A TRANSLATING COMPUTER INTERFACE FOR A NETWORK OF HETEROGENEOUS INTERACTIVE INFORMATION RETRIEVAL SYSTEMS

Richard S. Marcus
Massachusetts Institute of Technology
Electronic Systems Laboratory

ABSTRACT

The need for a network of heterogeneous interactive bibliographic information retrieval systems is projected from the facts of wide acceptance and growing demand for these systems and the limitations of their use caused by limited online data base size. Because of the established characters of the different I-R systems and unlikelihood that a standardized I-R system leading directly to the ultimate uniform network will soon be generally adopted, we propose an intermediate solution in which computer interfaces would provide the networking capability by translating and converting among the diverse languages and data bases of existing systems.

We have begun work on such a network in which the computer interface is based on the concept of a common language for commands, indexing vocabularies, and data base structures. In particular, the common language for commands and data base structures is based on identifying the basic or primitive I-R functions and bibliographic data base elements. The conversion among indexing vocabularies is based on the concept of a Master Index and Thesaurus containing the conglomerate thesaurus information from the separate data bases. In addition, the phrase decomposition and stemming of the individual words in the search request and subject index phrases are used as further techniques for automated conversion among diverse vocabularies. An initial experimental interface is described which interconnects the M.I.T. Intrex system, the MEDLINE retrieval system, and the TYMNET computer network using the ARPANET Terminal Interface Processors for intercomputer communication.

A. THE NEED FOR A NETWORK OF HETEROGENEOUS INFORMATION SYSTEMS

1. 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 to interactive retrieval systems from most points in this country at costs as low as from $6 to $50 per hour. These systems contain in the aggregate references to documents numbering in the millions in such subject areas as chemistry, physics, 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.

2. 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

*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.

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 are many more serious users who were from outside the five research areas 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. 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.

3. The Ultimate Uniform Network Solution

Ultimately, the solution to these problems may necessitate the construction of a large-scale, online 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 dozen or more 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 economics 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.

4. 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, nonstandarized retrieval systems. Lack of standarization 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 defacto standard.

5. The Computer Interface

In view of the foregoing obstacles to the immediate implementation of the ultimate network, we have 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 would achieve compatibility among systems of heterogeneous hardware and software components through translating and conversion algorithms.

B. THE COMMON INTERFACE/VIRTUAL SYSTEM CONCEPT

1. Alternatives to Interface Design

Having decided on a computer-interface approach to the network, we need to consider what design principles will best serve our objectives. In particular, what kind of switching mechanism is envisioned to bring the different I-R systems into compatibility. Perhaps the most obvious and conceptually simplest design would have the interface consist of a complete set of binary translation algorithms, one pair for each pair of different systems. Thus, if there were n different systems, there would be n(n-1), or of the order n^2 translation algorithms.

An alternate mode of interconnection would have a common node or interface, into which messages from each source I-R system would be translated before being translated into the target I-R system. Such a mode, as diagrammed in Fig. 1, requires only 2n translation algorithms and is the one we have chosen.

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

2. Development of the Common Interface

The format of the interface could be based on some existing system. However, even discounting the likely nontechnical objections by proponents of the other systems, we feel that no existing system has a comprehensive enough set of functions of flexible enough format to serve as a basis for the common interface. We plan, then, to construct a common interface with the requisite features including a common command language, data base structure, retrieval functions, and index vocabulary control.

The conceptualized version of the common interface will provide the intermediary basis for the translation algorithms. The features of this interface, taken together, also provide a virtual retrieval system, which is what U_C, a user engaging the network from the common node in Fig. 1, sees. A completely implemented interface physically realizes all the functions of the virtual system. It may be hoped that this common-interface/virtual-system, possessing as it should the comprehensiveness to encompass all existing I-R system functions in a convenient way, will be a precursor of the standard retrieval system that we are seeking for the ultimate network.

C. COMPONENTS OF THE INTERFACE

In this section we describe our general philosophy and plans by which the several components of the computer interface are being developed.

1. Physical Interconnections and Communications

Important considerations in the physical interconnection of heterogeneous I-R systems include (1) the means to transmit and receive properly formatted messages which will satisfy the code and baud-rate requirements at each computer, and (2), the ability to transmit the information in large enough quantities and at rapid enough rates to support the interactive requirements of many simultaneous users in an efficient operational fashion.

There are many projects and systems which have addressed these questions in considerable detail. We do not expect to make any major contributions to this aspect of networking, at least in the early stages of our work. Rather, our plan has been to establish telecommunications interconnections in a quick, convenient, and economical way so that we may concentrate on research into the logical aspects of translation among the I-R systems, including translation for command languages, index vocabularies, and data elements and structures.

A simple approach to intercommunications is to implement the communication links through the "communications-control unit" at each computer system. Since the communications-control unit is intended primarily to interface with terminals, this kind of arrangement amounts to each computer "thinking" that the computers it is connected to are terminals. While communication is at relatively slow terminal speeds, it can be supported by readily available voice-grade telephone lines and is sufficient for our research purposes.

2. Development of a Common Command Language

Our basic plan for the command language aspect is to develop a language in which all the functions for information retrieval operations can be expressed. The aim is to break the functions down into the smallest components that find any different application in any two systems so that any function in any language can be expressed as a combination of common language functions, i.e., a macro function in the common language.

It could be supposed that such an atomistic common language would provide the basis for a translation from any language to any other: first a command in the source language is decomposed into common language components and then the target language commands are synthesized by appropriate groupings of these components. There are several complications to this seemingly simple picture. There are theoretical questions --- e.g., ambiguity and context sensitivity in original and translated command strings ---- and practical concerns ---- e.g, efficiency. However, perhaps the chief stumbling block to straightforward conversion from one language to another is the simple fact that no I-R system provides a truly comprehensive set of retrieval functions. In particular, the functions that are available in one system are quite commonly impossible to accomplish in another, at least in the exact fashion found in the first system. For example, the algorithm which specifies what constitutes a match for a search command will provide automatic word stemming in some system, only user-specified stemming in other systems, but no stemming at all in still others.

Our general answer to this problem is to use the atomistic common language to help provide exact translations where these are possible and efficient and to provide partial or generalized translations in other cases. Users who wish to control the degree of faithfulness in translation will be given the mechanisms for doing so, including an estimation of the costs involved ---- e.g., in terms of increased processing time. Another answer is to emphasize the use of the common language as the user command language, thus avoiding some of the translation complexities.

Eventually, as experience is gained with different retrieval systems and their different capabilities --- including experience generated by the kind of network we are proposing, there will be a tendency for the different retrieval systems to adopt a more uniform set of capabilities and the translation problem will be reduced. Evidence that such steps toward uniformity are taking place was observed at a recent workshop* where it was noted that many systems had recently adopted functions found in others and that there was an even greater concensus among system designers on what functions were desirable.

A question exists as to the desired structure of the common command language. We are inclined toward a simple command-name/argument string type

* Workshop on Comparative Analysis of Interactive Retrieval Systems, organized by Thomas H. Martin, Institute for Communication Research, Stanford University, April 23-25, 1973.

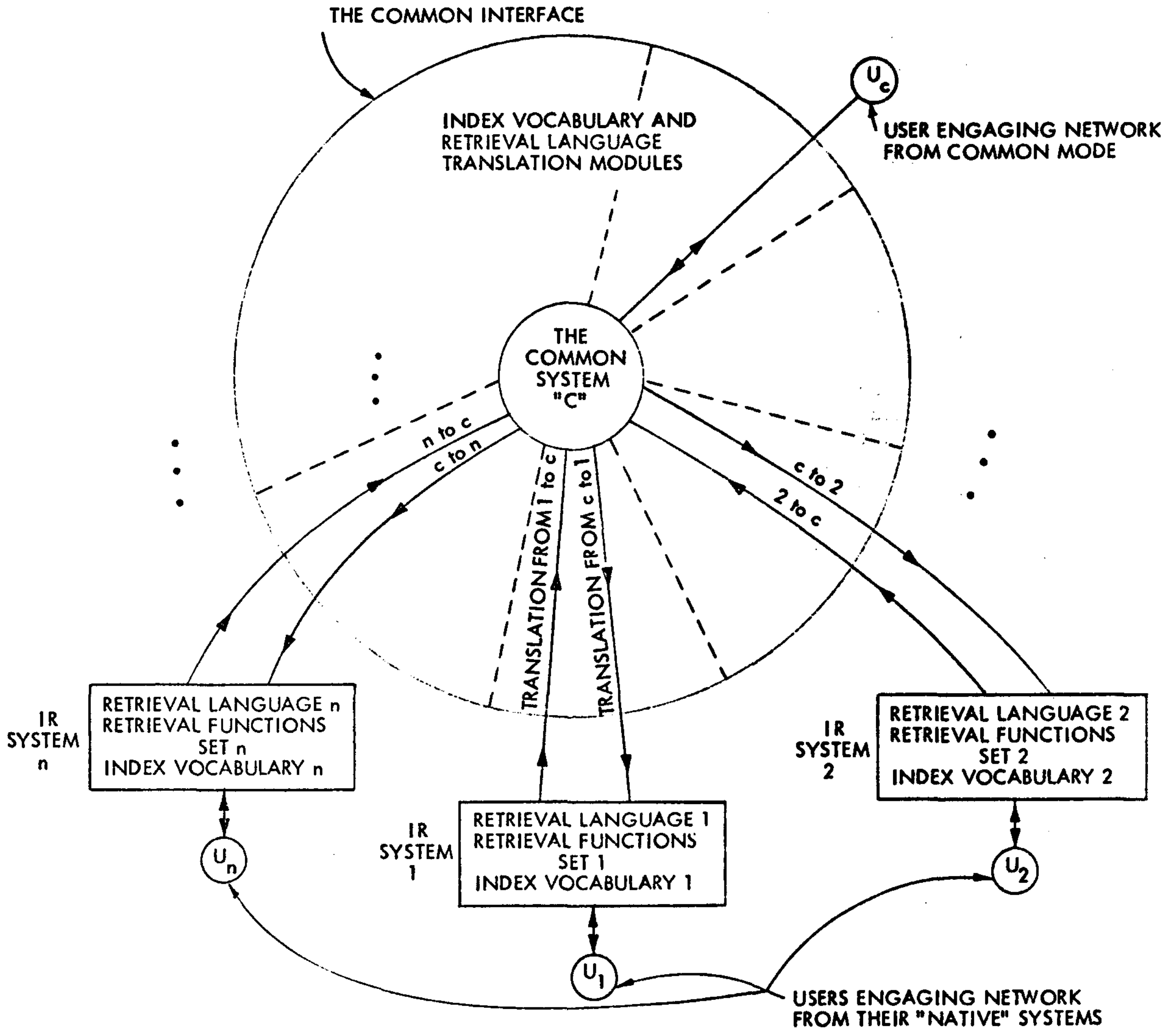

Fig. 1 The Logical Relations of an I-R Network Based on a Common Interface for Heterogeneous Systems

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 as a command language suffers from the twin defects of being of complicated structure and being very ambiguous; these defects make it hard to explain 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.

3. Indexing Vocabulary Conversion

Perhaps the most vexing problem confronting a prospective user of different data bases, even within a single system, is the different controlled-vocabulary techniques that may have been employed to index the subject matter of the documents. Our proposed mechanism for handling this problem, which has been dubbed the Master Index and Thesaurus, contains all the index and thesaurus elements of each of the data bases including a list of all the vocabulary terms used for indexing with the counts of the number of documents indexed by each and the thesaurus related terms for each. In addition, by the techniques of phrase decomposition (i.e., 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 inter-vocabulary relationships, even besides the obvious identity relationship. Thus, for example, the NASA term* electrical insulation is automatically found to be related to the following TEST** terms in the ways 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 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.) Searchers will be urged to express the subject matter of their request in ordinary English words, which are the basis of the common language index vocabulary. Search requests will be phrase decomposed and stemmed for matching against the expressions found in the Master Index and Thesaurus. If documents have been indexed by free vocabulary in the first place, matching reduces to intersecting those lists of documents with the appropriate word stems in their indexing. Project Intrex experiments have demonstrated*** that phrase decomposition and stemming of search requests and index expressions is an effective way to match requests and documents whatever indexing vocabularies are employed.

The index counts in the Master Index and Thesaurus can serve to help indentify terms that are appropriate to search under. Since these terms can come from different data bases, the counts may also serve to identify which data bases should be searched. The choice of which terms and data bases to search under can be performed automatically -- e.g., take all terms which match using phrase decomposition and stemming with the searcher's request in the synonymous or more specific relationships and that come from data bases that contain more than 10 documents with these terms. Alternatively, possible terms and data bases can be presented to the searcher for his final decision. As in other retrieval system considerations, our philosophy is to make the default condition --- i.e., where the user does not specify --- the automatic case but leave control with the user to be able to specify options to override the automatic default condition.

4. Data Elements and Structures

Another prime consideration in the development of means for users to interact conveniently with different data bases is the interrelation of the diverse data elements and structures from those data bases. Three ways in which the interrelations are important may be enumerated. In the first place, searching is done on one or more data elements: 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 (1) to identify when document references from different systems refer to the same document, (2) to establish common reference formats, and (3) to create common index (inverted file) and catalog data structures.

Our basic answer to this problem is to develop 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. 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 data element 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

* NASA Thesaurus Alphabetical Update, September, 1971.

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

***See, e.g., Project Intrex Semiannual Activity Report, March 15, 1972, Massachusetts Institute of Technology, pp. 20-41. PB 207 988.

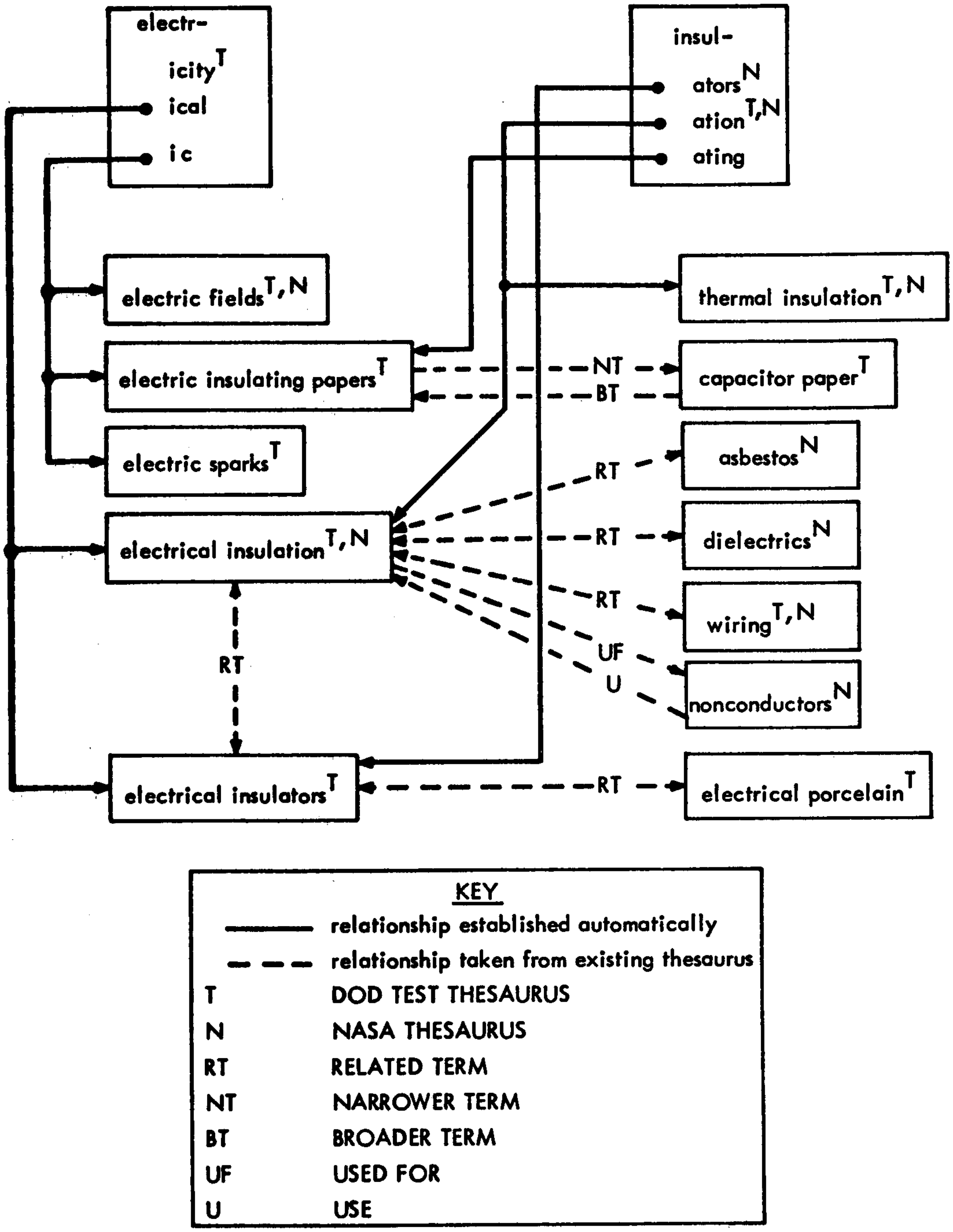

Fig. 2 Sample Relationship among Terms as Maintained in Master Index and Thesaurus

category, is shown in Fig. 3. There are 21 basic data elements identified in this category with 14 higher level hierarchical groupings. The other major categories which we have identified are Titles, Names and Relations, Related Documents, Library System Holdings and Shelving, Descriptive Document Features, and Control Fields. We have found it convenient to describe the common data structure in terms of the metalanguage for describing data structures as outlined by CODASYL.*

D. EXPERIMENTS UNDER WAY AND PLANNED

In order to test out the concepts detailed above in Section C we have undertaken an experimental program in which actual computer interfaces are to be constructed and used. Our initial interface system, which we call CONIT --- for Connector for Networked Information Transfer systems, is being constructed on the MIT MULTICS system based on the Honeywell 6180 computer. One advantage to using MULTICS is that it is a host system on the ARPANET computer network and provides certain advantages for network experimentation. In particular, the TIPs (Terminal Interface Processors) on the ARPANET allow for the simple kind of computer-to-computer interconnections which were described in Section C.1.

Our initial CONIT SYSTEM, called CONIT-1, interconnects the MEDLINE medical information retrieval system based on the National Library of Medicine IBM 370/155 computer, the M.I.T. Intrex system on an IBM 370/165, and the several systems that are accessible through TYMNET (the TYMSHARE Corporation Computer network used by many of the operational information retrieval systems to provide easier long-distance access from terminals). As Fig. 4 shows, connection to the remote systems are accomplished via ARPANET TIPs and interactive access to the systems is possible from anywhere with access to the ARPA network. A detailed description of the mechanisms by which these interconnections are achieved is given in a report by Therrien.**

The CONIT-1 user has a couple of basic commands by which to interconnect to the retrieval systems, including which system to search and what language to use, and a few others to perform some basic retrieval functions including simple versions of a search and output (print) commands. In addition, the user can save output from different systems in common files designated by the user. Another command will search the Master Index and Thesaurus and a final one will allow the user to rename commands and arguments and, more generally, to define macros combining commands and/or (variable) arguments. Although initially, the command language must be the simplified CONIT-1 language or the language of the system being searched --- i.e., a transparent mode ---, the macro capability will provide either user or system designer a flexible device for experimenting with various translating schemes and, indeed, testing what macro commands are most effective at the common language level.

Gradually, we expect CONIT will be extended to include more retrieval functions at the interface and to experiment with more effective intercomputer communications, including the use of ARPANET and TYMNET.***

E. CONCLUSIONS

Research on the coupling of interactive information systems has been described. 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 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.

* CODASYL Systems Committee, Feature Analysis of Generalized Data Base Management Systems, Technical Report, May, 1971. Association of Computing Machinery, New York, N.Y.

** Charles W. Therrien, Data Communications for an Experimental Information-Retrieval Network Interface, M.I.T. Electronic Systems Laboratory Technical Memorandum ESL-TM-515, August, 1973.

***For a recent report on project details see J. Francis Reintjes and Richard S. Marcus, Interim Report on Research in the Coupling of Interactive Information Systems, M.I.T. Electronic Systems Laboratory Report ESL-IR-533, February 15, 1974.

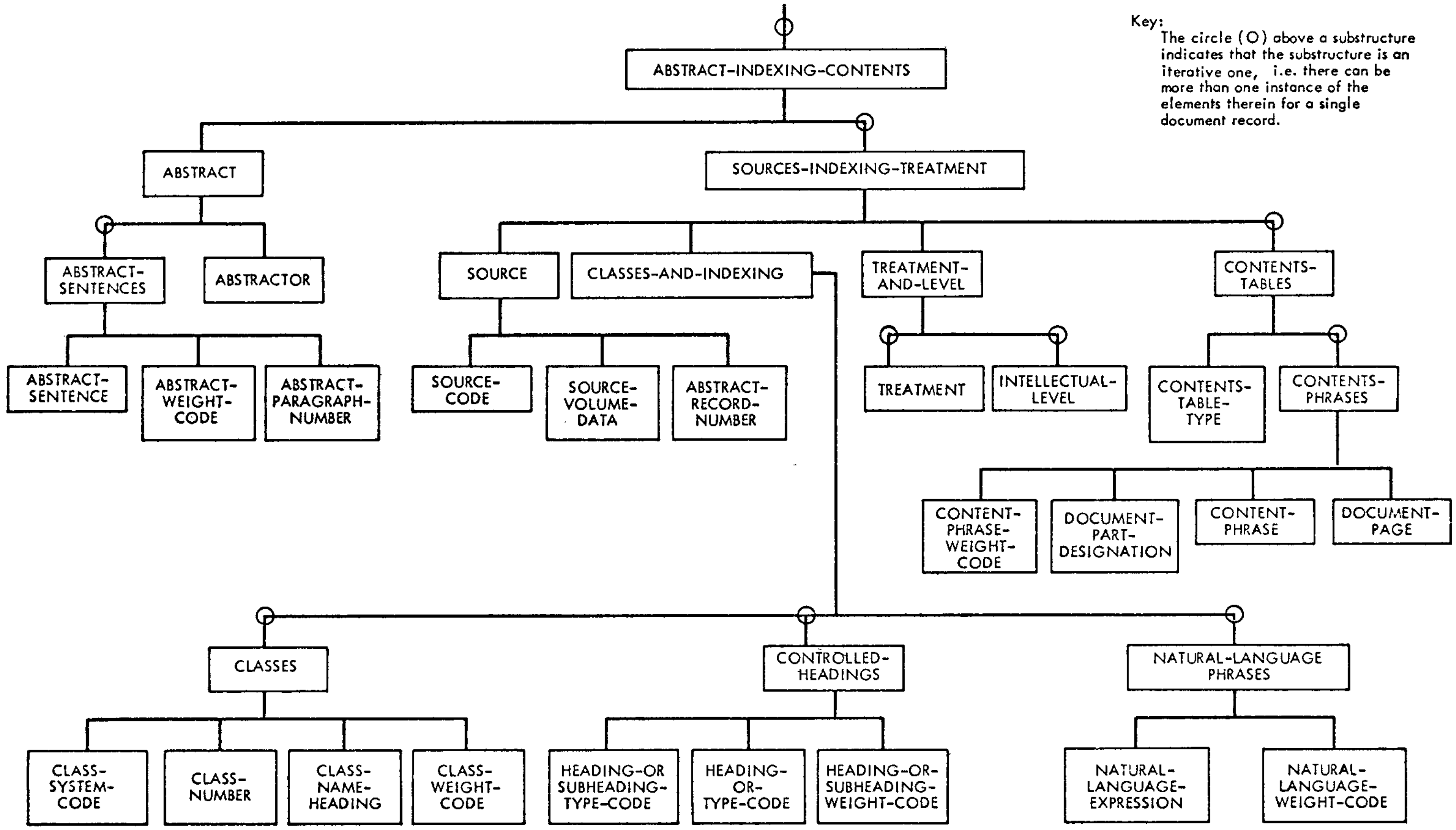
Key:
The circle (O) above a substructure indicates that the substructure is an iterative one, i.e. there can be more than one instance of the elements therein for a single document record.
ABSTRACT-INDEXING-CONTENTS
ABSTRACT
SOURCES-INDEXING-TREATMENT
ABSTRACT-SENTENCES
ABSTRACTOR
SOURCE
CLASSES-AND-INDEXING
TREATMENT-AND-LEVEL
CONTENTS-TABLES
ABSTRACT-SENTENCE
ABSTRACT-WEIGHT-CODE
ABSTRACT-PARAGRAPH-NUMBER
SOURCE-CODE
SOURCE-VOLUME-DATA
ABSTRACT-RECORD-NUMBER
TREATMENT
INTELLECTUAL-LEVEL
CONTENTS-TABLE-TYPE
CONTENTS-PHRASES
CONTENT-PHRASE-WEIGHT-CODE
DOCUMENT-PART-DESIGNATION
CONTENT-PHRASE
DOCUMENT-PAGE
CLASSES
CONTROLLED-HEADINGS
NATURAL-LANGUAGE PHRASES
CLASS-SYSTEM-CODE
CLASS-NUMBER
CLASS-NAME-HEADING
CLASS-WEIGHT-CODE
HEADING-OR SUBHEADING-TYPE-CODE
HEADING-OR-TYPE-CODE
HEADING-OR-SUBHEADING-WEIGHT-CODE
NATURAL-LANGUAGE-EXPRESSION
NATURAL-LANGUAGE-WEIGHT-CODE

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

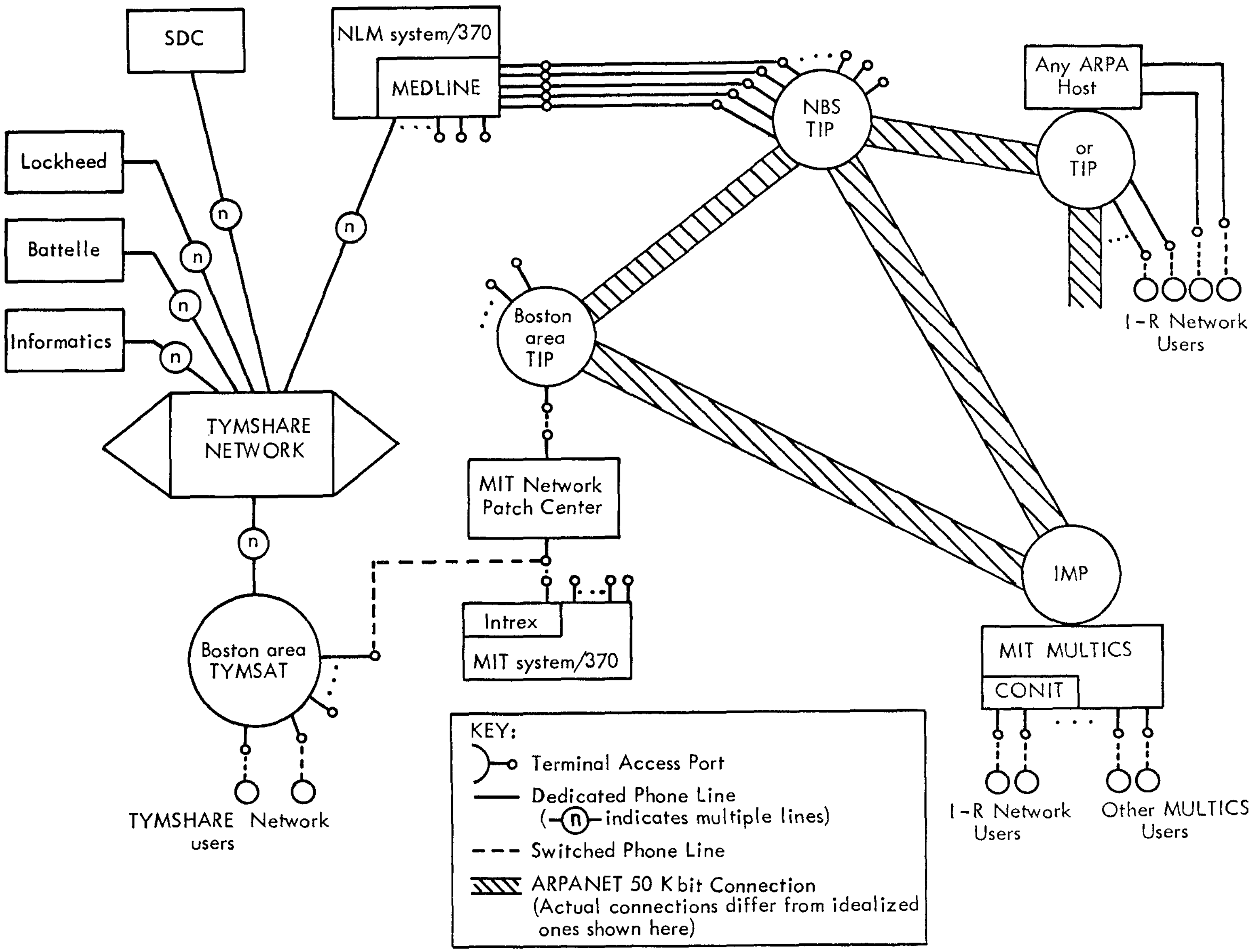

Fig. 4 Computer Interconnections for CONIT-1 Interface for Networked Information Retrieval Systems

QUESTIONS

Richard E. Nance:

What kind of language are you using for the translation from one system language to the common language?

Marcus:

As I mentioned our emphasis at least initially, is working from the common language to the other systems languages. At present we have a translator that searches for a particular bit string and makes a translation from that bit string to a function or series of functions represented from that bit string. It is not clear just how complicated a translator is needed for this purpose.

Lawrence L. Rose:

Is the translation language, the micolanguage, being set up so that someone else can control it at a later point in time?

Marcus:

It is being set up so that even the user could exercise control. The user may even have his own dialect for any system.

Tom Kibler:

You mentioned that you were using primitives. Should I assume that you are going from the query language to a set of primitives then to the common language or some other form of translation? Are you attempting to go from the macrolanguage directly to the common language?

Marcus:

We are experimenting with both. The paper illustrates translation from one macro to another. We have hopes of doing a more precise translation, going back to the primitives. There is a growing trend in systems these days to look similar. Perhaps this similarity is more conducive to an exact translation. We do use a common interface language; however, there is still the question of whether you attempt to go through the primitives or directly into the macrolanguage translation.

Frank Manola:

Could you give an illustration of how the different file structures might cause problems? Simply the inversion process being different from one system to another could cause problems. We could attempt to approximate a query from one system in the language of another but the question persists as to how well we can do this.

Gerald Salton:

You are not really doing this, are you? Isn't this somewhat of a fiction?
[This refers to one particular figure shown by Richard Marcus.]

Marcus:

This is a semi-fiction. We are currently working on that particular part; where the user engages another system through the common language.

Gerard Salton:

I could imagine that with specific commands the problems would not be too great. But in indexing vocabularies, the problem would be tremendous. I have been involved in such attempts.

Marcus:

We wholly agree that it is impossible to define a common source that brings together the different dictionaries, and this is not what we are doing. We really do two things: (1) We give information about the thesauri that have been used in the different vocabularies involved and that is useful in searching through systems using those vocabularies and (2) we try to use the natural language as the common part of the indexing language.

Jack Belzer:

Gerry's comment is well taken. When you begin to combine thesauri you get either inconsistency or ambiguity.

Marcus:

To reiterate the answer given to the last question, we are not saying that we want to create a common control vocabulary. What we are saying is that perhaps we should eliminate control vocabularies. In the meantime, however, we must find a way of translating from the natural language expression into the control vocabulary version.

Tom Kibler:

Cannot these thesaurus connections lead to the infinite expansion of search terms?

Marcus:

Absolutely. If you take every possible connection, soon you will have the entire data base. The "trick" is to be judicious about the level to which you go. It is very difficult to devise automatic algorithms to make these kinds of decisions, and that is why we are placing a great deal of emphasis on the intelligence of the user.

Tom Martin:

Would it not be proper to make the particular use of the data base dependent on the type of user? Should you not allow the search strategy to be dependent on the user and the data base combination?

Marcus:

Yes, and I think that is what we are doing.

Richard E. Nance:

One very detailed question. In your example, using the word "electrical" the stem that you chose had two characters in common with all the suffixes that remain. I could see how you would operate on a particular thesaurus where all the stems were a super set of the stems of your common thesaurus. Have you had to face the question where stems of one thesaurus might be a subset of your common thesaurus?

Marcus:

It sounds like a good question. We have done some study of a particular stemming algorithm, and these results are reported in a J.ASIS paper. You do not always get the particular stems you might like, and that is why you need to have user control over what is happening.