INFORMATION TRANSFER STANDARDS

A. Bornstein

System Development Corporation
Santa Monica, California

25 April 1966

The views, conclusions, or recommendations expressed in this document do not necessarily reflect the official views or policies of agencies of the United States Government.

This document was produced by SDC in performance of U.S. Government Contracts

TECH MEMO

a working paper

System Development Corporation/2500 Colorado Ave./Santa Monica, California 90406

TM- 2941/000/00

AUTHOR A. Bornstein

TECHNICAL

RELEASE S. L. Kameny

for J. I. Schwartz

DATE 4/25/66 PAGE 1 OF 22 PAGES

INFORMATION TRANSFER STANDARDS

ABSTRACT

Thi~~s~~e study is concerned with message description problems in intersystem communication. Involved is a search for a standard higher order language governing information transfer. A meta-descriptive approach is adopted. The concept of a standard representation language, the ARIADNE language, is introduced and developed. The requirements for this language are presented. The novelty of the approach and the recommended language resides in the requirement that intersystem messages be described at two levels of discourse, with only the higher-order level being in standard form. The difficulties involved in constructing the language, especially with respect to the implementation of a synonym- and homonym-resolving capability are offset by the advantages to be gained from adopting a meta-descriptive approach to the information transfer standards issue. ( )

22

CONTENTS

Section Page

1. INTRODUCTION

The study reported herein is exploratory. Involved is a quest for an Ariadne-type thread, whereby the labrynthian features of the intersystem communication environment can be dispelled. The problem on a philosophical level is to seek unity amidst diversity. Towards this end, the concept of the ARIADNE language (Abstract Representation of Information About Descriptor Nexus Equivalences) is introduced and developed. The requirements for and of the ARIADNE language are discussed.

The report is organized as follows: Section 2 defines the general framework and constraints governing the study. Section 3 presents alternative techniques for facilitating information transfer. Section 4 deals with the concept of the ARIADNE language and its distinctive features and problems. Section 5 discusses the development of a prototype ARIADNE language, alternative approaches to a morphological analysis of messages, and an implementation plan for demonstrating the feasibility of the ARIADNE classification technique.

2. PRELIMINARY CONSIDERATIONS

2.1 PURPOSE

The objective of this project is to explore and develop techniques for facilitating the exchange of information between data management systems. The basic motive underlying this effort involves an acknowledgement of the need for some form of standard governing information transfer. The guiding concept is that of a standard representation language. This would be a communications language that would convey, in a standard way, information about intersystem messages at some to-be-determined descriptive level.

2.2 THE STANDARDS ISSUE

The difficulties associated with determining future user needs has influenced the movement towards developing user-adaptable systems, involving such features as user-oriented languages and on-line programming. User-oriented programming languages not only relieve the user somewhat from mastering technical programming languages, but more importantly, such languages function as means for adaptation of system performance characteristics. This capability permits the user to be responsive to changing environmental conditions and/or objectives. Nevertheless, the trend towards user-initiated adaptations of operational systems, aimed at satisfying flexibility and timeliness requirements, raises serious questions as to the kind of descriptive standards that make sense in this environment. At the minimum, a distinction between rigid and flexible standards must be introduced. Where standards are rigid, system adaptations must leave them inviolable. This, in turn, represents a constraint on, and runs counter to, the desired flexibility in system use that called forth such adaptations.

Along a parallel line, the standards issue is complicated by the fact that large-scale data management systems deal with data which has many forms. The heterogeneity of this information prompts the search for a limited number of standard forms that adequately characterize and facilitate the transfer of such data. A data management system reflects both a domain of concern and a domain of discourse. Each system functions as a world view, in miniature, wherein a select set of entities, attributes, and relations are dealt with in various modes of system operation. Additionally, each system employs one or more languages and select sets of descriptive conventions to represent its own world view. Thus, the conventionality ascribed to descriptive procedures occurs at two levels, relating (1) to the relevant characteristics of data that a given system chooses to identify and to describe explicitly, and (2) to the specific techniques, formats and reserved-meaning terms used in describing the selected characteristics. These conventions affect both data input specifications and external message descriptions. Although referred to as "conventions," they are an integral feature of data management systems. Thus, any proposed standard that necessitates the abandonment of or a major modification to these descriptive conventions would probably meet with extreme resistance from within the communications environment. Accordingly, the study was directed towards a search for a common medium, a form of Ariadne thread, by means of which the diversity characterizing these descriptive conventions could be overcome without requiring their abandonment.

What is required is a standard that, although imposed and binding, does not create serious impositions to the interacting system. Obviously, this requirement is highly restrictive. Nevertheless, as with any labyrinthian problem, the possibility clearly exists that in the search for such a standard, more than one guideline would be available. This is the case. A number of options for introducing descriptive standards do exist. These potential approaches to the intersystem communication problem have to be distinguished, their presuppositions made explicit, their implications examined, and their relative advantages assessed. In Section 3, this examination is carried forth.

2.3 STUDY PERSPECTIVE AND CONSTRAINTS

The issue of transfer standards must be viewed within some perspective as well as with some perspective. Questions as to the type of standard required and the level at which standardization should be implemented must be formulated against the background of the intersystem communication problem. What became apparent from the inception of the study was that the intersystem communication problem was, in fact, not a single problem but a problem complex--a series of interrelated problems embracing such diverse topics as transmission system constraints, transmission code conventions, and message descriptive techniques, as well as data compatibility, message frequency, and message utility considerations. The scope of the study had to be restricted so that a coherent (if not a complete) picture could be presented. Accordingly, primary focus was placed upon the message descriptive aspect of the intersystem communication problem.

The selection of the message descriptive problem as the focal point of the study is based upon the integrative role of the data descriptive process in intersystem communication. The data descriptive process represents a conceptual bridge by which the data handling and the communication processes can be connected.

Thus, on the one hand, a functional relationship exists between the data descriptive features of a system and its data handling capabilities. Where this is the case, data that can be described and formatted in terms of these conventions can also be used by the system. Where this is not the case, the descriptive language employed is partly superfluous or else it is inadequate.

On the other hand, an epistemic relationship exists between the data descriptive process and the communication process. Data transfer functions are epistemically dependent upon the prior achievement of a structured and inter-subjectively based representation of data. This dependence is normally suppressed in natural language usage, but in information interchange between computer-based systems involving artificial languages it has to be made explicit. As a consequence, what can be initially affirmed is that if communication is to take place, at some level of language usage conformity to a prescribed and concurred-upon set of syntactical and semantical conventions has to be assumed and assured. Thus, it is essential to determine the descriptive level at which such agreement should be sought. Interrelated with this issue are relevance and sufficiency issues. What kinds of information about messages are necessary? How much descriptive information is required?

Even within the general area of message descriptive problems, neat boundaries were not established. The area was basically unchartered. The bounds of the problems encountered and uncovered became clear only as the study proceeded. Thus, it was soon apparent that the validity of any characterization of the nature and scope of the message descriptive problem and the feasibility of prospective solutions would be a function of the assumptions entertained about the communications environment. These relate to considerations such as the degree of data compatibility existing within a projected communication environment, the relative descriptive capabilities of the interacting systems, the frequency of data interchange among systems, the requirements for message variability, and finally, the diversity of descriptive conventions operative among the interacting systems.

As a result of the decision to restrict the inquiry to message description standards, a number of the related issues enumerated above had to be bypassed, while others had to be handled suppositionally. Thus, one area of the communication problem that has been bypassed concerns problems associated with character set compatibility and the development of a standard transmission code. The assumption has been made that either a standard character set will be available or code conversion can be effected. Thus, the various communication control characters, format effectors and message separators used in framing the heading and text of messages may be assumed to be properly understood by the receiving systems. If additional descriptive techniques were required for facilitating information transfer, such as self-described messages and meta-descriptive prologues to messages, then provisions would have to be made to correct and/or expand the standard transmission code.

The bracketing of these and other issues has been done to isolate more effectively some of the logical aspects of the message descriptive problem, and to permit intensified focus upon the requirements for and of a standard representation language.

Procedurally, the distinction between the requirements for and of a standard representation communication language is important. The justification for such a language is based upon its value within a wider context. The requirements for a standard language can be established only within the context of a postulated communication environment and in comparison with alternative techniques for facilitating information transfer. Thus, the rationale for the concept of a standard representation language is, in part, a function of assessments of and projection about the characteristics, objectives, and needs of present and future communication environments.

The requirements of a standard representation language embody the general and detailed specifications of the language. These are a function of the kinds of data and the types of message that have to be handled. These requirements can be formulated only within the context of a morphology of intersystem messages. The morphological analysis of messages and the specification of a standard representation language can be implemented by using the techniques of an experimental, an empirical, or an analytical approach singly or in combination. This aspect of the study is discussed in Section 5.

2.4 HARD DATA AND DESCRIPTIVE DIVERSITY

There is a multiplicity of meanings associated with the concept of data. Within an intersystem communication context almost anything serves as "data" insofar as it is incorporable within a message. Thus, natural language texts, programs, library routines, files, tables, arrays, lists, and data elements stored within a data management system all properly qualify as "data" for inclusion within a message. Obviously a standard governing information transfer has to be comprehensive enough so as to handle on the object language level any kind of data value or mode of data organization that might be involved in intersystem communication. Nevertheless, for the present study it has been necessary to restrict the use of the term "data" to hard or formattable data. Problems associated with the transfer of natural language text, programs, library routines, etc., have been excluded from consideration. However, even the concept of "hard data" is sufficiently broad to encompass diverse kinds, levels, and modalities of data.

How variable is the description of "hard data"? Is it not the case that there is differential descriptive treatment of data both within a given system and among systems? Different systems focus upon different characteristics of the same type of data, since the same content can be used for different purposes. Even where the expected use is identical, different data organizational techniques require differentiation as to the descriptive focus or as to what is considered to be the relevant characteristics of the data. Even where the descriptive focus is similar, different descriptors are used, so that on the terminological level there is no assurance of a one-to-one correlation between ostensively equivalent descriptors. Similarly, for the same or different types of data, both within a given system and between systems, the depth of descriptive detail and the explicitness of the descriptive material that is required varies. Thus, some characteristics of data have to be explicitly described for one system where, for another system, these same characteristics are either not attended to or are inferred.

Once descriptive diversity among systems is posited, an attendant difficulty relates to the possible relativity of relevance determinations among systems. This problem appears to be insurmountable, at least within the context of transfer standards. Where systems differ markedly as to the degree of descriptive explicitness or depth of detail that they require or in their use of specific characteristics of data, a fundamental asymmetry in descriptive capacity and a potential source of data incompatibility occurs. Consideration of these factors has also been bracketed and excluded from the study. They compound the information transfer issue too severely.

At issue is not merely the physical transfer of data, but the conveyance of meaning. Intersystem messages, if they are to have any functional value, must be intelligible to the receiving system. Formulating the message transfer problem as basically a problem of meaning involves all dimensions of the symbiotic process. Syntactic, semantic, and pragmatic considerations are involved. Thus, between any two interacting systems, the data flow can be characterized as a message stream wherein the constituent message elements are syntactically organized and semantically linked. Since messages are judged to be intelligible only insofar as they permit further action to be taken, pragmatic considerations are also involved. Nevertheless, the "utility" question and the related issues of descriptive symmetry and data compatibility had to be treated suppositionally. The information transfer process has been characterized as if each receiving system could utilize all of the data being transmitted provided that the data descriptions in the message could be properly interpreted. Thus, the central task is not to develop techniques that permit systems to handle types of data for which they were not designed. The task is much more limited. [illegible] that data compatibility and comparable descriptive capabilities exist among the inter-acting systems, but acknowledging the diversity of descriptive techniques [illegible] among these systems, the fundamental challenge is to devise techniques that [illegible] at some descriptive level, a standard description of messages.

2.5 THE POSTULATED COMMUNICATIONS ENVIRONMENT

The communications environment postulated for this study is one in which frequent two-way exchanges of messages of variable form and content are required between systems with diverse data descriptive conventions and formatting capabilities.

This environment is characterized by message variability, where "message variability" is to be understood both in a distributive and collective sense. It also applies to two levels of discourse. Thus, in the distributive sense, and on the object language level (i.e., the level at which the text of the message is represented) each sending system is assumed to possess the capability for generating messages of variable form and content. To differentiate among these messages, some form of descriptive prologue is required. This requirement may be expressed as the need for self-described messages. Similarly, on the object language level but in the collective sense, the sheer multiplicity of interacting systems imposes a requirement for differentiating among each of their respective messages. Again there is need for self-described messages. On the descriptive language level, i.e., the level at which descriptions about the text of the message are represented, the multiplicity and diversity of descriptive techniques and conventions used by each of the sending systems in describing messages results in message variability. Here the problem for each system in its message reception mode is to understand what the description means as well as what the text of the message means.

Understanding message variability in the above senses, a summary characterization of the postulated communication environment is that it is an environment in which each system in its message transmission mode can say "ever so much", whereas in its message reception mode, each system can understand "only so much".

2.6 GUIDING QUESTIONS

Within this environment, two features of the communication problem demand paramount attention. These relate to the crucial demand for message intelligibility and the complicating but interrelated demand for message variability.

Essential to any communication is the requirement that messages be successfully interpreted. They must be sufficiently understood so that the receiving system can interpret them, and if necessary, redescribe and reformat the data for subsequent use. This demand presupposes the existence of some degree of data compatibility among the interacting systems.

An obstacle to message intelligibility results from the need to convey messages of differing forms and contents. Hence, receiving systems must be capable of handling such differences. If communication is to take place between systems with variable message-generating capabilities, a trade-off is required between the responsibility to conform and to inform. The complicating factor associated with the responsibility to provide information about message variability is the existence of diverse descriptive conventions within the communications environment.

Thus, in the postulated communications environment, two interrelated questions require consideration.

1. Are self-described messages possible? Can intersystem messages be adequately described by the sending systems? A positive response to this question implies that the communication language available to each system is sufficiently rich and powerful in descriptive capability.

2. Are self-described messages understandable? Can the descriptive features of such messages be clearly understood by the receiving systems? Is the description itself self-explanatory? If so, then a receiving system, properly interpreting the description of the message, could redescribe and reformat the content of the messages.

The relations between these questions requires careful scrutiny. The questions refer to different focal points of study. The first question refers primarily to a capability possessed by a sending system, whereas the second question refers to the descriptive and interpretative capabilities possessed by the sending and receiving systems, respectively.

Also, it must be noted that the relevance of the second question is functionally related to the type of technique introduced for resolving the first question. If a standard descriptive language is used in describing messages, on the object language level the question of message intelligibility either does not arise or is trivially solved. Whereas, if the technique selected is one in which each system describes its messages in terms of its own descriptive conventions, then the question of message intelligibility becomes very real and the need for its solution very pressing.

Additionally, the pragmatic interdependence of both questions should be evident. If self-described messages can not be properly understood, their descriptive value in intersystem communication is negligible. Similarly, if messages cannot be described at the object language level, there is no immediate utility associated with the development of techniques for making the meaning of descriptors self-explanatory. Thus, from a practical viewpoint, if intersystem communication is to be possible within the postulated environment, both questions must be attended to and affirmatively resolved.

Nevertheless, immediate consideration of the first question can be deferred. By proceeding as if it were possible to make messages self-describing, focus will be placed upon exploring and developing techniques for making self-described messages self-explanatory.

3. AN EXAMINATION OF ALTERNATIVE TECHNIQUES

The primary task in intersystem communication is to provide techniques for assuring the intelligibility of messages of variable form and content. Two basic approaches are evident:

The Prescribed Message Route: Agreements are secured as to legal message forms and content. These agreements can be established either by imposition--through some forcing function, or by the mutual consent of the interacting systems. The net result is that the crucial aspect of the task is bypassed. Message intelligibility is assured, but the variable message-generating capabilities of the interacting systems are severely restricted.

The Self-Described Message Route: Each sending system is responsible for describing its messages and the data contained therein. This approach permits the transfer of messages of variable form and content, but founders upon the issue of message intelligibility. What techniques are available for providing assurance that the receiving system can understand the description? Direct translation techniques must be distinguished from standard form techniques and the implications of each examined.

3.1 DIRECT TRANSLATION TECHNIQUES

A message translation technique that is analagous to that used by meta-compilers in the translation of POL's is required. The descriptive features of messages would be directly translated from the language of the sending system into the language of the receiving system. A syntax specification metalanguage (e.g., META5 or BNF) as well as the syntax specifications for the message descriptive and data input languages of all of the interacting systems would be required. Separate translation programs would have to be written. If two-way communication were to be established between all systems within some postulated communications environment, the maximum number of translation programs required would be n(n-1). For an environment consisting of only 10 distinct data management systems, 90 such translation programs would be required. This figure contrasts sharply with the number of translation programs required by the next three techniques to be discussed, where values of 10, 20 and 20 obtain.

3.2 STANDARD FORM TECHNIQUES

Adoption of one of three techniques that transform self-described messages into a standard form. Associated with each technique is the use of some form of a standard representation language for describing messages. The advantage, feasibility, and negative consequences of each technique has yet to be determined, although a rough assessment can be made.

3.2.1 Use of a Standard Output Language (SOL)

Each sending system would describe its messages by means of a standard set of descriptive categories and formatting techniques. The form and content of the message would thus be intelligible to the receiving system. If a standard input language were also provided, translation or redescription of a message would not be required. Otherwise, each system would require a translation program that would reconvert incoming messages described in standard form into its own data input equivalents.

The prospects for the adoption and implementation of this technique are not promising. The comprehensiveness of the solution, the interface impact that it necessitates with respect to existing hardware configurations, software procedures, and special application areas suggest that the technique is subject to arguments similar to those directed against proposals for a standard programming language.

3.2.2 Use of a Standard Intermediate Language (SIL)

This language would perform the same functions associated with a SOL and would operate at the same descriptive level, except that instead of originally describing messages in standard form, a translator would be required that would process messages described in the sending system's own descriptive conventions. An equivalent but standard descriptor would be substituted for each native-language descriptive category. Each sending system's message-descriptive language would have to be equivalent to a proper subset of the SIL. For the receiving system, another translator would be required that reconverted the message described in standard form into its own input language-descriptive equivalents. Thus, in contrast to the SOL technique, twice as many translators would be required.

The practical impact of this technique is not so apparent. Each system would persist in using its own descriptive conventions for describing outgoing and incoming messages. Thus, the various interface requirements and the impact on existing intrasystem practices would be minimal. Unfortunately, the degree of consensus required to establish a specific equivalent in the standard intermediate language for each system's native descriptive categories appear to be overwhelming. Also, it would not permit the flexibility required for sending and handling messages of variable format and content, since any innovative technique would either have to match an existing SIL technique or the SIL would have to be expanded. In all probability, the language would be unwieldy, since an equivalent SIL descriptor would have to be adopted for whatever distinguishable features of data any given system chose to either describe explicitly or require a description of. This technique does not appear to be promising.

3.2.3 Use of a Standard Descriptor Language

The ARIADNE language (Abstract Representation of Information About Descriptor Nexus Equivalences) would be used to describe in standard form the descriptive features of messages. Messages would be self-describing on two levels of discourse: only the higher level would be in standard form. Thus, the ARIADNE language would operate on a descriptive level that was higher than the previous two standard representation languages. Similar to SIL, messages would initially be described in the sending system's descriptive categories. However, instead of substituting standard equivalents for the native language descriptors, the descriptors themselves would be described. The native language descriptors would remain an integral component in the message. The technique to be employed is as follows: An abstract representation of the descriptive features of a message would be provided. Standard information would be conveyed about the descriptors by identifying each descriptor as being subsumed within some general descriptor-class. Each descriptor feature of a message, including format specification techniques, would have to be properly identified and qualified with respect to descriptor nexus equivalences. It is necessary to determine descriptor nexus along two dimensions. The term "nexus" has two primary meanings:

(1) Connection or interconnection; tie, link

(2) A connected group or series

Both meanings must be remembered to understand the functional requirements of the ARIADNE classification technique. It is necessary to establish the interconnections between the various descriptive elements used in describing a message. This will require techniques for indicating syntactic and semantic connections. Additionally, it is necessary that the class-membership of every descriptor be properly qualified. Here, the use of role and linkage techniques associated with problems of information retrieval appears to be applicable. Descriptors are to be included within general descriptor-classes according to their descriptive function. Although each one will be part of a connected group of descriptors that constitute the membership of that class, it seems likely that the class itself will be characterized by a family of meanings. Although all will be functionally equivalent, semantically, only some of the members will be exact equivalents. An analogy to this situation is the problem encountered and the procedure adopted with respect to synonym listings in authoritative dictionaries. In defining terms, both the synonyms and the gradations of meaning among them are indicated. Similarly, the interconnections among descriptors within a given descriptor-class would have to be qualified and their similarities and differences specified.

If the ARIADNE language can be made sufficiently powerful and precise, the process of translating self-described messages into and from a standard form will be simplified. If the sending system's descriptive categories and their interrelations can be adequately characterized, since these in turn permit the actual data values to be interpreted, retranslation by the receiving system will pèrmit the data values to be redescribed in terms of an equivalent, amplified or abbreviated set of its own descriptive categories.

A sharp distinction has been drawn between two modes of intersystem exchange of information, either by means of prescribed messages or through self-described messages. This distinction can also be expressed in terms of two kinds of traffic between systems, namely, direct traffic requiring no intervening translation or higher level redescription of the message, and mediated traffic, involving self-described messages and the use of a standard meta-descriptive language, viz., the ARIADNE language. Some qualifications on these distinctions must be introduced. Where the interacting systems shared common descriptive conventions, if messages of variable format and content were to be interchanged, there would have to be self-described messages, but these would not require the use of the ARIADNE language. So not all self-described messages require explanation in terms of the ARIADNE language. Only where the descriptive conventions differ is the use of a meta-descriptive technique essential. As may be recalled, the postulated communication environment within which the concept of the ARIADNE language was introduced involved such descriptive diversity.

Another qualification relates to the sharp dichotomy that was drawn between prescribed and described messages. This holds for complete messages only. Within any given message, part of it may be in prescribed format, another part, representing an exception, may have to be described. This type of message would be treated as a self-described message, in which the ARIADNE language would have to be able to convey the difference between the two parts, proceeding with the described part in its normal mode of operation.

What the above discussion indicates is that there are degree of self-descriptiveness required for a message, both on the native language level and on the meta-descriptive level.

4. THE CONCEPT OF THE ARIADNE LANGUAGE

The ARIADNE language is to be used in intersystem communication. The language operates at the meta-descriptive level; its basic data is the descriptive features of the output and access languages of the interacting systems. In restricting the application area of the ARIADNE language to descriptors about hard data, in effect what is implied for information transfer is a second-order level of fact retrieval and transmission. The fact-retrieval and transmission statements involved are themselves composed of segments that include descriptive prologues on two levels of discourse as well as the object level data that is to be conveyed. The object language descriptive segment, expressed in terms of each system's own descriptive conventions, indicates the meaning of the data values and data definitions contained in the message. The meta-descriptive level segment, expressed in terms of the basic vocabulary of the ARIADNE language, functions as a descriptor-qualification statement that provides information about the object language descriptors. Thus, in the message reception mode of system operation, the descriptor-qualification segment of the message is to be interpreted as a descriptor retrieval imperative that directs the receiving systems to use the sender's descriptors if it can, or to supply equivalent or near-equivalent descriptors of its own. It is for this reason that the descriptor nexus must be fully specified and qualified by the sending system.

The primary function of the ARIADNE language is to represent, abstractly, information about equivalences and near-equivalences among data descriptors. Where a mapping among descriptors from diverse systems is exact, ARIADNE is to provide the directives for direct substitution. Where a mapping is not exact, the language is to represent the points of similarity and dissimilarity among descriptors.

Because the ARIADNE language operates on the meta-descriptive level, its descriptive requirements are reduced. The identification and representation of the descriptive features of varying output and access languages is a problem of lesser magnitude than that encountered in describing and representing object level events, processes, entities, attributes, and relations. Any proposed standard data specification language has to describe directly the diversity of data elements and modes of data organization characteristic of a communication environment of some complexity. Far less descriptive richness and specificity would be required of the ARIADNE language.

4.1 RELEVANCE AND SUFFICIENCY CONSIDERATIONS

In connection with the meta-descriptive approach to the information transfer problem, the general problem encountered in described data on two levels of discourse is determining what kinds of descriptive material must be supplied and how much descriptive information must be made explicit. Thus, relevancy and sufficiency issues occur: (1) with respect to the native language description of the data values, data definitions and the organization of the data contained within a message; (2) with respect to the ARIADNE language description and classification of the native language descriptors.

Nevertheless, only as a limiting concept is it a question of providing a complete description of these two levels of data. Generally, where the issue centers upon the selection of relevant aspects of the data to be described, the goal is descriptive sufficiency. But, with differential needs for descriptive explicitness, based upon the existence of systems with varying interpretative capabilities, what criterion of sufficiency will suffice?

The question of "how much" is tied in with the question of "what's important." At the heart of this difficulty is the fact that the sending system, in describing a message, selects the "relevant" descriptive categories and determines how full the description is to be, whereas, on a pragmatic basis, it is the redescriptive requirements of the receiving system that truly determine what is relevant and how much is enough.

In this connection, it is important to recognize that the mediating function of the ARIADNE language does not resolve this issue. A normative role has not been envisioned for it. The ARIADNE language does not function as a data specification standard, prescribing to each of the interacting systems what and how much information about the textual content of the message must be specified. On the contrary, the meta-descriptive approach embodied in the concept of an ARIADNE-type language merely provides a standard means whereby the interacting systems can understand the descriptive features of messages, whatever these may be and however adequate these may be. Thus, whereas an empirical test of the sufficiency of an ARIADNE-type language is determinable, only judgments about the relative descriptive sufficiency of the native language descriptors can be made. Such descriptions are either sufficient, insufficient or oversufficient, as determined by the capacity of each <u>specific</u> receiving system to interpret and redescribe the message. Thus, a message may convey enough information or less or more information than is needed.

With respect to the sufficiency of the sender system's description of a message, given that the ARIADNE messages can be used to identify and properly qualify the description, there remains the question as to whether or not the receiving system can retain the same meaning of the message in terms of its own conventions. Is there semantical loss or, worse still, semantical distortion involved in this transfer of information by means of a bi-level description of data? Additionally, if there is loss or distortion, is this primarily attributable to the insufficiency of the ARIADNE language, to the insufficiency of the description languages of the interacting system, or to a basic data incompatibility existing between the systems?

Data incompatibility issues have been by-passed in this study. Thus, the issue at hand is to determine the descriptive scope required of the ARIADNE language as well as that required of the descriptive languages of the interacting systems. Whatever their initial scope, obviously some anticipation of future software developments as well as an uncertainty factor must be provided for in both cases.

On both levels of discourse, the languages have to be made as complete in descriptive capability as possible, even to the extent that such capabilities exceed current processing capabilities and descriptor implementations. As a result, on the object language level, the languages envisioned should be able to describe types and structures of data that had not been implemented in their own or even on any other operating system. The same design goal would characterize the development of the ARIADNE language. It should be constructed so as to be able to describe kinds of messages and formats that were beyond the descriptive capability of systems within the current communication environment.

As a way of hedging against unanticipated developments, the languages in question must be capable of growth. The object level languages should be open-ended so as to be able to describe previously unanticipated kinds or characteristics of data or data structures. Similarly, the ARIADNE language should be expandable to enable handling previously unanticipated descriptive elements or techniques. In this sense, the ARIADNE language may also have to be self-describable, in that it should possess a capability for indicating, within the body of a message, changes to itself.

4.2 SYNONYM AND HOMONYM PROBLEMS

At any level of communication, synonym and homonym problems are encountered and provide a primary source of ambiguity. In intersystem communication, the synonym problem occurs when two or more systems use different names for the same descriptive function or for the same object-level variables. The homonym problem occurs when two or more systems use the same name or what is taken to be an equivalent name for different descriptive functions or for different object-level variables.

Thus, central to the concept of the ARIADNE language and its classification technique is the requirement for an effective synonym and homonym resolving capability. The technique of descriptor-class assignments is primarily directed towards overcoming synonymity confusions. The technique of using role and linkage associations to indicate and differentiate among near-equivalent descriptors serves as a qualification of and refinement upon descriptor-class assignments. It is equally valuable in resolving synonym and homonym difficulties.

Some cautionary notes are required. It may well be that, in focusing upon descriptor-class assignments, the synonym problem on the meta-descriptive level is easier to handle. But a corollary to this ostensive advantage is the inherent danger that precisely because of this focus, the existence and difficulties of homonymous relations among descriptors is either overlooked or de-emphasized.

A second cautionary note relates to the limited claim being made for the synonym and homonym resolving capabilities inherent in the ARIADNE approach. These capabilities apply only to the descriptive features of messages. On the object language level these problems would still exist. Clearly, some form of directory or dictionary would be required so that commonality of meaning among the interacting systems could be secured. In this connection, DOD directive 5000.11, dated December 7, 1964, relating to the establishment of standard data elements and data codes, represents an effort in this direction.

5. DEVELOPMENT OF A PROTOTYPE ARIADNE LANGUAGE

In exploring the concept of the ARIADNE language, the requirement that messages had to be adequately described on the object language level was handled suppositionally. Procedurally, this made sense. By assuming that each system could adequately describe its outgoing messages in terms of its own descriptive conditions, it was possible to focus upon the less obvious and more intransigent problems associated with assuring message intelligibility in the midst of descriptive diversity. The approach recommended (exemplified by the concept of an ARIADNE-type language) was to handle descriptive diversity by means of a standard representation language, operating at the meta-descriptive level, wherein the descriptive features of messages could be uniformly and adequately described.

5.1 SELF-DESCRIBED MESSAGES

Before an implementation plan for developing a prototype ARIADNE language can be set forth, previous suppositions must be cast aside. The question as to whether self-described messages on the object language level are possible must be raised anew and dealt with directly. The crucial task is to adequately describe the message forms of various systems. Any legal statement in the output languages of these systems may function as the content for an intersystem message. However, it is by no means evident that the descriptive capabilities embodied in the output and report generation programs of these systems would be sufficient for describing these statements, either to the users of another system or for automatic interpretation and redescription by another system. Even if it can be shown for a given system that it is possible to use the descriptive categories of its output language in order to adequately describe its legal message forms--thus demonstrating the possibility of self-described messages--the possibility and practicality of adopting a similar procedure for another system would still be in question. In all likelihood, selection from and augmentation of the descriptive conventions of most systems would be required if messages were to be adequately described on the object language level. This would entail a reconstruction of the output languages of each of these systems. How extensive this revision would have to be and how costly has yet to be determined.

The assertion that output languages will have to be reconstructed before they are adequate for the message descriptive function rests upon the specific relationship that holds between the class of legal message forms of a given system and the classes of legal statement forms in that system and in the receiving system. Any given message represents a legal statement in the output language of the sending system, whether or not that language overlaps any of the other access languages employed within the system. But, it is important to recognize that the set of legal statement forms used within two data management systems need not necessarily correspond to the lists of legal message forms for these systems. Thus, messages can be output in a form that would not be accepted as a legal statement for processing within the system. Similarly, many legal statements acceptable to the system would make no sense within an intersystem communication context. The interesting aspect of this problem relates to the requirement that a message must, whether or not it parallels a legal statement form in the sending system's access languages, be redescribable and reformattable so as to constitute a legal statement form in the receiving system's access language.

In Section 2.3 it was asserted that the requirements of a standard representation language could be formulated only within the context of a morphology of intersystem messages. Clearly, within a given communication environment, the descriptive conventions that are used in representing data will serve as the essential determinant affecting the specifications for and development of an ARIADNE-type language. These descriptors, in turn, reflect the kinds of data to be included within messages as well as the characteristics and modes of the organization of such data. It is precisely these data values, characteristics, relations, and modes of organization of data that serve as the evidential basis for establishing equivalences among descriptors. Thus, a survey of descriptive conventions that did not also involve a morphological analysis of messages and message components would be of limited value. Consequently, the implementation plan selected for developing a prototype ARIADNE language will entail some form of morphological analysis of messages.

5.2 ALTERNATE APPROACHES TO A MORPHOLOGICAL ANALYSIS OF MESSAGES

During the course of the study, several possible approaches to the development of the ARIADNE language were examined. These were characterized by differences in the degree of generality obtainable as well as by the ambitiousness and feasibility of the venture. Three main approaches were distinguished.

1. The experimental "benchmark" approach. This approach involves a step-by-step buildup of a prototype ARIADNE language. Two data management systems would be selected as "benchmark" systems, and their message descriptive features characterized and if necessary, reconstructed. Thus, the grammatical elements, syntactical rules and legal statement forms of access and output languages of each system would have to be examined. Next, variable kinds and formats of messages reflecting each systems data organizing and outputting capabilities would have to be specified. The kind, level, and modality of descriptive information appropriate to each message form would then have to be determined. Finally, a synthesizing effort would have to be initiated whereby descriptive classes would be constructed that subsumed the message descriptive features of each of these systems (that is, equivalence and near-equivalence relations would be established between the descriptors from each system.)

For two arbitrarily selected systems the possibility clearly exists that, as a limiting case, a perfect mapping could be effected. That is, within each general descriptor class constructed, equivalent descriptors from each system could be found. At the other extreme, the possibility exists that no descriptive equivalences could be established. Much more likely and of much greater interest are those cases in which, between two given systems, equivalences, near-equivalences, and distinct dissimilarities among descriptors could be found.

Two major difficulties are associated with the experimental approach. These relate (1) to the sufficiency and general applicability of the language to be developed, and (2) to the workability of the approach. The obvious limitation of the experimental approach is that it does not establish the sufficiency of the ARIADNE language, except in terms of the two systems that have served as the model for its construction. The general applicability of the language to messages from other data management systems would have to be established. At the minimum, a test with at least one other set of I/O reconstructed languages would be required, if some degree of confirmation were to be provided.

The second set of difficulties is more troublesome. The workability of the experimental approach is predicated upon the existence of convergences among the descriptors used in diverse descriptive languages. If this is so, then a process of condensation and concentration can be effected, whereby, starting with an analysis of two data description languages, the bulk of the descriptor-class categories required could be abstracted and enumerated. As a consequence, as experimentation with additional languages proceeded, their descriptors could be properly classified and qualified within the developed classification structure with only minor modifications and/or additions required. Obviously, this pattern of development presupposes that the relevant general descriptor-class categories can be abstracted from the selected "benchmark" systems and that powerful role and linkage associative techniques are available for articulating differences and for specifying interrelations among descriptors.

2. The empirical survey approach. A wide range of existing data management systems would be surveyed and the descriptive conventions of these systems would be catalogued. The instrumental value of a catalogue seems apparent. Taken collectively, these conventions would serve as a more reliable basis for selecting the basic descriptive categories of the prototype ARIADNE language. This approach would also insure that the categories selected would have general applicability. Additionally, a detailed breakdown of the specific features of each system's data descriptive language would reflect the specifications for translating messages into and from an ARIADNE-type meta-descriptive form.

Although this approach appears to be promising, during the course of this study efforts to produce a catalogue of descriptive conventions proved to be highly unsuccessful. An analytical tool for suitably characterizing, ordering, and comparing conventions was not available. This lack is not surprising. Upon reflection, it can be seen that the cataloguing effort itself was premature and the demand for an analytical tool excessive. The approach involves an implicit circularity of reasoning. Presupposed is the existence of the very classification technique that in effect represents one of the end products of the study.

3. The analytical "a priori" approach. This approach requires that a morphology of messages and message components be constructed and a prototype ARIADNE language be developed that is not tied to the specific message descriptive features of any select set of data management systems. A set of descriptor-classes would be identified according to their descriptive function within a possible set of message forms. This approach is obviously more general than the others in that it is not bound by empirical constraints. Various types of messages would have to be differentiated, the attributes of each type distinguished, and the common and individualized components of each distinct message form identified.

Each of the above approaches has its own set of distinct obstacles to overcome. Given the general complexity and ambitiousness of the overall venture, and given the singular lack of success achieved so far in proceeding along empirical and analytical pathways, the decision to adopt an experimental "benchmark" approach was not difficult to reach. The LUCID and COLINGO data management systems were selected as "benchmark" systems. It was felt that the diversity in message-generating capabilities and in data descriptive conventions of these systems would provide a suitable test basis for investigating the possibility of a prototype ARIADNE language.

5.3 IMPLEMENTATION PLAN

Using LUCID and COLINGO as "benchmark" systems, an implementation plan for developing a prototype ARIADNE language and a test plan for demonstrating the feasibility of the ARIADNE classification techniques has been formulated. The following key milestones were included:

- Specification of a sample set of legal message forms:
  List and analyze the legal statement forms of the various access languages and the output languages for each system.
  List the descriptive conventions used in the above languages.
  Construct a sample set of legal message forms for each system.

- Determination of the requirements for self-described messages in these systems: Given the list of descriptive conventions, select from and/or augment this list so that the data values and data structures contained within the sample set of legal message forms can be described. At this phase of the experimentation, separate sets of self-described messages on the object language level would be produced for each system.

- Specification of a prototype ARIADNE language: Examine and compare the descriptors employed in the self-described messages of each system. Develop a synonym and homonym resolving capability for the ARIADNE language by initially determining equivalence relations among the descriptors. Next, where near-equivalence relations exist, introduce role and linkage associative techniques so as to properly qualify the similarities and differences among these sets of descriptors. Based upon these determinations, construct a set of general descriptor classes that, at least for these systems, subsumed their respective descriptive features. Using these general descriptor-classes as the key descriptive elements of the ARIADNF language, specify the syntax and semantics of the ARIADNE language.

- Construct an experimental ARIADNE language based upon the above specifications.

- Construct two-way translation programs between LUCID and COLINGO: Output translation programs would have to be constructed so that any legal message from either system could initially be described on the object language.level and then, on the meta-descriptive level could be redescribed in terms of the descriptive functions of the ARIADNE language. Similarly, translation programs for message reception would have to be constructed. These would interpret the ARIADNE language description of the message, thereby permitting the substitutfon of equivalent and/or properly qualified near-equivalent descriptors in the receiving systems language for the descriptors of the sending system.

- Demonstration of two-way information transfer between LUCID and COLINGO.

- Extension of the prototype ARIADNE language: If the above demonstration is successful, apply the ARIADNE classification technique and language description to messages both from and to at least one other different system. If necessary, expand the ARIADNE language to include message descriptive features not subsumable under the descriptor-classes constructed upon the basis of the LUCID and COLINGO systems.

- Demonstration of data exchange between three systems.

- Analysis of experimental results and appraisal of the ARIADNE language and classification technique.

6. CONCLUSIONS AND RECOMMENDATIONS

The primary focus in this report has been conceptual, although it was not intended to be so restricted. Attention has been directed to the message descriptive aspects of intersystem communication. Information transfer has been viewed principally as a problem involving the search for a standard higher-order language. A meta-descriptive approach has been taken to the problem of facilitating information transfer between systems with diverse descriptive conventions. The major recommendation of the study is that a standard representation language--the ARIADNE language--should be constructed and tested.

The ARIADNE language will be a meta-language, designed to convey an abstract representation of information about descriptor-nexus equivalences. At issue is the general question of sameness and difference in meaning, although this issue is here raised to the descriptor-level.

The concept of an ARIADNE-type language is offered as a promising candidate for a standard representation language. It permits standardization to be introduced into a recalcitrant communication environment where the least amount of disturbance to the status quo is produced. Thus, while the concept of the ARIADNE language is radical, involving the description of messages at two levels of discourse, the motives governing its formation and the primary criterion used in assessing its prospects for acceptance are conservative. In the standards area, the guiding imperative is "Don't make trouble."

The concept of the ARIADNE language and the classification technique implicit within it has been developed to the point where detailed work with the "benchmark" systems is both possible and mandatory. Only in this way can a prototype ARIADNE language be constructed. Only in this way can the meta-descriptive approach to information transfer standards be substantiated. In the implementations area the guiding imperative is "Take the trouble ..." Someone has to do this!