning the utilization of automatic data processing and standardization.

(vi) Arrange for the assessment of the need, impact, benefits and problems related to the implementation of proposed and approved standards.

(vii) Coordinate requests for exceptions to and deferments on the implementation of approved Federal standards.

(viii) Arrange for and coordinate appropriate Federal representation and participation on voluntary industry committees.

(ix) Arrange for appropriate liaison with state, local and other governments on matters of mutual interest or concern relating to Federal development, implementation, and maintenance of standards.

(b) **Departments and Independent Agencies.** Each of these organizations will:

(1) Implement approved Federal standards that are announced under the provisions of this Part and assist the National Bureau of Standards in the assessment of the need, impact, benefits and problems related to the implementation of approved standards.

(2) Assume leadership and support of responsibilities for the development of Federal general and Federal program standards as may be mutually arranged by the National Bureau of Standards.

(3) Establish within their organizations, mechanisms for the development, implementa-

537-860 O - 74 - 28

tion and maintenance of agency and unit standards where such efforts will contribute to reduced costs or improved services.

(4) Establish appropriate procedures and mechanisms within their organizations for the dissemination and implementation of approved Federal standards.

(5) Review and provide information and comments on proposed standards that are being considered for Federal adoption. This includes the analyses necessary to assess implementation impact and potential savings or improved services.

(6) Prepare and submit selected registers of data elements and representations within the data systems of the department or agency as may be arranged by the National Bureau of Standards. These registers will be used as a source reference to avoid duplication in the design of new data elements and representations and to assist in determining possible subjects for future standardization.

(7) Provide participation on committees and task groups that may be formed to develop and maintain Federal general or Federal program standards.

(8) Provide participation, as requested by the National Bureau of Standards, on committees and task groups that may be formed to develop and maintain voluntary industry standards for use nationally and internationally.

(9) Designate an office or official to act as a single point of contact on matters related to this Part.

6.8. Exceptions, Deferments, and Revisions of Federal Standards

Requests for exceptions, deferments and revisions of standards will be forwarded to the National Bureau of Standards for consideration and/coordination. These requests will provide detailed justification for the exception, deferment or revision deemed necessary. These should be submitted at least forty-five days in advance of any exception or deferral action.

6.9. Effect On Previously Issued Standards

All standards that were issued under the provisions of Office of Management and Budget Circular No. A-86 prior to the effective date of this regulation remain in effect unless modified or supeseded pursuant to the provisions of this regulation.

Glossary of Terms

This Glossary includes definitions of terms used in this Part. Additional terms applicable to data standardization are provided for purposes of clarification. The terms and definitions are either from established vocabularies or have been defined for purposes of this Part.

Attribute Data Element—A data element that is used to qualify or quantify another data element (e.g., "Date of Birth" and "Mailing Address" would be attribute data elements in a personnel file where the primary element(s) is/are used to identify the person).

Character Type—An indication of the type of characters or bytes to represent a value (i.e., alphabetic, numeric, pure alphabetic, pure numeric, binary, packed numeric, etc.).

Alphabetic—A representation which is expressed using only letters and punctuation symbols.

Alphanumeric—A representation which is expressed using letters, numbers, and punctuation symbols.

Binary—A representation of numbers which is expressed using only the numbers 0 and 1; e.g., 5 is expressed as 101.

Numeric—A representation which is expressed using only numbers and selected mathematical punctuation symbols.

Packed Numeric—A representation of numeric values that compresses each character representation in such a way that the original value can be recovered; e.g., in an eight bit byte, two numeric characters can be represented by two four bit units.

Pure Alphabetic—A representation which is expressed using only letters.

Pure Alphanumeric—A representation which is expressed using only letters and numbers.

Pure Numeric—A representation which is expressed using only numbers.

Composite Data Element (Data Chain)—A data element that has an ordered string of related data items that can be treated as a group or singly; e.g., a data element named "Date of Birth" could have the data items, "Year", "Month", and "Day of Month".

Context Dependent Definition—A statement of meaning that relies upon a situation, background, or environment for proper interpretation.

Date Code—A coded representation used to identify a data item. Usually codes are designed according to established rules and criteria, and only by chance form a phonetic word or phrase.

Data Element—A basic unit of identifiable and definable information. A data element occupies the space provided by fields in a record or blocks on a form. It has an identifying name and value or values for expressing a specific fact. For example, a data element named "Color of Eyes" could have recorded values of "Blue (a name)", "BL (an abbreviation)" or "06 (a code)." Similarly, a data element named "Age of Employee" could have a recorded value of "28 (a numeric value)."

Data Element Abbreviation—An abbreviated form of the data element name.

Data Element Definition—A statement of the meaning of a data element.

Data Element Name—A name used to identify a data element.

Data Element Source—An identification of the source or provider of the particular data element; i.e., individual, organization, sensor, computation, etc.

Data Element Tag (Data Element Code)—A symbolic tag used to identify a data element,

Data Item—The expression of a particular fact of a data element e.g., "Blue" may be a data item of the data element named "Color of eyes".

Data Item Abbreviation—An abbreviated form of the data item name.

Data Item Definition—A statement of the meaning of a data item.

Data Item Name—A name used to identify a data item.

Dependent Code—A code that has segments which are dependent upon other segments in order to provide unique identification of the coded item. Usually, codes having classification significance are dependent codes.

Field—In a record, a specific area used for representing a particular category of data; e.g., a group of card columns used to express a wage rate.

Field Length—A measure of the length (size) of a field, usually expressed in units of characters, words, or bytes.

Field Length Type—An indication of whether the field of a record is fixed or variable in length.

Fixed Length Field—A field whose length does not vary.

Variable Length Field—A field whose length varies. Usually, the boundaries of this type of field are identified by field separators.

Field Separator—A character or byte used to identify the boundary between fields.

Filler Character—A specific character or bit combination used to fill the remainder of a field after justification.

Formatted Information—An arrangement of information into discrete units and structures in a manner to facilitate its access and processing. Contrasted with narrative information that is arranged according to rules of grammar.

General Definition—A statement of meaning that can be interpreted without regard to a specific situation, background, or environment.

Information Interchange—The transfer of data representing information between or among two or more points (devices, locations, organizations, or persons) of the same or different (dissimilar) information system or systems.

Justification—To adjust the value representation in a field to either the right or left boundary (margin).

Left Justify—Adjustment of a value representation to the left boundary (high order) of a field.

Right Justify—Adjustment of a value representation to the right boundary (low order) of a field.

Non-Significant Code—A code that provides for the identification of a particular fact but does not yield any further information; e.g. random numbers used as codes. Contrasted with significant code.

Numeric Value—The expression of a data item which denotes a measurement, count, or mathematical concept, usually represented by numerals and a limited number of special characters (i.e., plus (+), minus (−), decimal point (.), comma (,), asterisk (*), and slant (/)).

Padding—A technique used to fill a field, record, or block with dummy data (usually zeros or spaces).

Primary Data Element—A data element or elements that is/are the subject of a record. Usually the other elements, called attribute data elements, qualify or quantify the primary data element (e.g., in a personnel field, the element(s) that is/are used to identify the individual are primary; other elements such as "Date of Birth" and "Mailing Address" are attribute data elements).

Radix Point—A character, usually a period, that separates the integer part of a number from the fractional part. In decimal (base 10) notation the radix point is called the decimal point.

Record—A collection of related elements of data treated as a unit.

Record Index—An ordered reference list of the contents of a record together with keys or reference notations for identifying and locating the contents.

Record Layout—A description of the arrangement and structure of information in a record, including the sequence and size of each identified component.

Record Length—A measure of the length (size) of a record, usually expressed in units of characters, words, or bytes.

Record Length Type—An indication of whether the records of a file are fixed or variable in length.

Fixed Length Record—Pertaining to a file in which the records are uniform in length.

Variable Length Record—Pertaining to a file in which the records are not uniform in length.

Representation—A number, letter or symbol used to express a particular concept or meaning. It may be in the form of a name, abbreviation, code, or numeric value.

Rounding (Roundoff)—To delete the least significant digit or digits of a numeral, and to adjust the part retained in accordance with some rule.

Self-Checking Code—A code that is appended to another code to provide for validity checking. A self-checking code is derived mathematically from the characteristics of the base code.

Significant Code—A code which in addition to identifying a particular fact also yields further information; e.g., catalog numbers in addition to identifying a particular item also often indicate the classification of the item. Contrasted with non-significant code.

Truncate—To delete characters from a character string, usually from either end of the string.

Type of Code Significance—An indication of the type of significance that a particular code yields.

Collating Significance—A code designed in such a way that it facilitates ordering of the coded item.

Mnemonic Significance—A code designed in such a way as to facilitate the human recall of the name of the coded items.

Classification Significance—A code designed in such a way as to facilitate the classifying of the coded items into classes and subclasses.

Variable Name Data Element—A data element that identifies a set (array) of similar values (data items) By varying certain identifiers in the name the entire set (array) of values can be identified. For example, a set of values that give population by State and Year could be identified by the data element "Population of (State) in (Year)" where State and Year are variable names. The variable names are used to identify particular values in an array (e.g., "Population of (New Jersey) in 1970" was 7,168,164.) In this example "New Jersey" and "1970" are variable names used to identify a specific value "7,168,164" in an array.

BSR X3.35
X3L8/190

PROPOSED AMERICAN NATIONAL STANDARD
STRUCTURE FOR IDENTIFICATION OF ORGANIZATIONS
FOR INFORMATION INTERCHANGE

Sponsor

COMPUTERS AND BUSINESS EQUIPMENT MANUFACTURERS ASSOCIATION

Approved (Date)

AMERICAN NATIONAL STANDARDS INSTITUTE

TABLE OF CONTENTS

DESCRIPTION

FOREWORD

(This Foreword is not a part of American National Standard Structure for Identification of Organizations for Information Interchange.)

This Standard establishes Standard Identifiers for Organizations for the sole purpose of facilitating data or information interchange. It does not address the universe of historical organizations nor all organizations for which reference may be needed other than as outlined in this Standard.

The Standard describes the parts of the Standard Identifiers and provides the rules for their use.

Examples are provided in the text to illustrate the rules, but do not necessarily refer to any real situation.

Suggestions for improvement gained in the use of this Standard will be welcome. They should be sent to:

etc.

American National Standards Committee X3 - Computers and Information - had the following membership at the time this Standard was approved.

etc.

1. GENERAL

1.1 Scope and Purpose. This standard establishes a uniform structure for uniquely identifying organizations, specifying acceptable identifiers, for the purpose of facilitating all types of information interchange. Specifically, it is intended to:

- Reduce the time required to format the elements of identification and transmit them;

- Improve clarity and accuracy of identification through the discipline of properly using and displaying the standard identifier consisting of an Identification Code Designator (ICD), a specified identification code and a name;

- Minimize the amount of human intervention required for communicating unique identification;

- Reduce costs; and

- Enable immediate implementation.

1.2 General Concept. For purpose of standardized information interchange, name alone is not sufficient, nor is identification code alone. Name, associated with an identification code and Identification Code Designator (ICD) which uniquely identifies or distinguishes the named organization, is required.

1.3 Qualifications.

1.3.1 This standard does not prescribe procedures, file organization techniques, storage media, languages, etc., to be utilized in its implementation.

1.3.2 This standard is intended for application in cooperative environments; i.e., circumstances in which organizational entities will freely disclose their names and associated identification codes. This standard is not intended for environments in which the organization being identified is unaware of the identification.

1.3.3 This standard does not establish any requirement for organizations to disclose any information involuntarily. Although this Standard is intended to facilitate data interchange, great care must be taken by all users to prevent the unauthorized disclosure or release of information.

1.3.4 This standard does not provide for representation of self-employed persons.

1.3.5 The designation of the EIN or D-U-N-S as the code part of the Standard Identifier For Organizations (SIO) is not intended to establish any obligation or requirement on the part of the issuing organizations beyond that covered in law, regulation, or policy.

1.4 Character Set. This standard uses the ASCII (American National Standard Code for Information Interchange) coded character set (current version), including both upper and lower case alphabetic characters and other graphic characters (See Appendix C).

1.5 Related Standards.

- American National Standard Code for Information Interchange X3.4-1968

- American National Standard Specifications for Credit Cards X4.13-1971

  (The representations prescribed by this Standard are defined in such a way that they can be used in the coding for the Credit Card Standard.)

2. DEFINITIONS

2.1 Organization. An organization is a unique framework of authority within which a person or persons act, or are designated to act, toward some purpose. The kinds of organizations covered under this standard include:

2.1.1 A Corporation

2.1.2 A partnership, non-profit organization, cooperative or similar unincorporated body, in which ownership or control is vested in a group.

2.1.3 An unincorporated enterprise or activity providing goods and/or services.

2.1.4 A foreign, domestic or international government agency or instrumentality.

2.1.5 An organizational grouping of any of the above categories.

2.1.6 A subsidiary, division, branch or subdivision of any of the above having a need for separate identification for the purpose of external information interchange.

2.2 Standard Identifier. The Standard Identifier for Organizations (SIO) is a coded representation permitting a distinction between any specific organization and all other organizations. It consists of an Identification Code Designator (ICD), a code part and a name part, in that order.

2.3 Identification Code Designator. The Identification Code Designator (ICD) distinguishes between Identification Code Systems as set forth in the code part of the SIO (See Section 4 for further details).

2.4 Record Name. The name furnished by the organization or its representative, subject to the general syntax rules in 6.1, and designated as the official name of record.

2.5 Other Name. Trade name, trade styles and initials or acronym by which the organization is commonly known, subject to the general syntax rules in 6.1.

3. STANDARD IDENTIFIER FOR ORGANIZATIONS (SIO)

3.1 Parts of the SIO. The SIO consists of three parts, an Identification Code Designator (ICD), a code part and a name part, in that order.

3.2 Specifications. The SIO is displayed in the following manner:

The ICD and numeric code part are displayed first in a fixed field of thirteen positions. The high order position contains the ICD, followed by a space. The space is followed by a code containing nine numeric characters and two hyphens in varied display styles. The code part is separated from the name part by a space. The name part is displayed in a variable-length field in the format specified in Section 6.

EXAMPLES

1 NN-NNNN-NNN Univerix Corporation/
3 NN-NNN-NNNN Midwest Hardware Company/
9 N-NNNN-NNNN American Lethargic Association/

4. IDENTIFICATION CODE DESIGNATORS (ICD)

4.1 Concept. Identification Code Designators (ICD's) are intended to provide a reliable method for identifying or distinguishing code parts, thereby facilitating information interchange in those situations in which the records in a single file have not been assigned codes in a single code system. No single existing code system covers the full range of entities classifiable as organizations under the scope of this Standard, therefore,

the Identification Code Designators have been established to facilitate the interchange of information. Since there is more than one set of code systems in use, ICD's will facilitate differentiating among them.

4.2 Designations. ICD's have been assigned to two code systems in current widespread use in the United States (See Section 5) and one established in Canada. All selected ICD codes are odd code system numbers. Additionally, a separate user code is designated. All codes other than 1, 3, 5, and 9 are reserved for future use.

ICD	Name of Code	Abbreviation	Issuing Organization	Identification Code Styles: Issuance Style (right justified)	Identification Code Styles: Standard Display Style
1	Employer Identification Number	EIN	Internal Revenue Service	NN-NNNNNNN	NN-NNNN-NNN*
3	Data Universal	D-U-N-S	Dun and Bradstreet, Inc.	NN-NNN-NNNN	NN-NNN-NNNN
5	Statistics Canada Central Register	CDN-CRID	Statistics Canada	NNNNNNNN	NN-NNN-NN-N*
9	User Agreement Code for Organizations	UACO	-	Not prescribed	N-NNNN-NNNN

*EIN is reformatted by placing a hyphen in the 4th from the low order (last) position. This will facilitate human recognition of the numerics and conform to the eleven position format for identification codes. The CDN-CRID is also reformatted for display purposes using an extra hyphen to compensate for the missing ninth numeric character.

Note: All other ICD codes are reserved for future use.

5. CODE PART OF THE STANDARD IDENTIFIER FOR ORGANIZATIONS (SIO)

5.1 Code Part of the SIO. The code part of the SIO consists of one of the following codes:

5.1.1 Employer Identification Number (EIN) issued by the Internal Revenue Service, consists of nine numeric digits and one hyphen. When used, EIN's are to be preceded by ICD "1" and a space and communicated or displayed with two hyphens as follows:

1 NN-NNNN-NNN (See Appendix A for an explanation of the EIN.)

5.1.2 Data Universal Numbering System (D-U-N-S) issued by Dun and Bradstreet, Inc. consists of nine numeric digits and two hyphens. When used, the D-U-N-S code is preceded by ICD "3" and a space and communicated or displayed as follows:

3 NN-NNN-NNNN (See Appendix B for an explanation of D-U-N-S.)

5.1.3 User Agreement Code for Organizations (UACO). In the event that an EIN or D-U-N-S number is not sufficiently specific for the needs of the user, or has not been issued, or is not known, a UACO may be employed with prior agreement between users. When used, the UACO should be preceded by the ICD "9" and a space. Preferably the UACO should be 11 characters in length consisting of nine numeric digits and two hyphens, and be communicated and displayed as follows:

9 N-NNNN-NNNN (See Appendix D for an explanation of UACO.)

6. NAME PART OF THE STANDARD IDENTIFIER FOR ORGANIZATIONS (SIO)

6.1 General Syntax Rules.

6.1.1 Special Use Characters. For the transmission and display of names for information interchange purposes, the following conventions must be observed:

Period (.) - A period is not used even if part of the name.

Decimal Point (.) - When the decimal point (.) is part of the organization name, substitute "PNT" for the character itself.

Slant (/) - The slant is the terminal character which ends the name part of the SIO. If this character is part of the name, substitute "SLT" for the character itself.

Number Sign (#) - The number sign terminates the record name within the name part of the SIO when other names are displayed. If this character is part of the name furnished by the organization, substitute "NBR" for the character itself.

Space ( ) - A space is used as a separator only between the ICD and the code part, and between the code part and the name part of the SIO. Otherwise, it is used as it appears in the name.

Cent (¢) - When the cent sign (¢) is part of the organization name, substitute "CNT" for the character itself, since the character is not part of the ASCII set.

Semi-colon (;) - The semi-colon is only used in certain instances to distinguish organizational relationships as described in Section 6.2.2. Otherwise, it may not be used even if part of the name.

6.1.2 Other Characters. Characters other than those covered in 6.1.1 if part of the ASCII set are used only if a part of the name; any non-ASCII characters must be spelled out.

6.1.3 Articles. Do not use the articles "A", "An" and "The" when the article is the first word of the name. Use the word if it is not an article as in the case of an initial "A" in a person's name or when it is the trade style.

6.1.4 Formats. When formatting the name part of the SIO, abbreviations, capitalization, compound words, prefixes, titles, special symbols and numerals are to appear as provided by the organization:

EXAMPLES:

N J Grocery/
New Jersey Gro/
DeVinci Co/
D'Vinci Co/
D'Vinci Corp/
Van-Husen Iron Works/
Dr Doe Pain Killer/
NBR 17 St Louis Post/
3 PNT 2% Loan Company/
All Fuels SLT Natural Gas/

5 & 10 CNT Store/
Turnbull & Evans/
Independent Order of Ground Hogs/
American Lethargic Association/
Mid-West Hardware Co/
Midwest Hardware Company/
Cie Generale Transatlantique/
Luxury Island Hotel Corporation/
James W Smith Corp/
John Doe Supermarket NBR 11/

6.2 Specific Rules.

6.2.1 Recording of other names. The number sign (#) is used to separate the record name from other names, and both types of names will be formatted as stated in Section 6.1.4.

537-860 O - 74 - 29

EXAMPLES:

Smith & Jones#Smith & Jones Coal Co/
James W Doe#Doe Ford Agency/
Healy's Inc#Healy's Tours/
Healy's Inc#Healy's Tours#Healy's Travel Agency/

6.2.2 Names of divisions, subdivision, etc. To identify a specific branch, subsidiary, department, etc., to which a specific identification code has been assigned, transmit only the code and name associated with the specific branch, subsidiary, department, etc., to be identified. Assume that identification codes have been assigned to each of the following:

(No code part) Southwest Ohio Works
of
9 3-4567-8912 Metal Fabricating Division
of
9 5-6789-1234 All Purpose Manufacturing Corp
of
9 7-8912-3456 Universal Industries

To communicate regarding Metal Fabricating Division, transmit "9 3-4567-8912 Metal Fabricating Division/". Similarly, for any other line in the example use only the unique identifying code and its name.

When the sub-entity does not have a unique identifier a user must either assign a UACO or use the code of a higher order organization for which there is a numeric code. If the latter, use the name of the organization for which the code is provided, properly formatted followed by a semi-colon (;) then the name of the sub-entity properly formatted.

9 3-4567-8912 Metal Fabricating Division;Southwest Ohio Works/

APPENDIX A

EMPLOYER IDENTIFICATION NUMBER (EIN)

(This appendix is not a part of American National Standard Structure for Identification of Organizations for Information Interchange, but is included for information purposes only.)

To facilitate the keeping of Government records, every organization subject to taxes is assigned a nine digit identification number separated as follows:

NN-NNNNNNN

An organization may obtain Form SS-4 from any Internal Revenue Service office or the nearest district office of the Social Security Administration to apply for an identification number.

Each organization should have only one identification number. If an organization has more than one number and has not been advised which one to use, it should notify the Internal Revenue office where it files its return of the numbers it has, the name and address to which each number was assigned, and the address of its principal place of business. The Service will then advise it which number to use.

The EIN consists of nine numeric characters with the first two characters being numeric code for the district office in which the organization was located at the time of issuance. When recording the number, IRS places a hyphen between the second and third character from the high order position. In this standard, it is recommended that a hyphen also be placed between the sixth and seventh characters from the high order position, in order to facilitate transcription, provide a fixed-size in the code part of the Standard Identifier, and distinguish this code from other codes.

Excerpt from INTERNAL REVENUE MANUAL P-1200-26 (Approved 10-9-69):

"Certain assistance on a reimbursable basis to be given to non-Service activities using Employer Identification Numbers.

1. When an outside activity provides an EIN and name from its own records, the Service will inform such organizations whether these are consistent with data on Service records.

2. When the EIN is not known to the outside activity, but it can provide to the Service a fully-executed Form SS-4, initiated or certified by the organization requiring a code, the Service will determine whether an EIN had been previously assigned. If not, an EIN will be assigned if required for Federal tax purposes. Both the requesting activity and the organization being identified will be notified of the EIN's assigned or of the Service's negative action.

3. The work described above should be undertaken only when it does not interfere with regular Service processes and will ordinarily be on a reimbursable basis.

4. To avoid the disclosure of confidential information, the information furnished by the Service to non-Service organizations will be limited to that described."

Note: See Section 5.1.1 for EIN usage in this Standard.

APPENDIX B

DATA UNIVERSAL NUMBERING SYSTEM (D-U-N-S)

(This appendix is not a part of American National Standard Structure for Identification of Organizations for Information Interchange, but is included for information purposes only.)

The Data Universal Numbering System is an Identification Code System owned and maintained by Dun & Bradstreet, Inc., 99 Church Street, New York, N.Y. 10007.

In this system, D-U-N-S numbers are assigned to "establishments" where there is a sufficient community of interest among D-U-N-S users to merit the assignment of a D-U-N-S number. The word "establishment" in the D-U-N-S system context refers to a single physical location, either an operating location or a nonoperating headquarters address. Therefore, an organization may consist of just one "establishment" to which only one D-U-N-S number is assigned or several "establishments", each with its own D-U-N-S numbers. These may be located at one or multiple locations and may use only one or multiple names. When requested by an establishment, a D-U-N-S number is also assigned to each so called "payment address" - a special address such as a lock-box number to which checks in payment of invoices are mailed. D-U-N-S numbers have been assigned to over 3,000,000 "establishments" in the U.S., Canada and United Kingdom.

A D-U-N-S number is a randomly assigned nine digit number, the low order (last) digit being a "Mod 10, double-one-double-one check digit". D-U-N-S numbers are formatted as follows:

12-345-6789

The first 100,000 numbers (00-000-000X through 00-099-999X where the X is the check digit) have been reserved for the individual use of D-U-N-S subscribers (users). These numbers may be assigned internally to "establishments" to which Dun and Bradstreet has not assigned D-U-N-S numbers because of an insufficient community of interest or for other reasons.

While the D-U-N-S number is a completely random number, representing one "establishment" at one location, and has no built-in significance, Dun & Bradstreet does maintain it in its computerized D-U-N-S records "pointers" which enables it to furnish "family tree" type of data for all organizations to which more than one D-U-N-S number has been assigned to identify its various branches, subsidiaries, division, departments, etc. having a need to be separately identified.

An annually revised Alphabetical D-U-N-S Code Book is available to users in microfiche form. Supplements are published periodically throughout the year, with each containing cumulative additions. The constant updating is made possible by using Dun & Bradstreet national network of about 150 offices and about 12,000 reporters and correspondents.

Dun & Bradstreet makes no charge for assignment of a D-U-N-S number to an organization or its establishments and encourages each organization to print its D-U-N-S number on its checks, invoices, etc. Dun & Bradstreet does have a standard schedule of charges to its customers for use of the D-U-N-S Code Book and for performance of supplemental services desired by users when converting to D-U-N-S.

Identification Code Designator "3" has been assigned to signify the use of a D-U-N-S number as part of the Standard Identifier.

Note: See Section 5.1.2 for D-U-N-S usage in this Standard.

APPENDIX C

CONSIDERATIONS RELATING TO NAME PART OF THE STANDARD IDENTIFIER FOR ORGANIZATIONS (SIO)

(This appendix is not a part of American National Standard Structure for Identification of Organizations for Information Interchange, but is included for information purposes only.)

Dropping of Characters From Names. The dropping of characters from Names is not recommended. However, if users elect to reduce the length of such names, this should be accomplished through truncation rather than through abbreviation; i.e., by dropping off characters at the end of a name rather than by dropping characters from the beginning or middle of a name. When truncated records are interchanged, the transmitting organization should advise the receiving organization that truncation has occurred and the details of the truncation.

An analysis of the actual count of the number of characters in the names of all organizations in the State of Rhode Island as listed in Dun & Bradstreet's Reference Book is shown below. This provides a meaningful indication of feasible truncation points.

DUN AND BRADSTREET ANALYSIS OF ORGANIZATION NAMES IN THE STATE RHODE ISLAND

Number of Characters in Name	Frequency	Percent	Cumulative Percent
30 or less	15,743	95.73	95.73
31	148	0.90	96.63
32	112	.68	97.31
33	80	.49	97.80
34	82	.50	98.30
35	65	.40	98.70
36	43	.26	98.96
37	40	.24	99.20
38	28	.17	99.37
39	25	.15	99.52
40	25	.15	99.67
41	13	.08	99.75
42	10	.06	99.81
43	4	.02	99.83
44	6	.04	99.87
45	5	.03	99.90
46	5	.03	99.93
47	1	.01	99.94
48	1	.01	99.95
49	3	.02	99.97
51	2	.01	99.98
59	1	.01	99.99
65	1	.01	100.00
Total	16,443	100.00	

Summary:

95.73% of organization names had 30 or fewer characters.
4.27% had 31 or more characters.

Abbreviations. In recording a name caution should be taken to use abbreviations only when they are a formal part of the name.

Use of Name Formats. Some users may find it advantageous to utilize the Name Formats contained herein other than as part of the Standard Identifier for Organizations. The format rules were designed with this possibility in view.

Unavailability of Complete ASCII Character Set. If a user's system does not include the full range of characters contemplated by this standard, e.g., user's equipment has upper case capability only; he must substitute for unavailable characters. Transmitters in such cases, must advise receivers regarding details of the substitutions. It should be recognized that such substitution may result in a loss of information and should be avoided.

APPENDIX D

CONSIDERATIONS RELATING TO CODE PART OF THE STANDARD
IDENTIFIER FOR ORGANIZATIONS (SIO)

The potential universe of organizations as described in Section 2.1 of the proposed standard is quite large; covering at least 12 million entities. The number of organizations described in Section 2.1.6 could range from several hundred thousand to several million, depending on how deeply into the organizational structure the SIO will be applied.

Ideally, the code part of the SIO should be drawn from a single code system. However, it is presently impractical to design a new system and issue codes to all organizations coverable under this standard, because of the large number of organizations involved, the difficulties of finding them, and the attendant costs of code issuance, maintenance and updating. At this time, no government or private organization exists which could even potentially provide these services. Accordingly, the only realistic solution is to use an already-established system. Unfortunately, there is no single code system in widespread use which covers the full range of entities described in Section 2.1. Under the circumstances, it has been necessary to designate more than one identification code system as acceptable within this standard and to prescribe a standard method of using these systems. In some instances it may even be necessary to retain both codes in a file.

The use of existing systems subjects the standard to the limitation that 100 percent of the organizations potentially coverable are not assigned a code, since code issuance is based on the ground rules of each code issuer. However, the coverage of the EIN and D-U-N-S number systems is quite broad with respect to the categories of organizations described in Section 2.1 of the standard so that in many cases at least one of these codes has been assigned to an organization which must be identified in connection with information interchange. In instances in which no other code is applicable or available, formats for User Agreement Codes have been provided.

The suitability of existing code systems was evaluated on the basis of the following criteria:

- The system should apply to a broad segment of the potential universe.
- The code should be maintained on a current basis.
- The organizations involved should be aware of the code assignments.
- The code should be non-significant, in order to preclude the need for revisions in code assignments as characteristics change, and to avoid disclosures of confidential information.
- There should be the ability to check the accuracy of recording, storage and interchange of the code.

While no one of these code systems meets all of the above criteria, those designated do meet more than any other now in use or proposed for study, while at the same time having a large volume of numbers now in existence.

It is recommended that the UACO be 11 characters in length, but it is recognized that some user groups may find it necessary to use codes of more or less characters in length.

X3L84/104
72-02-09

DRAFT

AMERICAN NATIONAL STANDARD

STRUCTURE FOR THE IDENTIFICATION OF NAMED POPULATED PLACES AND RELATED ENTITIES OF THE STATES OF THE UNITED STATES

This draft standard has been accepted by the ANSI Board of Standards Review for public comment. It has been designated as X3.47 for reference purposes and will carry this number when approved.

SPONSOR

COMPUTER AND BUSINESS EQUIPMENT MANUFACTURERS ASSOCIATION

APPROVED (DATE)

AMERICAN NATIONAL STANDARDS INSTITUTE

TABLE OF CONTENTS

AMERICAN NATIONAL STANDARD
STRUCTURE FOR THE IDENTIFICATION OF
NAMED POPULATED PLACES AND RELATED ENTITIES
OF THE STATES OF THE UNITED STATES

1. PURPOSE AND SCOPE

1.1 Purpose. This Standard provides the structure for an unambiguous, concise code which will uniquely identify named populated places and related entities of the States of the United States for the purpose of information interchange among data systems.

1.2 Scope. The coverage of the standard includes codes for named populated cities, towns, villages, and similar communities, whether or not incorporated, and several categories of named entities that are similar to these in one or more important respects.

2. DEFINITIONS

Because of the large number (over 130,000) and varied character of the named places included, no single formal definition or criterion can be stated for inclusion in the standard. Instead, separate definitions are given for eighteen classes of entities included, and an entity meeting any one of these definitions is subject to inclusion in the standard (see 3.4 below and Appendices A and B).

In addition to incorporated and unincorporated named populated cities, towns, and villages, the standard provides codes for scattered rural communities, important military and naval installations, townships in the States where such units have governmental powers, Indian Reservations, national and State parks, named places that form parts of other places as defined, and named places with no permanent residents but important for transportation, industrial, or commercial purposes, such as unpopulated railroad points, airports, and shopping centers. The common characteristic of these varied types of entities is that all of them are recognized as named places by a significant segment of the public. In other words, for each class of entity there is an important group of users who would expect to find it included in a standard place code. The reverse is also true--there are users who will not wish certain classes of entitites included for their special purposes. Therefore the classification of entity by type is a very important adjunct to use of the standard, since it permits users to select those types of entities which fit their own particular conception of "populated places" (see 3.4 below).

3. SPECIFICATIONS

3.1 Format. Each named place within State is assigned a standard code consisting of five digits. To permit unique identification of a place within the United States, this five digit code is used in conjunction with the two-character State abbreviation or code as provided in American National Standard for Identification of States of the United States (including the District of Columbia) for Information Interchange, X3.38. (This Standard is identical to Federal Information Processing Standard 5-1, States and Outlying Areas of the United States, June 15, 1970).

The standard representation of the code is in the form AANNNNN or NNNNNNN (A = alphabetic character and N = numeric character) where the first two characters represent the State abbreviation or code and the last five numeric characters represent the place within a State. For example, the code for Jamestown, Virginia might be represented as VA44400 or 5144400 where VA and 51 are the State abbreviation and code and 44400 might be the place code for Jamestown.

A separator between state and place representations is not required when interchanging data among data processing systems in machine-sensitive form. However, if visual separators are needed to facilitate human understanding and readability, a hyphen (-) or a space may be displayed between the State representation and the place code (e.g., NN-NNNNN, NN NNNNN, AA-NNNNN or AA NNNNN).

The entities and specific codes provided for by this standard are not appended, but will be published separately.

3.2 Characteristics of the Code.

3.2.1 Uniqueness. A unique, one-to-one correspondence exists between each assigned code number and the named place to which it is assigned. For each place name there is only one code number and for each code number there is only one place name. Former or alternative names for the same place may be assigned their own unique codes, which are also cross referenced to the current standard name (see 3.3.1).

3.2.2 Conciseness. The code contains sufficient characters to achieve uniqueness and to provide for future expansion to accommodate a substantial number of additional entities without destroying the characteristic of alphabetic assignment.

3.3 Cross References.

3.3.1 Cross References to Former or Alternative Names. If a place is known by two (or more) names, either through the concurrent official or unofficial use of more than one name, or as a result of an official change of name, each of the names is assigned a unique code. However, the standard determines for each such place a current preferred name--the current name in case of a change of name, or one of the current official names when there is more than one, following specific rules (see 4. below). Codes for all names other than the current preferred name are cross referenced to the current preferred name by means of the latter's code, in a special cross reference column or field.

3.3.2 Cross References from "included" to "including" places. Many codes are assigned to places that meet the criteria for coding but that can also be determined to be parts of other places coded. Such places are assigned their own codes, but are also cross referenced to the place within which they are included, by means of the latter's code in a special "inclusion" column or field. This permits users either to recognize such "inclusions" as separate places or to combine them with their parent place, as best suits their particular requirements.

3.4 Class Designators. Not a formal part of this standard, but designed to be used in close conjunction with the code is a single-letter designator which serves to categorize the individual entities into one of a number of classes that collectively make up "named populated places and related entities". By indicating under which class a particular entity falls, the designator also indicates which of the class definitions the entity has met to permit it to have a code assigned. Attached is a list of the classes and their class designators (Appendix A), and definitions for each class describing the kind of entities it includes (Appendix B). The list of classes is annotated to indicate the extent to which each class is exhaustive of all entities meeting the definition.

Some classes (such as Incorporated Places, or National and State Parks) are determined by legal or official characteristics so that in principle all entities meeting the definition can be readily included. The other classes are designed to include all of the more important entities meeting the stated criteria. In practice, an entity meeting the criteria, but not included would be subject to addition on the basis of a demonstrated need for separate recognition by a significant portion of the public. An additional class (designated X) is comprised of the codes assigned to alternate and former names of entities (see 3.3.1).

4. RULES FOR DETERMINING THE CURRENT NAME

4.1 If the place or related entity contains a main post office, the post office name is used as current preferred name. If the corporate name, railroad name, or other names differ, they are treated as cross references to this name.

4.2 If the place does not contain a main post office, but is an incorporated municipality, the official corporate name is the current preferred name.

4.3 If the place does not contain a post office and is not an incorporated municipality, but does contain a named railroad point, the name used by the railroad is the current preferred name. Exceptions may be made in certain cases where it is clear that the railroad name is not the name in customary local use.

4.4 If a place does not meet any of the above criteria, the current preferred name is the name in most common current local use, as determined by the maintenance agent. In making this determination the maintenance agent will utilize official maps, responses to inquiries directed to postmasters and other local officials, and other names in present or past official use, such as branch postal names, names of discontinued post offices, and names of discontinued railroad points.

5. MAINTENANCE

5.1 Assignment. The assignment of code numbers will be administered by the maintenance agent, Rand McNally and Company, who will obtain information necessary to determine whether a given entity meets the criteria for one of the entity classes, and will interpret the definitions as may be required. The maintenance agent will also determine current preferred names and the status of former or alternative names. Appendix C provides guidelines which will be used in assigning codes to the original list of entities, maintaining an alphabetic sequence while allowing the maximum amount of space possible for the addition of future names in the same sequence.

As further entities qualify for inclusion, the maintenance agent will assign codes and class designators. The added codes will not necessarily be those exactly half way between the existing codes for the two adjacent names in the list, but will be positioned according to the best judgment of the maintenance agent as to where space for further additions should be left available, based on the spellings of the names involved and the known frequency of occurrence of specific letter combinations.

In the file as defined and described, saturation is unlikely to occur unless through the addition of large quantity of entities not now within the United States, or not now covered by any one of the defined classes of entities. Neither of these eventualities seems at all probable. However, in the unlikely event that saturation should occur in some portion of the sequence, each code assigned in that portion will be as close as possible to the one that would place the entity in its proper alphabetic position.

5.2 Dissemination. The maintenance agent will maintain a record of code numbers assigned. Specifically, it will either undertake or cooperate in an undertaking to publish a directory containing all the entities and their codes, to be updated between editions by frequent supplements. It will also either undertake or cooperate in an undertaking to prepare and maintain a tape record of the entities and codes. Both the published directory and the tape record will include complete cross reference codes (see 3.3). Both will also, subject to economic feasibility, include where appropriate other codes presently in use for named populated places and related entities, such as the Standard Point Location Code, the GSA Geographic Location Code, the 1970 Geographic Identification Code Scheme of the Bureau of the Census, and the ZIP Code. The maintenance agent will promote the use of this ANSI standard by encouraging the publishers of directories of named places to incorporate the code numbers in such directories.

APPENDIX A

CLASSES OF ENTITIES

CLASS DESIGNATORS. Not a part of the standard code, but designed to be used in close conjunction with it is a single-letter designator which serves to categorize the individual entities into one of a number of classes that collectively make up "named populated places and related entities". By indicating under which class a particular entity falls, the designator also indicates which of the class definitions the entity has met to permit it to have a code assigned. The following pages contain a list of the classes and their class designators. Definitions for each class describing the kind of entities it includes are included in Appendix B. The list of classes is annotated to indicate the extent to which each class is exhaustive of all entities meeting the definition.

Some classes (such as Incorporated Places, or National and State Parks) are determined by legal or official characteristics so that in principle all entities meeting the definition can be readily included. The other classes are designed to include all of the more important entities meeting the stated criteria. In practice, an entity meeting the criteria, but not included would be subject to addition on the basis of a demonstrated need for separate recognition by a significant portion of the public. An additional class (designated X) is comprised of the codes assigned to alternate and former names of entities (see 3.3.1).

It is stressed that the class designator is presented simply as a convenient adjunct, and is not a formal part of this standard nor an integral part of the code. Specifically, a code assigned to a place is a unique identifier independent of the class designator, and extensive applications of the code might be made without making any use of the class designators.

APPENDIX A

CLASSES OF ENTITIES

Class Designator	Class	Degree of Exhaustiveness (See key below)
C	Incorporated places	*
U	Unincorporated populated places	**
K	Seasonally populated places	***
R	Rural communities	***
L	Military/Naval installations wholly or largely within incorporated places	**
M	Military/Naval installations wholly or largely outside incorporated places	**
S	Unpopulated transport points	*
F	Unpopulated industrial points	**
G	Shopping centers (not parts of other places)	**
A	Airports	**
D	Indian Reservations	*
N	National and State Parks	*
P	Places which are parts of incorporated places	***
Q	Places which are parts of populated unincorporated places	***
V	Townships associated with a locality of identical name	*
W	Townships not associated with a locality of identical name	*
Y	Townships wholly comprised within an incorporated place	*
T	Urban townships	*
X	Alternate and former names	

Degree of exhaustiveness or completeness of classes

*Exhaustive: original file should include every entity meeting the definition, and future maintenance should add any new or corrected entities that meet the definition.

**Complete for major instances: the criteria for inclusion in the original file are intended to include all the more important entities of the class; additional entities should be added if (a) they meet the criteria and (b) there is a demonstrated need for users of the standard to recognize them separately.

***Not complete: application of the criteria should result in including most or all of the significant entities; additional entities can be added that meet the criteria if there is a demonstrated need for users of the standard to recognize them separately.

APPENDIX B

DEFINITIONS OF ENTITY CLASSES

Listed below are the eighteen classes of entities mentioned in Paragraph 2 and Appendix A above, together with their class designators and definitions:

Class Designator	Definition
C	Incorporated place. A place incorporated as a municipality under the laws of its State, but excluding a) the incorporated "towns" of eight States (Connecticut, Maine, Massachusetts, New Hampshire, New York, Rhode Island, Vermont and Wisconsin); b) the townships of any State; and c) any place that is part of another incorporated place under this definition. Note: This class specifically includes places incorporated as cities and villages, places incorporated as boroughs (except in Alaska), and places incorporated as towns except in the eight States listed above. "Towns" in these eight States, though they possess some or all of the corporate powers common to incorporated municipalities elsewhere, are really extensive units which are geographically more comparable to townships in other States. They frequently include separate population concentrations that are well recognized locally. Accordingly they are not regarded as "incorporated places" but as townships (classes T, V, W, Y; see definitions below).
C	Inactive incorporated place. An incorporated place with no active governmental organs. An inactive incorporated place is considered to be an incorporated place as long as it has the legal power to reactivate its government at any time. Otherwise it is considered to be an unincorporated place (class U).
P	Incorporated subplace. A place that would otherwise qualify as incorporated but which is part of another incorporated place. An example is the city of Berry Hill within the metropolitan government of Nashville-Davidson, TN.
U	Unincorporated populated place. A concentration of population which a) has at least some permanent residents; b) has a name that is in common use locally to refer to it; and c) is not part of any incorporated place, or of another unincorporated populated place. Note: A place is not considered "populated" if it has only daytime (working) population but no permanent residents, or if it has only seasonal population but no year-round residents. To qualify as a populated place, a community must generally have a population concentration of at least eight non-farm households or 25 permanent non-farm residents. However, some communities with fewer residents may qualify if they have a post office, a railroad station, or one or more stores. An unincorporated populated place has boundaries and an areal extent. These boundaries are delimited by the maintenance agent,

537-860 O - 74 - 30

on the basis of the verifiable extent of a concentration of residences, and local opinion as to the extent of the area known by the community name.

K — Seasonally populated place. A place that would qualify as an unincorporated populated place (see definition above) except that it has no year-round residents.

R — Rural community. A geographical community of scattered residences which

a) has a name that is in common use locally to refer to it;
b) does not qualify as an unincorporated populated place; and
c) is not part of any incorporated place as defined above.

Note: A rural community has an areal extent but frequently lacks any precise boundaries. It either has no concentration of non-farm households, or one of less than eight households or 25 residents.

L, M — Military/naval installation. A named base or similar facility of the Department of Defense or one of its branches, that is included in a current listing of major bases and installations issued for general circulation by the Department of Defense or one of its branches.

Note: There is apparently no official Department of Defense definition for the class of entities commonly referred to as military installations, but an ad hoc definition is provided by the existence of regularly published listings of what are described as "major military installations" or in similar terms.

L, M — Military/naval installation wholly or largely within an incorporated place. An installation as defined above most or all of whose headquarters structures are within the limits of an incorporated municipality. An installation not meeting this criterion is treated as wholly or largely outside an incorporated place.

S — Unpopulated transport point. A named point officially recognized by a water, rail, motor, or pipeline carrier for purposes significant to transportation, and not within the limits of any incorporated or unincorporated place.

Note: Most such places are named railroad points listed in railroad tariffs but located in open country away from any settlement.

F — Unpopulated industrial point. A named factory, quarry, industrial park, or similar industrial facility recognized as a point of origin or destination for transportation, but not qualifying as an unpopulated transport point and not within the limits of any incorporated or unincorporated place.

G — Shopping center (not part of another place). A "planned center" qualifying for listing in a standard directory of shopping centers, and not within the limits of any incorporated or unincorporated place.

Notes:

(1) Shopping centers are normally included in the list if they have a named postal facility, or if they have at least one department store or at least 100,000 square feet of area (sufficient to qualify them for the "regional center" or "community center" categories of the standard directory).
(2) Shopping centers that are within the limits of an incorporated or unincorporated place may be included in the list, but are designated as "parts of" entries (class P or class Q).

A — Airport. An area of land or water that is used for landing or takeoff of aircraft, and is so recognized by the Federal Aviation Administration, provided that it is used by a scheduled commercial carrier or is officially recognized by such a carrier for tariff purposes, and also provided that it is not within the limits of any incorporated or unincorporated place.

Note: An airport located within the limits of another place but otherwise qualifying under the above definition is categorized as part of another incorporated or unincorporated place (class P or class Q).

D — Indian Reservation. An area officially so designated by the Bureau of Indian Affairs.

N — National Park, National Monument, etc. Areas officially so designated by the National Park Service.

N — State Park (etc.). An area officially so designated by the relevant agency in its State.

P — Part of an incorporated place. A place within the legal limits of an incorporated place (class C).

Q — Part of an unincorporated place. A place within the delimited limits of an unincorporated place (class U).

T, V, W, Y — Township. A geographical-political entity recognized as a township, plantation, or "town" (in one of eight States), which

a) has a well-recognized name and boundaries;
b) qualifies as a local government under the laws of its State; or
c) formerly qualified as a township, plantation, or "town" government at some date since 1900, and is not now part of an incorporated place or another township or "town".

Notes:

(1) This category is comprised chiefly of townships but also includes "towns" in eight States (Connecticut, Maine, Massachusetts, New Hampshire, New York, Rhode Island, Vermont, and Wisconsin) in which the term "town" is used for areally extensive units similar to the townships of other States.
(2) Townships and "towns" are never wholly or partly included in other townships or "towns". However, they may overlap areally with one or more incorporated and/or unincorporated places as defined above.
(3) Townships which may exist as administrative subdivisions of counties in certain States, but which have never exercised local government powers, are not included. Examples are the townships of North Carolina and California.
V (4) Townships are subdivided into four classes: class (V), containing a locality (or part of a locality) which has the same name as the township and qualifies for some other category of entity, but not wholly comprised in the locality; these townships have the same names as localities also listed, but have different extents, so that the two
W entries must be carefully distinguished; class (W) containing no
Y locality with the same name as the township; class (Y) wholly comprised within an incorporated place; and class T, urban townships (see below).

T — Urban township. A township or "town" such that, in the judgment of the maintenance agent the customary criteria for delimiting the boundaries of a populated unincorporated place define a community that conforms substantially to the boundaries of the township or "town". Criteria for recognizing a township or "town" as an urban unit include

a) it must not contain any part of an incorporated city, borough, or village;
b) its area must be served or substantially served by a single post office; and
c) it must not have more than 10% of its population living on farms or in other scattered residences, outside a population concentration or concentrations.

Note: Usually the generally recognized name of the populated place is the same as that of the township or "town", but this not a requirement.

T Urban county. A county which would qualify as an urban unit under the above definition of urban township, if the word "county" were substituted in the definition for "township".

Note: The only examples are Arlington, VA, and Los Alamos, NM.

APPENDIX C

GUIDELINES FOR THE ASSIGNMENT OF CODES

In the original assignment, codes will be spaced and assigned according to the following rules:

1. The first code used in each State will be 00100.

2. Codes from 90000-99999 will be left unassigned and therefore available for special applications by individual users of the code.

3. Except for one State which has too many entities to make it possible, codes will be spaced in terms of quantities divisible by 5.

4. In twelve States that currently contain many townships bearing identical names (such as Union, Madison, Washington), codes for such townships will be assigned with a spacing of 5 or 10, in recognition of the relative unlikelihood of any need to add further names in the unassigned spaces between them. These twelve States are asterisked (*) in the list below. Consecutive townships with identical names will be spaced 10 apart except in Ohio, where the spacing will be 5. The same treatment will be used when a locality and a township of the same name occur in the same county.

5. With these limitations, the codes will be spaced as far apart as possible while still providing for coding of all entities (approximately 130,000) in the initial entity list.

Spacing every 8:	Pennsylvania
Spacing every 15:	California, New York, *Ohio, Texas, Virginia
Spacing every 20:	*Illinois, Kentucky, *Minnesota, *Missouri
Spacing every 25:	Alabama, *Indiana, *Iowa, *Michigan, North Carolina, Tennessee, West Virginia
Spacing every 30:	Arkansas, Florida, Georgia, *Kansas, Maryland, New Jersey, *Wisconsin
Spacing every 40:	Louisiana, *North Dakota, Washington
Spacing every 50:	*Maine, Massachusetts, Mississippi, Oklahoma, South Carolina, *South Dakota
Spacing every 60:	Colorado, *Nebraska, Oregon
Spacing every 80:	Arizona, Idaho, Montana
Spacing every 100:	Alaska, Connecticut, New Hampshire, New Mexico, Utah, Vermont, Wyoming
Spacing every 200:	Delaware, Hawaii, Nevada, Rhode Island
Spacing every 500:	District of Columbia

6. If at a future date American Samoa, Canal Zone, Guam, Puerto Rico, or the Virgin Islands should become part of the United States and therefore subject to inclusion in the code, this plan for assignment could be applied without difficulty to their named populated places and related entities. For Puerto Rico, the spacing between entity names would be about 50; for each of the other four territories it would be at least 100. (These are the five territories for which appropriate codes have been left unassigned in the ANSI Standard for States of the United States (Including the District of Columbia), X3.38.)

Appendix F

Data Element Management Symposium

Symposium Committee

Mr. David V. Savidge PROGRAM CHAIRMAN	DATRAN, 8130 Boone Boulevard, Vienna, Virginia 22180
Mrs. Hazel E. McEwen SYMPOSIUM COORDINATOR	National Bureau of Standards, Room B226, Technology, Washington, D.C. 20234
Mr. James W. Gillespie	Department of the Navy, Code 916E, Washington, D.C. 20350
Mr. Harry S. White, Jr.	Institute for Computer Sciences and Technology, National Bureau of Standards, Washington, D.C. 20234
Mr. Arthur Wright	Bell Laboratories, BISP Area, Room 4C1045, Box 2020, New Brunswick, New Jersey 08903

Session Chairmen

Mr. Don Hetzel	Computing and Software Services, McGraw Hill Incorporated, Highstown, New Jersey 08520
Mr. Charles A. Phillips	Coopers & Lybrand, 3003 Van Ness Street, NW, Apt. W428, Washington, D.C. 20008
Mr. Bernard Radack	Internal Revenue Service, 1111 Constitution Avenue, N.W., Washington, D.C. 20224
Mr. Arthur Wright	Bell Laboratories, BISP Area, Room 4C1045, Box 2020, New Brunswick, New Jersey 08903

Guest Speakers

Mr. William Knox Director, NTIS	U.S. Department of Commerce, Room 3859, Washington, D.C. 20230
Mr. David B. H. Martin Special Assistant to the Secretary	Department of Health, Education and Welfare, Room 5517, North Building, 330 Independence Avenue, S.W., Washington, D.C. 20001

Presenters and Authors of Papers

Mr. T. M. Albert	Logistics Management Institute, 4701 Sangamore Road, Washington, D.C. 20016
Mr. Ronald B. Batman	Sperry Univac Computer Systems, 251 West DeKalb Pike, King of Prussia, Pennsylvania 19406
Mr. Charles Bontempo	IBM Corporation, Federal Systems Division, 8100 Frederick Pike, Gaithersburg, Maryland 20760
Mr. Morey J. Chick	U.S. General Accounting Office, 9226 Federal Building, 6th & Arch Streets, Philadelphia, Pennsylvania 19106

Mr. Perry Crawford, Jr.	Advanced Systems Development, International Business Machines Corporation, Yorktown Height, New York 10598
Ms. Edith F. Curd	Headquarters, U.S. Army Materiel Command, 5001 Eisenhower Avenue, Alexandria, Virginia 22304
Miss Sherron L. Eberle	Office of Information Services, Office of the Governor, Box 13224, Austin, Texas 78711
Mr. Charles E. Emswiler, Jr.	Commonwealth of Virginia, Division of Automated Data Processing, 201 Eighth Street Office Bldg., Richmond, Virginia 23219
Mr. Dave England	Office of Information Services, Office of the Governor, Box 13224, Austin, Texas 78711
Mr. M. J. Gilligan	Western Electric Co., Finance Division, Information Systems Engineering, Gateway II - Floor 15, Newark, NJ 07102
Mr. Edward A. Guilbert	Transportation Data Coordinating Committee, 1101 Seventeenth Street, N.W., Washington, D.C. 20036
Mr. Murray A. Haber	Department of Transportation, 4103 Wexford Drive, Kensington, Maryland 20795
Mr. Robert R. Hegland	Naval Command Systems Support Activity, Washington Navy Yard, Code 70.3, Washington, D.C. 20374
Mr. Carter Paul Heitzler	Commonwealth of Virginia, 8th Street Office Building, Room 201, Richmond, Virginia 23219
Mr. Robert Landau	Science Information Association, 2480 - 16th Street, N.W., Washington, D.C. 20009
Mr. Robert A. Longenbaugh	Colorado State University, Department of Civil Engineering, Fort Collins, Colorado 80521
Mr. Norval E. McMillin	Colorado State University, Room 225, Old Chemistry, Fort Collins, Colorado 80521
Ms. Patricia McNamee	Celanese Corporation, Box 1414, Charlotte, NC 28201
Mr. James W. Pontius	Finance and Service Operation, General Electric Company, Schenectady, New York 12345
Mr. Merle G. Rocke	Catepillar Tractor Co., Merchandising Research, Admin. Building, Peoria, Illinois 61601
Mr. Hasan H. Sayani	University of Maryland, Department of Information Systems Management, College Park, Maryland 20742
Mr. Bernard H. Schiff	Office of Information Services, Office of the Governor, Box 13224, Austin, Texas 78711
Dr. E. H. Sibley	University of Maryland, Department of Information Systems Management, College Park, Maryland 20742
Miss Sheila M. Smythe	Blue Cross-Blue Shield, 622 Third Ave., New York, NY 10017
Mr. David L. Swanz	International Business Machines, Federal Systems Division, 8100 Frederick Pike, Gaithersburg, Maryland 20760
Mr. Helmut E. Thiess	Naval Command Systems Support Activity, Washington Navy Yard, Washington, D.C. 20374

Mr. Cliff Tierney	Whirlpool Corporation, Benton Harbor, Michigan 49022
Dr. Dik Warren Twedt	University of Missouri, St. Louis, Missouri 63121
Mr. R. E. Utman	Princeton University Library, Box 200, Princeton, NJ 08540
Mr. Carroll P. Weber	Bank of America, Box 37000, San Francisco, CA 94137
Mr. Harry S. White, Jr.	Institute for Computer Sciences and Technology, National Bureau of Standards, Washington, D.C. 20234
Mr. Arthur Wright	Bell Laboratories, Inc., BISP Area, Room 4C1045, Box 2020, New Brunswick, New Jersey 08903

Attendees

Mr. Richard G. Abbott	World Bank, 110 Church Place, Falls Church, VA 22046
Mr. Theodore T. Abromavage	U.S. General Accounting Office, 13023 Blairmore Street, Beltsville, Maryland 20705
Mr. N. Richmond Alexander, Jr.	Allendale Insurance, Box 7500 - Allendale Park, Johnston, Rhode Island 02919
Mr. Bruce H. Allen	Office of the Secretary, Department of Transportation, TAD-251, 400 7th Street, S.W., Washington, D.C. 20590
Ms. Suzanne Ambler	MTMTS-SYC, 5611 Columbia Pike, Baileys Cross Roads, VA 22041
Ms. Dawn Anderson	State of Minnesota, Information Systems Division, Centennial Building, 5th Floor, St. Paul, Minnesota 55155
Mr. Martin E. Anderson	Standard Oil (Indiana), 200 E. Randolph Drive, Chicago, Illinois 60601
Mr. William Roy Anderson	NAVCOSSACT (Code 90.2), 811 Oxford Cntr., Waldorf, MD 20601
Mr. Harold C. Andrews	General Accounting Office, 441 G Street, N.W. - Room 5840, Washington, D.C. 20548
Ms. Barbara Ansell	Wyeth Labs, P.O. Box 597, Paoli, Pennsylvania 19301
Ms. Margie Lee Armstrong	National Military Command Systems Support Center B245, Pentagon, Room BE685, Washington, D.C. 20301
Mr. Fred C. Azzara	Wyeth Labs, P.O. Box 597, Paoli, Pennsylvania 19301
Mr. Frederick W. Babbel	National Archives & Records Service, General Services Admin., 8th & Pa. Avenues, N.W., Washington, D.C. 20408
Mr. Doyle O. Bare	Naval Training Equipment Center, Code N017, Orlando, FL 32813
Mr. Robert Barger	International Bank for Reconstruction & Development (World Bank), 1818 H Street, N.W., Washington, D.C. 20433
Mr. George M. Barlow	Cincom Systems Inc., 2181 Victory Pkwy., Cincinnati, OH 45206
Mr. Anthony J. Barr	North Carolina State University, Raleigh, NC 27607

Major Thomas H. Baumgartner	DCA/JTSA, ATTN: Code J320, 11440 Isaac Newton Square North, Reston, Virginia 22090
Mr. Paul E. Beard	U.S. Navy, Fleet Material Support Office, Code 9641, Mechanicsburg, Pennsylvania 17055
Mr. Joseph E. Beltramea	Data Bank Staff, Caterpillar Tractor Co., 100 N.E. Adams Street - (B-3), Peoria, Illinois 61611
Mr. Robert S. Benjamin	National Security Agency, 11118 Mitscher Street, Kensington, Maryland 20795
Mr. Jack Bennett	Veterans Administration, 5300 Clifton Street, Springfield, Virginia 22151
Mr. Thomas J. Bergin	U.S. Veterans Administration, 810 Vermont Avenue, N.W., Washington, D.C. 20420
Mr. Charles Berry	Dun & Bradstreet, 2233 Wisconsin Avenue, N.W., Washington, D.C. 20007
Mr. Maynard Y. Binge	Air Force Systems Command, HQ AFSC (ACDO), Andrews Air Force Base, Washington, D.C. 20334
Cdr. Michael E. Bishop	U.S. Navy, Naval Ship Research and Development Center, Code 1806, Bethesda, Maryland 20034
Mr. Geoffrey B. Blood	USAMSSA, SSD, RMB, Pentagon, Room B6662, Washington, D.C. 20310
Mr. Lewis Boger	National Oceanographic and Atmospheric Administration-NOS, 5021 Seminary Road - Apt. 725, Alexandria, Virginia 22311
Miss La Von Boisen	NAVCOSSACT, 2000 F Street, N.W., Washington, D.C. 20006
Dr. Eugene W. Bold	National Security Agency, 51 St. Andrews Road, Severna Park, Maryland 21146
Mr. Joseph B. Bourne	National Aeronautics and Space Administration, Goddard Space Flight Center, Greenbelt, Maryland 20771
Mr. F. R. Bower	IBM Corporation, 10309 Logan Drive, Potomac, Maryland 20854
Mr. Oswald S. Boykin, Jr.	Naval Ordnance System Command, Central NOMIS Office, ORD-04N42, Indian Head, Maryland 20640
Mr. Douglas D. Bradham	Office of the Chief Medical Examiner, Box 2488, Chapel Hill, North Carolina 27514
Cdr. David W. Bradley	Fleet Combat Direction Systems Support Activity - Dam Neck, Virginia Beach, Virginia 23461
Mrs. Joyce E. Brady	Data Use & Access Laboratories, 1601 N. Kent Street, Suite 900, Arlington, Virginia 22209
Mr. Edward Brand	U.S. Geological Survey, 19th & C Street, NW - Room 1442, Interior Building, Washington, D.C. 20009
Ms. Alice E. Brass	DCSC-DSAO, 4000 E. Broad Street, Columbus, Ohio 43230
Ms. Gloria Brauer	Empoyers Insurance of Wausau, 2000 Westwood Drive, Wausau, Wisconsin 54401

Mr. G. T. Brebach, Jr.	Arthur Andersen & Co., 69 W. Washington Street, Chicago, Illinois 60602
Mr. David Bridge	National Museum of Natural History, Room 414, Washington, D.C. 20560
Ms. Janet Brooks	Defense Communications Agency, HQ, Code 205. Washington, D.C. 20305
Mr. Cyril Brosnan	Office of the President, Blue Cross of Greater New York, 622 Third Avenue, New York, New York 10017
Mr. David E. Brown	Xerox Corporation, 025 Xerox Square, Rochester, New York 14618
Mr. George J. Brown	Department of Treasury, Main Treasury - Room 5116, 15th St. & Pennsylvania Ave., N.W., Washington, D.C. 20020
Mr. Frank J. Bruno	Veterans Administration, 810 Vermont Avenue, N.W., Washington, D.C. 20420
Mr. William F. Bryan	Automation Industries (Vitro), 14000 Georgia Avenue, Silver Spring, Maryland 20910
Mr. Richard Buchner	Selling Areas Marketing Inc., 541 N. Fairbanks Center, Chicago, Illinois 60614
Mr. F. E. Bush	Bureau of the Census, Social and Economics Statistics Administration, Washington, D.C. 20233
Mr. Max A. Butterfield	Social Security Administration, Room 2B2, Operations, Baltimore, Maryland 21235
Ms. Ruth Calabrese	U.S. Coast Guard, Dept. of Transportation, 400 7th St., SW, Washington, D.C. 20591
Mr. James R. Calkins	Sun Oil Co., 240 Radnor-Chester Rd., St. Davids, PA 19087
Mr. Lawrence Callahan	Litton Bionetics, Box B, Frederick, Maryland 21701
Ms. Margaret B. Canty	Wayne State University, 5925 Woodward, Detroit, MI 48207
Mr. Albert W. Carson	Department of Navy, 5823 Oak Grove Street, Lorton, VA 22079
Mr. William J. Cassidy	Civil Aeronautics Board, 1825 Connecticut Avenue, N.W.. Washington, D.C. 20428
Mr. Gary R. Catzva	Applied Physics Lab, John Hopkins University, 8621 Georgia Avenue, Silver Spring, Maryland 20910
Mr. Ralph F. Cautley	The Procter & Gamble Co., Ivorydale Technical Center, Cincinnati, Ohio 45217
Ms. Susan Cavanaugh	LEAA/Department of Justice, 633 Indiana Avenue, Washington, D.C. 20530
Mr. Carman A. Cellucci	Small Business Administration, 1441 L Street, N.W., Washington, D.C. 20416
Mr. Paul J. Cervelloni	Social Security Administration, Room 2B2, Operations, Baltimore, Maryland 21235

Mr. John J. Chambers	Research Foundation of State University of New York, 217 Lark Street, Albany, New York 12224
Mr. Daniel L. Chicklo	Blue Cross & Blue Shield of Delaware, 152 Worral Drive, Newark, Delaware 19711
Mr. Thomas A. Chittenden	Department of Transportation, Box 14263, Washington, D.C. 20044
Mr. John B. Christiansen	Independence Computing & Software, 235 White Horse Pike, W. Collingswood, New Jersey 08107
Ms. Frances Clark	Management Information and Data Systems Division, Environmental Protection Agency, Washington, D.C. 20460
Mr. John A. Clark	New Jersey State Department of Higher Education, 225 West State Street, Trenton, New Jersey 08625
Mr. Norman L. Clark	U.S. Naval Security Group Command, 3801 Nebraska Ave., NW, Washington, D.C. 20390
Mr. Dan C. Clarke	IBM Corp., Monterey & Cottle Roads (052/D86), San Jose, California 95111
Mr. John D. Clyde	Rohm and Haas Co., Independence Mall West, Philadelphia, Pennsylvania 19105
Mr. Fred J. Cole	Department of Labor, GAO Bldg. - Rm. 2870, 441 G Street, NW, Washington, D.C. 20548
Mr. Reginald Creighton	Smithsonian Institution, 10th & Jefferson Drive, Washington, D.C. 20560
Mr. Edward H. Crim	Defense Mapping Agency, Topographic Center, 6500 Brooks Lane, Washington, D.C. 20315
Mr. Alan S. Crosby	Library of Congress, 10 First Street, S.E., Washington, D.C. 20540
Mr. Donald P. Cruse	N.E. Regional Data Center, University of Florida, Gainesville, Florida 32611
Mr. Robert C. Curry	University of Louisville, Belknap Campus, Louisville, Kentucky 40208
Mr. Richard G. Davis	Bechtel Power Corporation, P.O. Box 607, Gaithersburg, Maryland 20760
Ms. Delphine F. Day	Department of Justice, LEAA, 633 Indiana Avenue, N.W., Washington, D.C. 20530
Mr. Paul S. Day	Farm Credit Administration, 485 L'Enfant Plaza, SW, Washington, D.C. 20578
Ms. Linda E. Deiter	U.S. Geological Survey, 12201 Sunrise Valley Drive, Reston, Virginia 22092
Mr. George T. Denison	Xerox Corporation, 800 Phillips Road - Bldg. 205C, Webster, New York 14580

Mr. W. H. Dennett	Aerojet Electro-Systems, 1205 Indian Springs, Glendora, California 91740
Mr. William K. Devany	U.S. Civil Service Commission, 1900 E Street, N.W., Washington, D.C. 20415
Mr. Ernest DeWald	HQ, Defense Mapping Agency, Bldg. #56 Naval Observatory, Washington, D.C. 20305
Mr. D. F. DeWolfe	Administration Information Systems, National Cash Register, Main & K Streets, Dayton, Ohio 45479
Mr. Kenneth DeYoe	U.S. Geological Survey, Computer Center Division, 12201 Sunrise Valley Drive, Reston, Virginia 22092
Mr. Philip H. Diamond	Veterans Administration, 810 Vermont Avenue, N.W., Washington, D.C. 20420
Mr. Edward Digon	Bureau of Program Evaluation, Pennsylvania Department of Health, P.O. Box 90, Harrisburg, Pennsylvania 17120
Mr. James D. Dillon, Jr.	Research Foundation of State University of New York, P.O. Box 7126, Albany, New York 12224
Mr. Sol Dolleck	Systems Development Officer, ISPC, Bureau of the Census, Washington, D.C. 20233
Mr. Francis D. Donaghy	Naval Air Eng. Center, Naval Base, Philadelphia, PA 19112
Ms. Christine Dougherty	World Bank, 1818 H Street, N.W. - #N463, Washington, D.C. 20433
Mr. James P. Doyle	Bureau of Economic Analysis, 2400 M Street, N.W., Washington, D.C. 20230
Mr. Richard C. Dubuque	Information Systems Analyst, Central Intelligence Agency, Washington, D.C. 20505
Mr. Michel Dufresne	Quebec Health Insurance Board, 200 Chemin Ste-Foy, Quebec City, CANADA
Mr. George P. Duncan	Department of Transportation, Room 10317, Nassif Bldg., 400 7th Street, S.W., Washington, D.C. 20590
Mr. Hallet A. Duncan	Bureau of Hearings & Appeals, Social Security Administration, 801 Randolph Street, Arlington, Virginia 22203
Mr. Robert J. Dunn, Jr.	Florida Department of Education, Knott Building, Tallahassee, Florida 32301
Mr. William C. Dunn	The Rand Corporation, 1700 Main Street, Santa Monica, California 91335
Ms. Trinka Dunnagan	Conduit, 100 LCM, Iowa City, Iowa 52242
Mr. Eugene M. Dwyer	Agency for International Development, SA-12, Room 725, Washington, D.C. 20523
Mr. Cletus L. Eadie	General Services Administration, Automated Data & Telecommunications Service, CPPS, Washington, D.C. 20405

Mr. Howard Edels	Mathematica, P.O. Box 2392, Princeton, New Jersey 08540
Dr. Joanne M. Egan	Air Products & Chemicals, Box 538, Allentown, Pennsylvania 18104
Mr. James R. Egenrieder	AMP Incorporated, Box 3608, Harrisburg, Pennsylvania 17105
Ms. Alberta R. Eidman	Social Security Administration, BDP Systems, 6401 Security Blvd. - Room 3-L-30, Baltimore, Maryland 21235
Ms. Margaret Ann Eldridge	Central Intelligence Agency, #731, 4600 S. 4 Mile Run Drive, Arlington, Virginia 22204
Dr. Jess P. Elliott	Georgia State Department of Education, State Office Building, Atlanta, Georgia 30344
Mr. James H. Ellison	Social Security Administration, 6401 Security Blvd., Baltimore, Maryland 21235
Mr. Richard B. English, III	Research Foundation of State University of New York, Box 7126, Albany, New York 12224
Ms. La Verne Erwin	Department of Entomology, Smithsonian Institution, NMNH, Washington, D.C. 20560
Mr. Joel Farley	Xerox Corporation, 800 Phillips Road, Webster, New York 14580
Mr. Alfred Feldman	Division Biometrics, Walter Reed Army, Institute of Research, Washington, D.C. 20012
Mr. Jim Ferguson	U.S. Air Force, HQ USAF/ACDB, Washington, D.C. 20330
Mr. Wayne B. Ferrell	U.S. Army, Battletown Drive, Berryville, Virginia 22611
Mr. George B. Fineberg	Comptroller of the Navy (NCF-132), Room 424, Crystal Mall, Building 3, Arlington, Virginia 20376
Mr. Robert G. Fish, Jr.	National Ocean Survey, 6001 Executive Blvd., Rockville, Maryland 20851
Mr. Robert H. Follett	IBM Corp., 10401 Fernwood Road, Bethesda, Maryland 20034
Mr. Donald R. Fooks	Social Security Administration, 6401 Security Blvd. - Room 2-B-2, Baltimore, Maryland 21235
Mr. Andre Fournier	AMA & DPMA, P.O. 6600, Quebec, CANADA
Mr. James J. Fraher	Social Security Administration, 6401 Security Blvd., Baltimore, Maryland 21235
Mr. Kenneth R. Freeman	State of Louisiana, Office of Information Services, P.O. Box 44335, Baton Rouge, Louisiana 70804
Mr. Richard Fredette	Chief of Naval Operations, Code OP-916D, Washington, D.C. 20350
Mr. Thomas F. Fussell	Mallinckrodt Chemical Works, 2nd & Mallinckrodt, St. Louis, Missouri 63160

Mr. Eugene F. Gallagher	Naval Material Command Support Activity, (SA-14122), Washington, D.C. 20360
Mr. Walter L. Galson	HQ AMC, ATTN: AMCRD-EM, 5001 Eisenhower Avenue, Alexandria, Virginia 22304
Mr. John G. Garnish	Social Security Administration, Box 89A, Route 1, Oak Hill, Drive, Sykesville, Maryland 21784
Dr. T. Gary Gautier	National Museum of Natural History, Washington, D.C. 20560
Mr. Howard S. Geer	Montgomery College, 51 Mannakee Street, Rockville, MD 20850
Mr. Fred W. Gibb	Xerox Corporation, 800 Phillips Road - Bldg. 205C, Webster, New York 14580
Mr. John E. Giddo	Defense Supply Agency, Cameron Station, Alexandria, VA 22314
Ms. Concetta V. Gilbert	Corp. of Engr. (DAEN-DSU-DI), Forrestal Bldg., Washington, D.C. 20301
Mr. John S. Giltner	U.S. Army - MILPERCEN, 3593 N. Forestdale, Woodbridge, Virginia 22191
Mr. Charles M. Goldstein	Informatics Inc., 6000 Executive Blvd., Rockville, MD 20852
Ms. Jean W. Gordon	U.S. Civil Service Commission, 1900 E Street, N.W. - Room 6425, Washington, D.C. 20415
Mr. John Lawton Gore	S.C. Department of Social Services, P.O. Box 1520, Columbia, South Carolina 29202
Mr. David B. Gottlieb	Bureau of Labor Statistics, 441 G Street, N.W. - Room 2018, Washington, D.C. 20212
Mr. Laymon Gray	Department of Education, Florida, Knott Building, R #165, Tallahassee, Florida 32304
Mr. Leonard B. Greene	Informatics, Inc., Suite 100, 5501 Cherokee Avenue, Alexandria, Virginia 22312
Major J. P. Greeves	USMC Member, LOGDESMO, 8322 Fort Hunt Rd., Alexandria, Virginia 22308
Mr. R. J. Grill	Veterans Administration, 7 Dabney Center, Rockville, Maryland 20853
Ms. Etelle Grinoch	Sperry Div. Sperry Rand Corp., 15 Juneau Blvd., Woodbury, New York 11797
Mr. Richard E. Grove	University Computing Activities, Virginia Commonwealth Univ., 1015 Floyd Avenue, Richmond, Virginia 23220
Mr. William F. Grubb	J. P. Stevens & Co., Inc., P.O. Box 1566, Greenville, South Carolina 29602
Mr. D. C. Gurtner	U.S. Veterans Administration, 810 Vermont Avenue, N.W., Washington, D.C. 20420
Mr. Eric R. Haars	Westinghouse, 19116 Brooke Grove Court, Gaithersburg, Maryland 20760
Mr. Merrill F. Hadley	Dept. of Navy-NAVCOSSACT, 112 James Dr., SW, Vienna, VA 22180

Mr. Francis X. Hammett	Department of Navy, Code: NAVMAT 0142;CP #5, Washington, D.C. 20360
Mr. Romaine F. Harley	Civil Aeronautics Board, 1825 Connecticut Avenue, Washington, D.C. 20428
Mr. Donald F. Harrison	National Archives (NNPD), Washington, D.C. 20408
Ms. Janet M. Harryman	Social Security Administration, Room 612, East Building, 6401 Security Blvd., Baltimore, Maryland 21235
Mr. Clayton E. Hatch	National Highway Traffic Safety Admin., Department of Transportation, 2100 Second Street, Washington, D.C. 20590
Mr. W. Scott Haynie	Western Union, 82 McKee Drive, Mahwah, New Jersey 07430
Mr. Larry Heffel	Department of Agriculture, ASCS-DS Division, 14th & Independence Avenue, S.W., Washington, D.C. 20250
Ms. Joan Held	New York City Rand Institute, 545 Madison Avenue, New York, New York 10022
Mr. Clifford W. Hieta	Technomics Inc., 2936 Chain Bridge Road, Oakton, VA 22124
Mr. Everett H. Higgins	U.S. Army - DCSPER, 6816 Felix Street, McLean, VA 22101
Mr. Herbert J. Hirshenberger	DHEW-SRS-OIS, 3rd & C Streets, S.W., Washington, D.C. 20201
Mr. Aaron Hochman	Department of Defense, Room 7569, Hoffman Bldg. II, Alexandria, Virginia 22332
Mr. George M. Hodge	Western Electric, Department 31654, 3300 Lexington Road, Winston-Salem, North Carolina 27102
Mr. Ernest M. Hodges	Farm Credit Administration, 485 L'Enfant Plaza West, SW, Washington, D.C. 20578
Mr. William L. Hoerbert	Defense Mapping Agency, Topographic Center, 6500 Brooks, Washington, D.C. 20015
Mr. Robert R. Hohl	Code 532, Goddard Space Flight Center, Greenbelt, MD 20771
Ms. Jeannie Holdahl	NGS, NOAA, C515, WSC #2, 6001 Executive Boulevard, Rockville, Maryland 20852
Mr. Paul L. Holm	U.S. Department of Agriculture, Farmers Home Administration, 12th & Independence Avenue, S.W., Washington, D.C. 20250
Mr. Peter Hsueh	World Bank, 1818 H Street, Room N403, Washington, D.C. 20433
Mr. Michael A. Huffenberger	Chemical Abstracts Service, Ohio State University, Columbus, Ohio 43210
Mr. Douglas G. Hughes	Cincom Systems, Inc., 2181 Victory Parkway, Cincinnati, Ohio 45206
Mr. Luis Hurtado-Sanchez	Hewlett-Packard, 1501 Page Mill Road, Building 3L, Palo Alto, California
Mr. Harry R. Jacknow	Central Intelligence Agency, 12012 Greywing Square #C4, Reston, Virginia 22091

Mr. Robert R. Jacobus	Bureau of Naval Personnel, 107 W. Poplar Road, Sterling, Virginia 22170
Mr. Robert T. Janshego	Bureau of the Census, SESA, Washington, D.C. 20233
Mr. Ralph A. Jenkins	Computer Sciences Corporation, 1611 Simmons Drive, McLean, Virginia 22101
Mr. Edward J. Johnson	The Research Foundation of State University of New York, 411 State Street, Albany, New York 12203
Mr. Gary B. Johnson	Motorola Semiconductors, 6900 E. Camelback Road - CV901, Scottsdale, Arizona 85251
Mr. Joseph P. Johnson	Federal Deposit Insurance Corporation, 550 17th Street, N.W., Washington, D.C. 20429
Mr. Clyde E. Jones	USDA-Forest Service, Room 4230, South Building, 14th & Independence Avenue, SW, Washington, D.C. 20250
Mr. Michael T. Jones	Southern Services, Inc., 64 Perimeter Center East, Atlanta, Georgia 30346
Mr. Paul E. Jones, Jr.	Corporate-Tech Planning, Inc., 235 Wyman Street, Waltham, Massachusetts 02154
Mr. Michael A. Kahn	University of Michigan, 1225 S. University #35, Ann Arbor, Michigan 48104
Mr. Harvey P. Kaplan	FAA - (AMS-340), Department of Transportation, 800 Independence Avenue, S.W., Washington, D.C. 20591
Mr. Robert Kaunitz	Leasco Information Products, Inc., 1400 Spring Street, Silver Spring, Maryland 20910
Miss Emmy Lou Keeport	Army Member/LOGDESMO, Hoffman Building II, S769, Alexandria, Virginia 22332
Miss Prudence A. Kelly	Western Electric Co., Inc., Gateway II, Newark, New Jersey 07102
Mr. Andrew J. Kennedy, Jr.	NOAA-National Ocean Survey, 6001 Executive Blvd. (C516), Rockville, Maryland 20852
Mr. Ralph Kingsbury	University of North Dakota, Grand Forks, North Dakota 58201
Mr. Philip W. Kirsch	Bureau of Labor Statistics, 441 G Street, N.W. - Room 1257, Washington, D.C. 20212
Mr. E. G. Kirstein	Johns Hopkins Hospital, 601 North Broadway, Baltimore, Maryland 21205
Ms. Merrie B. Klotz	Civil Aeronautics Board, 1825 Connecticut Avenue, N.W., Washington, D.C. 20428
Mr. William F. Klugh	U.S. General Accounting Office, 8305 Winder Street, Vienna, Virginia 22180
Mr. Charles B. Korns	Computer Sciences Corporation, 26 Landsend Drive, Gaithersburg, Maryland 20760

537-860 O - 74 - 31

Mr. Paul D. Kosct	Software Planner, Weyerhaeuser Co., Tacoma, WA 98401
Mr. David Kuder	Hydrospace-Challenger, 2150 Fields Road, Rockville, Maryland 20850
Mr. Bennett Landsman	New Jersey State Department of Higher Education, 10 Bruce Park Drive, Trenton, New Jersey 08618
Mr. Gerald S. Lang	Code 152, Veterans Administration, 810 Vermont Avenue, NW, Washington, DC 20420
Mrs. Judith M. Lebowich	U.S. Department of Agriculture (OIS), 4200 Auditors Building, 14th & Independence Avenue, S.W., Washington, D.C. 20250
Ms. Anna K. Lee	A.I.D. - Department of State, 725 - SA-12, Washington, D.C. 20523
Mr. Raymond H. Lefurge	Bell Telephone Laboratories, Room 1B315, Holmdel, NJ 07733
Mr. Peter W. Leiss	University of Pennsylvania, 3471 Walnut Street, Philadelphia, Pennsylvania 19104
Mr. Robert L. LeRoy	Department of State, 21 & C Streets, NW, Washington, DC 20250
Mr. John T. Leslie, Jr.	Department of the Navy, 6509 3rd Street, N.W., Washington, D.C. 20012
Mr. Marshall Levitan	U.S. Naval Air Systems Command, 3618 34th Street, N.W., Washington, D.C. 20008
Ms. Pamela Libbert	Library of Congress, 3816 El Cerrito Place, Alexandria, Virginia 22309
Mr. Leonard Libster	U.S. Environmental Protection Agency, 401 M Street, S.W., Washington, D.C. 20460
Mr. James T. Lich	National Highway Traffic Safety Admin., Department of Transportation, 2100 Second Street. Washington, DC 20590
Mr. Allan R. Lichtenberger	U.S. Office of Education, 400 Maryland Avenue, S.W., Washington, D.C. 20202
Mr. Wayne Z. Lineburg	U.S. Army, Route 1, Box 122A, Middletown, Virginia 22645
Mr. Martin Lonnqvist	Statskontorct, Box 2106, Stockholm, SWEDEN 5-10313
Mr. William A. Losaw	ICA, Inc., Suite 710, 6110 Executive Blvd., Rockville, Maryland 20852
Mr. William G. Loveridge	United Aircraft Corporation, 400 Main Street, East Hartford, Connecticut 06108
Mr. Carl J. Lowe	Bureau of Labor Statistics, Room 2018, 441 G Street, N.W., Washington, D.C. 20212
Mr. Richard L. Lowe	Automation Industries, Vitro Laboratories Division, PSJ-16, 14000 Georgia Avenue, Silver Spring, Maryland 20910
Dr. Thomas C. Lowe	Informatics, Inc., 6000 Executive Blvd., Rockville, MD 20852
Mr. Paul L. Lujanac	97 Pleasant View Terrace, New Cumberland, PA 17070

Mr. James R. Macfadden	Westinghouse Management Systems & Services, % U.S. AEC Headquarters, Washington, D.C. 20545
Mr. Michael Mahoney	Goddard Space Flight Center, NASA, 4703 Marie Street, Beltsville, Maryland 20705
Mr. George Mandel	Aerospace Safety Research & Data Institute, NASA - Lewis Research Center, 21000 Brook Park Road, Cleveland, Ohio 44135
Mr. Charles R. Mandelbaum	U.S. General Accounting Office, 441 G Street, N.W. - Room 5840, Washington, D.C. 20548
Mr. Raymond H. Marcotte, Jr.	U.S. Securities & Exchange Commission, 500 North Capitol Street, N.W., Washington, D.C. 20549
Mr. Raymond J. Martin	NAVCOSSACT, 5643 Derby Court, Apt. 11, Alexandria, VA 22311
Mr. George G. Masin	Equitable Life Assurance Society, 1285 Avenue of the Americas, New York, New York 10019
Mr. Francis H. Mason	Mathematica Inc., P.O. Box 2392, Princeton, NJ 08540
Mr. Harry J. Mason, Jr.	U.S. General Accounting Office, 441 G Street, N.W., Washington, D.C. 20548
Mr. Paul B. Mayo	Central Intelligence Agency, Washington, D.C. 20505
Mr. Donald J. McCaffrey	PRC Information Sciences Co., 7600 Old Springhouse Road, McLean, Virginia 22101
Mr. John A. McCarty	Navy Command Systems Support Activity, Bldg. 196 - Washington Navy Yard, Washington, D.C. 20374
Mr. Joe R. McDaniel	On-Line Systems, 115 Evergreen Heights, Pittsburgh, PA 15229
Mr. Cliff McFall	Library of Congress, 7128 Dryburgh Court, Springfield, Virginia 22152
Mr. E. J. McLaughlin	Federal Trade Commission, 6th & Constitution Avenue, N.W., Washington, D.C. 20580
Mr. Thomas McNair	Department of Transportation, 19505 Worsham Court, Gaithersburg, Maryland 20760
Ms. Mary Ellen McNamara	IBM Corporation, 310 Lake Drive, Allenhurst, New Jersey 07711
Mr. Wallace R. McPherson	DHEW-OS-OMT-MIS, 330 Independence Avenue, S.W. - Room 3360N, Washington, D.C. 20201
Mr. Robert H. Menke	Securities & Exchange Commission, 500 North Capitol Street, Washington, D.C. 20549
Mr. D. R. Mereness	National Cash Register Co., Bldg. 10, 7th Floor, Dayton, Ohio 45479
Mr. Richard L. Merhar	University of Pennsylvania, Room 227, Franklin Bldg., Philadelphia, Pennsylvania 19174
Mr. John E. Metz	U.S. Civil Service Commission, 1900 E Street, N.W. (BRIOH), Washington, D.C. 20415

Mr. Benedict F. Milazzo	Data Base Dev., Celanese Corporation, Barclay Downs Drive, Charlotte, North Carolina 28210
Mr. Robert J. Miller	National Ocean Survey, Room 107, C515, 6001 Executive Blvd., Rockville, Maryland 20852
Mr. George J. Moldovan	Standard Oil (Indiana), 200 East Randolph, Chicago, IL 60466
Mrs. Carolyn D. Moore	Department of State - A.I.D., 1875 Connecticut Avenue, N.W. Rm. 721, Washington, D.C. 20009
Ms. Betty P. Morey	DATRAN, 8130 Boone Boulevard, Vienna, Virginia 22180
Mr. William C. Morgan	U.S. Coast Guard, G-FIS1/84, 400 7th Street, S.W., Washington, D.C. 20590
Mr. John F. Morris	Securities & Exchange Commission, 500 North Capitol Street, Washington, D.C. 20549
Mr. Jim L. Mouser	University Computing Co., P.O. Box 47911, Dallas, TX 75247
Mr. James L. Mueller	U.S. Geological Survey, 345 Middlefield Road, Menlo Park, California 94025
Mr. James C. Murdock	New York State Traffic Records, Swan Street Building, Empire State Plaza, Albany, New York 12228
Mr. Domenic J. Musotto	DHEW/SSA/BSSI, Route 2, Pin Oak Drive, Finksburg, Maryland 21048
Mr. Edward F. Nadrowski	Department of Army - CCSD, 4104 Dakota Center, Alexandria, Virginia 22312
Mr. Howard Nathanson	Federal Home Loan Bank Board, 101 Indiana Avenue, N.W., Washington, D.C. 20552
Mr. Joseph V. Natrella	Pers 351, BuPers, Department of the Navy, Washington, D.C. 20370
Ms. Diana L. Needham	State University of New York, College at Brockport, Brockport, New York 14420
Mr. Merle W. Nelson	Civil Aeronautics Board, 1825 Connecticut Avenue, N.W., Washington, D.C. 20428
Mr. Jack Newcomb	State of Tennessee, 203 Andrew Jackson Building, Nashville, Tennessee 37219
Mr. Frank R. Niedermair	NOAA-NOS, 6116 32nd Street, N.W., Washington, D.C. 20015
Mr. Clyde Nielsen	Social Security Administration - BRSI, Route 6, Box 565, Bartholow Road, Sykesville, Maryland 21784
Mr. Robert Norman	Federal Trade Commission, 177 Dry Mill Road, Leesburg, Virginia 22075
Mr. Ronald R. Notes	Federal Home Loan Bank Board, 101 Indiana Avenue, N.W., Washington, D.C. 20552

Mr. Michael P. O'Flaherty	Litton Bionetics, 35 Bradfield Drive, Leesburg, VA 22075
Mr. Samuel L. Oh	Chemical Abstracts Service, The Ohio State University, Columbus, Ohio 43210
Miss Ruth M. Oldfield	Data Base Administration, Girard Bank, 1 Girard Plaza, Philadelphia, Pennsylvania 19101
Mr. Wallace B. Oliver	Office of Preparedness, General Services Administration, 19th & F Streets, N.W., Washington, D.C. 20405
Mr. Joseph H. Ollivierre	DOD/Defense Intelligence Agency, AHS, 4000 Arlington Blvd., Arlington, Virginia 20301
Mr. Robert W. Orr	Air Force Member, LOGDESMO, DOD Logistics Data Element Standardization and Management Office, 12758 Captains Cove, Woodbridge, Virginia 22191
Mr. Yash Pal	J. P. Stevens & Co., Inc., 6300 Fairview Road, Charlotte, North Carolina 28209
Mr. Charles A. Palla	U.S. Department of Commerce, 14th & Constitution Avenue, NW - Room 5327, Washington, D.C. 20230
Miss Virginia P'an	Federal Reserve Bank of New York, 99 John Street (#8J), New York, New York 10018
Mr. George D. Parker	Union Carbide Nuclear Division, P.O. Box P, Oak Ridge, Tennessee 37830
Ms. Lucile A. Parks	AF Accounting & Finance Center, 3800 York Street, Denver, Colorado 80205
Mr. Michael A. Parmentier	USAF Data Services Center, 6525 Gildar Street, Alexandria, Virginia 22310
Mr. Norman Paull	Department of Commerce/NOAA, 14908 Flintstone Lane, Silver Spring, Maryland 20904
Mr. Joseph Penderghest	University of Pennsylvania, 215 New Road, Southampton, Pennsylvania 18966
Mr. Dave Pendleton	NOS, NOAA, 4910 Taft Road, Camp Springs, Maryland 20031
Mrs. Lynne K. Personius	Cornell University, 217 Rand Hall, Ithaca, New York 14850
Mr. Kort J. Peters	Texas Instruments, Inc., MS 2-17, 34 Forest Street, Attleboro, Massachusetts 02703
Mr. Alan Pflugrad	Planning Research Corporation, ISC, 7600 Old Springhouse Road, McLean, Virginia 22101
Ms. Lucille A. Phelps	Air Force Design Center (SYD), 3897 Cooke Drive, Montgomery, Alabama 36109
Miss Eugenia P. Pond	MTMTS, 12327 Colby Drive, Woodbridge, Virginia 22191
Mr. John W. Porter	IBM Corporation, Armonk, New York 10504
Mr. Joe L. Powell	Westinghouse NES, P.O. Box 355, Pittsburgh, PA 15230

Mr. D. Lee Power	Library of Congress, Washington, D.C. 20540
Mr. David J. Premer	Defense Mapping Agency (PPE), 2nd & Arsenal, St. Louis, Missouri 63118
Mr. John C. Prescott	U.S. General Accounting Office, 441 G Street, NW, Washington, D.C. 20548
Mr. E. D. Proudman, Jr.	U.S. Army Materiel Command, 5001 Eisenhower Avenue, Alexandria, Virginia 22304
Mr. James K. Pruitt	J. P. Stevens & Co., Inc., 6300 Fairview Road, Charlotte, North Carolina 28211
Ms. Sarah J. Pryor	Fort Ritchie, 64 W. Main Street, Waynesboro, PA 17268
Mr. Fernando Puente	U.S. Army - AMCALMSA, 210 North 12th Street, P.O. Box 1578, St. Louis, Missouri 63188
Mr. Kenneth M. Querry	Navy Fleet Material Support Office, Code 9641, Mechanicsburg, Pennsylvania 17055
Mr. William C. Rachau	Computer Data Systems, Inc., 8109 Phelps Place, Forestville, Maryland 20028
Mr. Mark A. Ramm	Fireman's Fund American, P.O. Box 3395, San Francisco, California 94119
Mr. Dewey E. Ray	Florida Public Service Commission, 700 South Adams Street, Tallahassee, Florida 32304
Mrs. Patricia Reed	National Institute on Drug Abuse, Rockwall Bldg., Room 8-68, 11400 Rockville Pike, Rockville, Maryland 20852
Ms. Judy E. Reems	DATRAN, 8130 Boone Blvd., Vienna, Virginia 22180
Ms. Clela B. Reeves	Amherst College, Station 2, Amherst, Massachusetts 01002
Mr. Copeland A. Reihl	SRS/DHEW, South Building - Room 2050, 330 C Street, S.W., Washington, D.C. 20201
Mr. James W. Rhodes	Naval Facilities Engineering Command, Code 011, 200 Stovall Street, Alexandria, Virginia 22332
Mr. Linwood A. Rhodes	Department of State, Agency for International Development, Washington, D.C. 20523
Mr. Charles W. Riggs	Frederick Cancer Research Center, P.O. Box B, Frederick, Maryland 21701
Mr. Donald A. Roache	DHEW, Social Rehabilitation Service, 330 C Street, S.W., Washington, D.C. 20201
Mrs. Elizabeth Roberts	NOAA, 2601 Woodley Pl., N.W., Washington, D.C. 20008
Mr. John Roberts	New York State Department of Motor Vehicles, Empire State Plaza, Albany, New York 12203
Mr. Paul M. Robinson, Jr.	Assistant for ADP/Telecommunication Standards, Navy OP-916E1 Washington, D.C. 20350

Miss Marian E. Rowe	Charles County Community College, P.O. Box 910, LaPlata, Maryland 20646
Mr. Harvey L. Rubinstein	Johns Manville Corporation, P.O. Box 5108, Denver, Colorado 80217
Mr. Herman H. Rugen, Jr.	National Archives, NNPD, Washington, D.C. 20408
Mr. Albert Sacca	Naval Air Engineering Center, B-76 MS-NP021, Philadelphia, Pennsylvania 19112
Ms. Sonia St. Clair	State of Minnesota, Information Systems Division, Centennial Bldg. - 5th Floor, St. Paul, Minnesota 55155
Mr. Sidford F. Sand	National Institute of Mental Health/ADAMHA/DHEW, Parklawn Building - Room 7-102, 5600 Fishers Lane, Rockville, Maryland 20852
Mr. Kenneth R. Sanford	National Defence Headquarters, Department of National Defence, Ottawa, Ontario DIA OK2, CANADA, ATTN: CCS/DCTS6
Mr. Karl W. Sanger	O/ISO U.S. Department of State, 21st and C Street, N.W., Washington, D.C. 20250
Mr. Frank Sauber	NAVCOSSACT, U.S. Navy, Washington Navy Yard - Bldg. 126, Washington, D.C. 20374
Mr. Richard Schafer	Penn State University, 120 Ridge View Drive, Dunmore, Pennsylvania 18512
Mr. Daniel B. Schneider	ADP, Telecommunications Policy, U.S. Department of Justice, 10th & Constitution Avenue, NW - Room 6307, Washington, D.C. 20530
Mr. Neils Schulz	General Foods Corp., 250 North Street, White Plains, New York 10603
Mr. Leighton R. Scott	NOAA/NESS, 9466 Pinecone Row, Columbia, Maryland 21043
Mr. George R. Seiler	Procter & Gamble Co., Hillcrest Towers, 7162 Reading Road, Cincinnati, Ohio 45237
Mr. Harry C. Shafer	United States Coast Guard, Department of Transportation, 400 7th Street, S.W., Washington, D.C. 20590
Mr. James E. Shaffer, Jr.	U.S. Coast Guard, Department of Transportation, 400 7th Street, S.W., Washington, D.C. 20590
Mr. William L. Shaffer	Dun & Bradstreet, 2233 Wisconsin Avenue, NW, Washington, D.C. 20007
Mr. Harvey N. Shapiro	NAVCOSSACT, 2530 Q Street, N.W. - #24, Washington, DC 20007
Mr. Fletcher H. Shaw	SAC 6-144, Automation Industries, Inc., Vitro Laboratories Division, 14000 Georgia Avenue, Silver Spring, Maryland 20910
Mrs. Rama L. Shaw	Computer Systems Administrator, P.O. Box 762, Fort Huachuca, Arizona 85613
Mr. L. L. Shields, Jr.	M. Bryce & Associates, Inc., P.O. Box 15459, Cincinnati, Ohio 45215

Mr. Robert B. Shives	ACC-CONUS, MISO, Fort Ritchie, Cascade, Maryland 21719
Mr. William G. Shope, Jr.	U.S. Geological Survey, National Center Mail Stop 437, 12201 Sunrise Valley Drive, Reston, Virginia 22092
Mr. Randy Sigite	USAMC-ALMSA, 210 North 12th Blvd., P.O. Box 1578, St. Louis, Missouri 63188
Mr. Frederick Simmonds	Department of the Army, 2212 Emporia Street, Woodbridge, Virginia 22191
Mr. Barry R. Sims	U.S. Army Coastal Engineering Research Center, Kingman Building, Fort Belvoir, Virginia 22060
Ms. Janet S. Spitzer	IBM, Monterey & Cottle Roads - D86-052, San Jose, CA 95111
Mr. Merle E. Sprague	Operations Directorate, U.S. Army Computer Systems Support Evaluation Command, Nassif Building, Falls Church, Virginia 22041
Ms. Loren Stafford	Burroughs Corporation, P.O. Box 299, Detroit, MI 48232
Mr. James R. Stear	NOAA, 3300 White Haven Street, Washington, D.C. 20008
Lt. Col. John C. Stoob	USAF, HQ AFLC/ACTCU, Wright-Patterson Air Force Base, Dayton, Ohio 45433
Mr. George A. Strasser	Government Printing Office, North Capitol & H Streets, NW., Washington, D.C. 20401
Mr. Gunnar Sundblad	Swedish Standards Institution, Box 3295, Stockholm, SWEDEN S-10366
Mr. James R. Swallow	NAVCOSSACT LANTCOMDET, % CINCLANTFLT, Norfolk, Virginia 23511
Mr. Lincoln W. Talbot	Veterans Administration, 810 Vermont Avenue, N.W., Washington, D.C. 20420
Ms. Bonnie C. Talmi	Oak Ridge National Laboratory, P.O. Box X, Oak Ridge, Tennessee 37830
Mrs. Evelyn B. Taylor	USCG - GFIS1/84, Department of Transportation, 400 7th Street, S.W., Washington, D.C. 20591
Mr. J. Taylor	Veterans Administration, 810 Vermont Avenue, N.W., Washington, D.C. 20420
Professor Daniel Teichroew	Department of Industrial Engineering, University of Michigan, 241 W. Engineering, Ann Arbor, Michigan 48104
Mrs. Mary C. Thomas	Federal Deposit Insurance Corp., 550 17th Street, N.W., Washington, D.C. 20429
Mr. E. Joseph Thompson	National Civil Defense Computer Facility, P.O. Box 256, Olney, Maryland 20832
Mr. Arthur J. Tonelli	Northrop Corporation, 934 Washington Street, Norwood, Massachusetts 02062

Mr. C. Ronald Trueworthy	U.S. Civil Service Commission, 1900 E Street, N.W. - Room 6455, Washington, D.C. 20415
Mr. Walter J. Utz, Jr.	Hewlett-Packard Corp., 11000 Wolfe Rd., Cupertino, CA 95014
Mr. Gerry Varsoke	IBM Corp., 18100 Frederick Pike, Gaithersburg, MD 20760
Mr. Charles J. Venturi	National Highway Traffic Safety Admin., Department of Transportation, 2100 Second Street, S.W., Washington, D.C. 20590
Mr. Frank T. Verdone	Grumman Data Systems, Department 137, 1111 Stewart Avenue, Bethpage, New York 11714
Mr. Dale Vetter	University of North Dakota, Room 411, Twamley Hall, Grand Forks, North Dakota 58201
Mr. Laymon H. Vinson	U.S. Army - AMC - ALMSA, ATTN: AMXAL-MB, 210 N. 12th Street, St. Louis, Missouri 63188
Mr. K. E. Virta	U.S. Naval Oceanographic Office, Suitland, Maryland 20373
Mr. Conrad S. Voegler	Comptroller of the Navy (NCF132), Room 424, Crystal Mall Building 3, Arlington, Virginia 20376
Mr. R. K. Vredenburgh	New York Times, 229 West 43rd Street - Room P53, New York, New York 10036
Mr. J. Neil Wallace	Central Intelligence Agency, HQ Building, Room 4E12, Washington, D.C. 20505
Mr. Marvin G. Wallis	NASA, Code DSC, Washington, D.C. 20546
Mr. R. E. Ward	New York State Department of Motor Vehicles, Swan Street Bldg., Empire State Plaza - 5th Floor, Albany, New York 12228
Mr. John Watson	Occupational Safety & Health Admin., Department of Labor, 4914 Yorktown Blvd., Arlington, Virginia 22207
Mr. Benjamin H. Weiner	U.S. General Accounting Office, 441 G Street, N.W., Washington, D.C. 20548
Mr. John Wellons	City of Madison, 210 Monona (City County Building), Madison, Wisconsin 33709
Mr. Irving Werner	U.S. Geological Survey, 19th & C Streets, NW - Room 1442, Interior Building, Washington, D.C. 20009
Mr. George W. West	World Bank (IBRD), 1818 H St., NW (N441), Washington, DC 20433
Mr. Thomas J. West	Insurance Co. of North America, White Horse Road, Somerdale, New Jersey 08083
Mr. L. G. Whitener	Air Force, 5621 Gale Wind Avenue, San Antonio, TX 78239
Mr. Fred S. Williams	Small Business Administration, 1441 L Street, N.W., Washington, D.C. 20416

Mr. John R. Williams	OO Branch Dairy Division, Agricultural Marketing Service, U.S. Department of Agriculture, 14th & Independence Avenue - Room 2965, Washington, D.C. 20250
Ms. Roxanne Williams	U.S. Department of Agriculture, 4200 Auditors Building, Washington, D.C. 20250
Mr. Charles J. Williamson	Defense Intelligence Agency, Department of Defense, 5009 North 25th Road, Arlington, Virginia 22207
Mr. Rance R. Willis	Engineering Information and Data Systems Office, Room 5E058H, Forrestal Bldg., Washington, D.C. 22314
Capt. S. Withers	Computer Centre, Department of National Defence, National Defence Headquarters, Ottawa, Ontario CANADA
Mr. Ronald S. Wood	Emory University, Computing Center, Atlanta, Georgia 30329
Mr. Ronald S. Woolwine	Centralina Council of Government, 1229 Greenwood Cliff, Charlotte, North Carolina 28204
Mr. Dennis W. Wright	Institute for Law and Social Research, 1125 15th Street, NW - Suite 625, Washington, D.C. 20005
Mr. Frank S. Yevcak	Allendale Insurance Co., Allendale Park, P.O. Box 7500, Johnston, Rhode Island 02919
Mr. Horace R. Zibas	Procter & Gamble, 7162 Reading Road, Cincinnati, Ohio 45237

FUTURE DATA ELEMENT MANAGEMENT CONFERENCES

TO: Harry S. White, Jr.
Associate Director for ADP Standards
Institute for Computer Sciences and
Technology
National Bureau of Standards
Washington, D.C. 20234

FROM: (name, address, and telephone number)

___ Please add my name to your mailing list for announcements and information relating to future data element management conferences.

___ I wish to present a paper on the following subject: _____

_____

_____

___ I would like to participate on a panel or in an open forum on the following subject(s): _____

_____

_____

I suggest the following subjects or subject areas for consideration at the future conferences: _____

_____

_____

Other recommendations: _____

_____

_____

_____

_____

NBS-114A (REV. 7-73)

U.S. DEPT. OF COMM. BIBLIOGRAPHIC DATA SHEET	1. PUBLICATION OR REPORT NO. COM 74-10700	2. Gov't Accession No.	3. Recipient's Accession No.

4. TITLE AND SUBTITLE

Management of Data Elements in Information Processing (Proceedings of the First National Symposium, 1974 January 24 and 25)

5. Publication Date: *1974 April*

6. Performing Organization Code:

7. AUTHOR(S): *Various: Editor: Hazel E. McEwen*

8. Performing Organ. Report No.: *NBSIR 74-466*

9. PERFORMING ORGANIZATION NAME AND ADDRESS

NATIONAL BUREAU OF STANDARDS
DEPARTMENT OF COMMERCE
WASHINGTON, D.C. 20234

10. Project/Task/Work Unit No.: *6009580*

11. Contract/Grant No.:

12. Sponsoring Organization Name and Complete Address (Street, City, State, ZIP)

American National Standards Institute Committee X3L8 and the National Bureau of Standards

13. Type of Report & Period Covered: *Final*

14. Sponsoring Agency Code:

15. SUPPLEMENTARY NOTES

16. ABSTRACT *(A 200-word or less factual summary of most significant information. If document includes a significant bibliography or literature survey, mention it here.)*

Recent technological advances in computers and communications make possible the integration of data systems and the exchange of data among them on an expanding scale. However, the full effect of these advances cannot be realized unless the need for uniform understanding of the common information (data elements) and their expression in data systems is recognized and a means provided to effectively manage this information. The increasing interrelationships among the data systems of Federal, State, and local governments, and with industry and the public add emphasis and dimension to the need for the improved management of data elements in information processing.

These Proceedings are for the first Symposium on the Management of Data Elements in Information Processing held at the National Bureau of Standards on 1974 January 24 and 25. Over 400 representatives of Federal and State governments, industry and universities from 30 states, from Canada, and Sweden were in attendance. 34 speakers discussed data element management in the fields of health care, water resources, state government information systems, transportation, libraries, market research, manufacturing, banking, information retrieval systems, military systems, computer programming and software systems, and motor vehicle registration.

17. KEY WORDS *(six to twelve entries; alphabetical order; capitalize only the first letter of the first key word unless a proper name; separated by semicolons)*

American National Standards; American National Standards Institute; data; data base systems; data elements; data management; data processing; Federal Information Processing Standards; information interchange; information processing; information systems.

18. AVAILABILITY

- ☐ Unlimited
- ☐ For Official Distribution. Do Not Release to NTIS
- ☐ Order From Sup. of Doc., U.S. Government Printing Office Washington, D.C. 20402, SD Cat. No. C13
- ☒ Order From National Technical Information Service (NTIS) Springfield, Virginia 22151

19. SECURITY CLASS (THIS REPORT): UNCLASSIFIED

21. NO. OF PAGES: *490*

20. SECURITY CLASS (THIS PAGE): UNCLASSIFIED

22. Price: *$9.75 paper copy $1.45 microfiche*

USCOMM-DC 29042-P74

U. S. GOVERNMENT PRINTING OFFICE : 1974 O - 537-860