DOCUMENT RESUME

ED 025 273 LI 001 129

By- Libbey, M. A.; Blum, A. R.

A Study of Information Elements for the National Information System for Physics.

American Inst. of Physics, New York, N.Y.
Spons Agency- National Science Foundation, Washington, D.C.
Pub Date Jun 68
Grant- NSF-GN-686
Note- 62p.
EDRS Price MF-$0.50 HC-$3.20
Descriptors- *Automation, *Cataloging, Data Analysis, Data Sheets, Documentation, Indexing, *Information Processing, *Information Storage, Information Systems, *Physics, Records (Forms), Standards

The identification of information elements can provide an important tool for the systematic development of an information system design. A state-of-the-art survey reveals mounting recognition and interest in the problem, a considerable history of prior efforts, but no well-defined methodology. A study in the context of a national information system is reported. A "trial structure" has been developed andis described. (Author)

A Study of Information Elements for the National Information System for Physics

by
M.A. Libbey
and
A.R. Blum

ED025273

Information Division
AMERICAN INSTITUTE OF PHYSICS
335 East 45 Street, New York, New York 10017

Report on work supported by the National Science Foundation under Grant No. NSF-GN 686

ED 001129

ABSTRACT

The identification of information elements can provide an important tool for the systematic development of an information system design. A state-of-the-art survey reveals mounting recognition and interest in the problem, a considerable history of prior efforts, but no well-defined methodology. A study in the context of a national information system is reported. A "trial structure" has been developed and is described.

Preface

One of the efforts undertaken by the American Institute of Physics under National Science Foundation Grant No. GN-686, "Additional Prerequisites for Development of a National Information System for Physics," was a study of appropriate information element structures. The effort, reported in this paper, was intended to provide a basis for considered decisions as to the nature, extent, and priority of any future information element standardization effort by AIP.

TABLE OF CONTENTS

APPENDICES

I. Introduction

To design an information system in which, in general, the flow of information (e.g., printing, writing, acoustic or electrical waveforms) is to be mediated by humans, it may only be necessary to note, tautologically, that "information flows in an information system." If, however, any significant role is to be assigned to digital computers, this will normally not suffice. Where a computer is involved, any elements of "meaning" which it is to receive from a human source by any means, and which it is to retain in fact within itself, must be utterly explicit. It is also a truism that in the present state of the digital computer art, if one machine (or automated system) is to interchange information with another, these elements being exchanged must be standardized explicitly and in extenso. (It need not concern us at the moment whether or not this would apply to "Perceptron type" machines or systems or sophisticated translation or parsing programs.)

Present plans of the American Institute of Physics call for use of computers in the process of producing primary physics publications by photocomposition. It is also anticipated that the byproduct tapes will be processed by computer for various purposes. To these extents, the eventual selection, definition and detailed format specification of units of information which are to be handled (input, processed, output) in the system becomes a sine qua non. This requirement was not the original reason for AIP's considering the information element standardization problem at this time. It did, however, assume increasing importance as the plans for a computer implementation of the system became more concrete.

Section II of the report will discuss the usefulness of a standardized information element structure as a tool for systems analysis and development. This was the original objective of the study here reported.

Since the general problem of information element standardization is not familiar to most people, some effort is made to put it in perspective by discussing its background in Section III. Section IV deals with definitions, a problem that has plagued efforts in this field. Section V presents alternatives which could be considered for an information element standardization program, and Section VI, on that basis, describes the philosophy of the approach that had been planned for AIP in getting started in this area. The actual implementation of the plan is outlined in Section VII. The subsequent sections contain the description of the resulting tentative data element set: its structure, sources, and contents.

Anyone considering the information element standardization problem should be warned that it seems to create more than a reasonable number of cases of confusion in inter-personal communications. The problem seems to be mainly due to confusion of levels of terminology--undoubtedly a major occupational hazard in any undertaking in which words must be used to talk about other words.

Another terminology difficulty results from the fact that many of the terms that must be used routinely are themselves vague and ambiguous. Such terms include "information," "data," and "element." The definitions of combinations of such terms are often even more vague and may be downright misleading.

II. A Tool for Systems Analysis and Development

Basic to any systematic approach to the design of a system is the identification and characterization, quantitatively if possible, of whatever will flow in the system and of whatever will determine or affect the flow in the system. In some cases this also helps to identify all of the nodes and interconnections between them which constitute the system. It is to be noted that the fact that a system is being designed in no way implies that the design will ipso facto be systematic.

Ideally such identification of the flow of information would take into account information of all forms, including when appropriate the full text of conversations, books, journal articles, etc. However, the state-of-the-art of information system design, expecially, but not only, when the system is to be automated, does not permit such a thorough treatment. For now the system must be conceived of as handling discrete, demarkable and identifiable elements of information and data, especially data.

The topic of definitions of information and/or data elements requires special attention and will be discussed in Section IV. In the meantime, an information or data element can be thought of in operational terms as a concept which is particularized by a character, sequence of characters, sequence of sequences, etc., which appear in some particular location in space or time, such as a field on a punched card or in a magnetic tape or disc file record, a particular place on a piece of paper, the nth item of a formatted message, etc. For example, "WEST" appearing in the appropriate location would particularize the information element "Last name of author," while in a different location it might particularize the element "East or West longitude."

Typical of the kinds of information elements that might be expected to be important to a national information system for physics are the following:

Author	Descriptor	End of Message
Co-Author	Classification Number	Patent Number
Title	Query	Method
Journal	Query Number	Classification Symbol
Volume Number	Logical Connective	Target Nucleus
Inclusive Pages	Count of Citations	Bombarding Particle
		Emitted Particle
		Energy Range

To this list can be added elements needed by those concerned with the operation and financial support of the system. These might include elements that pertain to records of system usage as well as the more obvious ones of cost, computer time, telephone line time, etc.

However, to identify or standardize such elements in order that they might actually be used in an operating system some day is not the only purpose of the project here discussed. Rather, the original purpose was to develop

a consistent and structured language which would assist the systems development staff to proceed with its task in a systematic and scientific way. For this a considerably more sophisticated structure and a considerably larger number of elements was required. Typical of such additional elements are the following:

Most used journal
Preferred information channel
Source of awareness of existence of stated journal article
Percent of person's time spent in physics research and development
Months in current sub-specialty
Journals subscribed to
Journals scanned or skimmed personally
Attitude toward preprint centralization proposal
Effect on own work of responses from recipients of own documents
Whether or not a contributed paper later appeared as a journal article

It can be seen that these are far more complex than the preceding list and that the relations between these can be quite intricate. Yet the data named and identified by such elements are essential to the systematic design of a national information system for physics. To avoid major sins of omission (by not utilizing existing data) or of commission (by conducting research that was not needed) the systems designers need a comprehensive and internally consistent means of identifying, organizing, naming and classifying all such data found in files and documents whether at AIP or elsewhere and whether generated by their own efforts or those of others. The paramount purpose of the development of an information element structure was to fulfill--or at least indicate how it was possible to fulfill--that need.

III. Background

The recognition of the need to identify and standardize elements of information and data that occur within and are handled by a system or group of systems can be traced to the experience of the Army and Navy in World War II. The fantastic growth in intrinsic complexity and in the numbers of items carried in both material and personnel logistics systems compelled the development of new operational logistics, inventory, and communications techniques. In addition the sheer cost of military operations made it impossible to tolerate situations where a single physical item might actually appear as more than a hundred different inventory items, under different stock numbers, and requiring separate procurement, accounting, stocking, requisitioning, etc. Even more intolerable were situations in which military weapons and equipment such as anti-aircraft guns might be out of action simply because it was not realized that a missing part was available at hand under a different stock number than the one shown in the repair manual.

Another factor favoring standardization was the realization that the communication protocols and circuit disciplines developed by different organizations were simply not compatible. The increasingly global and inter-service nature of military operations caused these difficulties in inter-organization communication to compound the difficulties in logistics. Both were in turn further compounded by the attempt to solve them by applying computer technology. Where before human data processors might be able to solve a difficulty-- or might at least recognize that there was a difficulty--the electronic data processors could only cope with those difficulties that had been foreseen and programmed for.

All these factors got worse in the post-war period. By the early 1950's all of the military services, and, in addition, other government agencies such as the Census Bureau, were putting computers to work to process the masses of management, personnel and logistics data which they had to handle. In general, however, all of these systems, including the large military systems, were designed and developed only in the context of their individual needs. Therefore, the selection, definition, and detailed format specification of units of information which could be handled (input, processed, output) were done independently. It is therefore difficult or impossible for such systems to communicate with one another as is increasingly required by considerations of economy, efficiency, flexibility and, sometimes, even survival.

Recognition of the need for standardization of elements of information has, since then, spread rapidly. Element standardization programs which originated within a DOD department were picked up, or in some cases overridden, at the Department level. The Department-level programs were in turn overridden in 1964 by a program at the DOD level. And this has since found itself subordinated to a government-wide program directed by the Bureau of the Budget.

The tool thus developed has more recently spread outside of Government. While the concept undoubtedly occurred independently in some companies, there was certainly some direct transference to industry from its development in the

military. For one, the Sutherland Co., Peoria, Ill., extended the techniques and methods used in this area in parts of the U.S. Air Force and, as a management consulting firm, introduced them in many client companies.

In 1966 a group was organized in the American Standards Association (now the USASI (U.S.A. Standards Institute) at the instigation of commercial, rather than Federal agencies, to identify and standardize elements of information required in the interchange of information in industry, business and Government. This group is now Subcommittee X3.8, "Data Elements and Their Coded Representation." More recently, in response to the rapidly increasing rate of application of computers to scientific and technical information and to library problems, another group within the USASI (Z 39 SC 2) has addressed itself to the identification and standardization of the elements of information occurring in the preparation, description, and interchange of bibliographic information such as citations.

Many other organizations have more recently expressed an interest in the data element standardization problem. These include COSATI and SATCOM. In particular, the Task Group for Interchange of Scientific and Technical Information in Machine Language (ISTIM) recently established by the Office of Science Technology, recognized this as a basic and important problem.

IV. Definitions

The problem of arriving at a standard definition for an information or data element has given extraordinary difficulty. For all practical purposes the terms "information element" and "data element" have been used to refer to the same things. The definition that would appear to have the greatest consensus of agreement at the moment would be the following one by USASI Subcommittee X3.8:

> "Data Element- A grouping of informational units which has a unique meaning based on a natural or assigned relationship and subcategories (data items) of distinct units or value."

It is impossible--or at least unwise--to isolate the definitions given for information and/or data elements from certain closely associated concepts. In the case of the X3.8 definitions are:

> Data Item- A unit of distinct information or value classified under a data element which cannot be logically subdivided and retain significance of the data element grouping.
>
> Data Use Identifier- A name or title given to the use of a data element.
>
> Data Code- A number, letter, symbol or any combination thereof used to represent a data item.
>
> Data Group Identifier- A name or title given to the use of a combination of two or more related data elements.
>
> Data Element Reference- A number, letter, symbol or any combination thereof used to represent a data element.
>
> Data Use Reference- A number, letter, symbol or any combination thereof used to represent a data use identifier.
>
> Data Group Reference- A number, letter, symbol or any combination thereof used to represent a data group identifier.

The following examples are used to illustrate the above definitions:

Data Element	Year
Data Item	1967
Data Use Identifier	Model Year
Data Code	1967
Data Group Identifier	Date of Action
Data Elements	Year Month Day
Data Group Identifier	Date of Purchase
Data Elements	Year Month Day
Data Items	1967 April 24
Data Codes	1967 04 24

Probably the definitions which have the most influence at the moment

are those of the DOD:

> "Data Element- A grouping of informational units which has a unique meaning and subcategories (data items) of distinct units or values.
>
> Data Item- A subunit of descriptive information or values classified under a data element.
>
> Data Chain- A name or title given to the use of a combination of two or more logically related standard data elements, use identifiers, or other data chains."

The definition of "data element" given at the highest Federal level, the BOB, is identical to that given above X3.8 except for use of the plural form of the last work, "values." However, for "data code," BOB gives:

> "A data code is a number, letter, symbol or any combination thereof used to represent a data element or a data item."
>
> Subcommittee X3.5 of USASI has submitted the following definitions:
>
> "Data Element- The name for a class or category of data based on natural or assigned relationships that can be used to denote a set of data items.
>
> Data Item- The name for an individual member of set denoted by a data element.
>
> Data Code- A structured set of characters used to represent the data items of a data element.
>
> Data- Any representations such as characters or analog quantities to which meaning is or might be assigned.
>
> Information- The meaning assigned to data by known conventions.
>
> Data Chain- See macroelement.
>
> Macroelement- An ordered set of two or more elements used as one data element with a single data use identifier.

The following definitions were given by the information element standardization program of the MITRE Corporation in 1964:

> Information Element is a definable entity whose values, when determined, convey knowledge in an information system.
>
> Value is the smallest piece of information that may be used to communicate intelligence."

All of the above represent consensuses, and therefore compromises. The following definitions were proposed by one of the authors (Libbey) in 1965 on the basis of a study of relevant material in linguistics and logic:

> Information Element- A concept selected, defined and distinctively symbolized for efficient communication which collects, designates, gives meaning to and is extensionally defined by a specified set of its instances.
>
> Data Element- An information element which deals with data, i.e. with

that kind of information which is usually thought of as being tabulated, listed or otherwise formatted.

Value- One of the specified set of instances of the concept designated by an information element; also, the symbolization thereof.

It is still submitted that these represent the picture more accurately, although expediency may require subscription to one of the previously given definitions, presumably that given by X3.8.

V. Alternatives

Selection of an approach to the information element standardization program in AIP is rendered difficult by that fact that: (1) the information element standardization methodology itself is far from well-developed and understood and (2) AIP's needs for such a tool have not yet been adequately explored. It is not even certain just what variables would be most critical in this case. Several will be discussed in this section.

A. Objectives

For one thing, the approach taken can be expected to vary with the objective. Various distinguishably different possible objectives for a study or project concerned with information elements and their standardization are listed below. Although the objectives listed cannot be thought of as related in any simple linear fashion, in general they are listed in a "least to most" order.

1. Problem knowledge and definition. Any exploration at all of the problem will contribute to overall knowledge, problem definition, identification of parameters, magnitude estimates, etc. Any of these should be useful to AIP in its systems development efforts.

2. Information Interface Description Tool. The identification and definition of information elements in a standard manner is an absolute essential for describing information interfaces between automated systems.

3. Systems Analysis. By identifying and defining information elements throughout a system instead of only at an interface and by paying more attention to document description and the delineation of the flow of documents throughout the system a useful and sometimes indispensable tool for systems analysis can be developed.

4. Systems Network Analysis Tool. Extension to additional systems of any information element structure developed as a tool for analysis of a single system would be very useful in analyzing a network of inter-connected systems.

5. Basis for an Operational Information Retrieval System. Any standardization system pitched at the information element level necessarily implies going through much detailed labor to identify and define elements. The product of such detailed labor should be directly applicable to the establishment of an operational retrieval system, whether manual, semi-automated or automated, since any such standardization system. as presently envisioned, should supply at least the following basic elements of an information retrieval system: classification system, controlled vocabulary of index terms, relation of synonyms, and presumably a capability to relate index terms appearing on specific documents.

6. Inputs for Developments Based on Linguistic Theories. Further refinement should enable the same detailed labor to produce another by-product if desired, the kind of data needed for development of modern linguistic theories or their actual application to a national physics information system. The principal extension required for this objective would be coverage (or more coverage) of non-substantive types of terminology and of syntactic relations.

7. Manual Operational Translation Facility. The necessary procedures, files, etc., could be developed to establish a manual translation capability in, for example, an information analysis center to translate information ... expressed in the terms and formats of one system into the terms of the standardized information element structure and if desired from this into the terms and formats of other systems.

8. Semi-Automatic Operational Translation Facility. By automating the procedures and files needed for the facility described in the preceding paragraph, but retaining humans to perform some of the more sophisticated decisions and processes, a semi-automatic operational translation facility could be developed to link two or more automated information systems. The delays inherent in such a facility should be small enough to make it practicable.

9. Fully Automated Operational Translation Facility. In principle it should be possible to extend the process mentioned in the last two paragraphs and develop a fully automated facility which would receive information messages, reports, etc., in the terms and formats of one system and produce as outputs the same information in terms of the standardized system and/or in the terms and formats of any desired other system. There would be little doubt as to the usefulness of such a facility. However, it would be very costly and extremely difficult to justify on economic grounds.

B. Products and Coverages

For any one of the objectives listed above there would still be a variety of possible products or coverages which could be chosen. The following list is intended to be at least indicative of this spectrum of choices. Each possibility listed is itself capable of further variation.

1. Dictionary. The most modest coverage would be that required to produce a simple dictionary of the terms to be used in a national information system for physics. This would serve the same function with respect to standardizing the physics information system's language as does the conventional dictionary with respect to standardizing a natural language. This would involve merely the identification and the conventional definition of terms used with no attempt at inter-relation or classification. Such an effort would, of course, make maximum possible use of the AIP glossaries and of any useable products of AIP classification and indexing efforts.

2. Dictionary with Adjuncts. A considerably more useful tool would result if, in addition to a simple dictionary, the terminology structure of the physics information system's language was further elaborated by means of

such devices as cross-references in the dictionary, a thesaurus, a classification scheme, etc. Hardly any extension of this coverage or of any of those listed after this would be needed to provide for language purification: confusing or misused terminology, unnecessary synonyms and ambiguously used terms could all be recognized and documented.

3. A Standardized Information Element Structure. This would involve the establishment of standardized information elements meeting rigid requirements of uniqueness, exhaustiveness, and explicitness. Combinations of such basic elements which were found to be needed in the conduct of the systems operations would also be established.

4. Value Recording. In some cases it might be desirable to identify and to record the various possible values (sets of numeric, alphabetic, alphanumeric, special, etc., characters that can be used to represent each given information element allowable throughout the system). In such cases explicit rules may, if appropriate, be specified which must be satisfied for a value to be accepted as valid.

5. Concordances. In carrying out any of the foregoing a concordance can be developed--i.e. an "inverted file"--linking information elements to their uses in the system. Such a capability would make it possible to locate all places where information on a given topic existed in the system.

6. Formatted Message Description. This could consist of merely identifying the information elements in formatted messages which were to be used in the system (and probably the sequence in which they occur) or, more usefully, could go on to specify any additional constraints on information element values that might be imposed by particular message formats.

7. General Document Description. The foregoing formatted message description could be extended to the more general case of describing documents in general. How far this could usefully be carried is impossible to say out of the context of specific requirements and other detailed information.

8. Information Processing Rules and Algorithms. This would involve the explicit description and standardization of the rules and algorithms according to which various input information elements are processed, operated upon, combined, output, etc.

9. Other System Information. Actual implementation of any of the above listed alternatives would involve amounts of detailed labor varying from the considerable to the staggering. It will almost always be possible in the process to generate extremely valuable by-products at little additional cost. Examples: place (or frequency) of usage of terms, rate and volume figures for information flow, tracing of the paths followed through the system by information either in terms of documents or in terms of documents or in terms of information elements, identification and description of nodes in the information processing network in terms of the information transformations they effect, etc.

VI. Method of Approach

Since AIP's program is still in an exploratory stage, it is appropriate to choose an exploratory approach to its information element aspects. A heuristic, first-approximation information element standardization structure has been developed. Herein lies the key to an appropriate approach for AIP at this time. There are many different problems in the actual process of developing such a structure. Experience has shown that it is all too easy to err seriously by going into one, or a few, of these problems in depth, and at considerable expense, only to find later that other problems needed either concurrent or prior attention. The best--perhaps the only--way to avoid such errors is to head as quickly as possible for a first-round "trial structure." This will be intentionally inadequate and perhaps wrong. It will, however, have as its principal merit that some recognition will have been given--or attempted--to identifiable problems, or aspects, of the task of developing such a structure. A few of the more obvious of these problems are:

1. What will be considered to constitute an "information element" (or "data element") for this purpose?
2. Will there be different kinds of elements? For example, should data elements that might later be needed in either operations or production subsystems of an eventual physics information system be distinguished from those established primarily for the use of the systems development activity? If so, why? And how?
3. Will ranges or domains for the values (or "data items") be specified? If so, how?
4. What "files" will be established?
   a. Information element definitions?
   b. Document/message/file descriptions?
   c. Term list?
   d. Classification schedules?
   e. Etc.?
5. To what extent can structures already established, such as those of X3.8, Z 39, DOD, MITRE, etc., be adopted or adapted?
6. To what extent will the initial structure be made to be amenable to conversion to machine readability?
7. What provisions will be made for concatenating or combining two or more elements to form an entity that will act as an element in its own right at times? (E.g., "Date" composed of the basic elements "year," "month," and "day of month.")
8. At what intellectual/semantic level will variation of concept be considered as being different, with respect to what is called out as one element of several, from variations of use?
9. What relations between elements will be noted on work sheets?

10. Which of these will be carried over onto the formal element definition entries?

The foregoing are only representative of the problems which arise, and they are not even the most fundamental. Obviously, the "trial structure" is intended to be changed, very possibly in toto. Therefore it should be constrained to staying small enough so that human inertia will not become a factor. Similarly the amount of effort put into any classification scheme or schemes must be limited. And finally, no detailed attention is given to formatting for machine processing.

An approach with these limitations is described in the remainder of this report.

VII. Implementation

It is apparent from the above discussion that data element standardization is a semantic approach to controlling the information transfer process. Its purpose is to facilitate the use and exchange of information, particularly information contained in the data that appear in machine-readable form. The orientation of this study is based on the assumption that the standardization of meaning and the units used to represent meaning is fundamental to proper control of the use and interchange of data. Obviously, such data will have to be communicated among the human and machine components of the National Physics Information System.

The approach of this study has been exploratory and heuristic. It was hoped that two results might be yielded by it. First and foremost, a method of identifying, developing and using a data element structure would be found. Second, the mechanism for an operational system would be proposed.

The strategy of the study was first to survey the physics information world about which a data element language must speak. The main sectors of this world pertinent to vocabulary control were surveyed and explored. Both the documented and operational information resources offered by promising relevant subject areas and the various divisions of the American Institute of Physics were tentatively identified. It was among these sectors that significant reiterated terms occurred. Only those terms considered eligible for inclusion in a controlled standard data element vocabulary were chosen for display in an experimental prototype data element file.

The next phase went on to structure the terms. Structuring the data elements was the first step, after identification and selection, in organizing the various terms. The data element structure assumes that the data element is the name or generic designation for certain items. Thus, the data element "Type of Equipment" might have data items such as "Keypunch," "Verifier" and, perhaps even more specifically, "IBM 1401 Computer." The items are named and hence may be organized and recalled by this name.

In addition, when various names can be interrelated within the structure, by being grouped or linked together, we have a design for a language that can represent the information world at hand. Classification and a number of other organizational techniques were applied to the data elements, including definitions whenever needed. Attention was given only to the design of the semantic system. Significant coding and formatting problems were left to future stages of standardization of the operational system. These include questions of standard character types, allowable character set extensions, control characters, modes of representation, message and field formatting and sizes (although free fields are strongly wanted), standard media, and common codes for data elements and data items. Nevertheless constant endeavor was made to keep the design to a level of simplicity which could, without loss of descriptive or discriminative power, effect the efficient handling, transfer and exchange of blocks of information between man and machine, to some extent between man and man and, at a later stage using codes, between machine and machine.

At least three distinct sources of data elements emerged. The first sector coincides with the traditional published physics literature. It contains principally the detailed biographical units used in referring to published documents (such as author, title, journal name, classification and indexing terms, citations, etc.). Precise and unambiguous designation of these units can assume special importance in the planned computer-based photocomposition of the primary journals in physics, as well as the subsequent development of a bibliographic data base and further byproducts from that base. The second sector comprises general management files and their contents as well as special collections relating to physicists -- both individually and collectively -- and to events in physics. Data element control of such files could help provide day-to-day and line supervisory records and their functional interpretation for the organizational and evaluative needs of decision makers. Control of the information contained in the special collections and archives would, if deemed feasible, enable improved services and products to be developed (e.g. from AIP resources such as historical biographical and special bibliographical repositories, institutional records, etc.). The third and final sector covers the system analysis, design and development activities, where a distinctive information metalanguage is used as a vocabulary to talk about information.

VIII. The Data Element Structure

A. Structuring of Terms

The brief discussion of definitions given in Section IV indicates certain shortcomings in currently accepted conceptual definitions of data elements and related concepts. Operational definitions, particularly those of the type used by USASI X3.8 in its Technical Guidelines, may be somewhat less objectionable. Such definitions should take into consideration the components of a data element structure, their functions and interrelations. A deeper examination of the interrelations between the components of a data structure (i.e. data elements, data items, data use identifiers, data groupings, etc.) necessarily involves consideration of data processing theory, particularly the theory underlying the practices of computer programming, information retrieval and construction of artificial languages. Although the method of this study, starting with empirical exploration and heuristic problem-solving approach, was carefully chosen so as to preclude excessive involvement in the intricacies of these vast subject areas, the need to understand the problems clearly made it helpful to turn to theory occasionally.

The prodigious growth of computer technology in recent years has been accompanied by interest in the theory underlying computer and computer programming processes. Searches for theoretical foundations of information processing are legend. A brief description of a number of outstanding developments in this field with regard to computer applications may be found in the article by William C. McGee.[1] One of these studies may be cited here.

Quite relevant to the concept of data elements is the work conducted by the Share Committee on Theory of Information Handling (TIH). A definitive report of the TIH Committee in 1959 established certain basic data processing concepts that are both extremely influential and highly relevant to the process of structuring terms. Among the fundamental concepts were entity (an object, person, or idea capable of being described for data processing purposes); property (characteristics in terms of which entities are described); and measure (value assigned to properties). A datum was defined as the smallest unit of information, consisting of the triple Dij where D is a measure, i is the index of an entity and j is the index of a property. The unit record was considered on the basis of this structure as a one-dimensional array of datum triples, in which the index i is fixed and the index j ranges over all properties being represented; the file was conceived as a two-dimensional array, with index j varying over all properties and index i varying over all entities. Essentially, this represents a specialized case of the general classification Rij, where R is any relation, and i and j are any tabular indices. Later work of the Committee developed the concept of a generalized array whose elements have the general form of an ordered pair (v,x) in which v is a data value corresponding to an argument x. Arguments are expressed as a set of ordered pairs

$$[\ (v_1, p_1), (v_2, p_2), \ldots (v_n, p_n)\ ]\ ,$$

where p_1 indicates the name of the array dimension and v_1 a value of the corresponding dimension. A two-dimension generalized array in which one

dimension is property and the second is entity is equivalent to the original two-dimensional array. Here, the data value v in an element corresponds to the datum Dij.

The concept of data element, originally advanced by O.Y. Evans,[2] is equivalent to the notion of property advanced here. "Property" is also the same as the USASI X3.8 concept of data element and coincides with the X3.8 contention that the data element names the kind of data or data items (entities) which make up its class. However, each standard data element represents one unique and defined property which ranges over a set of entities (data items). The entities may, in addition, have certain properties of their own which are either independent of or subordinated to the data element property. The data element "Corporate Author" has, for example, the property (explicitly stated by its definition) of being an organizational source responsible for the writing and generation of a publication or document.

The data items or entities which are subordinated to this element in the hierarchical structuring may be: "Name of Organization - Largest Unit"; "Name of Organization - Smallest Unit"; "Location"; etc. There may be in turn a superordinate data grouping which includes a number of data elements. The term "Author Entries" is, for instance, the data grouping which includes the elements "Personal Author Entry," and "Corporate Author Entry." It also includes the information null class of "no author given"!

Proper structuring of data elements consequently supplies a hierarchical arrangement of concepts in which the basic one-dimensional value(s) v, or the original TIH datum or data Dij, are subsets of a set (called data element) which may itself be a member of a family of subsets (data grouping). The data element vocabulary or data structure is therefore the set of all subsets.

Structuring presumes a law of types which prevents the properties and data entities from being confused with the classes of which they are members. The data can thus be named without mistaking the name of the kind of data with the name of each example of the data, or with the individuals represented by the data. This structuring enables different - higher and lower - levels of meaning to be designated. Once designated and fixed, once ranked according to appointed levels of generality and specificity, the data couched in natural language can readily be manipulated in machine-readable form as well as translated, if need be, and interchanged from system to system.

B. Term Levels

The discussion of structuring above tried to demonstrate that the terms used in the data element vocabulary may be considered generics and ranked in fixed positions relative to one another in a hierarchical scale. Uniformity and consistency are achieved by standardization. An example is given in Section IX. C. below, showing how this assignment of relative ranking and fixing is performed on two data elements "Symbols" and "CODEN." The data element "Symbols," the generic, denotes or contains the data item "Codes (General)." On the other hand, a whole code structure is named by the data element "CODEN," the name of a specific kind of code. Its data items in turn are the peculiar characters that indicate the titles of particular journals.

Terms appear at various levels to which they are assigned by convention or by the standardizing consensus principle. But a word of caution must be said lest standardization unwittingly restrict our handling of these semantic tools. Data processing and transfer must continuously respond to the needs of our vocabulary, not vice versa. The possibility always exists that because of some fixed structured approach, perhaps such as tree structuring, certain meaningful groupings of terms or retrieval strategies will remain unused. It is probably quite simple to avoid such rigidity by being aware of these pitfalls. For instance, the case may be imagined where we have a conceptual cluster structured in the following tree pattern:

Kinds of conceptual expression

- Signals
  - Gestures
  - Semaphore
- Symbols
  - Sound
    - Speech
  - Alphanumeric Characters
    - Codes (General)
      - CODEN
  - Punctuation

Standardization of "Codes" under "Alphanumeric Characters" exclusively may reduce our ability to retrieve "speech pattern recognition" as, for example, a code used to identify individuals.

One method used to differentiate between the data item, data element, and data grouping levels is to identify the level of each type of data collected into arrays (defined as n-dimensional collections of elements, all of which have hierarchically identical attributes) or other structures. Identification of level by means of a simple numbering device, the use of a two-digit numeric sequence, was tried out in the experimental data element file.

It is thus possible to display the hierarchical ranking of the components of a data element structure by simply counting the hierarchical level. The most general term is numbered "level 1"; and the next "level 2" and so on. We could have, for example, the data grouping "Author" (level 1), the data grouping "Member of the American Physical Society" (level 1), composed of data elements "Forename" (level 2), "Middle Name or Initial" (level 2), "Surname" (level 2), which denote the data items "John Q. Smith (level 1), "Mary J. Jones" (level 1). Another possibility is to have a subordinate data grouping under "Author" (level 1), say, "Corporate Author" (level 1), "Individual Author" (level 1). Then the data element "Surname" is (level 3), the data item "John Q. Smith" is treated as (level 4), etc.[3]

C. Classification

In addition to the natural or assigned structuring of a data element whereby certain attributes are generically arranged or attributed to its constituent data items, the device of classification was assayed.

As we have seen above, the structuring of a data element implies a classification. For it requires a two-dimensional matrix, where an entry in cell Rij indicates that the relation R holds between the entity, property or term i and that of j. But the data element also requires the process of standardization which tags the particular data element as a generic. It must by rule remain in a generic relation to its subordinate constituents or list items. That is to say, the data element is by fiat confined to a tree structure.

But the classification is not necessarily subject to this same restriction. Hierarchical structuring or levels are not mandatory.

To enrich the data element vocabulary, and to explore different organizational syntaxes, various classification schemes were developed and applied to the collection.

The first classification was radically pragmatic. It was simply a breakdown of the subject field as required by an overview of all the data elements on hand:

1. Subject Field

SYSTEMS ANALYSIS AND DESIGN CATEGORIES:

A. Bibliographies and bibliographic components (including document description)
B. Authorship, generation, production
C. Collection, formal literature, storage (incl. all forms of computer storage and file structures)
D. Intellectual organization (incl. content analysis: indexing, classification, abstracting, annotation, reviews; excludes management functions, systems analysis and design)
E. System analysis, design, planning
F. Equipment (hardware and software)
G. Processing
H. Retrieval
I. Dissemination, communication, publication, products (primary, secondary and tertiary)
J. Transfer, translation, conversion
K. Services, system structures, installations, centers, organizations, missions
L. Use, users, ends, needs, plans, channels

MANAGEMENT, ADMINISTRATION AND AIP FILE CATEGORIES:

M. Management, control and processing command
N. Personnel
O. File
P. Biography
Q. Historical event

R. Manpower, education
S. Subscription fulfillment
T. Doctoral program
U. Legal requirements

The main merit of these schemes was that they worked better on the data elements at hand than the others. But the schedule was cumbersome, inelegant, arbitrarily designed, and neither systematic nor transparently exclusive. However, simply abbreviated versions of the subject field division failed to be sufficiently discriminatory when applied to indexing and retrieval functions.

One of the classification and coding schemes which was suggested, and is now used on the data element description sheet shown in Appendix II, is as follows:

Simplified Classification

Field 1	Field 2	Field 3
1. Systems		
	1. Components	
		1. Hardware
		2. Software
	2. Aspects	
		1. Subject
		2. Physicists
		3. Users
		4. Institutions
2. Files		

Alternative classifications were developed or borrowed. The traditional classification systems, such as Colon Classification, Universal Decimal Classification, or even the Perry and Kent semantic code were found to be too general or non-explicit for the designation of the data elements in our three sectors.

Methods describing elementary items according to the following characteristics were considered:

Class (alphabetic, numeric, alphanumeric)
Size (number of characters or digits)
Sign (signed or unsigned)
Justification (left or right)
Synchronization (correspondence between item value and computer words)
Punctuation (position of editing symbols, etc.)

The classification scheme worked out at the Mitre Corporation for data element regulation of a command and control system was taken into account:

Presentation Form

R - Representation
N - Name
V - Value
E - Either (a name = alphabetic
or a value numeric)

Document Description

I - Implied
P - Punctuation

Attributes

L - Location
I - Identification
Q - Quantity
C - Condition
Ø - Not Applicable

Basic Group

T - Time
S - Space
A - Mission
W - Weapon
L - Plan
R - Provisions
E - Equipment
F - Facility
P - Personnel
M - Environment
V - Event
N - Natural Feature
D - Document
O - Organization
X - More Than One
Ø - Not Applicable

The following simple and clearcut classification was applied together with the more complicated schemes:

Systems	Files
1. Components	
Hardware	
Software	
2. Aspects	
Subject	
Physicists	
Users	
Institutions	

Adaptations of the Mitre scheme were made for functional purposes (ignoring the non-exclusive and mixed type character of the classes):

II. TYPE OF FORM OR CONTENT

A. Document, message or component
B. Sign, symbol or their conventional representation, including punctuation
C. Code
D. Documentation file

E. Event, process, product or doer
F. Fact (concept)
G. System or system component
H. Value

Ø Not applicable

III. ATTRIBUTE FACETS

N. Name or identification
S. Space, geography, location
Q. Quantity
L. Quality
C. Condition, status, development, state of affairs, state of existence
T. Time (duration or point of time)
E. Evaluation, analysis, conceptual synthesis

Ø Not applicable

A somewhat less objectionable, although less useful version of the Mitre approach was advanced:

PRESENTATION FORM	ATTRIBUTE
R - Representation of 1) a name of a series of digits 2) conventional symbols and signs	N - Number P - Percent T - Type, kind L - Location Q - Quantity C - Condition, incl. quality I - Identification Ø - Not applicable
G - Graphic	
N - Name	
V - Value	
M - Mixed	
C - Code	
A - Abbreviations, acronyms	
P - Punctuation	
I - Instruction symbol	

For the less than three hundred data elements in the experimental file, the facets for type of form or content and the attribute facets seemed superfluous. Naturally, it was recognized that eventual expansion of the number of elements in the data element control device could make a classification most useful and possibly even rely on the descriptive value of these facets for retrieval purposes.

CLASSIFICATION TAG

One device to aid in the identification and recognition of the data element was the use of a classification tag. The tag formed part of the code attached to each data element. The data element "Number of Man Hours," for example, was given the tag EEQ to signify: Position I: E = System analysis, design, planning; Position II: E = Event, process, product or doer; Position III: Q = Quantity. A mnemonic abbreviation was employed as an alphabetic code, yielding NOMHR for the example. The combination of tag and mnemonic produced a unique code for "Number of Man Hours" - EEQNOMHR.

DISCUSSION

The classification schemes that were developed and tried out (by comparing the existing data element file with possible new entries) proved interesting and successful as far as discrimination and sorting were required. But one faced the dilemma that each scheme entertained was either rationally faulty and useful or elegant and worthless.

An objection in principle to an extraneous classification scheme might be made in behalf of the data elements themselves. Each data elements is the unique name of a quality or relation. Each utilizes this uniqueness to describe and order the specific set of data items which it denotes. Further capitalization on classifying or ordering schemes might, even if they do not totally confuse the question of exclusive description, prove unnecessarily redundant.

So much for the descriptive capacity of classification. Other horizons are still open, however, and may merit further exploration than was possible in this study. The use of classification to interrelate and organize data elements into syntactic patterns useful for retrieval purposes might be investigated. The tree structuring required by the data elements bound to their data items, when combined and matched with non-hierarchical strings, may yield interesting correlations between the data items in different sets. Such combinations could prove interesting in analysis of management files.

IX. THE THREE AIP SECTORS

Let us now turn to a discussion of the three sectors identified in Section VI.

A. Bibliographic Data Elements

Following the analysis of physics journals made by Inforonics, Inc., in its attempt to derive unambiguous and retrievable bibliographic items, a list of "information items" (data elements) was compiled:[4]

- Journal title
- Volume
- Number
- Date
- Issue Title
- Article Title
- Abbreviated Article Title
- Author(s)
  - Forename
  - Middle Name or Initial
  - Surname
- Author Affiliation(s)
- Place of Presentation of Paper
- Date of Manuscript Submission
- Page Number
- Abstract
- Body of Text
  - Heading
  - Sub-heading
  - Sub-sub-heading
- Figure
- Figure Caption
- Table
- Table Caption
- Table Footnote
- Equation
- Footnote
  - Author (present author address)
  - Title (sponsor)
  - Text
  - Citation
- Non-alphabetic Symbols and Symbol Sequences
- Text Structure Data
  - Section
  - Paragraph
  - Sentence
  - Word
  - Character

Additional material not related to the primary journal product (copyright statement, information for contributors, and indexes) were omitted from the original Inforonics item identification list.

Certain identification requirements created by secondary use of data are noted in the Inforonics report and differentiated according to the type of use. The first type involves the creation of bibliographic reference tools, such as author, title, and subject indexes. The second type is required for text extraction, such as abstract journal entries, announcements of publication or compressions required for review. Analysis was made of the contents, forms, formats and procedures required for the different types of indexes and text presentation. The present study accepted with slight modification the basic list of data elements suggested by the report. However, further analysis of individual elements resulted in somewhat different structuring. Rather than data element "Author," for example, the experimental file contained data group identifier "Personal Author" which was a composite of data elements "Forename," "Middle Name," and "Surname." The data use identifier for "Personal Author" may be, e.g. for the bibliographic units "Journal Article," "Textbook," "Paper in Conference Proceedings," etc. A later listing of the data elements required for journal article presentation which was developed by Inforonics, Inc. and Vance Weaver Composition, Inc. was adopted in the experimental file and is reproduced in Appendix I.

Numerous groupings of potential data elements presented themselves as this area was further explored. Over two hundred plausible candidates for the experimental file were suggested by the invaluable work by Ann T. Curran and Henriette D. Avram, The Identification of Data Elements in Bibliographic Records.(5)

On the other hand, the actual operational file used at AIP during the study contained only seven bibliographic data elements. The file was the machine-readable store of physics journal literature maintained in the Technical Information Project (TIP) store in the MAC system at Massachusetts Institute of Technology. The seven data elements made up the 'unit record', containing journal, volume, page(s), title, author, affiliations, and citations. Later additions to the unit record will include physics subject classification number(s), index terms, and possibly abstract(s). Comparison between the TIP data elements and those used, planned or contemplated by Physics Abstracts, published by the Institute of Electrical Engineers (IEE) in London disclosed several other candidates such as language of original publication, paper number (including CODEN), corporate author, PA classification code, etc. In addition, certain highly specialized collections at AIP, such as the Bio-Bibliographic Collection of the Center for the History and Philosophy of Physics maintain such extraordinary data elements related to manuscripts, tape recordings, apparatus. that a separate division of elements seemed warranted.

All of the candidates were noted. Many that were synonymous or nearly synonymous or translatable were noted as such. But certainly not all possibilities were incorporated into the file.

A number of options or criteria for selecting the data elements were available. One could use a stockpiling approach and enter all possible, usable or unusable, elements into the file. Or one could apply canons of use. If the first approach was taken, about a thousand - mostly unusable - elements would be selected for bibliographic purposes. If the latter was chosen, it would be necessary to require that the elements be actively used at AIP. In such a

case, only seven elements could be allowed. The more practical criterion, requiring that the elements be used in interchange between system interfaces, a method recommendable for later operational implementation of a data element structure, was even less adoptable. For no active interchange between systems was occurring - except for the work of IEE, the details of which were not yet available.

Selective adoption of elements was the course elected among the alternatives. One of the criteria for selection was whether the elements would probably continue to be used for journal input to a primary data base. Elements from TIP, Physics Abstracts and several AIP primary and a few other secondary journals, (e.g. Nuclear Science Abstracts, Science Citation Index) seemed likely to remain significant for primary input as well as bibliographic purposes. The actual keyboarding operation when journals are being inputted to the computer for later photocompostion would undoubtedly require still finer differentiations, especially for certain mixed entries appearing in footnotes and citations. Naturally, the assemblage of bibliographic data elements in the experimental file was considered tentative. The present collection will unquestionably differ from the later operational file. The full testing of the data element control structure to determine its efficacy in facilitating data interchange and perhaps even in performing its secondary data retrieval function can be properly made when the system is fully operational. The present study has addressed itself principally to exploring a method of handling such a control structure.

Listings of sample bibliographic data elements from the experimental file may be seen in Appendix I. Appendix V shows a comparison of sample data elements used in various systems at AIP.

B. AIP Files, Records and Resources

The second sector comprises AIP management and special services files. It forms a crossroads for many of the Institute's services, services which have accumulated considerable amounts of information. Most of this information is organized, much of it can be represented in a data element vocabulary. Consequently, the AIP files in the Subscription Fulfillment Division, the resources of the Center for the History and Philosophy of Physics, the cumulative author index of AIP journal contributions the Education and Manpower Studies publications, the Directories of Physics and Astronomy Faculties and numerous other sources were examined for possible data elements of interest to the AIP Information Program.

Generally, two classes of elements appeared in this survey: those of possible interest from an information management perspective, and those which dealt with physicists - individually and collectively - and with events in the subject matter and history of physics.

Each division and service was looked at without preconception or design. The analysis of each area was carried out without consideration of possible requirements of existing AIP operations. No thought was given to any changes in existing practices. Only the possible data element vocabulary which each could yield wasconsidered. In this manner, it was possible to characterize the classes of data as they now appeared. Their possible

utility for management decision making or for special tabulations, listings and products could be considered on the basis of the elements alone.

1. Subscription Fulfillment

The Subscription Fulfillment Department has been converting its automated punched card system to a magnetic tape file for operation on a Univac 9300. The data elements required in the three major records have been clearly identified and coded. The three major records consist of a master history record file, a society record and a journal record.

A few samples of the data elements used in these records may be instructive:

Field Number	Field Name (Code)	Content (Data Element)
1	CC	Card Record Code
2	TRAVC	Transaction Code
3	ACCT	Account Number
4	RECT	Record Type
5	ZIG	ZIP or Geographic Code
6	FNAME	First Name
7	MNAME	Middle Initial
8	LNAME	Last Name
9	LINE 2	2nd Line of Address
10	TITLE	Title
11	STAC	State Code
12	LINE 3	3rd Line of Address
13	LINE 4	4th Line of Address
14	DTSUS	Date of Suspense

Field numbers 6,7,8 and possibly 10 contain information that appears as bibliographic data elements.

The Society Record contains certain biographic information of possible interest to historians of science. This is indicated by the data elements "Date of Election (to membership)," "Date of Promotion to Highest Class," "Date of Birth." The Journal Record contains an important code structure for the primary physics journals published by AIP. Another field of possible use in the dissemination of information is indicated in the Society Record by the data element "Area of Interest."

Out of the 41 data elements listed in the experimental file for the subscription fulfillment function, ten elements are of possible significance for conventional information purposes. Varieties of management statistics may be obtained by means of other data elements. Such statistical reports might conceivably cover various segments of the journal subscribers, classified according to, say, areas of interest or society membership. Circulation

and business data needed for publication control as well as day-to-day operations are readily identified by the classes of data elements. It may conceivably become possible some day to perform statistical forecasting without excessively complex new programming based on the data organized by the data elements. One might, for example, not only be able to identify the individuals and organizations interested in a new class of services, but predict the volume of subscription according to past performance in other media.

2. Center for the History and Philosophy of Physics

One of the nation's significant repositories of information about physicists and events in physics, this Center is based on a nucleus of several manually tended collections. They are comprised of the National Catalog of Sources, the Bio-Bibliographic Collection, the files of the Oral History Project (made up of tape recordings and transcripts of interviews) and the Niels Bohr Library - which houses about 5,000 volumes and a unique historical archive containing manuscripts and documents.

As might be expected, the high degree of organization of these collections was instrumental in supplying many data elements. The cataloging for the Niels Bohr Library reflects standard national practices. The manuscript collection, for instance, follows the prescribed data sheet for the National Union Catalog of Manuscript Collections recommended by the Manuscripts Section of the Library of Congress.

Data elements, data groupings or data use identifiers suggested by this cataloging process include, e.g. "Name of Repository,""Principal Name around Which the Collection Is Formed," "Occupation of Principal Person, Family or Corporate Body," "Form of Manuscript Reproduction" with data items: "Handwritten Transcripts," "Typewritten Transcripts," "Positive Photocopy," "Negative Photocopy," "Positive Microfilm," "Negative Microfilm," "Number of Microfilm Reels," etc. The data element "Types of Papers" have data items: "Correspondence," "letters," "diaries," "documents," etc.

The National Catalog of Sources for the History of Physics maintains a card catalog from which information may be retrieved concerning historical source materials such as manuscripts, diaries, experimental notebooks, interviews, correspondence and apparatus. Data elements or Data use Identifiers "Name of Physicist," "Name of Organization," and general subject terms, e.g. "Nuclear Physics," "Accelerators," "Administration of Science," etc are used here.

The Oral History Collection on Twentieth Century Physics maintains records that contain the following data elements and data use identifiers: "Tape Number(s)," "Name of Interviewee," "Date of Interview (or speech)," "Number of Hours of Interview," "Date Sent to Transcriber," "Transcriber's Name," "Date Transcript was Received from Transcriber," "Number of Pages of Transcript," data element "Corrections" with data use identifier, "Completed Date of Corrections Made on Transcript Against Tape," etc.

Many of the data elements derived from other divisions coincided with common bibliographic elements or groupings: such as "Name of Organization," "Surname," "Address," etc. Some data elements, such as "Press Release" and

"Visitor's Evaluation of the Institution" from the Visiting Scientist Program require a full text as their data item. Others, particularly from Education and Manpower Studies, presented contexts which required considerable analysis to derive a standardized data structure. For example, the concept and data expressed in a table entitled "Size of secondary school attended by doctorate-holders" was treated as follows. Data element: "Size of Secondary School," its data items: "School Enrollment Size 1-19;" "School Enrollment Size 20-59," etc. through "School Enrollment Size 500 or more." Data Element: "Percentage of Physics Doctorate-Holders," data items: the values "5.0," "31.2," etc. for both, the data use identifiers: dates "1960-1," "1961-2," etc. For comparison, there were the data elements "Percentage of Chemistry Doctorate-Holders," "Total Doctorate Holders," etc. The importance of such elements lies in their potential use for storage, manipulation and perhaps retrieval, probably in the handling and control of questionnaires of similar multiple forms.

Finally, study of the Institute's needs for a management information system and its feasibility might be considered at some future time. A centralized data element file will probably help to implement such an undertaking. Naturally, before designing a data element file for such a purpose, the desirability and need as well as the details and magnitude of AIP's requirements would have to be specified by prior study.

C. System Analysis Vocabulary

The language needed to perform system analysis, design and development activities in an information program is different from the vocabularies of the first two sectors in several ways. The subject matter is necessarily different; the lists or data items named by the element generally resemble inventories; and the function of the entire structure seems more to serve internal "housekeeping" purposes than act as an interchange language between systems.

The following five types of terms have been adopted as system analysis data elements: 1) Names for manifolds such as files, catalogs and inventories; 2) Relatively self-contained concepts or systematic overviews that have familiar, conventional or readily assignable sub-divisions, such as an integral classification scheme, for example, the new hierarchical for physics; 3) Names of key activities, functions, processes or operations in an information handling establishment, which can be broken down into a finite number of steps (data items), e.g. "Retrieval," items: subject analysis, question formulation, encoding, search of system, etc.; 4) Criteria or evaluations that require a checklist for their data items, e.g. the hardware capability of a "Processor": data items - number in line, speeds, parity check - I/O failure checking, computer circuity, etc.; and 5) A category of miscellania somewhat less institutionalized that 1) that describes sundry items which belong to a real world of less than coherent objects, bric-a-brac, rules, commands. Codes would be named by such a category, the code units would form its data items.

To illustrate the first type we can cite the data element "Name of System" which appears as a member of the data grouping "General Description of a System" which appears as a member of the data grouping "General Description of a System." Some data items for this element are "National Physics Information

System," "Clearinghouse for Federal Scientific and Technical Information," "Science Information Exchange," etc. The data element "AIP Information Files" has data items: "Information Division Library Card Catalog," "TIP Store," "Education and Manpower Department Records."

Data elements illustrative of the second type of system analysis terms are "Universal Decimal Classification," "Colon Classification," "Newton's Laws of Motion." The contents of each classification system have been construed as the data items, as have the descriptions and series of equations for each of the three laws of motion formulated by Newton in 1687. The data grouping for the first two elements is "Classification Schemes" or "Classification Schedules," depending upon what is meant.

The third type names activities or processes, such as "PERT Analysis," or "Benefits from Data Element Standardization." The latter might have the data items: 1) "Facilitation of interchange and compatibility of data among different data processing systems," 2) "Reduction in total number of data elements and codes," 3) "Reduction in processing costs by using standard codes instead of full descriptors," 4) "Facilitation of the development of standard information and data systems by standardizing the elements and codes," 5) "Facilitation of systems integration and direct computer-to-computer information transfer."

The fourth type supplies those criteria or canons for evaluation that require lists which every 'expert' should always have on the tip of his tongue or at least at his fingertips, e.g. data element "Capabilities of a Software Master Control System." Data items: 1) "Static or dynamic storage allocation," 2) "Controls," 3) "Interrupt Handling," 4) "Task Scheduling," 5) "Multiple Processor Capabilities," 6) "System/Operator Interface," 7) "Debugging Features," 8) "Accounting Capability," 9) "Programming and Data Protection," 10) "Time-Sharing (Conversational)," 11) "Foreground/Background Processors," 12) "Processors under Control of Master Control System," 13) "Device Independent."

The fifth type contains a category of somewhat arbitrarily chosen terms. Many of the bibliographic terms could be data elements in this sense. For instance, take the composite term or data use identifier "Unit Records for Physics Information" used by Physics Abstracts. Its component data elements are, inter alia "Paper Number," "Chapter Code," "Author's Affiliation," etc.

The system data element "Symbols" might have data items "Characters," "Signs," "Numerics," "Punctuation," or even "Codes." However, "Codes (General)" will require a special data use identifier. For, due to the standardization process, the name of a particular code, say, "CODEN" for serial title abbreviations, may stand for a data element, the data items of which are the specific codes within it that designate each periodical title. The latter course for structuring CODEN was followed in the experimental file.

This fifth type occurs more typically in such data elements as "Number of Respondents (to a questionnaire)," "Type of WIC [Written Informal Communication]."

Data elements of the type used in systems analysis may be used to index tables. For example, the headings "Media" and "Percent Selecting" can be

regarded as data elements in the following table. The data items of "Media" are then "Journal and Abstract Indexes," "Regular Scanning of Literature," etc. The data items of "Percent Selecting" are "21," etc. When the data elements are matched in the same array and their data items are correlated in a specific matrix, the full table is reproduced:(6)

Media	Percent Selecting
Journal and abstract indexes	21
Regular scanning of literature	45
etc.	etc.

The system data element vocabulary stock could turn out to be extremely large even when the terms are selected with restraint. Before incorporating this vocabulary, consideration must be given to the resources and requirements of the system. Appendix I presents a few additional examples of the data elements from this sector, accommodated by the experimental file.

X. THE EXPERIMENTAL DATA ELEMENT FILE

A. Overall Features

An experimental manually operated file, a possible prototype of an operational system, was set up in the AIP Information Division. Its purposes paralleled the study. The file structure was exploratory. It was an attempt to raise questions and find solutions at a microcosmic level which could be applied to the larger system of which it was a model.

The function of the file was to accommodate the data element structure, to integrate all of its relevant parts, especially the "building blocks" of the information system, the data elements. The data elements were identified, processed, and then recorded by the mechanism of the file.

The principal recording instrument was a standard Data Element Description Sheet. Several versions of the sheet were drafted and employed. Appendices II and III reproduce two of these versions. Additional forms were used for cross-references and qualifiers.

B. Detailed Components

Data Element Description Sheet

It is essential to the standardization process that each data element be identified, registered, defined, classified, and coded in a uniform manner. To make this possible within the framework of an experimental data element file, a principal recording instrument, the standard Data Element Description Sheet, was drafted.

Data Element Name

Each data element is given a unique and unambiguous name. Its meaning must be clearly distinct from that of every other data element. Generally, the name consists of a noun or noun phrase, common segments of which may be regarded as qualifiers of the more unique portion. In our example, "Number of Man Hours," "Number" is a qualifier of "Man Hours." In exceptional cases, where ambiguity can arise, a term qualifier or scope note (entered in parentheses after the name) may be used to distinguish this element from another.

Qualifiers

Qualifiers are entered on a separate qualifier sheet, and are alphabetically interfiled with the Data Element Description Sheets. Synonyms of qualifiers are noted.

Data Element References

Names or abbreviations differing in form from the data element name are entered on the line appropriate for AIP usage or that of an outside organization with which the data element is identical.

Data Element Code

The unique coded reference to the data element used by AIP and/or other organizations is recorded. The mnemonic "NOMHR" is given in the example.

Type

Indication of whether the entry under data element name is a data element (basic) or a data element grouping (composite). Basic data elements are, e.g. "Year," "Month," "Day of Month," The composite or data grouping is "Date." In the latter case, the data elements "Year," etc. are entered on the line Data Items.

Synonyms

Synonyms of the data element are entered on this line and on a separate Reference to the Data Element sheet. See Appendix II.

Data Group Identifier

The name or designation given to a composite or combination of two or more related data elements.

Data Group Reference

A number, letter, code, or other symbol used to represent a data group identifier.

Definition

An acceptable and distinct definition of the data element is entered in cases of ambiguity between two data elements that are not resolved by parenthetical term qualifiers or scope notes.

Data Items

Each data item classified under a data element will have a unique name and meaning different from any other data item classified under that data element. A data item may be given a unique abbreviation or code, entered under code. Each data item classified under a data element must have a homogeneous characteristic that allows it to fit within the data element grouping. A data item cannot be logically subdivided and retain significance of the data element class.

Data Item Code

Existing codes should be considered at the implementation stage to minimize conversion. The length of the code should be short to conserve storage space and transmission time. Mnemonic abbreviations should be considered as codes to facilitate human use and understanding. For machine to machine interchange numeric codes may be preferable. In any case, the code should be designed to provide high reliability in interchange. The code should allow for adding or inserting new members without having to recode or expand the code length. Redundancy may be considered, when appropriate . Where applicable, the code should provide for sorting so that, if a sort operation is performed on the code the members are ordered in the desired sequence.

Data Use Identifier

Each data use identifier is different from any other data element or related feature. A data use identifier may be given a unique mnemonic abbreviation, entered under Data Use Reference. Data use identifiers apply

to the data items of the data element from which they are derived. Two or more data use identifiers can be chained to each other in a prescribed sequence and used as data group identifiers. For example, two data use identifiers called "City of Birth"and"State of Birth" could be grouped together to form a data group identifier called "Birth Place."

This arrangement is based on the notion that, when the data items of a data element appear in a system, they are used in specific contexts and have specific connotations. These uses do not change the class, the data items of the basic definition of the data element. Such applications are named by data use identifiers. For example, consider the data element, "States of the United States." The system may require "State of Birth." In the system design, the terminology "State of Birth" could be used to name a file, and would be designated as a standard data use identifier. Subsequently, whenever it becomes necessary to use a data use term for "Birth State" or other designation with the same meaning, the standard data use identifier "State of Birth" should be used. Other examples of data use identifiers for the data element, "States of the United States" might be "State of Residence," "State of Legal Residence."

Classification

Two experimental sets of classification terms were tested. The question of classification is discussed in Section VIII. C -Classification.

Array

A two-digit code discussed in Section VIII B - Term Levels.

Filing

The experimental Data Element File is thus a manually handled collection of Data Element Description Sheets, References and Qualifier Sheets, filed alphabetically. In some cases, additional Data Element Description Sheets are entered into the file under the names of significant data items and data use identifiers.

C. Criteria

The definitions and guidelines used to specify and record the data element as a class of data, whose members are data items, indicated initial criteria which can be followed to identify the data element structure. Once the terms have been identified and interrelated according to this structure, the next stage is that of standardization.

Let us refer to the recommended procedure in the USASI X 3.8 - Technical Guidelines to demonstrate the procedure followed in accommodating these essential terms.

(1) To assure the broadest scope and application of the resulting standard data element, research existing data systems, publications and forms where similar or equivalent data elements are likely to exist. The first thing to establish is the tentative list of data items which determines the true meaning of the data element by establishing its homogeneous class characteristic. As far as possible, lists of all data items involved should be assembled or compiled.

(2) As research proceeds, data elements should be identified, and recorded to show: the title each time it appears; any data element definitions which appear; any additional data items, and the publication, form citation, or system in which they appear. The breadth and depth of research and recording must be tailored to the situation. Obviously, any data element with limited data items such as "Month" does not require an extreme depth of research into systems before all the possibilities are exhausted. Nor will it be necessary to record its use each time. On the other hand, a more complex data element, such as organizational entity, exists in so many different forms, in so many different contexts, in so many different permutations and combinations that extremely broad and deep research must be undertaken before the requirements and the existing solutions to these requirements are known, which is the basis for development of a standard data element. (In other words, all data elements are not the same in essential nature and therefore the rules cannot be arbitrarily applied. The nature of the words in the language which have been required and selected for use in data systems varies widely. Contrast "Month" with twelve data item possibilities to organizational entity which first requires extended conceptual development to determine exactly what it is , how it should be defined, whether it needs to be divided into several data elements, etc. long before the list of data item possibilities can be examined in any detail toward standardization.)

(3) Survey the known and foreseeable data system requirements of the various participants and interested organizations, in terms of the data element(s) under consideration, and list them by tentative title. Again the depth and breadth of the survey must be tailored to the situation.

(4) Study the elements of data extracted during the research. Select and develop the one which most nearly meets each anticipated data requirement. Name and define the element and its related features in accordance with the criteria set forth above.

The experimental file proved effective in regulating the usage of class terms used to refer to bibliographic data. As more records were received, the existing data elements and their components in the file needed less and less to be changed. More data use identifiers were required for the same data elements, rather than new elements. Unfortunately, the amount of data encountered during the identification phase was still relatively too small to allow any quantification for testing purposes. The use of a larger data base for comparison against the present existing file of around three hundred terms will probably allow measurement and some consideration of cost factors related to the efficiency and maintenance of the data element file.

XI. RECOMMENDATIONS AND CONCLUSIONS

A standard data element structure to control the use, transfer and interchange of records, particularly from machine-readable files, was examined and found viable. Three major areas of application for the file were discovered at AIP with differing types of data elements and varying requirements. Considerably more work is needed to clarify the formal definitions of the basic terms, although analysis has made it possible to understand and work with the essential data element structure. Standardization is also required at a number of levels. The standard data element file is highly adaptable and corrigible and can adjust to innovative standardization at the national level by controlling data conversion as well as identifying common elements and codes.

Operational implementation of the experimental prototype file is recommended. It is also proposed that eventual automatic processing of the operational file be considered.

The manually operated experimental data element file could be automated in two sequential stages. The first stage could begin at a semi-automated level. The beginning would be marked by studies and decisions relevant to coding and formatting requirements. The open questions that were raised with regard to standardization would have to be practically resolved. These include the establishment of common codes for data elements and data items, standard character types, allowable character set expansion, message and field formats, flags and field size. Questions relating to media would of course be determined by available equipment. The prototype file could then be converted to automatic processing. Human judgment would at least at the outset be needed to identify, define, and structure the terms. In addition, human comparison of the model file data elements and the elements used by other systems in machine processing of their data bases would have to be performed at the initial stage.

If the method proves successful in automatically regulating data flow, perhaps during the input stage for photocomposition or the output for bibliographic retrieval, thought might be given to automation of the data element file on a still larger scale. It could at such a more advanced stage serve as a machine-controlled authority file on vocabularies. It is imaginable that the value of such a file could ultimately outweigh the programming problems that this approach would entail.

APPENDIX I

SAMPLE LIST OF DATA ELEMENTS IN EXPERIMENTAL FILE

1. BIBLIOGRAPHIC DATA ELEMENTS AND DATA GROUPINGS

- Personal Author(s)
  - Forenames or Initials
  - Surname
  - Titles or Identifiers
    - If Appear in Index Entry
    - If Precede Forename in Natural Order
  - If Corresponding Author
  - If 1st, 2nd, 3rd Author
- Corporate Author
- Title of Article
  - Subtitle
- Author's Position
- Author's Affiliation (Present)
  - Name
  - Location
  - If 1st, 2nd, 3rd Author
- Author's Affiliation-- Name, Location
- Organization Where Work Was Done (If Different from Affiliation)
  - Name
  - Location
  - If 1st, 2nd, 3rd Author
- Manuscript Received (or Submitted) Note
  - Date
- Sponsor Note
- Presented of Conference Note
- Miscellaneous Note
- Abstract
- Text of Article
- Subheading
- Table Caption
- Figure Caption
- Equation
  - Equation Number
- Summary or Conclusion
- Numbered Footnotes
  - Text Only (No Reference)
  - Reference(s) Included
    - Text
    - Citation(s)
      - Book
        - Author
        - Title
        - Edition
        - Place Published

Publisher
Date
Pages (or Volumes)
Comment (Any Data Other Than Elements Listed Above)
Part of Book
    Author of Part
    Title of Part
    Author
    Title
    Edition
    Place Published
    Publisher
    Date
    Pages (or Volumes)
    Commentary (e.g., "In," "Edited by")
Journal Article
    Author of Article
    Title of Article
    Periodical Title
    Volume Number
    Beginning Page
    Date
    Commentary (e.g., suppl.)
Patent
    Author (Inventor)
    Country
    Patent Number
    Date (Year)
    Commentary
Miscellaneous
    Author
    Title
    Date (Year)
    Commentary
References (at End of Article)
    Periodical Title
    Periodical CODEN
    Periodical Abbreviation Non-USASI
    Country of Publication
    Series
    Volume Number
    Issue Number
    Part Number
    Supplement Number
    Season
    Month
    Day
    Year
    Article Sequence Number
    Beginning Page Number

Ending Page Number
Subject Terms
Subject Codes
Short Title (Index Annotation)
Descriptive Phrase Annotation
UOC Number
AIP Classification Number
Type of Bibliographic Form
  (e.g., letter, research note)
Type of Work
  (e.g., experimental, theoretical)
Language
Language of Summaries
Acceptance Oate
Keyboarder
Date Keyed
Input Keyboard Number
Translator(s)
    Forenames or Initials
    Surnames
    Titles or Identifiers
        If Appear in Index
        If Precede Forename
    If Not, Ed., Translator
Title of Article (Original)
    Subtitle
Periodical Title (Original)
Periodical COOEN (Original)
Periodical USASI Abbreviation (Original)
Periodical Abbreviation, Non USASI (Original)
Periodical Translated Abbreviation (Original)
Country of Publication (Original)
Series (Original)
Volume Number (Original)
Issue Number(Original)
Part Number (Original)
Supplement Number (Original)
Season (Original)
Month (Original)
Day (Original)
Year (Original)
Beginning Page Number (Original)
Ending Page Number (Original)
Submission Oate (Original)
Acceptance Oate (Original)
Language (Original)

2. MANAGEMENT DATA ELEMENTS

Education and Manpower Qualifications

- ›nal Data
  - "ame
    - Surname
    - Middle Name
    - Forename
  - Address
  - Birthdate
  - Marital Status
- Education
  - Graduate
  - Undergraduate
  - Secondary School
- Employment Record
- References
- Title of Theses, Principal Research and Publications
- Years of Training and Experience
- Preferred and Acceptable Positions
  - Industrial Research
  - Undergraduate Teaching
  - Academic Research
- Minimum Acceptable Salary
- Professional Affiliations
- Date of Availability
- Geographic Limitations

3. SYSTEMS DATA ELEMENTS (Including Data Groupings and Data Use Identifiers)

SYSTEM DESCRIPTION FORM *

- Name of Organizational Unit
- Address of Organizational Unit
- Formal or Functional Name of System
- Name and Title of Person to Whom Your System Manager Reports
- System Use of Software and Hardware Devices:
  - Uniterm Cards **
  - Peek-a-boo Cards **
  - Edge-Notched Cards **
  - Standard Tabulating Cards **
  - Microimage Searching Devices **
  - Computers or Other Devices Using Paper Tape or Magnetic Media **
- Specific Missions or Functions for Which System is Operating Major Activities
  - Research and Development **
  - Production and Quality Control **
  - Marketing **
  - Design and Planning **
  - Others **
- Medium of Storage of Documents That Are Retained
  - Full-Size Hard Copy **

Microcards (Opaque) **
Roll Microfilm **
Aperture Cards **
Microfiche, Sheet or Strip Microfilm **
Punched Cards **
Magnetic Tape **
Magnetic Disc **
Magnetic Drum **
Others**

Devices or Techniques used to Establish Relationships, Contexts of Subject Concept Terms at Time of Input or Indexing

Fixed Order of Subject or Concept Terms or Headings **
Role, Cause and/or Effect Indicators or Interfixes **
Partitioning of Document Via Links **
Indexing Classification, or Posting to Show Generic, Specific, Coordinate or Collateral Relationships, Including Cross-referencing**
Functional Group Relationship Indicators (e.g. chemical element indicators) **
Logical Connectives **

* From National Science Foundation, System Description Form for Nonconventional Scientific and Technical Information Systems in Current Use, No. 4, Washington, D.C., 1966, pp. xviii-xxvii.

** Data Items

SYSTEMS AND OPERATIONS ANALYSIS

Innovation Required
Lack of Knowledge of Operational Requirements
Number of Organizational Users
Number of ADP Centers
Complexity of Program System Interface
Response Time Requirements
Stability of Design
On-Line Requirements
Total Object Instructions Delivered
Percent Delivered Object Instructions Reused
Total Nondelivered Object Instructions Produced
Percent Source Instructions Written in POL=Procedure-Oriented Language
Percent of Total Object Instructions Discarded
Percent of Total Source Instructions Discarded
Number of Conditional Branches
Number of Words in the Data Base
Number of Classes of Items in the Data Base
Number of Input Message Types
Number of Output Message Types
Number of Input Variables
Number of Output Variables
Number of Words in Tables, and Constants not in Data Base
Percent Clerical Instructions
Percent Mathematical Instructions

Percent Input/Output Instructions
Percent Logical Control Instructions
Percent Self-Checking Instructions
Percent Information Storage and Retrieval Functions
Percent Data Acquisition and Display Function
Percent Control or Regulation Function
Percent Decision-Making Functions
Percent Transformation Functions
Percent Generation Functions
Average Operate Time
Frequency of Operation
Insufficient Memory
Insufficient I/O Capacity
Stringent Timing Requirements
Number of Subprograms
Programming Language
POL Expansion Ratio
Support Program Availability
Internal Documentation
External Documentation
Total Number of Document Types
Total Number of External Document Types Written During a Programming Step
Total Number of Internal Document Types Available From Previous Step
Total Number of Internal Document Types Written During a Programming Step
Type of Program
- Business
- Scientific
- Utility
- Other

Compiler or Assembler Used
Developmental Computer Used
First Program on Computer
Average Turnaround Time
ADP Components Developed Concurrently
Special Display Equipment
Core Capacity
Random Access Device Used
Number of Bits per Word
Memory Access Time
Machine Add Time
Computer Cost
Percent Senior Programmers
Average Programmer Experience with Language
Average Programmer Experience with Application
Percent Programmers Participating in Program Design
Personnel Continuity
Maximum Number of Programmers
Lack of Management Procedures
Number of Agencies Concurring in Design
Customer Inexperience

Computer Operated by Agency Other than Program Developer
Program Developed at Site other than the Operational Installation
Different Computers for Programming and Operation
Closed or Open Shop Operation
Number of Locations for Program Data Point Development
Number of Man Trips
Program Data Point Developed by Military Organization
Program Data Point Developed on Time-Shared Computer
Complexity of System Interface with Other Systems
Security Classification Level
Number of Sources of System Information
Accessibility of System Information
Degree of System Change Expected During Development
Degree of System Change Expected During System Operations
Number of Functions in the System
Number of System Components
Number of System Components -- Not Off-the-Shelf
Percent Senior Analysts
Quality of Resource Documents
The Availability of Special Tools
Degree of Standardization in Policy and Procedures
Number of Official Reviews of Documents
Personnel Turnover
Output Volume
Input Volume

APPENDIX II

AIP DATA ELEMENT DESCRIPTION SHEET
(CURRENT)

TYPE:
BASIC [x]
COMPOSITE [ ]

ENTRY:
NEW [x]
CHANGE [ ]
ADD. [ ]
DEL. [ ]

DATA ELEMENT NAME: NUMBER OF MAN HOURS

DATA ELEMENT REFS: AIP /E/E/Q/ N O M H R (MNEMONIC / ABBREV) OTHER REFS:

QUALIFIERS

DEFINITION: Amount of staff time per man required to perform a given job or task, expressed in hour units.
Range=0 to n hour units

SYNONYMS: Man hour consumption

	NAME:	ABBREV.:	CODES:	EXPLANATION:
DATA ITEMS: (1)	Sample; 5 hours (3)			
(2)				
(3)				
(4)				
(5)				
	IDENTIFIER:			
DATA USE: (1)	Programming			
(2)	Job/Task Analysis			
(3)				
(4)				
(5)				

(CONTINUED [ ])

DATA GROUP: Job Time Requirements (1)

OTHER SYSTEMS (NOTES)

NOTES (SOURCE, REMARKS)

CLASSIFICATION: /1/1/2/ (CODE) /0/2/ (ARRAY)

PREPARED BY ARB DATE 30 Jan. '68

REVIEWED BY ARB DATE 1 Feb. '68

APPENDIX III

DATA ELEMENT DESCRIPTION

DATA ELEMENT NAME: FORENAME

TYPE:
Basic ■
Composite □

DATA ELEMENT REFERENCES: AIP: FORENM

CODE:

OTHER:

SYNONYMS: First Name, Given Name, Christian Name

CLASSIFICATION

SYSTEMS □
1. Components
Hardware □
Software ■
2. Aspects
Subject □
Physicists □
Users □
Institutions □

FILES □

DATA GROUP IDENTIFIER: Name (1)

REFERENCE:

DEFINITION: Designation that differentiates one individual from other members of his family; the first name, a distinctive designation or appelation which precedes all other words that comprise the full name.

DATA ITEMS: Sample: John in the grouping "John Q. Smith" (3)

CODES:

DATA USE: Full alphabetic or initials: used in Data Grouping "Author"

DATA USE REFERENCE:

AUXILIARY CODES

TAG [A][A][N]

ARRAY [0][2]

Prepared By ARB Date 19 Feb. '68

Reviewed By _____ Date 20 Feb. '68

APPENDIX IV

Structuring Display of Bibliographic Data Elements

Structured

Data Grouping				Data Element	Sample Data Items
Level 1	Level 2	Level 3	Level 4	Level 5	Level 6
Entries	Main	Author	Personal Author	Surname.....	Smith Jones
				Middle Name or Initial (s).	Quincy Q.
				Forename or Initial (s)....	John Albert J.

			Data Element		Data Items	
Level 1	Level 2	Level 3	Level 4	Level 5	Level 6	Level 7
			Corporate	Name of Organization	Main Body.....U.S. House of Representatives	
					Subdivision(s)...Committee on Science and Astronautics	
					Sub-Subdivision(s)..Subcommittee on Science Research and Development	

Hierarchical Display Micron Ignored

APPENDIX V

OCCURENCE OF COMMON BIBLIOGRAPHIC DATA ELEMENTS

BIBLIOGRAPHIC ENTRY (DATA ITEM)	DATA ELEMENT	DATA GROUPING OR USE IDENTIFIER
Journal Article		
Jones, John Q.	1. Surname	A. Name of Author
Physics Laboratory, Ohio	2. Middle Name or Initial	
State University, Columbus, Ohio	3. Forename	
Conversion of Light to Ultrasonic Energy	4. Name of Organization a. smallest unit b. largest unit of organization	B. Affiliation of Author
Annals of Physics	5. Location of Organization	
Vol. 26, No. 3, March 1967	6. Title of Article	
pp. 369-374	7. a. Name of Journal b. CODEN (Journal code)	
	8. Volume Number	
	9. Issue Number	
	10. Month	C. Date of Publication
	11. Year	
	12. Page number a. First b. Last	
Book		
Smith, Harry J.	1.	A. Name of Author
Department of Theoretical Physics	2.	
Rockefeller Univ., New York, N.Y.	3.	
Atomic Collision Processes	13. Title of Book	
McGraw-Hill, New York, N.Y.	14. Name of Publisher	
1968	15. Place of Publication	
	11.	C. Date of Publication

BIBLIOGRAPHIC ENTRY	DATA ELEMENT	DATA GROUPING OR USE IDENTIFIER
Patent		
Brown, B, B.	1.	D. Name of Inventor
American Institute of Physics	2. (Initial only)	
Photoelectric scanning device	3. (Initial only)	
Cl. 250-239 1 Jan. 1968	4. (Largest unit only)	
Filed 31 Dec. 1967	16. Patent Title	F. Name of Assignee
9,999,999	17. Patent Classification No.	
	18. Day	C. Date of Issuance
	10.	
	11.	
	19. Country of Issuance	
	20. Patent Number	
OTHER ENTRIES		
Subscription Records	1.	G. Neme of Subscriber
	2.	
	3.	H. Name of Subscribing Organization
	4.	
	21. Address	I. Address of Subscribing Organization
Society Records	1.	J. Name of Society Member
	2.	
	3.	
	21.	K. Address of Society Member
	4.	L. Name and Location of Organizational Member of Society
	5.	
	21.	M. Address of Organizational Member of Society
Oral History Interviews	1.	N. Name of Interviewer
	2.	
	3.	
	4.	O. Organizational Affiliations of Interviewee
	5.	
	6.	P. Publications of Interviewee
	13.	

OTHER ENTRIES	DATA ELEMENT	DATA GROUPING OR USE IDENTIFIER
National Manpower Register	1. 2. 3.	Q. Name of Registrant
	21.	R. Address of Registrant
	4. 5.	S. Affiliations of Registrant

TABLE: APPLICATION OF STRUCTURED STANDARD DATA ELEMENTS

WHERE USED	DATA STRUCTURING OF ENTRIES
AIP Primary Journals	A(1,2,3),B(4,5),6,7(a,b),8,9,C(10,11),12 For Book reference citations, also 13,14,15
Secondary Journals (e.g. Physics Abstracts)	A(1,2,3),B(4,5),6,7,8,9,C(10,11),12,13,14,15,16,17,18,19,20
Special Bibliographies and Critical Reviews	A(1,2,3),6,7,8,9,C(10,11),12,13,14,15
Subscription Fulfillment	G(1,2,3),H(4,5),I(21)
AIP Societies Membership Records	J(1,2,3),K(21),L(4,5),M(21)
National Register of Physicists	Q(1,2,3),S(4,5)
Center for the History and Philosophy of Physics	
Oral History Collection	N(1,2,3),O(4,5),P(6,13,16)
Neils Bohr Library	A(1,2,3),7,8,9,C(18,10,11),12,13,14,15,16,17,19,20

Legend: The key is composed of a Data Grouping or Use Identifier (Alpha) plus Data Element (Numberic), e.g. Name of Author (A) = Surname (1) + Middle Name or Initial (2) + Forename (3) = A(1,2,3). The name of an AIP society member is composed of the same data elements = J(1,2,3)

	DATA ELEMENT	DATA GROUPING OR USE IDENTIFIER
OTHER ENTRIES (cont.)		
	1.	Q. Name of Registrant
	2.	
	3.	
	21.	R. Address of Registrant
	4.	
	5.	S. Affiliations of Registrant

Table: APPLICATION OF STRUCTURED STANDARD DATA ELEMENTS

WHERE USED	DATA STRUCTURING OF ENTRIES
AIP Primary Journals	A(1,2,3),B(4,5),6,7,a,b,8,9,C(10,11),12 For book reference citations, also 13,14,15
Subscription Fulfillment	G(1,2,3),H(4,5),I(21)
Center for the History and Philosophy of Physics	
Secondary Journals(e.g. Physics Abstracts)	A(1,2,3),B(4,5),6,7,8,9,C(10,11),12,13,14,15,16, 17,18,19,20
AIP Societies Membership Records	J(1,2,3),K(21),L(4,5),M(21)
Oral History Collection	N(1,2,3),O(4,5),P(6,13,16)
Special Bibliographies and Critical Reviews	A(1,2,3),6,7,8,9,C(10,11),12,13,14,15
National Register of Physicists	Q(1,2,3),S(4,5)
Niels Bohr Library	A(1,2,3),7,8,9,C(18,10,11),12,13,14,15,16,17,19,20

Legend: The key is composed of a Data Grouping or use Identifier (Alpha) plus Data Element (Numeric), e.g. Name of Author (A)= Surname (1) + Middle Name or Initial (2) + Forename (3) = A(1,2,3). The name of an AIP society member is composed of the same data elements= J(1,2,3)

Appendix VI

STANDARDIZATION OF DATA ELEMENTS

A more detailed examination of standardization efforts in the area of data elements, their codes and formats required for information interchange may be helpful in understanding the background of this report. Two aspects of the problem are considered. The first addresses itself to standardization generally and to data element standardization on the whole. The second looks at the recent history of data element standardization as it relates to the AIP work.

Standardization

The following definition of standardization has been offered by the Standing Committee for the Study of Scientific Principles of Standardization (STACO) of the International Organization for Standardization (ISO), and adopted by the ISO Council in 1962:

> "Standardization is the process of formulating and applying rules for an orderly approach to a specific activity for the benefit and with the cooperation of all concerned, and in particular, for the promotion of optimum overall economy taking due account of functional conditions and safety requirements.
>
> It is based on the consolidated results of science, technique and experiences. It determines not only the basis for the present but also for future development, and it should keep pace with progress."

Dr. N.A.J. Voorhoeve in his paper "International Documentation in the Domain of Standardization" points out a further specification by STACO:

> "A standard is the result of a particular standardization effort, approved by a recognized authority. It may take the form of a document containing a set of conditions to be fulfilled..."

On this basis Dr. Voorhoeve draws the conclusion that "standardization of documentation, to mention only one example, is certainly included in STACO's definition."

We have seen above how the set of conditions to be fulfilled by data element standardization applies to the identification of the names, meanings and relationships of certain concepts of groupings of data items that are interchanged and communicated between systems. Certainly the documentation required to record what was identified and compiled is standardization

in Dr. Voorhoeve's sense. But this describes only one instance of the standardization process. Actually, at least four levels of data element standardization with different topical considerations can be identified. The first applies to basic agreement about what constitutes the general data element structure and requires the formulation of relevant definitions. The second encompasses common agreements among the different users of specific data elements with regard to the meanings and ways of representing the meanings of these elements. Agreement may extend beyond the elements to the data items on the one hand, and/or proceed in a different direction toward the definition, form and users of the elements themselves. This process must start at a local data system level. It may then rise to more general, perhaps national and international levels of agreement. The third area of standardization includes the whole thorny range of coding and formatting problems: standardization of character types, character set extensions, control characters, modes of representation (binary, octal, decimal), message and field formatting and size, preferred media for interchange (tapes, cards...), common codes for data elements, data items, etc. The fourth and final area is perhaps the most difficult: The standardization of standardization practices treats the question, how does one go about getting other people to agree on things to be agreed on? How does the realization that standardization is needed ever begin? Synonymy is one way. Ambiguity of data terms is another.

Recognition of the need for a common and economical language arose in the Department of Defense (DoD) with the development of high speed digital date transmission systems. As the computer systems have become bigger and faster and centrally controlled decision making increasingly important, the obstacles to systems integration due to linguistic factors became ever more apparent. The same datum could be established as an element in several data systems, with a different name or identification (synonomy), a partly or totally different meaning and, almost invariably, a different code, either in structure, size or both.

The National Military Command System, a group planned at the highest operational level, entered its implementation phase by the middle of 1962, announcing at that time a program to standardize the data elements and

codes feeding from one data system or level to another. The hardware capability to establish and maintain the multipurpose data files and integrated systems necessary for centralized management was simultaneously made available. Thus a hardware environment was provided which was to be dependent on standard data elements and codes.

Historical survey

The need to facilitate data interchange and systems integration required for high-speed data transmission led to a determined effort to standardized the data elements used within the DoD.

DoD data standardization became operational with the establishment of the DoD data standards organization in the Office of the Assistant Secretary of Defense (Comptroller) on June 10, 1964, and that of the Data Standards Division in 1964. (7)

Informed guesses have estimated that the number of data elements in DoD data systems total upwards of 200,000. DoD has already standardized a a number of fields, including geographical areas around the world, and the states of the United States of America. It is currently working upon the Military Standard Contract Administration Procedures (MILSCAP) data processing system (Project 60), which will be based on a fully standardized data element vocabulary.

The DoD has developed the largest on-going data standardization operation . However, DoD is only one of the Federal Agencies being coordinated in the Bureau of the Budget (BoB) effort to integrate the data systems used throughout the entire Federal Government. The BoB has recently issued a circular (BoB Circular, A86) which defines specific policies and responsibilities, together with procedures by which the recommendations of its Task Forces will be developed and adopted as Federal standards for Data Elements and Codes. At present, there are seven Task Forces covering business, individual, time, government agency, state and country and place codes as well as countries of the world.

Technical advice and the maintenance of Federal registers required for this centralized program will be provided by the National Bureau of Standards.

A number of professional and industrial organizations have been concerned with general standardization endeavors to promote data interchange capabilities. Among these are: the World Meteorological Organization, the Air Transport Association and a number of international and national voluntary standards organizations.

The United States of America Standards Institute (USASI) is the principle organization for United States data element standardization work at both the national and international levels. At the national level standardization efforts are under the cognizance of the United States of America standards Instutute (USASI) Subcommittee X3.8 Data-Elements, Codes and Formats, organized in 1966 and currently working under the chairmanship of Mr. David V. Savidge of UNIVAC and the vice-chairmanship of Mr. Harry S. White, Jr of the National Bureau of Standards. The mission of Subcommittee X3.8 is to develop standards and related understandings in the area to facilitate information interchange. The work will attempt to develop, in addition, a standard method of describing and designating data formats for data interchange. Detailed work is handled by six task groups, administrative and special tasks by two ad hoc groups and a steering committee. The task groups are as follows:

X3.8.1 Standardization Criteria:

This committee is responsible for definitions, criteria, methodology, glossary.

X3.8.2 Time Designations:

This work area includes both macroscopic and microscopic time periods. It now appears that the first standard proposal will be for calendar dates in data systems and will recommend the order YEAR-MONTH-DAY.

X3.8.3 Individuals and Organizations:

This work is organized into two areas--

a) Personal Identifiers - One of the proposals under consideration is to use the Social Security Number as the identifier of individuals. However there are existing legal restrictions that must be clarified or changed before this proposal can be adopted. A further need is verification of number/name combinations. The major questions of cost, organization, and feasibility of a national verification system are now under study.

A second study project covers the procedure for representing individual names and the use of extraction systems such as Soundex Code.

b) Organization - This study area covers identifiers for organizations, both governmental and private. Severe questions exist, for example, as to the problem of widely diversified branches or divisions of large organizations such as multi-division firms holding companies, school systems, and the like.

X3.8.4 Geographic Units

This Working Group is maintaining close coordination with a similar Task Force of the Budget study (described below) which is studying geographic units. The Committee will primarily look for and deal with situations which will be important to private industry and state and local governments.

X3.8.5 Data Structures:

This work area includes arrangemnt of data into formats and the necessary syntactical rules necessary to separate elements of data.

X3.8.6 Quantitative Values in Data Systems:

This work includes:

a) The problem of specifying quantitative data,

b) Error detection and correction (self-checking) codes. Check characters can be added to a base number or code so that at given points in the processing it will be possible to check whether the number is correct. This check character is determined by mathematical formulas using the characters in the base or code.

F-NSF

REFERENCES

(1) McGee, William C., The Formulation of Data Processing Problems for Computers, in Advances in Computers, edited by Franz L. Alt and Morris Rubinoff, Vol. 4, Academic Press, New York, 1963, pp 1 - 52, esp. pp 38 - 49

(2) Evans, O.Y., Advanced Analysis Method for Integrated Electronic Data Processing, General Information Manual F 20 - 8047, International Business Machines Corp., New York, N.Y. (no date)

(3) cf. IBM Systems Reference Library, IBM System/360, PL/1 Reference Manual, File No. S 360 - 29, IBM, New York, N.Y., 1967

(4) Machine Recording of Textual Information during the Publication of Scientific Journals, Report on Work Done on National Science Foundation Contract 305, during Period January 1, 1963, to May 30, 1965, Prepared by Lawrence F. Buckland, May 30, 1965, Inforonics Inc., Maynard, Massachusetts, 1965

(5) Curran, Ann T. and Henriette D. Avram, The Identification of Data Elements in Bibliographic Records, Final Report of the Special Project on Data Elements for the Subcommittee on Machine Input Records (SC-2) of the Sectional Committee on Library Work and Documentation (Z-39) of the United States of America Standards Institute New York, N.Y., May 1967

(6) Libbey, Miles A. and Gerald Zaltman, The Role and Distribution of Written Informal Communication Theoretical High Energy Physics. New York, N.Y.,... American Institute of Physics, August 1967. Available from AIP as Report No. AIP/SDD-1 (REV.), also USAEC Report No. NYO-3732-1 (REV.)

(7) Toward a Disciplined Data Systems Language, in Data Processing Yearbook for 1966, pp 107-116

(8) United States of America Standards Institute, Subcommittee Z 39 SC 2 - Machine Input Records, Proposed U.S. Standard for a Format for the Communication of Bibliographic Information in Digital Form, Z 39 SC 2 (1968(2)), New York, N.Y., 1968

(9) Licklider, J.C.R., et al., "Report of the Office of Science and Technology Ad Hoc Panel on Scientific and Technical Communications", Washington, D.C., Office of Science and Technology, 8 February 1965

(10) Executive Office of the President, Office of Science and Technology, Task Group for Interchange of Scientific and Technical Information in Machine Language (ISTIM), Reporting to the Executive Office, Final Report, Washington, D.C., April 3, 1968

(11) Avram, Henriette D., John F. Knapp, and Lucia Rather, The MARC II Format, Information Systems Office, Library of Congress, Washington, D.C., January 1968