Terminological Aspects of Data Elements

R. A. Strehlow
Oak Ridge National Laboratory[1];
Chair, ASTM Committee on Terminology

W. H. Kenworthey, Jr.
Department of Defense
Chair, ANSI Committee X3L8, Data Classification;
Convener, ISO/IEC-JTC-1,SC14, Working
Groups 3 and 4, Data Element Coordination and Terminology

R. E. Schuldt
Martin Marietta Aerospace Systems,
CALS Industry Steering Group;
Chair, Data Classification Methodology Team

ABSTRACT

The creation and display of data comprise a process that involves a sequence of steps requiring both semantic and systems analysis. An essential early step in this process is the choice, definition, and naming of data element concepts and is followed by the specification of other needed data element concept attributes. The attributes and the values of data element concept remain associated with them from their birth as a concept to a generic data element that serves as a template for final application. Terminology is, therefore, centrally important to the entire data creation process. Smooth mapping from natural language to a database is a critical aspect of database design, and, consequently, it requires terminology standardization from the outset of database work.

In this paper the semantic aspects of data elements are analyzed and discussed. Seven kinds cf data element concept information are considered and those that require terminological development and standardization are identified. The four terminological components of a data element are the hierarchical type of a concept, functional dependencies, schematas showing conceptual structures, and definition statements. These constitute the conventional role of terminology in database design.

A different problem exists for more complex or more abstract entities that change in time. Complex entities such as "product" or "department" are comprised of sets of data elements associated with one or more databases. For such entities new data can be created by the very act of using existing data as well as normal additions to the data collection. The use of a taxonomic approach at the outset can offer aid to the database designer. Sound terminology obtained by standardization beginning with each data element concept, followed by comprehensive definition of the data elements throughout the process, should provide appropriate links to a data system and promote definability of complex entities.

INTRODUCTION

The role of terminology in data management has been an often unspecified, but central need in information management. The role of standard terminology in the process, however, has rarely been examined. Both technical and other business data may be found in structured form in databases or unstructured in documents. We assert that standard terminology is important in both of these, and is crucial in the former.

To retrieve a piece of data reliably, e.g., a tensile modulus from a material properties database or a person's salary from a personnel database, one **must** know that the data is called *tensile*

"The submitted manuscript has been authored by a contractor of the U.S Government under contract No. DE-AC05-84OR21400. Accordingly, the U.S. Government retains a nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or allow others to do so, for U.S. Government

*modulus*of the material or *salary*of an employee, respectively, and **nothing else**.[2,3] Terminology plays a different role in situations where a piece of data is embedded in a document. In this case one is obliged to use the intrinsically probablistic techniques of document retrieval. Therefore, the role and usefulness of terminology and the role of terminology standardization are different for these two applications. However, standardization is important in each.

This paper attempts to systematize the role of terminology in data element concept and design and to describe the need for standardized terminology in a database's documentation. The use of structured sets of data element concepts to define complex and abstract time varying concepts is described in terms of a taxonomic approach. The role of standard terminology in data modeling, management of, and retrieval from documents is not treated.

DATA ELEMENTS

A database is a formal organization of data and includes data elements as well as other entities needed for use of the database. An early step in the process of creating data is the definition of data element concepts.

A data element is defined by first composing a data element concept. The various necessary attributes of the concept are then specified. After this specification is standardized or otherwise approved for use, the data element may be called a generic data element. This sequence is seen in Figure 1. The data element concept and the generic data element are more general (more abstract) than an application data element. For example, *date of admission* would be modelled after a generic data element, *date of event.*

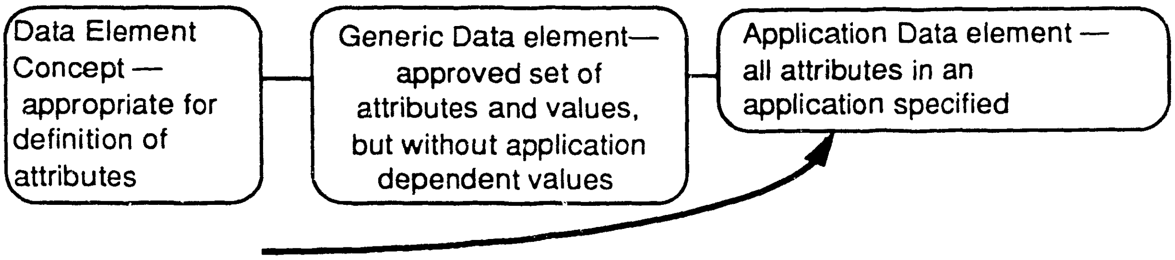

Data Element Concept — appropriate for definition of attributes
Generic Data element— approved set of attributes and values, but without application dependent values
Application Data element — all attributes in an application specified

Fig. 1. The development of a data element from concept to a defined generic data element and then to an application of the data element. The arrow illustrates the streaming of attributes from the concept to the generic data element and, finally, to the application data element.

Data element concept definition includes terminological activities at the very beginning of the definition process; again, when the data element concept is "fleshed out" with all the necessary attributes that the data element requires; and, finally, for some databases, in the use of a value thesaurus for values of data element attributes. An example of this last-named role of terminology would be the use of diagnostic terms for a patient record in a medical application, or lists of synonyms and aliases in a materials property data record. In any event, the user of a database must have access to the definition of the data element including the statement of meaning.

The process of defining a data element concept is identical with terminological practice for defining any term. This includes the selection of a genus and a listing of characteristics (attributes). Some of those that are selected as necessary attributes in a data element are also included in a statement of meaning.

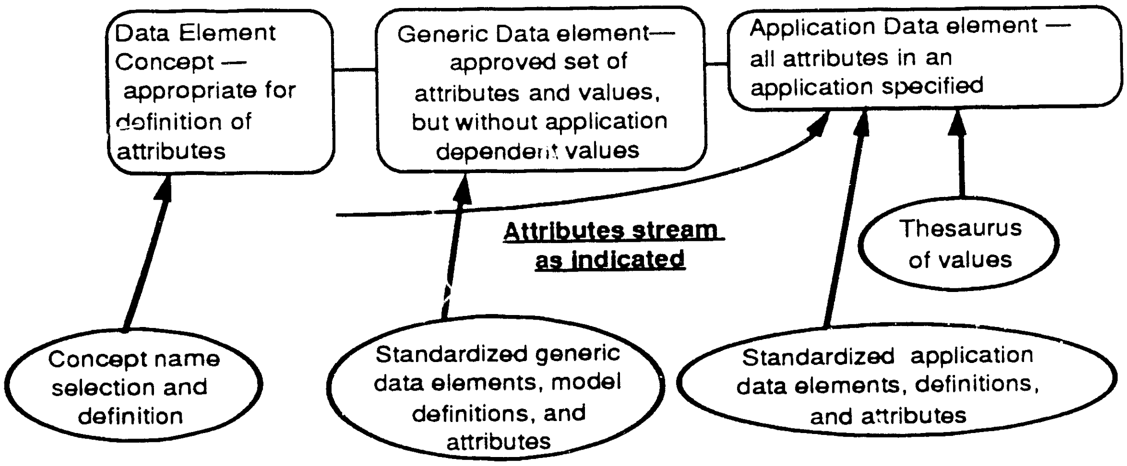

Data Element Concept — appropriate for definition of attributes
Generic Data element— approved set of attributes and values, but without application dependent values
Application Data element — all attributes in an application specified
Attributes stream as indicated
Thesaurus of values
Concept name selection and definition
Standardized generic data elements, model definitions, and attributes
Standardized application data elements, definitions, and attributes

Fig. 2. Terminological components in the process of data element definition.

A term and statement of meaning may be fully equivalent to a desired data element concept. Accordingly, standardized terminology can serve as a source list of data element concepts. The defined generic data element with a statement of all necessary attributes is closely similar to a fully developed standard definition. Of course, standard terminology also serves to supply standard definitions of other entities as well.

To compose a statement of meaning for a data element concept, a first step is in the selection of an appropriate genus. This can be assisted by considering the four basic functions of a data element: the qualitative ones of identification and description, and the quantitative ones of counting and measuring. Recognition of the function can assist the database designer in choosing an appropriate genus for a definition statement of the data element concepts. As examples, for qualitative data elements, one may draw more on the common language. For quantitative elements, scientific terminology is generally desirable.

The selection of attributes is facilitated by considering the type of knowledge that is required for database semantics, considered in the next section. However at this point in the process, suitable standardized generic data elements may exist that can be used.

To illustrate the flow or streaming of attributes we may display the data elements as frames, such as in Fig. 3. The name of the data element concept is listed at the top of a frame. Beneath it is shown the set of attributes and their values.

Using this representation and as one example of the process, we may look at a data element for *date of event*. This element might be used in a hospital, school, or more broadly as date of shipment, etc. The first step is the definition of the concept. Recognizing that a date is an identification of a particular year, month, and day leads to an easy selection of genus:

date of event—a statement of the year, month and day of a specified action, process, or happening. (The genus is underlined in this example.)

Concept name

ρ_1	:	α_1
ρ_2	:	α_2
ρ_3	:	α_3
ρ_4	:	α_4
. . .	:	. . .
ρ_i	:	α_i
Attributes — either objects or relations	**:**	**Values**

Fig. 3. General frame display of a concept.

Figure 4. shows the concept, *date of event*, including a parsing of the concept name into head word and modifier. In a database for linguistic purposes one might parse with finer structure down even to the morpheme level. Here the display shows the separate definition of the head and the modifier (differentia), although one could include the entire definition statement as simply the definition for *date of event* instead. This is basically a bare statement of meaning for the concept.

Date of Event

head	:	**date**
modifier	:	**event**
Definition:		
...of head	:	**a statement of the year month, and day**
...of modifier(s)	:	**a specified action, process, or happening**

Fig.4. Data element concept for *date of event.*

The next stage is the definition of the generic data element. This might include the format of the date, the range of acceptable dates,and possibly other information information. At this level a standardized format would be desired. The general frame representation for a generic data element is shown in Fig. 5. At this level one has a structure not unlike that of a standard definition.[4]

Generic data element

head	:	head
(modifiers)$_i$	:	M$_i$
units	:	U
range	:	R
dimensions	:	D
entry no.	:	E

Fig. 5. Frame for a generic data element. (Many more attributes and links to other data elements could be included.)

At this point of the data element development stream for date of event could be as shown in Fig. 6.

Date of event (Generic data element)

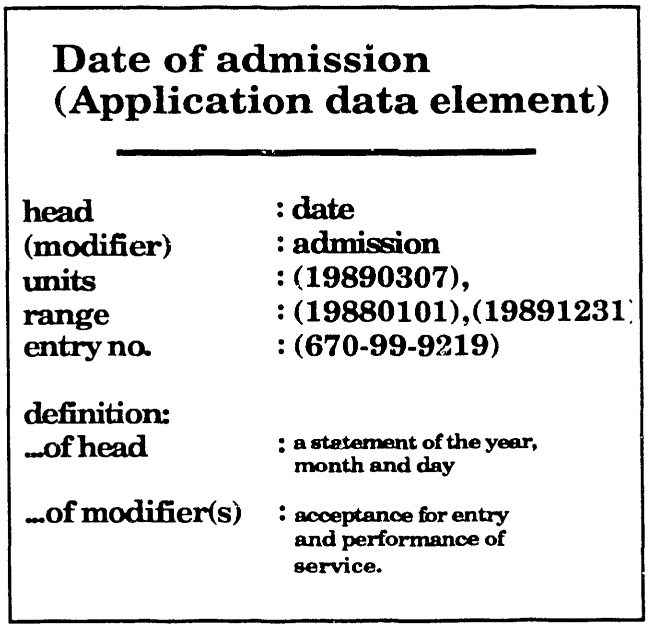
head	: date
(modifier)	: event
units	: (YYYYMMDD)
range	: (YYYYMMDD), (YYYYMMDD)
definitions:	
...of head	: the time of an event specified to the day
...of modifier(s)	: a specified action, process, or happening

Fig 6. Generic data element for *date of admission.*

The role of data element standardization is to agree in advance on the various attributes that constitute a specific generic data element. Numerous standards exist and are being developed to aid this objective. The generic data element might be a national or international standard and would include all of the agreed-upon data necessary for the data element, *date of event.*

The final stage of the streaming process, the application data element, involves the application of this data element to a particular application, such as *date of admission to a* school, hospital, etc., or *date of shipment* for a business application. Here the genus of the original data element concept *date of event* has flowed to the application and is presented along with a specified differentia. At this point, the data built into the fully defined data element still including terminological data are and must be part of the documentation for the database. This is shown in Fig. 7.

At this final stage, additional terminological consideration pertains to the values that are chosen for entry of values in an application data element.[5] An application data element might include attributes that require a thesaurus of values. For example, the choice of name or designation for the data elements, *eye color* of a person or *diagnosis* of a patient, would be based on their definitions or descriptions in a such a glossary or other listing. In most fields conventional usage develops over time and ultimately can produce a defacto standard terminology. In some cases individual authorities are used, but because variant usage can lead to communication problems, a standardization effort is desirable.



Date of admission
(Application data element)

head	: date
(modifier)	: admission
units	: (19890307),
range	: (19880101),(19891231)
entry no.	: (670-99-9219)
definition:	
...of head	: a statement of the year, month and day
...of modifier(s)	: acceptance for entry and performance of service.

Fig. 7. Application data element for *date of admission*. showing some attributes with possible values. (The entry number would serve as a link.)

The data element concept is defined in a general and abstract way. The definition carries the structure that can lead to numerous applications. The generic data element serves to augment this definition with additional detail, such as units, dimensions, domain, and documentation. At this point it can be the equivalent of a term entry in a standardizaed compilation of generic data elements or as an entry in a termbank, and it is ready for application in numerous ways.

KNOWLEDGE REQUIRED FOR DATABASE SEMANTICS

There is a profound difference in the level of detail needed for retrieving data from a database and in retrieving documents using keywords or index terms.[6] Terminology control is needed in both cases. In document retrieval by keywords or index term, the search efficiency or effectiveness is related to the match between one's choice of terms and those of the authors or abstract writers of sought documents. There is no theoretical or practical limit to the number of index terms which might be used.[7] The effectiveness of retrieval is expressed probabilistically.

In data retrieval from a database, however, either the data is that which is desired or it is not. Therefore, every data element must include a sufficiently detailed definition statement that will allow unambiguous entries and searches to be made. The definition statement for a data element concept is generally expressed in *canonical form*. This is a statement of a type or *genus* along with a selection of differentiating characteristics, the *differentia*. These have been discussed by many authors.[8]

A definition statement for each data element concept in a database must be made by the designer, in order to minimize ambiguities by the data entry and user communities. How one achieves an adequate definition statement is still much of an art. Sowa[9] lists seven types of semantic data that are needed in a data element definition that are helpful in this regard. Although the intension or full meaning of the data element concept might require all of these seven types, four have clearly identifiable terminological components that can assist in selection of distinguishing characteristics. These are the first four listed below:

1. The type hierarchy or level of generality.

For example, the data element of *status of employee* in a personnel file, may be defined as entity, living thing, animal, person, employee, chemist, organic chemist, electro-organic chemist, etc. Identifying the level of abstraction is a fundamental terminological activity that is normally also reflected in the concept definition. The level of abstraction is specified by choosing a genus in a definition statement. This was described above.

2 Functional dependencies.

A data element definition requires a conceptual analysis of the element to identify independent and dependent variables that may be found in a database. The concepts associated with the data element may or may not be reflected in the written definition, but will include keys that permit a query to be formulated when searching the database. The construction of a conceptual analysis is a basic terminological activity. How a data element concept is related to the other data elements in a database is best shown by one or more of the differentia in a definition statement.

3. Schemata.

The fundamental terminological tool in definition is a clear display of the relations and entities constituting a concept. Graphical display of concepts is well illustrated by Sowa.[10] A well-developed conceptual graph showing the relationships among all of the data elements in a database can be a useful tool in writing definitions for data element concepts, as well as defining a generic data element. The schemata for a particular data element is a display of all the perceived differentiating characteristics of that element. The differentiating phrases in a definition statement for a concept should relate to at least a portion of this relationship map.

A schema shows not only the relationships selected for the type definition, but also incidental differentia, such as the fact that an employee has a number, works in a department, etc. In technical areas the data element for modulus of elasticity would be schematically expressed showing characteristics of strain at which it was measured, temperature, and numerous other aspects of the concept. The extended set of schema relations is drawn on in selecting the attributes for the generic data element.

4. Type definitions.

We have been discussing the writing of type definitions, considering the elements associated with them, and giving some hints as to the issues of genus differentia selection. The reasons for this terminological activity need to be kept in mind always. A statement of the meaning or definition of a data element concept is needed to communicate to the data entry personnel and users of the database those essential aspects of the conceptual analysis sufficient for use. Definition writing is quintessential terminology activity. It is classically done in genus differentia form. The genus may be selected from the type hierarchy that was established. For example, an employee is a **person** (who works for an organization for pay).

We see that these four types of semantic data are terminological–they are involved with concept analysis and definition. In addition, there are derived semantic data that are part of a data element definition. When composing a definition for a data element concept, these may occasionally be used by the terminologist. These include:

5. Domain rcles.

Some data elements are functionally related, such as a person's current age, the age when hired, date of birth, and present date. The relationships are dominantly logical rather than terminological, but they do derive from and are associated with the terminological conceptual analysis. This sense of domain is not the sense that terminologists use the word as it relates to a field of application for a concept defined in a term entry.

6. Procedural attachments.

Databases may require description of means of calculation of various relations among the elements. The employee age referred to above is of this type. Again this kind of knowledge, although semantic, is not predominantly terminological.

7. Inference generators.

Description of means of deriving implications from the data are a final necessary part of database semantics.

The type hierarchy, definition, schemata construction, and functional dependencies are thus seen to be the four specifically terminological parts of database semantics. Domain roles, procedural attachments, and inference generators are additional types of database metadata that are involved with data element definition. They are conceptually based, but are not specifically terminological in character.

TAXONOMIC ANALYSIS FOR TIME-VARYING COMPLEX CONCEPTS

We accept as fundamental the fact that data should be created only once. If a process generates new data to be associated with that previously existing, the data elements should be keyed or related to the older data as an enhancement of the database. In a database of an employee's data, the employee's name should not need to be entered more than once. As new data are generated, new data elements may be needed, but simple keying to the employee's record should link the data appropriately.

Most data is fixed at the time of original entry. However, some important concepts have a history and change in meaning with time. Frequently these complex concepts are needed by an enterprise over a long period of time. These concepts pose special problems, because the data elements that constitute them are required at different stages of development. A uniform and coherent approach is needed. We propose a taxonomic approach that can be designed at the outset to organize the anticipated data elements.

Many types of complex data have different stages. The particular set of stages will depend on the subject. Human data, scientific data, product data, capital data, control data, and geopolitical data are some types of complex data that show this development with time.

As an important example, consider the complex concept, *product*. This concept is not, itself, a data element, but its full meaning comprises a set of data elements. The set changes in time, but at any instance, the extensional meaning of *product* is given by the set of data elements. Data about any product have become dominant contributors of value to the product. This is represented in Figure 8. Products always include information, knowledge, and services as part of any offering. Their importance in the value of the product is increasing.

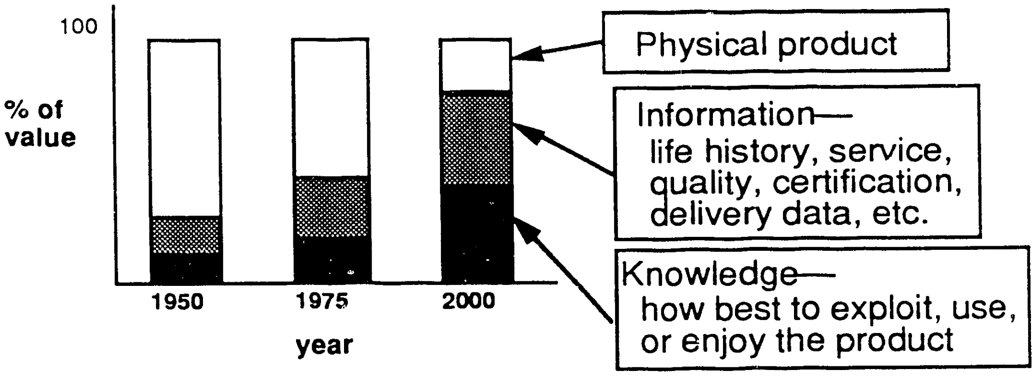


Fig. 8. Estimated part of product value represented
by information and knowledge. Adapted from E. Mahler[11]

As a consequence of this increasing information component of the concept, we need to examine in greater detail the time-dependent meaning of the concept and its representation.

The full intension or meaning of the concept, *product*, is given by only using the full set of data elements associated with it. Performing a conceptual analysis requires introducing both a time coordinate and classifying the data elements that may be called for at different stages of the product's history. For the concept, *product*, we identify 11 kinds of data classified by use, each of which falls into three categories produce a total of 33 types of data elements for the concept. The proposed classification is presented here to illustrate the taxonomic approach.

Basic categories include physical, functional, and programmatic data. From the outset of a product development, there are several specific uses to which data will be put as the work continues. This grouping forms a sequence of data types arranged by use and more-or-less sequentially. Different organizations would normally have different sequencing. Table I lists one set of types of data elements associated with the concept, *product*, with use as the organizing principle.[12]

TABLE I. Data types–by use:

Market
Requirements
Decision
Acquisition
Description
Production
Test
Operation
Support
Disposal
History

The three broad categories of physical, functional, and programmatic data can be examined for any of the 11 use categories. As an example, we may consider one of the types, *requirements data*. The physical requirements might include size, shape, and material properties such as mass, specific heat, reflectance, etc. The functional requirements would include data on purpose or objective, usage profiles, reliability, testability, etc. Programmatic data include the usual managerial considerations–resource allocation, scheduling, etc.

Test data, as a second example, might include test procedures, environment, and results as part of the functional data. Data associated with test equipment, facilities, specimens, etc. are part of the physical product test data. Programmatic data associated with testing would include the usual managerial information, but would include the capability to certify particular tests and demonstrate conformity.

Product *disposal data* is a matter of increasing concern to firms. Recognizing the need for disposal data from the outset and incorporating it as part of a taxonomic analysis is an essential part of information resource management. A structured taxonomic approach such as the one described serves this need.

The general terminology to be associated with the product would be included in the functional requirements area. This approach to a complex concept consisting of a time-ordered series of data elements is similar to the definition of a single data element. The data from early stages of the product creation process is part of the data stream for all subsequent stages. With this structuring of the data, the significant goal of creating data only once can be met.

CONCLUSIONS: THE ROLE OF STANDARDIZED TERMINOLOGY

Data elements are developed through three stages. Each involves terminological considerations and can profit from standardization. The first stage defining a data element concept is precisely the same as developing a definition. Indeed, terms in a standardized or otherwise authoritative term bank may be used as a prime source for pre-crafted data element concepts. In the absence of consensus standardized lists, the database designer must maintain these for each database as part of its documentation.

A data element concept is presented here as an abstract, general concept that may be implemented in many applications. The definition of a data element concept is its principal attribute. This definition streams to the generic data element, which serves as a model for various applications. The generic data element includes all necessary attributes that should be included in an application. Standardized generic data elements may be used at this stage, but a terminological entry may sufficiently present the necessary attributes.

In an application data element, the meaning of the data element concept streams as a model for the definition statement. In addition, standardized application data elements may be used to provide values for needed attributes. The data entry user of the database should expect that the definition associated with the data element is complete and correct. A standardized or specified authority for the terminology is needed to ensure that the terms used are those that will communicate to subsequent users of the data. If a data base is to be most useful to a wide group of data users, its documentation must include a thesaurus or similar listing. The use of specified standard terminologies is a preeminently practical way to achieve this.

Terminology standardization is thus seen as a needed activity at the beginning of design of a data element concept and is reflected in the definition of each data element concept. This definition is an integral part of the documentation for any further development or application of the concept. It is called onto assist in the definition of application data elements.

In the process of generating a data element, seven types of semantic data are listed. Four of these require specifically terminological development. These can assist in the defining of concepts. The use of standardized terminology assist in the maintenance of databases and in the retrieval of data from them. A taxonomic approach to concepts in the form of a thesaurus of concepts and standard definitions may be employed.

Complex concepts consisting of data elements that are added over a period of time were examined using a structured set of parameters for one specific complex concept, *product*. Standardized terminology for data element concepts and standardized generic data elements can serve to decrease the ambiguity for even such complex and time-varying concepts.

NOTES AND REFERENCES

1. Operated by Martin Marietta Energy Systems, Inc., under U. S. Department of Energy contract DE-AC05-840R21400

2. D. C. Blair, *Language and Representation in Document Retrieval*, Elesevier Science Publishers, Amsterdam, 1990, p 5.

3. We recognize that automated thesauri and synonym lists might provide access to information if the synonym list is complete, precise, and accurate. However, we have with this technique added the task of maintaining an updated thesaurus as an additional requirement for database administration.

4.Part E, *Form and Style for ASTM Standards*, ASTM, 8th Ed. September, 1989, pp. 27-34.

5. These include ASTM standards from Committee E-31 on health care services and E-49 on Computerization of Material properties data. In addition standard thesauri, such as that of Committee E-6 also serve to provide a common lexicon for specifying values of attributes in databases.

6. D. C. Blair, op. cit.

7. D. C. Blair, *op cit.*,p 156.

8. See, for example, H. Felber, "**Basic Principles and Methods for the Preparation of Terminology Standards,**" *Standardization of Technical Terminology: Principle and Practices, ASTM STP 806*, C. G. Interrante and F. J. Heymann, Eds., American Society for Testing and Materials, 1983, pp 3-13.

9. John F. Sowa, "*Conceptual Structures· Information Processing in Mind and Machine,*" Addison-Wesley Publishing Company, Reading MA, 1984, p. 304.

10. Sowa, op. cit.

11. Ed Mahler , "Knowledge Based Systems: The Competitive Imperative of the 90's," Proc. The Sixth Conference on Artificial Intelligence Applications, Vol. 2, March 5-9, 1990," IEEE Computer Society Press, Los Alamitos, CA 90720-1264, 1990, p. 1

12. R. A. Strehlow, "**The Varieties of Compound Terms and Their Treatment,**" *Standardization of Technical Terminology: Principles and Practices, ASTM STP 806*, C. G. Interrante and F. J. Heymann, Eds., American Society for Testing and Materials, 1983, pp. 26-33.

END

DATE FILMED

01/16/92

I