DOCUMENT RESUME

ED 094 758 IR 000 958

AUTHOR Reintjes, J. Francis; Marcus, Richard S.
TITLE Research in the Coupling of Interactive Information Systems. Final Report. ESL-FR-556.
INSTITUTION Massachusetts Inst. of Tech., Cambridge. Electronic Systems Lab.
SPONS AGENCY National Science Foundation, Washington, D.C.
REPORT NO ESL-FR-556; MIT-OSP-80720
PUB DATE 30 Jun 74
NOTE 62p.

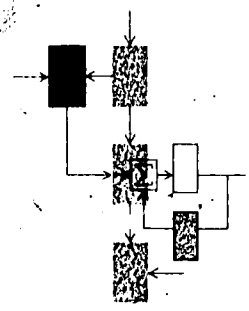
EDRS PRICE MF-$0.75 HC-$3.15 PLUS POSTAGE
DESCRIPTORS *Computers; Data Bases; *Information Networks; *Information Retrieval; Information Science; Information Services; *Information Systems
IDENTIFIERS ARPANET; Compatibility; MEDLINE; Network Interfaces

ABSTRACT

This reported research centered on development of the concept of a translating computer interface by which the networking of heterogeneous interactive information systems may be achieved during the period in which information retrieval system and network standards are evolving. The particular concepts and techniques investigated are the virtual system concept, a common command language, a master index and thesaurus, and a common bibliographic data structure. In addition to the theoretical study of the problem, an experimental interface has been developed that connects the MEDLINE and Interex retrieval system via ARPANET communication links and that performs some of the networking functions of the virtual system. (Author/WH)

June 30, 1974

Report ESL-R-556
MIT-OSP Project 80720
NSF Grant GN-36520



RESEARCH IN THE COUPLING OF INTERACTIVE INFORMATION SYSTEMS — FINAL REPORT

J. Francis Reintjes
Richard S. Marcus

Electronic Systems Laboratory

MASSACHUSETTS INSTITUTE OF TECHNOLOGY, CAMBRIDGE, MASSACHUSETTS 02139

Department of Electrical Engineering

June 30, 1974 Report ESL-FR-556

RESEARCH IN THE COUPLING OF INTERACTIVE INFORMATION SYSTEMS

FINAL REPORT

J. Francis Reintjes
Richard S. Marcus

The research reported herein was made possible through the support extended by the National Science Foundation through Grant GN-36520.

Electronic Systems Laboratory
Department of Electrical Engineering
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139

ACKNOWLEDGEMENT

The authors would like to acknowledge the efforts of several project co-workers: Dr. Charles W. Therrien and Mr. Don Cantor in the area of computer systems; Mr. Alan Benenfeld in the area of bibliographic data structures; Dr. Peter Kugel in the area of retrieval language analysis; and Messrs. Charles E. Hurlburt, Leonard Goodman, and Hsin-Kuo Kan in the area of computer programming and analysis.

We would also like to acknowledge the support and cooperation of Mr. Davis McCarn and Mrs. Barbara Sternick of the staff of the National Library of Medicine in our efforts to couple, experimentally, MEDLINE to our local interface.

ABSTRACT

This report describes results of an 18-month research effort in the coupling of interactive information systems. The research has centered on development of the concept of a translating computer interface by which the networking of heterogeneous interactive information systems may be achieved in the period during which I-R system and network standards are evolving. Particular concepts and techniques which have been investigated are:. (1) the virtual system concept by which users perceive the network as a single homogeneous system; (2) a common command language synthesized from a basic language of primitive I-R functions; (3) a master index and thesaurus which stores the vocabularies of the separate data bases along with index term interrelationships and counts; and (4) a common bibliographic data structure in which the data elements for bibliographic information are hierarchically structured and interrelated among different data bases. In addition to the theoretical study of the problem, an experimental interface has been developed that connects the MEDLINE and Intrex retrieval system via ARPANET communication links and that performs some of the networking functions of the virtual system.

I. INTRODUCTION

This report describes results obtained under National Science Foundation Grant GN-36520, entitled 'Research in the Coupling of Interactive Information Systems'. The period of the grant was January 1, 1973 through June 30, 1974.

The research was motivated by our concern that the information systems that are beginning to appear in operational environments will not be effectively utilized because of an inability of users and information specialists to master them. Degradation in the quality of service is likely to occur because of the many differences that exist among the systems and the time required to gain an understanding of their specialized features.

Under the present state of affairs, in which each information system has its own unique features and is accessed in accordance with its own set of special rules, it is out of the question to expect users themselves to make efficient use of the various systems. There are just too many subtle procedures to master in a short time. We have even observed that considerable training of information specialists is required to bring them to a high level

of proficiency and we foresee an upper limit to the number of disparately designed systems the specialists will be able to handle.

It is clear that until such time as the user community becomes highly proficient in online-access procedures, information specialists must be on hand to serve user needs. Although it may turn out that a certain percentage of users will always want to delegate search responsibilities to the specialist, one would like to move in the direction of self-service as a means of getting the user involved directly in the solution of his own informational problem, thereby reducing his costs and improving personal satisfaction. This will not be possible unless system access is made simple and straightforward.

Hence, future courses of action become obvious; either uniform standards must be adopted for all systems, or computerized interfaces must be developed which accommodate nonuniformities and re-present them to the user in a single, standardized form. It is along the latter line that we have been working.

II. DISCUSSION: THE NEED FOR A NETWORK OF HETEROGENEOUS INFORMATION SYSTEMS

A. Recent Advances in Interactive Retrieval Systems. A number of interactive bibliographic information retrieval systems have been developed in recent years.* This type of online computer system has been widely acclaimed by users for rapid and easy access to large data bases of bibliographic references. The economic viability of these systems is attested to by their continued growth and by the fact that a number of commercially sponsored systems are currently available.** In fact, it is now possible to gain access from most points in this country at costs ranging from about $6 to $100 per connect hour to these retrieval systems. These

---

* A collection of descriptions of several of these systems is found in Walker, Donald E. (ed), Interactive Bibliographic Search: The User/Computer Interface, AFIPS Press, Montvale, N.J., 1971

** Economic viability is discussed in the following two papers:

C.W. Therrien and J.F. Reintjes, Modeling of Information Systems; Proceedings of the Sixth Annual Princeton Conference on Information Sciences and Systems, Princeton University, March, 1972

Davis B. McCarn and Joseph Leiter, On-Line Services in Medicine and Beyond, Science, Vol. 181, No. 4097, 27 July 1973, pp 318-324

systems contain in the aggregate references to documents numbering in the millions in such subject areas as chemistry, aeronautics and astronautics, education, agriculture, nuclear science, toxicology, medicine, engineering, and environmental studies as well as data bases covering several subject areas for such document types as journal articles, government-sponsored reports, Library of Congress cataloged monographs, and news articles.

B. Limitations of Present Systems. A major limitation of current systems is the size of the data base that can be stored online. A data base containing bibliographic information for a million documents is about the maximum size for effective online operation with current hardware/software environments for single computer systems. However, a collection of this size represents a very few documents when measured against the total amount of published literature. In particular, a million documents will cover the literature of a single discipline --- for example, chemistry --- for only a very few years.

One might argue that most researchers work only in a fairly narrow area and could be adequately served by a data

base of the size of a million documents. There are several problems with this kind of argument. In the first place, the information needs of users are often most critical in areas outside their own specialty. Thus, for example, when starting out in a new aspect or when seeking an experimental device for instrumentation, a researcher may have more need for information than when working strictly in his own specialty. In Project Intrex we found* that there were many more serious users who were from outside the 5 specialties for which the data base was selected than there were from within those specialties. Also, there seems to be a growing trend toward interdisciplinary activity with the concomitant need for multiple data bases. Even users of systems with many hundreds of thousands of documents regularly ask for a broader coverage. Then, too, much of the use of these systems is by information specialists acting as delegated searchers; these specialists may have many clients from different subject areas and, hence, a need for a multiplicity of large data bases.

Another limitation on current systems is their capacity in terms of number of simultaneous online users.

* See Project Intrex Semiannual Activity Report 15 September 1971, Massachusetts Institute of Technology, pp 6-8. PB 202 860.

This capacity is usually numbered in the tens whereas there is a potential for thousands of simultaneous users, even if only the United States is considered.

C. The Ultimate Uniform Network Solution. Ultimately, the solution to these problems may necessitate the construction of a large-scale, on-line information retrieval network made up of many similar --- preferably identical --- computer nodes, each node being associated with an online data base of a million or more documents on a separate set of topics. For maximum efficiency users connected to each node --- there might be several hundred online users at any one time --- would make requests in a common retrieval language. Such requests would lead to parallel searching of the appropriate data bases which would be organized within a standard file structure. Intercommunication among computer nodes would be accomplished over high-speed communication lines for which data-concentrator techniques would be employed to gain further efficiencies and to reduce response times.

Thus, in order to achieve economies of scale, with data bases created only once and used many times by a large user community, and to provide easy transfer of

information among this large community, the ultimate solution appears to be a uniform network of standardized parts. The telephone and railroad (standard gauge) networks would seem to provide good analogies.

D. Obstacles to Immediate Implementation of the Ultimate Network. For the next several years, however, the degree of standardization required for the ultimate network is unlikely in view of the already heavy investments that have been made in existing heterogeneous, nonstandardized retrieval systems. Lack of standardization is a pervasive barrier to intercommunication. A potential user of different retrieval systems is faced with a series of obstacles right from the start: the necessity to discover these systems in the first place, to make separate procedures to gain access and account for costs, and, quite possibly, to make actual access via different terminals and separate locations. Other obstacles face the user once access is made: different command languages, retrieval functions, indexing vocabularies, and output formats. If the programmers of one system wanted their system to communicate directly with another, they would face problems of different operating systems, hardware, programming languages,

character codes, word/byte/bit organization, file organizations, and most directly for the majority of I-R systems, no established computer-to-computer communication links.

Because of the established character of the different I-R systems, their environments, and their user clientele, and the cost of remaking data bases in different file organizations --- even if permission to do so were granted, it seems unlikely that any existing I-R system and environment will soon become a de facto standard.

E. The Computer Interface. In view of the foregoing obstacles to the immediate implementation of the ultimate network, we decided on a course of action for the intermediate term which seeks to approximate the effectiveness and efficiency inherent in the ultimate network as best as possible through a currently achievable network based on computer-interface techniques. Such an interface achieves compatibility among systems of heterogeneous hardware and software components through translating and conversion algorithms. It is this kind of interface that we have investigated and are reporting upon here.

The over-all objective of our work was to establish through analysis and experiment the feasibility (or infeasibility) of a computer-stored common language interface for disparately designed information systems. Our program was conducted under three major headings:

* Study of I-R Systems and Research Planning
* Advanced Network Research
* Experimental Interface Design, Implementation and Analysis

III. RESULTS OF THE PROJECT

A. Study of I-R Systems and Research Planning. Our effort called for an examination of several online I-R systems from the viewpoint of the requirements they would impose on a computer interface design. The purpose was to find at least one I-R system outside M.I.T. which would be suitable for our network experiments and whose administrators and technical staff would be willing to cooperate in these experiments.

We reviewed many of the important I-R systems. We participated in the feature analysis of twelve interactive systems performed by Dr. Thomas Martin of Stanford* and attended a three-day seminar for this purpose at Stanford University. Our review of these systems led us to the conclusion that our original ideas for a network interface merited detailed development. In addition, we identified the MEDLINE retrieval system of the National Library of Medicine as the best one for us to choose as the first system remote from M.I.T. with which to experiment. Our reasons for choosing MEDLINE include

---

* Dr. Martin, now at the University of Southern California, and Dr. Edwin Parker, of the Institute for Communication Research at Stanford, are completing a report on this comparative analysis.

(1) NLM staff was very cooperative in aiding our research efforts. They provided information about MEDLINE and ready access to the system.

(2) MEDLINE uses controlled vocabulary indexing which provides a good contrast for experimental purposes with the free-vocabulary techniques of the Intrex retrieval system of M.I.T.

(3) MEDLINE has established an experimental connection to the ARPANET, the ARPA network, which would prove helpful in furthering our own network experiments.

In the course of our work on this task we broadened the review effort to include the review of existing computer networks which could be used in our network I R experiments. This review proved rewarding in that we discovered how we could effectively use in our experiments both ARPANET and TYMNET, the computer network of the TYMSHARE Corporation, through which most of the important operational online bibliographic retrieval systems are accessible. ARPA and ARPANET users at Bolt, Beranek & Newman (BBN), Computer Corporation of America (CCA) and M.I.T. proved very helpful in assisting us in setting up our experimental network described more fully below.

In our reviews of various I R systems we also received strong expressions of support and cooperation

from such groups as TYMSHARE, Lockheed, Systems Development Corporation (SDC), the Atomic Energy Commission, Battelle, Stanford University, and NASA.

B. Advanced Network Research. The second task under our grant was a broad-based study of the problem of interconnecting diverse and geographically separated information-retrieval systems. A paper* describing the early results has been prepared. Partial results were also presented at the Network Interconnection Panel at the annual meeting of the American Society for Information Science on October 24, 1972 at Los Angeles, California. Highlights of these results are given below.

1. The Common Computer Interface

Our basic philosophy for interconnecting

* Richard S. Marcus, "A Translating Computer Interface for a Network of Heterogeneous Interactive Information Retrieval Systems," Proceedings for the Interface Meeting for Programming Languages and Information Retrieval, November 4-6, 1973, Gaithersburg, Maryland. (Publication in Preparation). Jointly sponsored by the Association for Computing Machinery Special Interest Groups on Programming Languages (SIGPLAN) and Information Retrieval (SIGIR) and the Institute for Computer Sciences and Technology, National Bureau of Standards, U.S. Department of Commerce.

heterogeneous information retrieval systems may be termed the common computer interface -- a computer system for effecting the translations necessary to allow a user to access multiple diverse I R systems in a common framework (see Fig 1). Our studies and experiments have tended to confirm the efficacy of this general approach while pointing to a need for its elaboration and extension. In particular, the common interface should have the property of a virtual system, that is, it should appear as a single system to the user, with all the necessary capabilities of a retrieval system. Details are given under the various components of the interface listed below.

1. Physical Interconnections and Network Communications

Interconnection of several computers was implemented experimentally in a limited way through an elaboration of the pseudo-terminal type links conceived early in our efforts. These connections are described more

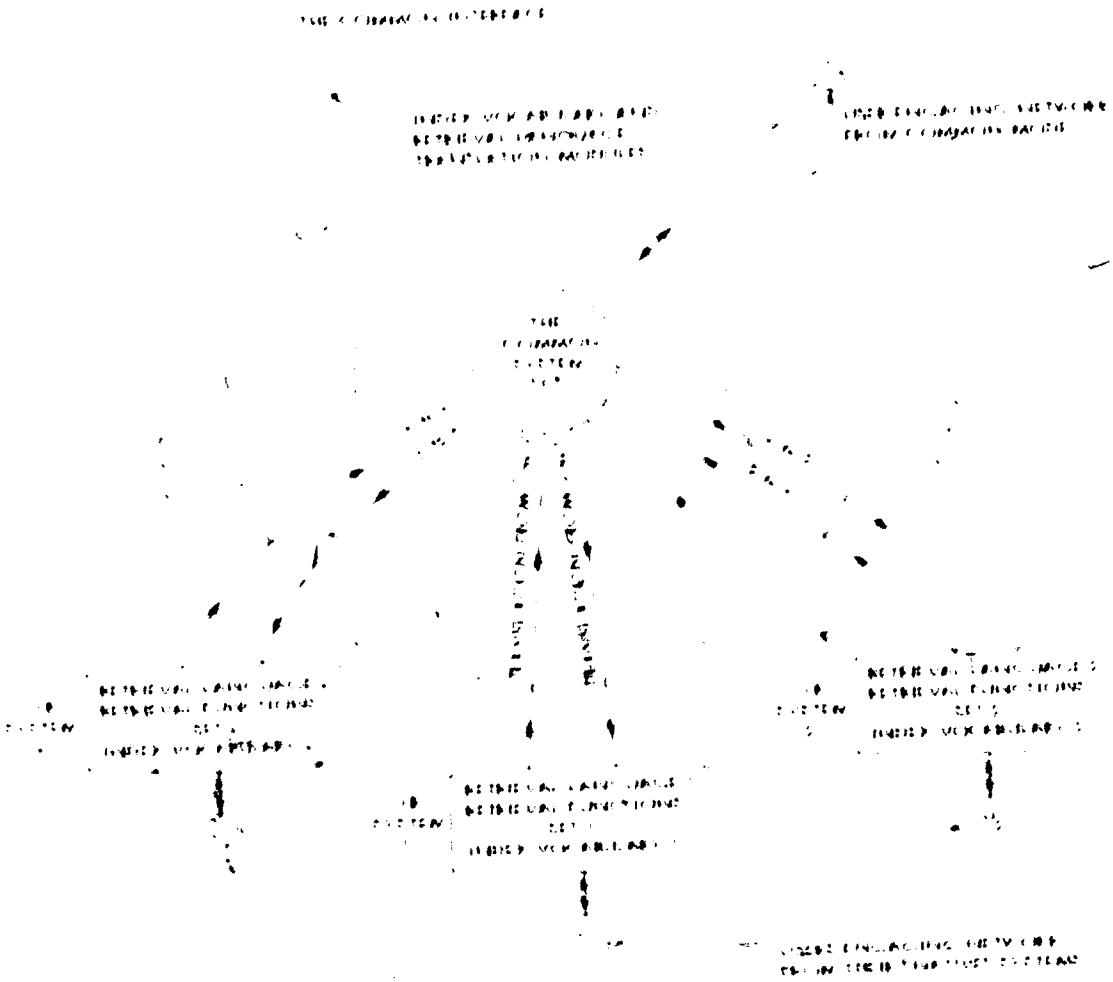


Fig. 1 The Logical Relationship of an Intelligent Network Based on a Common Interface for Heterogeneous Systems

fully in Section C. The development of more advanced communications channels awaits a more complete exposition of functions to be performed by the interface.

3. Common Command Language

Our study of the command languages for several information retrieval systems and for a common interface language led us to a number of important, if still tentative, conclusions:

a. Retrieval commands generally comprise bundles of individual functions.

b. To analyze the commands adequately, then, we must enumerate the many primitive functions from which ordinary commands can be synthesized as macros.

c. However, for user convenience, it is necessary that commands be macros rather than individual primitive functions.

d. Existing command languages, and the systems for which they are intended, are generally far from complete or comprehensive in the number of functions they can perform.

e. Because of the above, it is generally impossible to translate exactly from any given command in language A to one or more commands in language B.

f. Existing systems and their command languages are continually undergoing revisions. These revisions tend toward greater comprehensiveness and commonality among systems.

See Section C below for further analysis.

g. It is a common misconception that the difficulty in translating among command languages results from the different names given to the functionally similar commands and their arguments in the different systems. As suggested above it is, rather, the diversity of "function bundles" and the incompatibility of exact translation that are the prime difficulties. It is, however, convenient for the user to be able to reassign command names into more familiar or easy-to-remember labels. This is one requirement of an interface language which, in effect, allows many dialects within the common-language framework.

h. Therefore, it is prudent to research command languages along the following lines:

i. Analysis of retrieval commands in order to dissect them into their primitive functions.

ii. Development of a common command language consisting of macros of these primitive functions.

iii Emphasis on techniques for assisting users in situations of general incompatibility and inexactness of translation, as well as on general user aids for easier system use.

iv Initial emphasis on access to the I-R network through the common language rather than through the languages of any of the several I R systems themselves.

v Open endedness, flexibility and modularity in the design of the command language is desirable in the dynamic environment of today's I-R systems.

4 Interfacing Between Diverse Indexing Vocabularies

In our early efforts we investigated the notion

of using the Intuck free vocabulary techniques of phrase

decomposition and stemming to alleviate the difficulties of searching data bases that have been indexed under different controlled vocabularies. During the research, this notion was extended to comprise the concept of the Master Index and Thesaurus (MIT). The MIT contains all the index and thesaurus elements of each of the data bases, including an ordered list of all vocabulary terms used for indexing together with the counts of the number of documents indexed by each and the thesaurus relations for each. In addition, through use of the techniques of phrase decomposition (that is, breaking a phrase down into its individual words) and stemming (dropping word endings so as to consider only the word stems) we can automatically identify most intervocabulary relationships in addition to the obvious identity relationship.

The mechanism for readily storing and referring to these relations is to provide in the Master Index and Thesaurus references to all index vocabulary terms under each word stem that appears in that term. (See Fig. 2) Thus, for example, the NASA term* electrical insulation is

* NASA Thesaurus Alphabetical Update, September, 1971

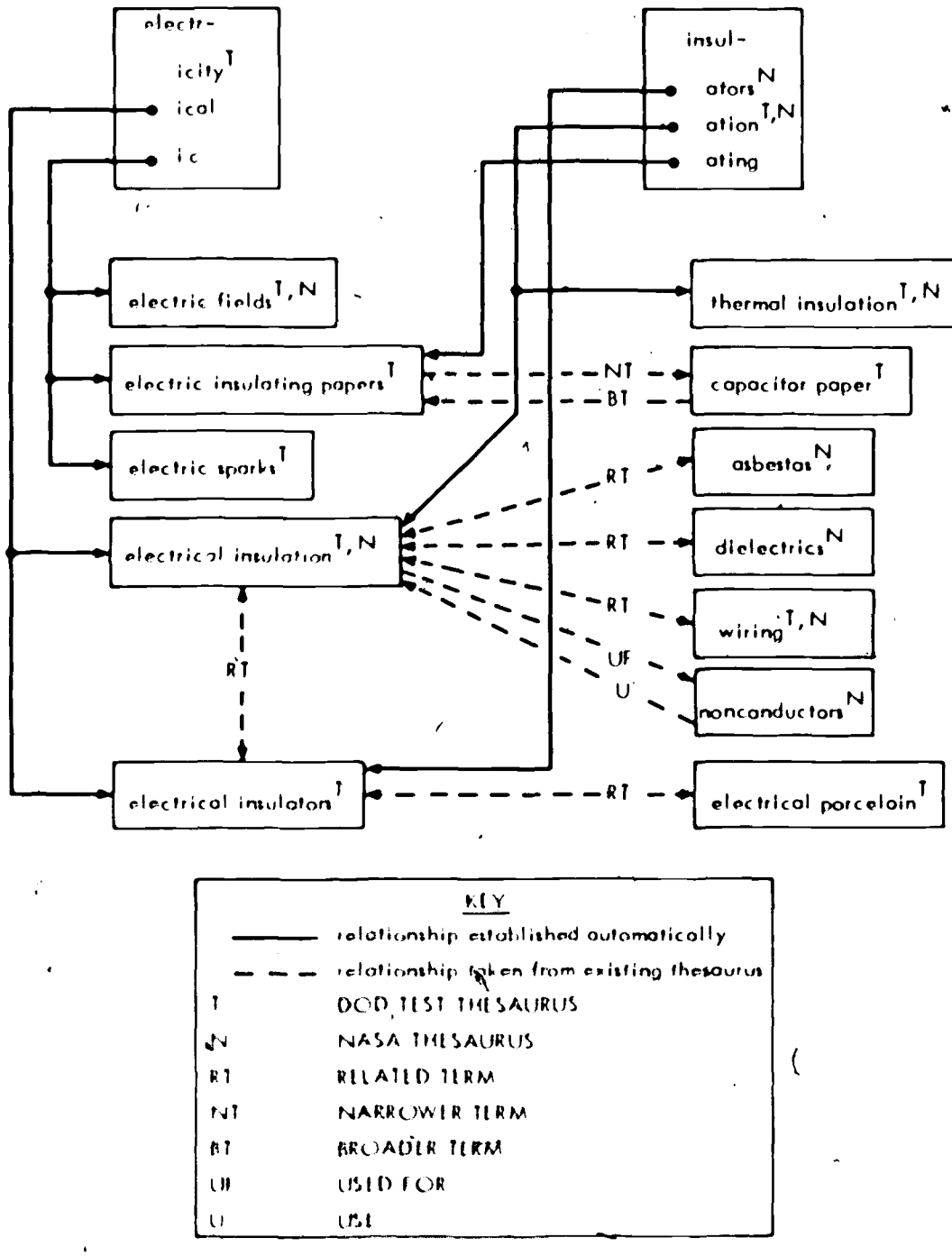


Fig. 2 Sample Relationship among Terms as Maintained in Master Index and Thesaurus

is automatically found to be related to the following TEST* terms in the way specified: specific to electricity and insulation, generic to electric insulating papers, synonymous with electrical insulators and, of course, electrical insulation, and otherwise related to electric fields, electric sparks, and thermal insulation.

The MIT can be used to help determine which terms to search under and, indeed, which data bases to search in. The MIT can be used either in a purely automatic mode or in a manual or prompting mode in which the user is given a display of terms to search under. Interestingly, the MIT should be a useful adjunct even for a single system with a single data base having a controlled vocabulary. In short, the MIT concept clearly has important potential in the development of I-R networks**.

5. Bibliographic Data Elements and Structures

In our research we have determined that another prime consideration in the development of means for users to interact conveniently with different data bases is

---

* Thesaurus of Engineering and Scientific Terms, prepared for U.S. Department of Defense by Office of Naval Research Project LEX, 1967

** See Section C.6 and the appendices for additional details.

the interrelation of the diverse data elements and structures from those data bases. Three ways in which the interrelations are important may be enumerated. First, searching is done on one or more data elements; in order to translate a search done on one system into another, the correct correspondence of data elements must be found. Similarly, user output requests require the specification of combinations of data elements from the catalog records. Finally, in order to combine retrieved document sets from different data bases and to create searchable document sets from separate data bases, we need to: identify when document references from different systems refer to the same document: establish common reference formats: and create common index (inverted file) and catalog data structures.

Our basic solution to this problem is the concept of a common data structure based on the identification of data primitives or basic data elements analogous to the basic component functions of the common command language. Compound data elements in any system can then be translated into, or composed from, combinations of basic data elements in the common data structure. The

basic data elements would be hierarchically arranged into a data structure and, typically, the compound data elements of a system would be equated to a higher level node of the common data structure.

At the highest level we have subdivided the common data structure for bibliographic data into seven major categories. An initial breakdown of one of these, the Abstract-Indexing-Contents category, is shown in Fig. 3. There are 21 basic data elements identified in this category with 14 higher level hierarchical groupings. Note that the abstract sentences, which are included under the abstract grouping, are separated out individually to make it easier to use this information in subject indexing.

Three additional major categories which have been similarly analyzed are Titles, Names-and-Relations, and Related-Document-Citations as shown in Figs. 4-6, respectively. The other three major categories which were identified --- but not fully analyzed --- are Descriptive Document Features, Library Systems Holdings and Shelving, and Control Fields.

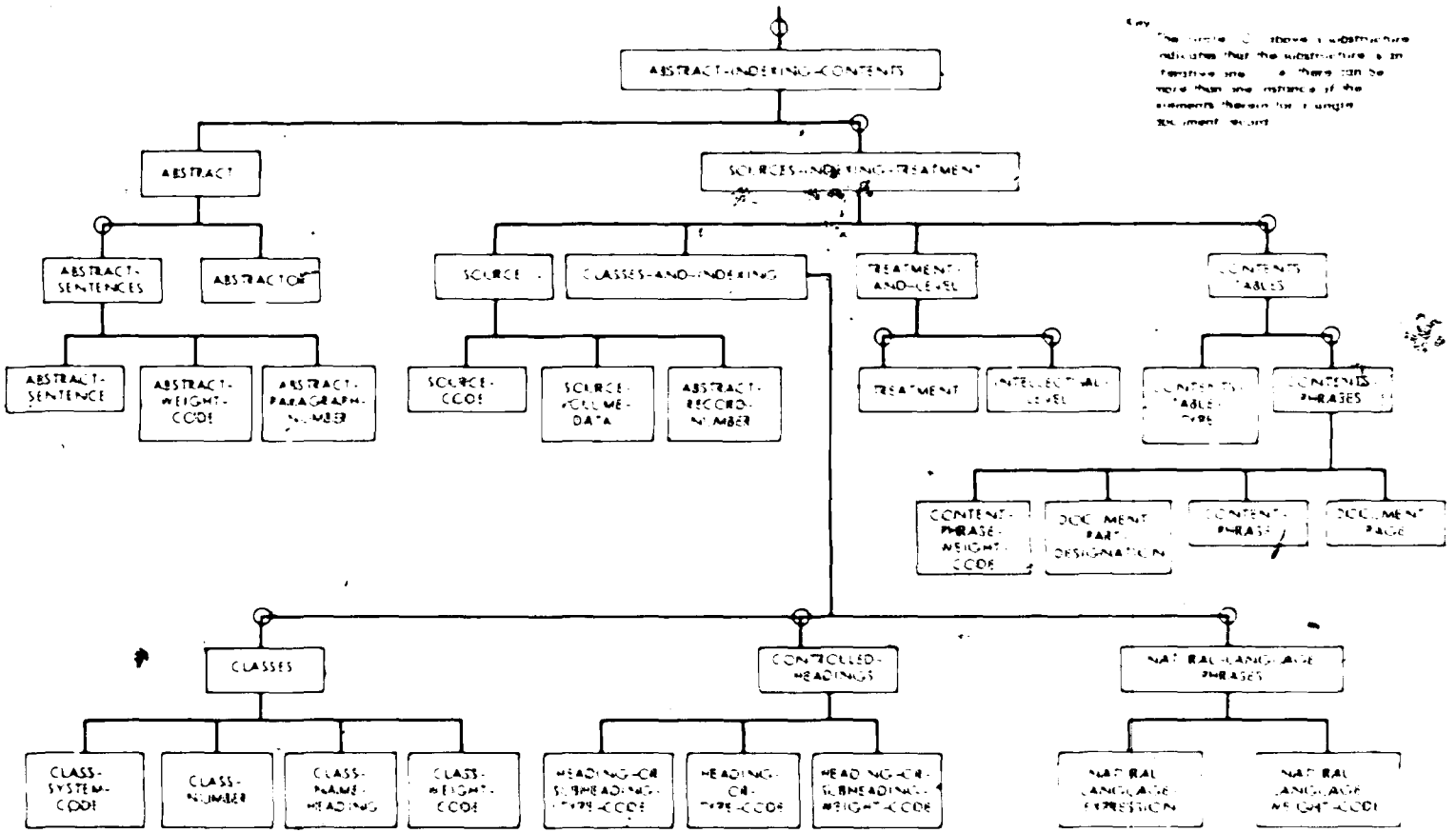


Fig. 3 Common Bibliographic Data Elements and Structure for the Indexing Category (Initial Version)

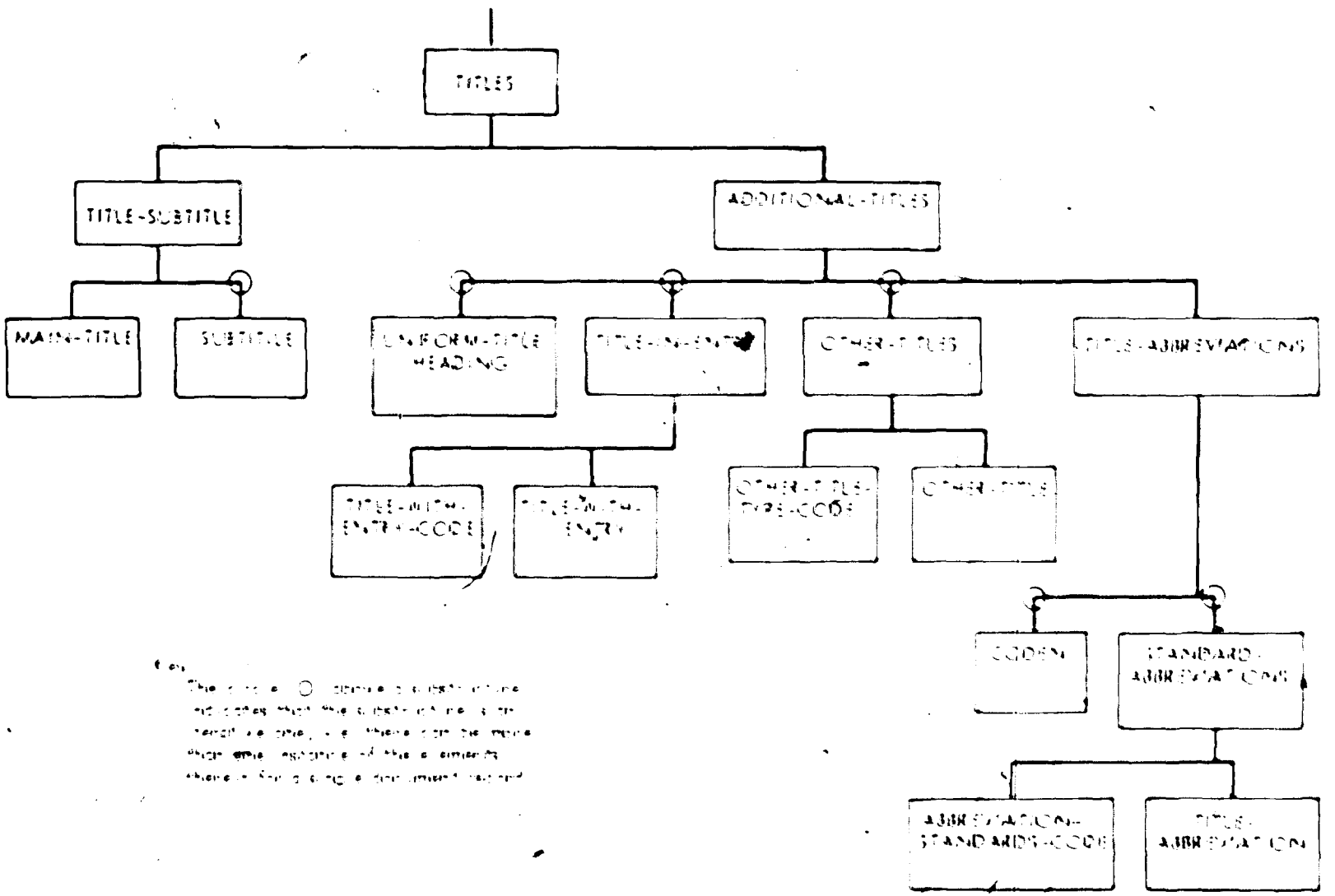


Fig. 4 Common Bibliographic Data Elements and Structure for Titles Category (Initial Version)

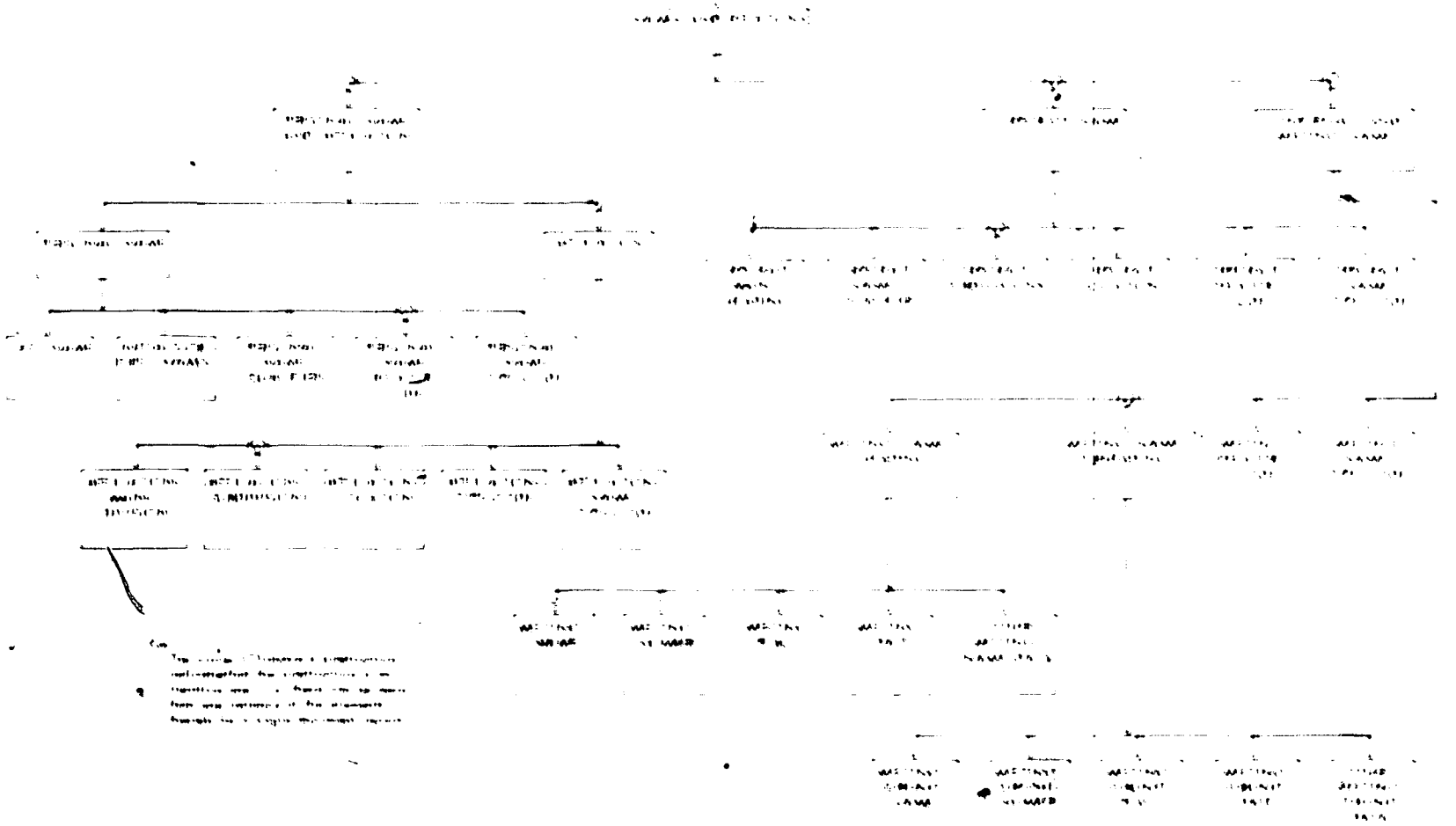


Fig. 5 Common Interchangeable Data Elements and Structure for the Names and Addresses Category (Patient Version)

Fig. 6 Common Bibliographic Data Elements and Examples for the Selected Document Characteristics Categories (continued)

c. Experimental Interface Design, Implementation and Analysis. This is the third task included in our research program. We completed an initial design of the experimental interface and each of the several components of the interface underwent various stages of partial or complete implementation, testing, analysis, redesign etc. as described below.

1. General Design of Experimental Interface

The following criteria were established as guidelines in the design of our initial experimental interface.

a. Demonstration of specific networking functions including:

i. physical interconnection of two or more retrieval systems

ii. user requested connection to either system

iii. elements of a common command language and a means for translating from this language to either system

iv. techniques for aiding in the translation of index vocabulary terminology

v. simple index file search and catalog output in the common language

vi. A collection of search results from different systems in a common file

vii. Ability to use either system to the full extent of its current capabilities using its own command language, if so desired

i. Provision of a test bed for experimenting with the communication techniques mentioned above, and others

j. Simple and limited enough design so that construction is achievable within one year, or less

k. A flexible, modular design so that development beyond this basic core could follow easily and be consistent with our common (or virtual) system philosophy.

With these guidelines in mind, our initial interface design, termed CONIT-1, specified that the following operations could be performed using the interface

a. Select one of two currently connected I R systems for searching (call this System A)

b. Select either the common CONIT language or the language of System A in which to issue commands.

c. If System A language is used, make any requests of System A that are possible within that system.

d. If CONIT language is chosen, permit the following commands.

1. Make a simple search request.

ii. Make a simple print request to see catalog information derived from documents found in the previous search request.

iii. Name a file into which the results of the above print request can be stored for subsequent viewing

iv. Perform viewing of files previously saved as in (iii).

v Request portions of the Master Index and thesaurus for aid in determining appropriate search terms

vi Release the commands given in (a) and (b) so as to select and search the other I R system which is currently connected.

Note that through appropriate combinations of these commands the results of searching in different systems can be collected in a common file.

The detailed status of our implementation of this CONIT I design is given below:

2. Physical Interconnections and Network Communication

Our original intention was to make a connection of the interface to the Intrex retrieval system at M.I.T. and one remote retrieval system. Establishing the communications links to do this required more effort than we originally anticipated. However, we finally overcame several difficulties and we did achieve a communications base that considerably surpassed our original goals. In particular, our interface program termed CONIT (COnnector for Networked

Information Transfer, was installed on the M I T MITTIC's computer. Since MITTIC is a host computer on the ARPANET, the CONIT interface is readily accessible to users from locations throughout the U S served by this network. The interface can communicate with at least six bibliographic information retrieval systems, none of which is based on an ARPANET computer. The first of these is our Intrex system which is currently resident on the M I T IBM 370/165 computer. The second system is MEDLINE which is resident on an IBM 370/155 in Bethesda, Maryland. In addition, we established the means to connect our interface with TYMNET which gives us the ability to communicate with the Systems Development Corporation ORBIT system, the Lockheed DIALOG system, the Battelle BASIS 70 system, the Informatics RECON system, as well as an alternate mode of connection to the MEDLINE system.

A detailed description of the mechanisms by which these interconnections are achieved is given in a recent report.*

* Charles L. Thetton, Data Communications for an Experimental Information Retrieval Network Interface, M I T Electronic Systems Laboratory Technical Memorandum ESL TM 515, August 1, 1974.

A brief summary of the nature of the communications is given below.

As shown in Fig. 7 connections to the I P systems are made through ARIANT TIP (Terminal Information Message Processor). Through appropriate setting of TIP parameters the external I P system is made to look like a terminal to the TIP. Conversely, the TIP is made to look like a terminal to the I P system. The connection to MEDLINE is regularly made through a port on the National Bureau of Standards TIP. The National Library of Medicine has given us use of one of five such connections that MEDLINE maintains on the NBS TIP so that we may pursue our network experiments. The connections between the MEDLINE computer and the TIP ports are by dedicated phone lines.

The connection between the Interface and Interactive MEDLINE is achieved through a port on the TIP located at [illegible] Corporation of America. (Actually, by the same mechanism any port on any TIP could be used.) The connection is initiated by first establishing a connection from the TIP to one data set and

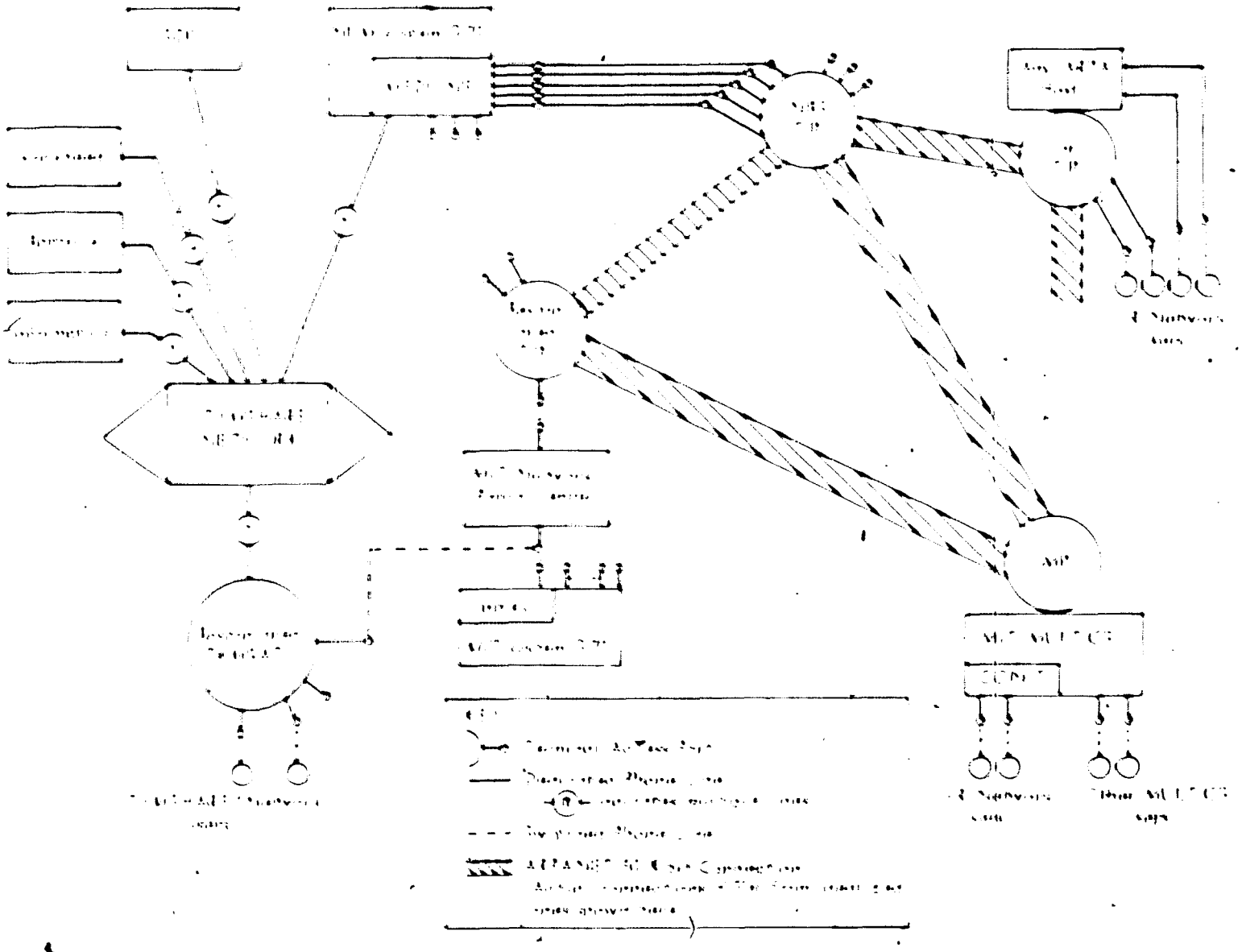


Fig. 7 Computer Interconnections for CONSAT-1 Interface for Networked Information Retrieval Systems

setting certain parameters in the TIP for the connection. A terminal is then connected through a second data set to the M.I.T. IBM 370/165 in which INTREX is located. Finally, a direct connection is established between the two data sets, thus completing the communication link between the ALPANET TIP and Intrex. The electrical and mechanical hardware by which the terminal is switched from TIP to 370 computer, and by which the two data sets are connected together, has been termed the "network crossover box" and was designed and constructed at the Electronic Systems Laboratory. (See schematic diagram in Fig. 8).

The same mechanism is used to establish connection between the TIP and the TYMNET network by merely making the second terminal call be to the local TYMNET satellite computer rather than to the M.I.T. 370 computer. It should be noted that the network crossover-box mechanism has the flexibility to be connected (at different times) to different computers. This flexibility is of importance in experimenting with different I-R systems.

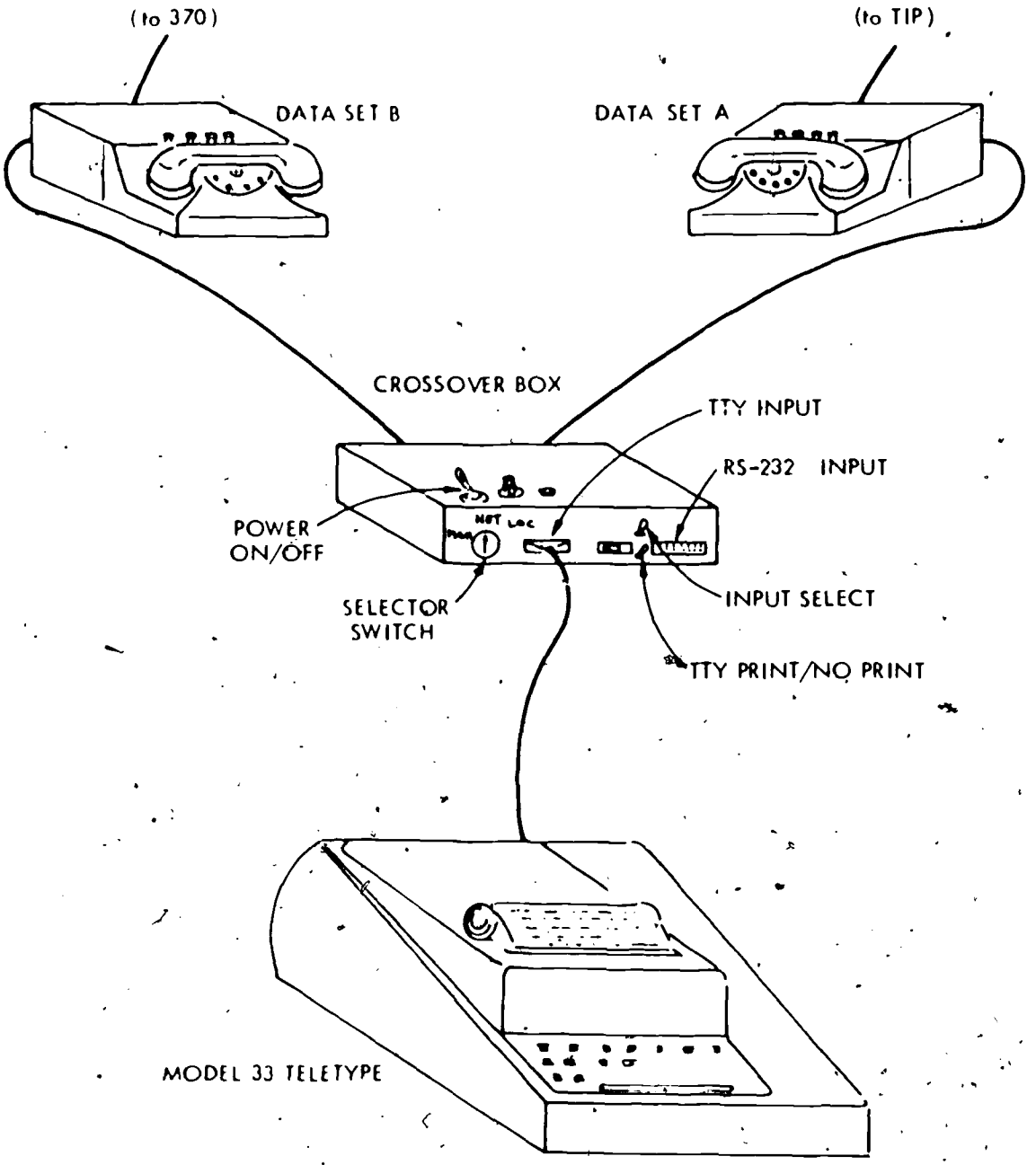


Fig. 8 Hardware Components Used in the TIP to System/370 Connection

In any case, once any TIP has been connected to a non-ARPANET computer, the CONIT interface program must establish access to the specified TIP port through appropriate calls on ARPANET software. This having been done, CONIT can send commands to either of the two currently connected I-R systems through the respective TIPs and receive responses from these I-R systems over these same full duplex connections.

Thus, at any one time, CONIT can be communicating with either of the two I-R systems currently connected: one being the MEDLINE system through the NBS TIP, the other being either Intrex or one of the TYMNET systems through the CCA TIP. The different TYMNET systems can be connected to CONIT sequentially by having CONIT send the commands to logout the currently connected system and then logon a second TYMNET I-R system. Finally, CONIT can switch from a TYMNET system to Intrex, or vice versa, by redialing the connection to the CCA TIP from the network crossover box.

3. The Basic Interface Programs

The CONIT-1 system contains as a nucleus a set of program modules that enable a user to have an

interactive dialog with the computer. These modules include, for example, routines that provide for message transfer between a user terminal and CONIT, the parsing of user requests into individual command and argument strings, transfer of control to routines that execute commands, and the storage and selection of messages sent the user in full or abbreviated formats.

Another set of routines enables and disables the ARPANET TIP connections, as mentioned in the previous section.

A third set of routines allows for the translation of user commands into commands in CONIT or external retrieval systems. These routines are based upon the concept of translation tables. CONIT automatically selects the appropriate translation table when, for example, the user is speaking the CONIT language to MEDLINE or Intrex. Special purpose routines also perform other functions required for proper translation to external systems; for example, MEDLINE must receive characters in upper case, and user commands to MEDLINE are so converted.

4. CONIT-1 Command Language

In the design of command language for CONIT-1 the question naturally arose as to the structure and characteristics of such a language. We have inclined toward a simple command-name/argument string type format with common English words for the vocabulary (abbreviations allowed), very simple punctuation (e.g., spaces) for delimiters, and general independence of command and argument ordering. This structure seems preferred for ease of use --- especially compared to a more complicated programming-type language. English, used as a common language, suffers from the twin defects of being of complicated structure and being very ambiguous; these defects make it hard to explain to users what precisely can be accomplished in the system at hand and even more difficult for the system to parse the user's request syntactically and semantically.

These views on language structure have been supported by our own previous work in Project Intrex* and the direction

* See, Marcus, R.S., Benenfeld, A.R., and Kugel, P., "The User Interface for the Intrex Retrieval System" in Interactive Bibliographic Search, D.E. Walker, (ed), AFIPS Press 1971; pp. 159-201.

we see existing systems taking as evidenced, for example, by the previously mentioned Stanford seminar. With these views in mind, we began the design and implementation of the CONIT command language. The initial set of commands for CONIT-1 that were implemented are listed in Table 1. Note that in several cases --- e.g., FIND, PRINT --- only the simplest default conditions were implemented. In addition to these regular CONIT commands, if a language other than CONIT is being spoken, any command of that language may be given; i.e., a "transparent" mode is possible. It can be seen by reviewing the explanation of the commands that --- with one exception --- all the goals for CONIT-1 as set forth in the initial design given in Section C.1 above have been met including the translation of a couple of commands from the common command language into the languages of two target systems. (The status of the one exception --- display of the Master Index and Thesaurus --- is described below.)

The details of the CONIT command language, while following the general principles described above, were developed in only a very preliminary stage; should our work be continued, they are subject to change. Nevertheless, the

Table 1. List of Commands Executable in CONIT-1

COMMAND	EXPLANATION
SELECT x:	Select a system x to search in and enable TIP connections.
SPEAK x:	Select command language x.
LIST CONDITIONS:	List current language and system being searched.
VERBOSE:	Have CONIT give full-form, instructive responses.
BRIEF:	Have CONIT give abbreviated responses.
SET TABLE x:	Pick translation table named x.
REPLACE $x=y:	In current translation table, add the rule that the string x is replaced by y.
DELETE RULE $x:	Delete from translation table that rule with x as the string to be replaced.
LIST TABLE:	Print out rules in current translation table.
DISCONNECT x:	Disable TIP connection to system x.
(T)* FIND x:	Do a simple search on x in system currently selected.
(T) PRINT:	Print out standard catalog information on documents in current list.
NAME FILE x:	Select file x (create a new file if x does not already exist) to be used to save future catalog output.
FILE:	Add response of next request to currently selected system to save file.
EXIT:	Leave CONIT and return to MULTICS command level.

* (T) indicates that a translation to the selected system is required.

decision-making process for choosing certain of these details may be of some interest. The use of the command name FIND for the basic search request deserves some comment. It is different in form from both of the systems it now translates into. In Intrex there are three different search commands (SUBJECT, AUTHOR, TITLE) to specify the different inverted files or inverted file subsets (title words are a subset of the subject inverted file) to be searched. In MEDLINE the absence of a command name implies a search request (just the term to be searched is given), although one may use, or need to use in certain situations, the form "FIND x".

Thus, both Intrex and MEDLINE may be characterized as having, in general, the search command being a default situation; i.e., being unspecified as such. Our Intrex experience has suggested that users find the command language easier to learn and remember when it is more consistent. The default command upsets the simple rule: command name first then arguments. Leaving off the command name does not save much since an abbreviation --- as short as one letter --- may be used. It also makes parsing and error checking more difficult. Thus an explicit command, one in common use, was chosen.

It should also be noted that even in this rudimentary interface it can be seen how a number of the complexities of

use of multiple diverse systems can be mitigated. For example, the CONIT user need not know the intricacies of logging on to the remote systems --- only the SELECT command need be given; the rest is taken care of for him by CONIT. In the case of regular MEDLINE use through the ARPANET, there is the special problem of accessing through one of 5 TIP ports. CONIT automatically transmits the proper protocols and even cycles through the 5 ports, one at a time, until it finds one that is available. Similarly, CONIT gives the proper log off, or QUIT, message to system x whether because of a DISCONNECT x or implicitly required through an EXIT from CONIT. Proper termination of a user from a system is important not only for the individual user (e.g., for proper charging) but also for other users who might otherwise face problems in logging on.

5. Implications for Command Language Design

In considering how to design a common command language that could serve as an interface to existing retrieval systems, we came to a rather paradoxical conclusion: it is both impossible and not too difficult to translate from such a language to the languages of the existing systems. It is impossible in the sense described above in Section B.3

in that existing systems are usually rather limited in the functions they can perform and simply cannot perform an arbitrary request. On the other hand, if an approximate translation is allowable, it is often not too difficult to make a reasonably good one. For example, the simple PRINT command in CONIT can be translated to the "PRINT" command in MEDLINE or the OUTPUT command in Intrex with reasonably close results even though the catalog information printed might occasionally vary slightly in the default modes of these commands.

Two ways to aid in the translation process are described elsewhere in this report: the Master Index and Thesaurus (MIT) and the common bibliographic data structure. Two other elements we have found important in the interface problem are the identification of search statements and the disambiguation of named entities (commands, arguments, sets, etc.), as discussed below.

Firstly, it should be recognized that the names given entities in the target systems can be isolated from those names given in the common language, as long as an unambiguous translation is possible from common-to-target language. Thus the main task is to provide a mechanism for insuring an

unambiguous set of named entities at the common system level. (Of course, the naming of entities, such as commands, at the common level should consider such factors as the advantage of using already established widely used terms --- e.g., PRINT.) The common interface then needs to keep account of all entities that are named a priori in the common command language or that may be named in the course of a given user's session. Such entities can be generically classified as:

1. command names
2. names of arguments to commands
3. names given to retrieval search sets
4. new names given by a user (as through a REPLACE function) to any one or combination of entities in the above 3 categories.

Argument names cover such entities as names of systems, data bases, saved filed, catalog elements, vocabularies, vocabulary elements, modes of command operation, and so forth. The common interface would insure against ambiguities by checking names a user might initiate against the current list of named entities and request that the user choose again if a duplication is detected. It is possible, of course, to disambiguate at times on the basis of context, but it seems simpler and

easier to explain system use to a user if all ambiguities are avoided in the first place.

The identification and recording of search statements and their results appears to be an especially vital task where multiple systems and data bases are involved. The elements associated with a search that need to be recorded include:

1. the search statement number (assigned by system);
2. an alternate search statement name given by user;
3. the search statement itself;
4. the system(s) and data base(s) for which intended;
5. the number of references (or documents) retrieved;
6. the name of search given in any target system;
7. the actual list of references retrieved;
8. for any subsearches necessitated by this search*, either at level of common system or the target system, all the above items of information.

All items of information above --- except, for some systems, item (7) --- could be obtained and maintained at the common interface level without requiring modification of the target system.

---

* See the discussion in Section C.6 below of the automatic use of the MIT for an example of how subsearches may be generated.

6 The Master Index and Thesaurus

The Master Index and Thesaurus (MIT) concept described in Section B.4 was developed for inclusion in the experimental interface. A detailed design was prepared describing the specific information to be contained in the MIT and the logical interrelations among the various parameters so contained. Appendix A describes this initial design.

How the MIT could be displayed under manual control is described in Appendix B. How the MIT could be used in an automatic way has also been studied. Assume a user makes a search request of the form FIND A B C, where A, B, and C are ordinary English words. The system could then search the MIT for all words or phrases that had a word stem that matched the stem for the word A. Call these terms A1, A2, A3.... Then a search consisting of the union of searches on all these terms would be found; i.e., A_S = FIND A1 OR A2 OR A3.... Similarly, B_S and C_S are found. The search set answering the initial request would then be the intersection of the three sets A_S AND B_S AND C_S. Previous Project Intrex

work* suggests that this resulting set would generally give very effective search results.

Of course, many other modes of use of the MIT may be desirable in particular contexts including automatic selection of synonymous and specific related terms and user selection from lists of terms found automatically, as above.

Programs were written to transform information from the data bases of our two main experimental systems --- MEDLINE and INTREX --- into the Master Index and Thesaurus organization shown in Appendix A. The Intrex data base** was generated from INSPEC tapes containing bibliographic data from three separate data bases: Science Abstracts, Electrical Engineering Abstracts, and Computers and Control Abstracts. The MIT formed from these data bases contained both the free and controlled-vocabulary information that could be gleaned from both the INSPEC catalog records as reformatted in the Intrex system and the Intrex inverted file resulting from inverting certain of the INSPEC fields. In particular, the controlled-vocabulary terms --- both classifica-

---

* See, 1) Project Intrex Semiannual Activity Report PR-12, 15 September 1971, Massachusetts Institute of Technology, pp. 29-47.

2) Overhage, C.F.J., and Reintjes, J.F., "Project Intrex: A General Review," to be published in Information Storage and Retrieval.

** It is of interest to note that the catalog records for the Intrex INSPEC data base were organized in the hierarchical data structure as described in Section B.5.

tion terms and controlled index terms --- and their index counts are captured as are all free-vocabulary word stems and their index counts from titles, abstracts, and free subject expressions. In addition, the listing of controlled terms under each of the word stems they contain was generated.

Similarly, programs were written to generate MIT information from MEDLINE data bases in the new MEDLARS-II format. This information included the thesaurus information given in the MEDLARS vocabulary files.

IV. CONCLUSIONS

A. Work Accomplished. Research on the coupling of interactive information systems has progressed in a significant way. The research has focussed on the concept of a translating computer interface by which the networking of heterogeneous interactive information retrieval systems can be achieved. This concept appears to be a viable approach to the development of I-R networks in the interim period during which I-R system and network standards are gradually evolving. Particular concepts and techniques which appear --- through analysis and the design, implementation, and testing of experimental interfaces --- to be especially useful in developing

the over-all interface concept include: (1) the virtual system concept by which users perceive the network as a single homogeneous system; (2) a common command language synthesized from a basic language of atomic I-R functions; (3) a master index and thesaurus which stores the vocabularies of the separate data bases along with index term interrelationships and counts; (4) a common bibliographic data structure by which the data elements for bibliographic information may be enumerated, hierarchically structured, and interrelated among different data bases.

B. Recommendations for Further Work. While a basis for research into the coupling of retrieval systems has been laid, much additional work is obviously needed including the further elaboration of the techniques listed above; their implementation in additional demonstration systems which connect several systems and several data bases and cover most retrieval functions; the testing and evaluation of these systems with real users; and the development of more effective computer-to-computer communications. Also, we see the need for additional study of the relationship of the computer interface to the network of I-R systems including such

questions as:

1. How many of the I-R functions should be performed within the interface as distinct from being performed by the separate I-R systems?

2. What are the technological and economic reasons for treating the separate I-R systems within the network as inviolate "black boxes" (the assumption we have made in our research to date), as contrasted with the alternate concept that they should be modified to interact more effectively with the interface?

3. To what extent can and should the common-interface concept become a de facto standard toward which existing systems may evolve in order to take maximum advantage of networking potentials?

APPENDIX A

DATA COMPONENTS FOR THE MASTER INDEX AND THESAURUS (MIT)

Each entry in the Master Index and Thesaurus contains up to five fields. These fields with their data components are listed below.

A. Header

1. Is the entry for a single word stem or a phrase? (1 bit)

2. Entry type (3 bits) (e.g., subject, author, other)

3. Name (variable length A/N string) (e.g., "magnet", "magnetic resonance", "Smith", "Van Pelt")

4. (a) Number of English words in phrase (6 bits). (phrase entry only)

   (b) Number of affixes (stem entry only)

5. Counts (document and reference counts for each collection --- i.e., data base, for which counts are being kept in MIT)

B. Affixes (This field needed only for word-stem entry).

This field contains the set of affixes for a word stem.

For each affix there would be:

1. Relative (to this entry) affix number (6 bits) (sort endings alphabetically)

2. Absolute affix code (12+bits) (at present this would be the Intrex 12-bit suffix code; however, we should allow for future expansion to permit prefixing, as well).

3. Document and Reference counts (i.e., $2n_a n_c$ counts, where n_a = number of affixes, and n_c = number of collections as in A.5).

C. Controlled Vocabulary Terms (may be null)

This field lists all the controlled vocabulary terms for this entry.

For each term there would be:

1. Term number (relative to this entry) (6 bits)

2. Controlled vocabulary code (6 bits) (i.e., what vocabulary the term is in). (sort by ending, this code, and then type).

3. Type of term (4 bits) (e.g., classification or index terms).

4. Affix code (5 bits) (needed only for single word --- code comes from B.1).

5. Status (2 bits). Is this a currently used term? (If not, we will eventually include dates when it was used --- also date of initial use for recent term).

6. Reference counts for each collection, i.e., $n_c n_t$ counts where n_t = number of controlled vocabulary terms; note n_t includes separate item for each vocabulary - type - affix combination.

D. Phrases With This Stem (For word-stem entries only; may be null).

For each such phrase there would be:

1. Phrase number (6 bits).

2. Term number of phrase (from Field C.1 of entry for phrase) (6 bits)

3. Type of term of phrase (from Field C.3 of entry for phrase) (4 bits).

4. Word number of word stem in phrase (4 bits).

5. Controlled vocabulary code (6 bits).

6. Phrase name (variable-length A/N string) (sort alphabetically, then by vocabulary and term type).

7. Reference counts for each collection.

E. Thesaurus Relations (may be null)

The standard thesaurus relations are given here. For each relation there would be:

1. Term code for given term (6 bits) (from Field C.1).

2. Type of relation (4 bits) (e.g., broader term, narrower term, synonym, use, used for, related term, morphological variant).

3. Controlled vocabulary for related term (6 bits). (Could conceivably be different than for given term --- note could also relate "free vocabulary" to controlled vocabulary).

4. Name of related term (variable-length A/N string). (sort alphabetically).

5. Is related term a word or a phrase? (1 bit).

6. Automatic search expansion? (2 bits) --- should the related term be automatically added to a search when user asks for given term.

7. Remarks (variable-length A/N string) (probably null). (Remarks in free-form English on nature of the relation and when it should be applied).

8. Reference counts for related term for each collection.

APPENDIX B

DISPLAYING THE MASTER INDEX AND THESAURUS

Preliminary specifications for the display of the MIT under user control are given below.

The command RELATE is used to look up and display information from the MIT. The syntax is

RELATE A B C X

The last argument (X) is that word or phrase to be looked up in the MIT. The first argument (A) specifies the type of relation to be displayed as follows:

ALPHA--terms that surround X alphabetically

PHRASE--terms having word stems in common with X

THESAURUS--thesaurus relations for X

Eventually, some combination of relation types may be permissible on one display. For now, the default option should probably be ALPHA.

The second argument (B) specifies that only terms and relations in a particular vocabulary be considered. Sample values for B might be MESH or INSPEC or INSPEC PHYSICS etc. Several vocabularies could be specified at once, so that the actual syntax is B1 B2 B3...; i.e., the connector OR is implicit. The default condition would mean all vocabularies considered.

The third argument (C) specifies, analagously to B, the particular collection(s) to be considered; e.g., MEDLINE, SDILINE, INTREX/ INSPEC, etc. Again, the default condition means all collections considered.

These specifications will be further illustrated through examples of display output for some RELATE commands. Note the default options for arguments B and C. The index terms are not meant to be actual terms from the given vocabularies.

A. SAMPLE DISPLAY OF ALPHABETICAL RELATIONS

READY

relate alpha radiation

The index terms alphabetically near the term radiation[7] are listed below.[1] <MRA> [6]

POSTINGS[2]	TERM NO.[3]	: TERM[4]	: VOCABULARY[5]	: COLLECTION
220	TY	rabid dogs	MESH; HEADING	MEDLINE
15	TZ	rachet wheels	INSPEC EE; TERM	INTREX
20			INSPEC COMP; TERM	INTREX
305	TA	rad-	(English stem)	INTREX
21				SDILINE
107	TA1	radical	(English)	INTREX
5				SDILINE
203	TA2	radiation	(English)	INTREX
17				SDILINE
1276			MESH; HEADING	MEDLINE
45				SDILINE
410			INSPEC EE; HEADING	INTREX
13	TA3	radius	(English)	INTREX
192	TB	radiation damage	INSPEC; HEADING	INTREX[8]
217			MESH; HEADING	MEDLINE
130	TC	radiation effects on tissues	MESH;HEADING	MEDLINE[8]

If you want to see terms that follow alphabetically, type relate more. To see phrase thesaurus relations type relate phrase X or relate thesaurus X, where X is the name or term no. (see above) of the term whose relations you want to see. <MRAZ> [6]

READY

NOTES FOR SAMPLE DISPLAY OF ALPHABETICAL RELATIONS

(1) This is the VERBOSE mode message: in TERSE mode no message would be given. If the argument X did not correspond to any term in the MIT the following message would be given:

There is no index term radiation. [However, there are terms with the same stem: rad-.] Terms alphabetically near radiation are listed below. <MRA3> (6)

The sentence in brackets would be inserted when the stem for (the single word) X does exist in the MIT, although the full word X does not. In the former case, the display would be:

TA radiation (No Such Term)

(2) The document (or reference) counts are given in this column when available. (A stem where no free vocabulary word exists would not have a count.)

(3) The term number is assigned by CONIT relative to the term for argument X, which is TA. The two terms coming alphabetically before TA are given first and labeled TY and TZ. As many terms after TA are given so as to have at least 20 lines of term display. Note full words under word stems are given numbered final characters. If a continuation display is requested (relate more), then numbering is continued from previous display.

(4) Full words under word stems are indented one space. If a term is repeated (for a different vocabulary or different collection), its printing is not repeated. For each full word in Field B of

the MIT, first the free-vocabulary counts would be given (from Field B.3) and then the controlled counts (from Field C).

(5) The vocabulary is given followed by a semicolon and then an indication of the type of vocabulary term (e.g., class heading or index term). The vocabulary and type term parameters come from Field C.2 and C.3, respectively, of the MIT (see Appendix A).

(6) These are message names. They could be arguments to a SUPPRESS command which thereafter causes their deletion (or replacement by TERSE form) or the EXPLAIN command.

(7) Some emphasis on the playback of argument X is desirable. Here, for example, radiation could be in a different color or capitalized.

(8) If information under a given column is too long to fit in that column, some convention such as suggested here will be desirable.

B. SAMPLE DISPLAY OF PHRASE RELATIONS

READY

relate phrase radiation damage

The index terms with words whose stems match the stem(s)[1] of rad-iation (and damag-e)[1] are listed below: <MRP>

POSTINGS	TERM NO.	: STEM 1[4]	: VOCABULARY	: COLLECTION
305	TA	rad-	(English Stem)	INTREX
21				SDILINE
107	TA1	radial	(English)	INTREX
...[2]				
13	TA3	radius	(English)	INTREX
		PHRASES WITH STEM[4]		
50	TB	polarized radiation	INSPEC EE: HEAD	INTREX
192	TC	radiation damage	INSPEC EE: HEAD	INTREX
217			MESH: HEAD	MEDLINE
		: STEM 2		
	TD	damag-	(Stem)	INTREX
...	...	...	...	...
		PHRASES WITH STEM		
18	TE	damaged goods	BUSINESS	INFORM

* see term TC above[3] radiation damage

To see thesaurus relations type relate thesaurus X, where X is the name or term no. (see above) of the term whose relations you want to see. <MRP2>

READY

NOTES FOR SAMPLE DISPLAY OF PHRASE RELATIONS

(1) Delete parenthetical parts if only one word in X. Null result message:

There are no index terms with words whose stems match the stem(s) for rad-iation (or damag-e). <MRP3>

(2) The full display for this stem, as in Section A, goes here.

(3) Duplicate phrases are indicated in this fashion.

(4) Stem information is given in addition to multi-word phrases as a convenient way to indicate what the stemming is for the given words in argument X as well as to show single-word "phrases".

C. SAMPLE DISPLAY OF THESAURUS RELATIONS

READY

relate thesaurus radiation damage

The index terms with thesaurus relations to the term radiation damage are given below:(1) <MRT>

POSTINGS	TERM NO.	TYPE(2)	: RELATED TERM	: VOCABULARY	: COLLECTION
NONE	TA	CODE	6.21	INSPEC	
NONE	TB	CODE	H.50.31	MESH	
2005	TC	BROADER	radiology	MESH: HEAD	MEDLINE
75	TD	NARROWER	radiation dosage	MESH: HEAD	MEDLINE

READY

NOTES FOR SAMPLE DISPLAY OF THESAURUS RELATIONS

(1) Null messages:

There is no index term radiation damage <MRT2>

There are no thesaurus relations for the term radiation damage <MRT3>

(2) From Field E.2 of MIT (see Appendix A)