Center for Advanced Computation

UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN
URBANA, ILLINOIS 61801

CAC Document No. 213

RECOMMENDATION OF COMPUTER SYSTEMS FOR OPERATION OF THE HABITABILITY DATA BASE

by

Karl C. Kelley
and
James A. Gast

September 30, 1976

CAC DOCUMENT NO. 213

RECOMMENDATION OF COMPUTER SYSTEMS FOR OPERATION OF THE HABITABILITY DATA BASE

Prepared for the

U. S. Army
Construction Engineering Research Laboratory

under Contract
DACA88-76-M-0291

by

Karl C. Kelley
and
James A. Gast

Center for Advanced Computation
University of Illinois
Urbana, Illinois 61801

September 30, 1976

Table of Contents

RECOMMENDATION OF COMPUTER SYSTEMS FOR OPERATION OF THE HABITABILITY DATA BASE

1. Summary of Recommendations

The results of this investigation suggest a number of courses of action which will be in the best interests of the Construction Engineering Research Laboratory and their implementation of the Habitability Data Base. We divide the recommendations into short and long term based on the amount of time and effort which will be involved to achieve a measure of satisfaction of the goals of the HDB effort.

1.1 Short Term Recommendations

a) Reactivate the HDB implementation at the Computer Services Office of the University of Illinois. This holding action will make something available until mid-1977, while work is underway to provide a longer-term solution.

b) Implement the HDB files and versions of the AND, OR, DOCAX, and BIBAX programs on the Michigan Time Sharing system.

c) Implement SMART on MTS, but do the rest of the tasks on the UNIX system at CAC.

1.2 Long Term Recommendations

a) Expand the CELDS work at CAC into a form suitable for both the Environmental IAC and the Habitability IAC. Pool efforts with the CELDS group to acquire appropriate hardware to be used for both systems.

b) Convert the HDB to a form suitable for the Lockheed "DIALOG" system and use it remotely through a Telenet port in Chicago.

2. Foreword

The recommendations of this report are based on a short term investigation which precluded hands-on trial of most of the systems. The reader is cautioned not to base a long-term expensive information system implementation solely on the results of examining the features of systems which are available. A more advisable course of action would be to take steps to contact vendors of the suggested commercial systems directly to discuss the suitability of their system for the HDB. It is also advisable to use their system on existing data bases to gain a feeling for the kind of response times typically available, the ease with which a search can be made, and the convenience of gaining access to the system.

It is not possible to determine the real cost, speed, or reliability of a system just by looking at the price sheets and the vendor's description of the system. Surveys and literature searches can only guide one in deciding which systems deserve a closer look. They should never be used in the place of hands on experience or a carefully documented benchmark to determine the choice of a system which will be used over an extended period of time.

No claim is made that all possible systems have been included in our considerations. Work currently underway at various research centers may not be in the current literature. It is suggested that CERL may want to avail themselves of some of

the automated information systems accessible from local sources to make their own search of the literature to obtain bibliographies relating to automated information retrieval. Some of the sources listed in the bibliography of this report themselves contain extensive bibliographies which probably should be examined in greater detail.

3. Approach

The search for an appropriate recommendation included a literature survey and an examination of locally known systems which have the potential to meet the needs of the HDB. The literature survey included data base management systems and large-scale interactive information systems. The similarity of the HDB project to another project currently under investigation at the Center for Advanced Computation led to a more thorough comparison of the needs and features of the two systems to determine whether a recommendation to combine the two efforts was warranted.

The initial literature survey was directed at data base management systems, since they would form the backbone of any information retrieval system which involved managing a new set of data (the HDB). The essential needs of the HDB are a data base management system, some type of interactive editing capability, and the ability to recreate the interactive programs which provide the AND and OR functions available in the earlier version of the HDB. The ability to do the AND/OR functions are taken as

given in all systems considered. This task is simply not sufficiently complex to be a meaningful consideration in any recommendation. Furthermore, since CERL is usually restricted to using time sharing services, it was thought that the choices of computer hardware were restricted to an IBM 360/370 system, a large CDC system known to be available to CERL, or possibly one of the larger DEC-10 systems, none of which is reknowned for its built-in capability to manage files effectively. For any of these, a sophisticated data base management system would be desirable as a prerequisite to implementing an HDB. The requirement for managing a large data base was thus taken as a controlling factor and led to the consideration of data base management systems.

Further experience with the existing HDB documentation and discussions with the technical monitor and programmers involved in the implementation of the SMART system reinforced the notion that the HDB is not conceptually different from a bibliographic system, either in the way that data is stored or in the kind of programs which would be required to respond to the customer's query. The HDB is essentially a collection of statements, each having been indexed in a special way, and each referring by some means to the original document from which it was taken. This is conceptually similar to a collection of abstracts, indexed by key words or other searchable fields, and each pointing or referring to the original bibliographic citation. Section 4 discusses the

needs of the HDB as an information retrieval system.

This conceptual similarity widened the literature search to include a closer look at the very specialized type of data base management systems with query languages which can be called bibliographic systems or interactive information systems. Through a series of literature sources we were led to a survey done by the National Bureau of Standards in 1973 (published in 1974). This source constitutes a reference to the technical features and operational status of interactive information systems, that is, those providing a 'conversational' usage mode to a 'non-programmer' through a data terminal. In addition to technical information about some 46 systems, it provides guidance in the use of the index to narrow the field of choices in selecting an interactive information system for a particular application. Section 5 discusses the approach suggested by this reference and goes through a first [illegible] selection of systems which meet the needs of the HDB.

Section 6 discusses the appropriateness of SMART as a choice for implementation of the HDB, as well as the implications of continuing with SMART for the next phase of HDB development. Section 7 discusses some of the issues inherent in abandoning the SMART program and replacing it with some other system, whether developed anew or adapted from existing systems.

4. Needs of the Habitability Data Base

We separate the notions of conceptual needs of the HDB from

the current needs in the following way: Conceptual needs are based on the problem itself, that is, the problem of storing the HDB and retrieving the information stored therein on the basis of user requests. Conceptual needs reflect the end user of the HDB. Current needs, on the other hand, deal with the more practical immediate concerns of the HDB effort, making the service available to the current set of users in a cost effective manner as quickly as possible. The current need seems to be primarily for a system which will run the ANDOR programs and allow SMART requests to be submitted in a batch mode. We include in current needs any system which could do the user's end function as effectively as SMART, without extensive reprogramming or reformatting of the data base.

4.1 Conceptual Needs of the Habitability Data base

Conceptually, the HDB consists of a set of statements drawn from an appropriate literature. These statements are formulated by trained specialists who not only condense the information in the literature, but also classify the information by indexing the statement. This process is conceptually equivalent to abstracting a document and providing an index classification of the document. Whereas most bibliographic retrieval systems keep the information about the document in the same record as the abstract (and possibly key words in addition to author, title, etc), in the HDB the only information directly linking the entry in HDB with the original source of the information is a document number encoded

as part of the sequence number field of the HDB statements. The index of the document is a multidigit string of codes which is prepended to the first card image of a particular statement.

The user of the HDB wishes to formulate a simple request to retrieve information which is of immediate concern to him. With the HDB as originally designed, this request is stated in terms of the classification of the statement as represented by the index. This index is comprised of 10 coded values: [5]

```
FUNC...a three digit functional area code
TRFC...a 5 digit training facility code
PHYS...a 1 digit physical setting code
AENV...a 2 digit "A" environmental descriptor
BENV...a 2 digit "B" environment descriptor
OCCU...a 1 digit occupant code
PSTR...a 1 digit code for posture of people
INVM...a 1 digit code for involvement of people
ORGF...a 1 digit code for organizational functions
SFCN...a 1 digit code for function of the statement
```

A more complete description of the classification and indexing scheme is given in [1].

The primary programs for selecting statements interactively on the basis of the indexes are the AND and OR programs which run interactively on the DEC-10 system as part of the Prototype HDB [4][5][3]. These programs use a rather forced dialog to input the appropriate fields which are to be searched on and the values to be searched for. The interactive response is a set of statements, along with the appropriate document number and the number of this statement with respect to the source document. There is no capability to get a count of documents which meet one

criteria or set of criteria and then determining whether that set should be further limited by ANDing with another set. The entire request is made at the outset, and the entire set of documents which match the request is printed as output. There seems to be no capability of saving the numbers of these documents for later refinement by further search requests.

Two other programs exist in the Prototype HDB system which reflect both conceptual and current needs. The function of the programs is to allow the user to see the bibliographic citation of a document if he knows the document number and to see the text of the document if he knows the number. Note that ANDOR returns the statement and the number of the document. (The document is not really there, it is just the collection of all statements which came from that document)

This technique of finding statements is perhaps appropriate to some potential users of the HDB. In particular, the person who wishes to write a criteria manual for design of a certain training facility might want to retrieve what is available and related to that kind of facility. However, information specialists responding to a submitted query, and to some extent the end customer himself, might find that a better way of expressing the inquiry and conducting the search is needed. The HDB does not contain keywords which can be used to characterize content of statements. (In its present form, content is only characterized by the index digit string). Thus a retrieval

system based on full-text search of the statements, preferably with natural language input, is a second conceptual need of the HDB.

At the present time this need is met by the collection of programs known as the SMART system. This system operates in batch mode on the IBM 360 system. It has been implemented at the Computer Services Office (University of Illinois at Urbana) as part of the prototype HDB effort. "The system takes documents and search requests in English, performs a fully automatic content analysis of the texts, matches analyzed documents with analyzed search requests, and retrieves those stored items believed to be most similar to the queries. Among the language analysis procedures incorporated into the system are word suffix cutoff methods, thesaurus lookup procedures, phrase generation methods, statistical term associations, syntactic analysis, hierarchical term expansion, and others."[6]

As a part of the SMART user interface for the prototype HDB, a program on the DEC-10 computer accepts input of query submittals and formulates batch jobs for the 360. These jobs are submitted across a link to the batch machine. The user returns later to see if his job is done and retieves his output (responses to his query) by running another program on the DEC-10. The time lag between request and response has not been satisfactory with the present implementation.

What is needed and is missing in the current implementation is an interactive on-line version of SMART. Salton recognized this as a need [6]. To run SMART interactively would require a different operating system on the 360, namely one that allows for time-shared user interactive terminals. There have been no major updates of SMART since the library was obtained from Cornell for the prototype HDB. At latest report, no interactive version of SMART is available in release form, although some effort was expended at Cornell in implementing an interactive version under the IBM TSO operating system. Even if that were successful, it would be of little value to any solution which proposes using the 360 at CSO, since that system will stay batch until its eventual retirement.

If the conceptual need for natural language processing of a query is artificial, some of the systems to be mentioned in Section 5 would probably well serve the needs of the HDB.

4.2 Current Needs of the Habitability Data Base

The current need of the HDB is a system which provides for the conceptual needs outlined above as well as the more immediate concerns of finding an appropriate operating system and computer to run it on. The scope of work for this contract lists five definitions of the needs of the HDB. Two of these fall into the class of conceptual needs:

1) the types of programs currently in use must be available

2) the system has the capability of handling summary data as well as bibliographic and textual data

These have been examined in the section on conceptual needs. The other needs are current needs discussed in this section.

The text-editing capability is desirable so that corrections and changes can be made to the HDB statements, and so that new statements can be added as the collection grows. Any system which will be capable of the interactive access required for the AND/OR programs will, without exception, have text-editing capability. So long as the HDB statements are part of a non-specific text file, they be accessible and editable with the editors on most systems.

However, the capability to edit HDB statements which are already included in a data base which has undergone some degree of inversion might be somewhat of a problem. The typical retrieval system requires that the data and the fields which will be searched be made ready for a large inversion process which is run against the data base to get it properly organized for faster retrieval. In some organizations this data base inversion process is very time consuming. The capability to access the statements independent of indexes to the statements is thus a requirement for on-line correction to the HDB statements. Similarly, in order to keep the data base updated, it should be possible to input new statements in text form. This is not a

problem. However, it is quite likely that before new statements can be used as an integral part of the HDB, the inversion process must be run again. This would restrict updating to periodic updates of perhaps once a month. This is the norm rather than the exception in data base systems of the capability described.

Commercially available systems will automatically be able to take care of control and billing of outside users (they make a living doing it). University computer centers sometimes have more difficulty with this in that their process for establishing user accounts is sometimes rather cumbersome. However, the systems under consideration and outlined in the accompanying recommendations all meet the criterion that outside users can be admitted to the system and billed directly. Similarly, the capability for remote low-speed access from terminals should be taken as given in all of the systems under discussion here. The only systems for which this is not the case are systems for which access is restricted to remote batch, and such a system cannot meet the editing and interactive requirements. Where necessary, submission to batch systems should be accomplished via an interactive system, similar to the technique used between the DEC-10 and the 360 in the prototype HDB work. This should always, however, be considered as clumsy and not conducive to the kind of immediate feedback to be obtained with interactive

systems such as those commercially available.

5. NBS Survey of Interactive Information Systems

The National Bureau of Standards has already anticipated the need for government agencies to consider the choice of an interactive information system. A report published in 1974 constitutes a reference to the technical features and operational status of such systems available at the time [2]. From the introduction to that report:

> "This report is written for the purpose of providing Federal ADP customers with information on a certain class of computer systems which are capable of handling scientific and technical information. The report attempts to show what is available and to characterize these systems in such a way as to answer questions which naturally arise prior to selecting such a system for a particular installation. The report is written at a level of technical detail which is aimed at information specialists rather than programming experts. It is intended to be informative and instructive, and not critical or evaluative."

> "We have reviewed for inclusion in this index over 200 systems which came to our attention from various published and unpublished sources as well as from word-of-mouth. The systems which were selected conform to the following definition: "Information Retrieval" or "Data Management" packages or services which are available to any Federal ADP installation, and which offer an interactive query and search capability that is geared for use by non-programmers."

They eliminated from consideration systems which: 1) are batch systems, 2) have query languages not for use by non-programmers, 3) are in research or development, 4) a.e no longer supported, 5) are no longer in business or locatable, 6) are

subject to legal or security problems in the way of releasing the system, or 7) were not documented.

It seems at least strongly suggestive that these systems meet the basic needs of the US Army CERL, if one of them meets the specific conceptual needs of the HDB.

The intent of this section is to examine the organization of that report and to frame current concerns in terms of the selection criteria outlined therein. Table 1 is a list of the systems which met the criteria for inclusion in this survey. Table 2 is the questionnaire which was used to characterize the features of the various systems. Included in the report is a summary of the features of each of the examined systems, listed in a manner similar to the format of the questionnaire.

However, before examining each of the systems reported, the suggestion is made that the needs of potential users of the system be classified in order to make a first cut at system selection. Their recommendation for a first elimination is based upon potential usage and estimated cost first, then on the availability of a given main-frame, and in the case of a requirement for a specific data base, on the availability of that data base as a service. In the case of the HDB investigation, several choices of main-frame are available, and it has not been determined whether a package should be put up on one of these mainframes or a service bureau should be used. Since a decision can be made on these choices at a later time, we can proceed

Name
BASIS
CDMS
CIRCOL
(Data/Central)
DIALOG
DMARS
DML
DRS
DS/3
EMISARI
ENFORM
FLEXIMIS
GIM
GIPSY
IMARS
IMS(OEP)
IMS/360
IMS/8
INQUIRE
INSYTE
LEADERMART
MARK IV
MARS III

Name
MARS VI
MASTER CONTROL
MICROTEXT
MINIDATA
MIRADS
MUSE
NASIS
N.Y.TIMES
OLIVER
ORBIT III
PIRETS
QUERY UPDATE
RAMIS
RECON
RFI
RIQS
SHOEBOX
SOLAR
SPIRES II
STAIRS
SYSTEM 2000
TICON
UNIDATA

TABLE 1. SYSTEMS INCLUDED IN THE NBS SURVEY

TABLE 2. QUESTIONNAIRE FROM THE NBS SURVEY

A. GENERAL DESCRIPTION	
1. SYSTEM NAME	Additional names are given for alternative versions. System originator is developer and implementor.
ORIGINATOR TELEPHONE	
2. SOFTWARE AVAILABILITY	Filled in if the computer program package itself is offered for customer installation.
a. FOR PURCHASE AT WHAT COST b. FOR LEASE AT WHAT COST	
3. SERVICE AVAILABILITY	Filled in if remote service is offered.
a. SEARCH SERVICE AT WHAT COST b. DATA BASES c. TAKES CUSTOMER DATA BASES	
4. HISTORY OF SOFTWARE	Filled in where the package itself is offered.
a. FIRST INSTALLATION b. SIGNIFICANT INSTALLATIONS	
5. HISTORY OF SERVICE	Filled in where service is offered.
a. SERVICE INITIATION b. PRESENT USAGE	
B. COMPUTER ENVIRONMENT	
1. MAIN FRAME	
2. OPERATING SYSTEM	
3. SOURCE LANGUAGE	Used to implement the software.
4. TERMINAL TYPES	
5. TRANSMISSION RATES	
6. RE-ENTRANT FOR MULTIPLE USERS	Identifies systems which can serve multiple users simultaneously through re-entrant code, thus conserving storage.
C. FILE DEFINITION	
1. USER DEFINABLE	Can the remote user define and implement his own data and file structures?
2. VARIABLE LENGTH TEXT	Do individual records accomodate text of any length, or at least as much as a typical bibliographic abstract?
3. REPEATED FIELDS	Is it possible to have a variable number of identically named fields in a record, e.g., keywords or authors?
D. FILE MAINTENANCE	
1. ALLOWS ON-LINE a. CREATION b. UPDATE	"On-line updating" refers to change of the logical content of a record, while "on-line editing" refers to an associated capability to selectively indicate partial changes of the record content.
c. DELETION d. EDITING	
2. PREPROGRAMMED DATA VALIDATION CHECKS	Are there functions to check the "correctness" of the incoming data?
3. OPTIONAL ORDERING OF ENTERED DATA	Can incoming data elements and records be optionally ordered with respect to pre-existing elements and records in the file?
4. AUTOMATIC OR MACHINE-AIDED CONTENT INDEXING	Is there a preprogrammed automatic or machine-aided process to scan raw text and develop concept-indicating keywords and phrases?

TABLE 2., CONTINUED

E. QUERY

1. USER-SYSTEM INTERACTION	
a. MULTIPLE OPTIONS AT ANY POINT	During a session with the system, does the user have the freedom to specify various actions at any time? For example, after entering a search term, is the user returned to a state where he may choose more terms, review previous choices, or execute the search?
b. ENGLISH-LIKE PHRASING	Do the system commands have English or near-English names indicating their function?
c. SYSTEM-FORCED DIALOGUE	Is the user corpletely "led along" by the system in a dialog completely controlled by the system?
2. CONTENT SEARCHING	
a. CONTROLLED VOCABULARY	Is the user required to use a pre-established set of search terms?
b. FULL TEXT INVERSION	Can the user use any word expected in text (excepting perhaps an excluded list of stop words like "a", "of", "the"?)
c. STEMMING PERMITTED	Can the user specify a set of search terms by using a root expression such as "comput*", where * has all the values "e", "er", "ers", etc.
d. SYNONYMS	Does the system have the capability of specifying a class of semantically equivalent terms either in its vocabulary or else by definition?
e. DISPLAY OF RELATED TERMS	Can the system display terms from its vocabulary which are equivalent to a given term or are more or less specific, but related to it?
f. CHECKING TERM IN CONTROLLED VOCABULARY	Is there a command which checks a given input term for inclusion in a controlled vocabulary and perhaps gives immediate alphabetic neighbors?
3. SEARCH SPECIFICATION	
a. MUST NAME FIELDS	Is it necessary to always specify fields in query formulation, e.g., "Author=Freud"?
b. MAY LIMIT FIELDS	Is it possible to specify fields in query formulation to limit or control the search?
c. BOOLEAN AND	Is there an explicit capability for taking two valid search expressions and connecting them with AND or OR so as to construct a new expression with the usual Boolean interpretation?
d. BOOLEAN OR	
e. BOOLEAN NOT	Is there an explicit capability for negating a valid search expression by preceding it with NOT?
f. NESTING OF BOOLEAN EXPRESSIONS	Is there an explicit capability for embedding search expressions within other search expressions?
g. PHRASE AND DISTANCE SEARCHING	Is it possible to use multi-word strings as search terms? Is it possible to use a query which specifies that two terms must occur within some stated distance in the text from each other?
h. NATURAL ENGLISH	Can the user phrase his search objective in natural English sentences and phrases?
i. RANGE SEARCHING	Is there a capability for specifying a range of numeric values in a search expression, e.g., "...published since 1970"?
4. TUTORIAL FEATURES	
a. "HELP" COMMAND	Is there a command which gives assistance to a user on what options are available to him at various points in a session?
b. DOCUMENTATION ON-LINE	Can the user get at documentation on-line which explains the system, the data bases, etc.?

TABLE 2., CONTINUED

5. SEARCH STRATEGY	Refering to the file accessing technique used in the implementation, is an individual record uniquely located: a. from an index, b. by consecutive search of each record in the file, c. within a group of records?
a. RANDOM b. SEQUENTIAL c. INDEXED SEQUENTIAL	
F. REPORT GENERATION	
1. LANGUAGE TYPE	How are alternative output formats determined?
a. STANDARD OUTPUT b. SELECT AMONG OPTIONS c. OWN CODE	Can the user write a program to specify his output format?
2. MEDIA FLEXIBILITY a. OFF-LINE PRINTING b. DISPLAY OF GRAPHS	Do programs exist to display simple charts and graphs?
c. SPECIAL OUTPUTS	Are there special output forms (e.g., microfilm, etc,) available?
3. SPECIAL CAPABILITIES a. SORTING	Can output be selectively sorted on chosen fields?
b. SEQUENCING	Refers to a capability for sorting or "ranking" by relevance to a query.
c. COUNTING	Of number of hits.
d. ARITHMETIC	On field values.
G. SECURITY PROTECTION	
1. TERMINAL	Refers to the use of passwords or other identifiers to inhibit use of a terminal or access to data elements by non-qualified users.
2. DATA BASE 3. RECORD 4. FIELD	

immediately to elimination of unsuitable choices based on the technical features.

The NBS report suggests drawing distinctions in three broad classes of system applications: formatted data processing, structured text searching, and personal text handling. The needs of the HDB fall into the class of structured text processing. In the following excerpt from the NBS report, the items in parentheses refer to the characteristic features listed in the questionnaire.

> Structured text searching is conceived as representative of bibliographic information searching, legal text searching, and similar uses where the file records consist of prescribed segments of text, (1000 characters or more). Examples of text segments would be report titles, abstracts, patent claims, paragraphs, statute sections. To identify a text record for selection there must be a technique for abbreviated content description, since requiring an exact match to all the text in a segment would be inconceivable and inconsistent with the intended function. Content description may be provided by indexing each file record by a set of keywords from a controlled vocabulary (E.2.a) which the user can inspect to check his desired term for acceptability (E.2.f). Or else, any significant word occurring in text may be provided as a valid search term (E.2.b). Because these systems are specially aimed at users unaccustomed to programming encoded forms, English-like phrasing is deemed essential as well. Because a search may select voluminous text records that would be exceedingly long to print on the usual 10 or 30 character/second terminals, off-line printing (F.2.a) at high speed is also essential. Moreover, these systems should present a count (F.3.c) of the records that would be selected by a proposed search so that a user can judge the desirability of continuing the search.

> The data files of structured text searching systems would be expected to be unchanging in content and very large in volume. It would be expensive to reorder or restructure them as new data is received, so it would be desirable for the system to accept new data in any order (D.3). Other desirable features would extend content searching capability, for example by giving a synonym facility (E.2.d) or a presentation of other terms that are conceptually related (E.2.e). As in formatted data processing, tutorial aid is desirable. In contrast to that application however, full Boolean capability, optional report formatting, and optional ordering are suggested here as desirable rather than essential. Only a Boolean AND, allowing the conjunction of distinct search terms, is imperative for user convenience, to avoid a tedious selection from record subsets found by individual terms. Optional formatting and ordering may not be used often for such simple structured output records as bibliographic citations. A standard output presentation then is generally sufficient, unless text fields become numerous and frequently of marginal importance, requiring more selectivitty to be given the user.

The chart from the NBS report for categorizing systems is reproduced here as Table 3. Figure 1 shows just the entries which have an x in the feature row corresponding to structured text processing systems. This figure shows in a compact format the choices which on the face of it would be suitable for the HDB application. Those systems marked with a "+" are listed specifically as allowing customer data bases to be added to a "service" system. Also, two systems are included on this table which are not mentioned in the NBS report. These are the CELDS system and the EUREKA system currently in some stage of development at the Urbana campus of the University of Illinois. The SMART system is also indicated on this chart, though it does

1. Applications

System / Category	BASIS	CDMS	CIROOL	(Data/Central)	DIALOG	DMARS	DML	DRS	DS/3	EMISARI	ENFORM	FLEXIMIS	GIM	GIPSY	IMARS	IMS(OEP)	IMS/360	IMS/8	INQUIRE	INSYTE	LEADERMART	MARK IV	MARS III	MARS VI	MASTER CONTROL	MICROTEXT	MINIDATA	MIRADS	MUSE	NASIS	N.Y.TIMES	OLIVER	ORBIT III	PIRETS	QUERY UPDATE	RAMIS	RECON	RFI	RIQS	SHOEBOX	SOLAR	SPIRES II	STAIRS	SYSTEM 2000	TICON	UNIDATA
Formatted Data Processing	X	X				X	X	X	X			X	X	X	X	X	X	X	X			X	X	X	X		X	X	X	X		X			X	X			X					X	X	X
Structured Text Searching	X		X	X	X									X		X			X		X	X			X	X		X		X	X		X	X			X	X	X	X	X	X		X	
Personal Text Handling	X			X												X			X						X	X		X		X			X	X					X			X		X	

2. Main Frame Series

CDC 6000	X										X										X			X	X		X								X				X					X		X
DEC PDP-10						X									X																	X		X												
IBM 360/370		X	X	X	X			X	X				X	X			X		X			X				X				X	X		X			X	X	X		X	X	X	X	X	X	
UNIVAC 1100	X						X			X			X			X		X									X																X			
Other								X				X								X			X		X				X																	

3. Type of Offering

Service	X	X	X	X	X	X	X	X				X		X	X	X					X				X	X	X		X	X	X	X					X	X	X	X				
Package	X	X		X	X			X	X	X	X		X	X			X	X	X	X	X	X	X	X	X	X	X	X	X			X	X	X	X	X	X	X	X	X	X	X	X	X	X	X

4. Ownership

Public										X				X		X					X				X	X		X		X				X			X	X	X	X	X	X					
Proprietary	X	X	X	X	X	X	X	X	X		X	X	X		X		X	X	X	X		X	X	X			X		X		X	X	X		X	X						X	X	X	X

TABLE 3. CATEGORIZATION OF SYSTEMS FROM THE NBS SURVEY

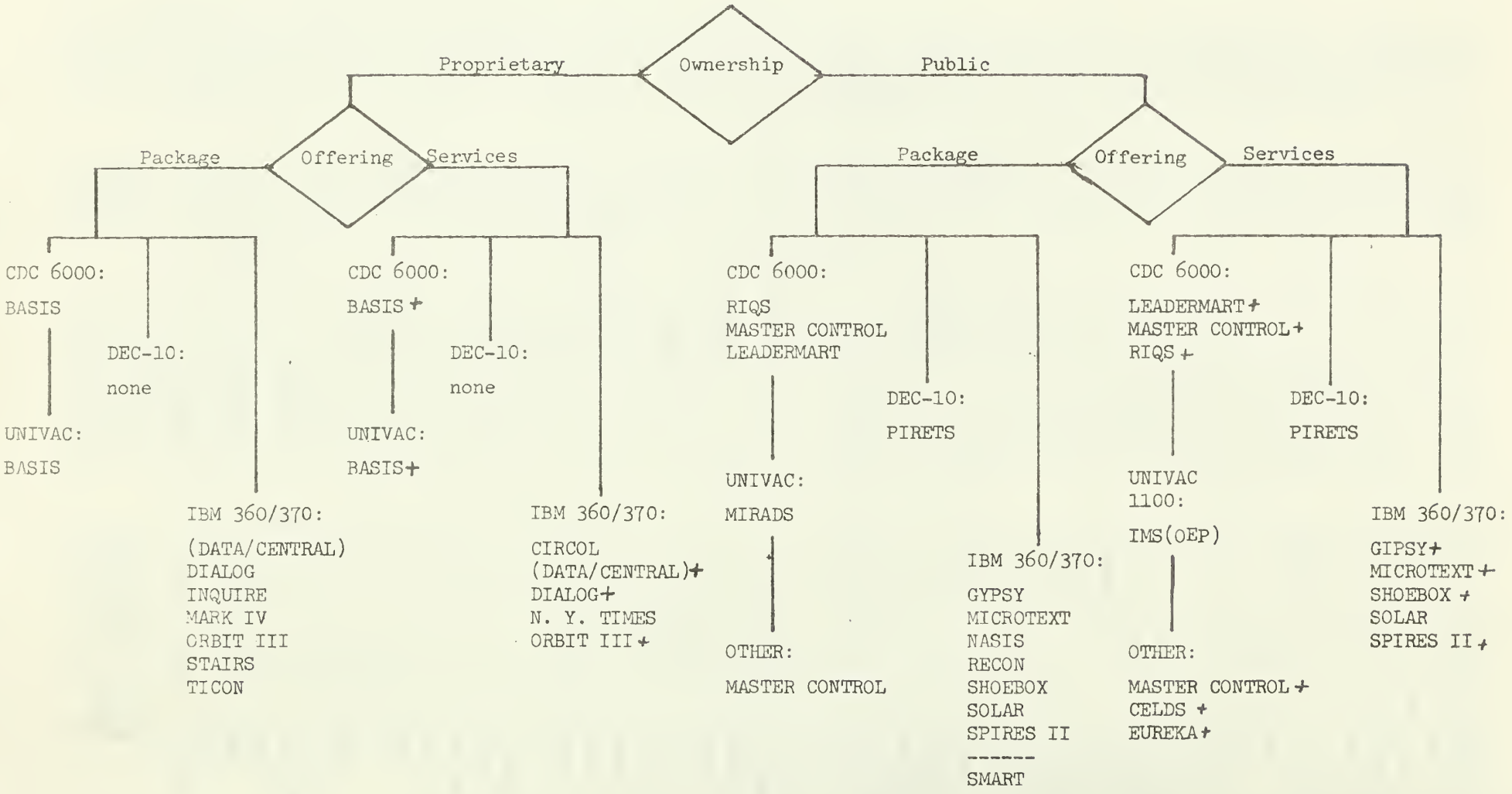

FIGURE 1. STRUCTURED TEXT SEARCHING SYSTEMS

not meet the criteria required for inclusion in the NBS survey. Tables 4 and 5 characterize the SMART and CELDS system features, while Table 6 characterizes the "ideal" system for an HDB.

6. The Choice of an Appropriate HDB System

The process of finding an appropriate information retrieval system to be used for the HDB reduces first to the decision of whether or not natural language queries are necessary and then, if they are necessary, to deciding whether SMART is satisfactory. If it comes close to being satisfactory, then one can consider how it can be implemented so that it is available to CERL.

6.1 The Choice of Natural Language Query

There is a seemingly unanswered question of whether or not natural language inquiry and fully automated language analysis procedures are effective in a document retrieval environment such as the HDB. Note here that we are assuming that the HDB task is equivalent to document retrieval in the sense that the statement content is similar to abstracts. However, in the techniques now being used for the HDB the title, author, and other citation type of information is not used. The only thing used is the text of the statement itself, the index is only used in the ANDOR approach.

TABLE 4. QUESTIONNAIRE FROM THE NBS SURVEY

A. GENERAL DESCRIPTION	
1. SYSTEM NAME	SMART/ANDOR/DOCAX/BIBAX
ORIGINATOR	CERL
TELEPHONE	(217) 352-6511
2. SOFTWARE AVAILABILITY	
a. FOR PURCHASE AT WHAT COST b. FOR LEASE AT WHAT COST	Owned.
3. SERVICE AVAILABILITY	
a. SEARCH SERVICE AT WHAT COST b. DATA BASES c. TAKES CUSTOMER DATA BASES	
4. HISTORY OF SOFTWARE	
a. FIRST INSTALLATION b. SIGNIFICANT INSTALLATIONS	
5. HISTORY OF SERVICE	
a. SERVICE INITIATION b. PRESENT USAGE	
B. COMPUTER ENVIRONMENT	
1. MAIN FRAME	IBM 360/75 and DEC-10, both must be used.
2. OPERATING SYSTEM	OS/MVP on 360.
3. SOURCE LANGUAGE	Fortran on DEC-10, no source for SMART.
4. TERMINAL TYPES	All TTY-compatible ASCII terminals.
5. TRANSMISSION RATES	300 baud.
6. RE-ENTRANT FOR MULTIPLE USERS	No*
C. FILE DEFINITION	
1. USER DEFINABLE	No
2. VARIABLE LENGTH TEXT	Yes
3. REPEATED FIELDS	No
D. FILE MAINTENANCE	
1. ALLOWS ON-LINE	
a. CREATION b. UPDATE c. DELETION d. EDITING	Full editing.
2. PREPROGRAMMED DATA VALIDATION CHECKS	No
3. OPTIONAL ORDERING OF ENTERED DATA	Yes
4. AUTOMATIC OR MACHINE-AIDED CONTENT INDEXING	No
E. QUERY	
1. USER-SYSTEM INTERACTION	
a. MULTIPLE OPTIONS AT ANY POINT	No
b. ENGLISH-LIKE PHRASING	No
c. SYSTEM-FORCED DIALOGUE	Yes, within any program.
2. CONTENT SEARCHING	
a. CONTROLLED VOCABULARY	No, in SMART - Yes, in ANDOR, but vocabulary is digits.
b. FULL TEXT INVERSION	No
c. STEMMING PERMITTED	No
d. SYNONYMS	Yes, in SMART
e. DISPLAY OF RELATED TERMS	No
f. CHECKING TERM IN CONTROLLED VOCABULARY	No

*Insufficient information to be certain.

TABLE 4., CONTINUED

3. SEARCH SPECIFICATION

a. MUST NAME FIELDS	Yes, in AND/OR
b. MAY LIMIT FIELDS	Yes, in AND/OR
c. BOOLEAN AND	Yes
d. BOOLEAN OR	Yes
e. BOOLEAN NOT	No
f. NESTING OF BOOLEAN EXPRESSIONS	No
g. PHRASE AND DISTANCE SEARCHING	No
h. NATURAL ENGLISH	Yes
i. RANGE SEARCHING	No

4. TUTORIAL FEATURES

a. "HELP" COMMAND	No
b. DOCUMENTATION ON-LINE	No

5. SEARCH STRATEGY

a. RANDOM	That is, part of SMART uses
b. SEQUENTIAL	consecutive search.
c. INDEXED SEQUENTIAL	

F. REPORT GENERATION

1. LANGUAGE TYPE

a. STANDARD OUTPUT	Yes
b. SELECT AMONG OPTIONS	
c. OWN CODE	

2. MEDIA FLEXIBILITY

a. OFF-LINE PRINTING	Yes
b. DISPLAY OF GRAPHS	
c. SPECIAL OUTPUTS	

3. SPECIAL CAPABILITIES

a. SORTING	None
b. SEQUENCING	
c. COUNTING	
d. ARITHMETIC	

G. SECURITY PROTECTION

1. TERMINAL	Logon password.
2. DATA BASE	
3. RECORD	
4. FIELD	

TABLE 5. QUESTIONNAIRE FROM THE NBS SURVEY

A. GENERAL DESCRIPTION	
1. SYSTEM NAME	CELDS
ORIGINATOR	CAC/CERL
TELEPHONE	(217) 333-3568 or (217) 352-6511 Ext. 363
2. SOFTWARE AVAILABILITY	
a. FOR PURCHASE AT WHAT COST	Public Domain.
b. FOR LEASE AT WHAT COST	
3. SERVICE AVAILABILITY	
a. SEARCH SERVICE AT WHAT COST	
b. DATA BASES	
c. TAKES CUSTOMER DATA BASES	Customers develop their own data basee.
4. HISTORY OF SOFTWARE	
a. FIRST INSTALLATION	Naval Shipbuilding Research and Development Center.
b. SIGNIFICANT INSTALLATIONS	Center for Advanced Computation, Urbana, Illinois.
5. HISTORY OF SERVICE	
a. SERVICE INITIATION	November, 1974
b. PRESENT USAGE	
B. COMPUTER ENVIRONMENT	
1. MAIN FRAME	PDP-11/50
2. OPERATING SYSTEM	Unix
3. SOURCE LANGUAGE	'C'
4. TERMINAL TYPES	All TTY-compatible ASCII terminals.
5. TRANSMISSION RATES	Up to 9600 baud by arrangement, typically 300 baud.
6. RE-ENTRANT FOR MULTIPLE USERS	Yes
C. FILE DEFINITION	
1. USER DEFINABLE	No
2. VARIABLE LENGTH TEXT	Yes
3. REPEATED FIELDS	Yes
D. FILE MAINTENANCE	
1. ALLOWS ON-LINE	
a. CREATION b. UPDATE c. DELETION d. EDITING	Full editing, updating, error checking on-line.
2. PREPROGRAMMED DATA VALIDATION CHECKS	Yes
3. OPTIONAL ORDERING OF ENTERED DATA	Mandatory.
4. AUTOMATIC OR MACHINE-AIDED CONTENT INDEXING	No
E. QUERY	
1. USER-SYSTEM INTERACTION	
a. MULTIPLE OPTIONS AT ANY POINT	Yes
b. ENGLISH-LIKE PHRASING	Yes
c. SYSTEM-FORCED DIALOGUE	No, but prompting suggests parameters.
2. CONTENT SEARCHING	
a. CONTROLLED VOCABULARY	Yes
b. FULL TEXT INVERSION	No
c. STEMMING PERMITTED	No
d. SYNONYMS	No, use manual thesaurus.
e. DISPLAY OF RELATED TERMS	No, thesaurus is manual.
f. CHECKING TERM IN CONTROLLED VOCABULARY	Yes, but does not give alpha neighbors.

TABLE 5., CONTINUED

3. SEARCH SPECIFICATION

a. MUST NAME FIELDS	Yes
b. MAY LIMIT FIELDS	Yes
c. BOOLEAN AND	Full boolean, including
d. BOOLEAN OR	"Exclude".
e. BOOLEAN NOT	
f. NESTING OF BOOLEAN EXPRESSIONS	Yes, with parentheses and precedence.
g. PHRASE AND DISTANCE SEARCHING	Yes, phrase No, distance.
h. NATURAL ENGLISH	No
i. RANGE SEARCHING	No

4. TUTORIAL FEATURES

a. "HELP" COMMAND	Yes
b. DOCUMENTATION ON-LINE	No

5. SEARCH STRATEGY

a. RANDOM	Index, Random access.
b. SEQUENTIAL	
c. INDEXED SEQUENTIAL	

F. REPORT GENERATION

1. LANGUAGE TYPE

a. STANDARD OUTPUT	Yes
b. SELECT AMONG OPTIONS	
c. OWN CODE	

2. MEDIA FLEXIBILITY

a. OFF-LINE PRINTING	Yes
b. DISPLAY OF GRAPHS	
c. SPECIAL OUTPUTS	

3. SPECIAL CAPABILITIES

a. SORTING	None
b. SEQUENCING	
c. COUNTING	
d. ARITHMETIC	

G. SECURITY PROTECTION

1. TERMINAL	Logon password.
2. DATA BASE	Data base secured by group.
3. RECORD	
4. FIELD	

TABLE 6. QUESTIONNAIRE FROM THE NBS SURVEY

A. GENERAL DESCRIPTION	
1. SYSTEM NAME	HDB Desired
ORIGINATOR	
TELEPHONE	
2. SOFTWARE AVAILABILITY	
a. FOR PURCHASE AT WHAT COST	
b. FOR LEASE AT WHAT COST	
3. SERVICE AVAILABILITY	
a. SEARCH SERVICE AT WHAT COST	
b. DATA BASES	
c. TAKES CUSTOMER DATA BASES	
4. HISTORY OF SOFTWARE	
a. FIRST INSTALLATION	
b. SIGNIFICANT INSTALLATIONS	
5. HISTORY OF SERVICE	
a. SERVICE INITIATION	
b. PRESENT USAGE	
B. COMPUTER ENVIRONMENT	
1. MAIN FRAME	Advantages to in-house, otherwise important that it be available to users outside CERL.
2. OPERATING SYSTEM	
3. SOURCE LANGUAGE	
4. TERMINAL TYPES	ASCII
5. TRANSMISSION RATES	300 baud minimum.
6. RE-ENTRANT FOR MULTIPLE USERS	Yes
C. FILE DEFINITION	
1. USER DEFINABLE	Not required.
2. VARIABLE LENGTH TEXT	Yes
3. REPEATED FIELDS	Yes
D. FILE MAINTENANCE	
1. ALLOWS ON-LINE	
a. CREATION	Yes
b. UPDATE	Yes
c. DELETION	Yes
d. EDITING	Yes
2. PREPROGRAMMED DATA VALIDATION CHECKS	Valuable.
3. OPTIONAL ORDERING OF ENTERED DATA	Not required.
4. AUTOMATIC OR MACHINE-AIDED CONTENT INDEXING	Valuable.
E. QUERY	
1. USER-SYSTEM INTERACTION	
a. MULTIPLE OPTIONS AT ANY POINT	Yes
b. ENGLISH-LIKE PHRASING	Valuable.
c. SYSTEM-FORCED DIALOGUE	No
2. CONTENT SEARCHING	
a. CONTROLLED VOCABULARY	Valuable for interactive use.
b. FULL TEXT INVERSION	Desirable for batch use.
c. STEMMING PERMITTED	Desirable.
d. SYNONYMS	Valuable.
e. DISPLAY OF RELATED TERMS	Valuable.
f. CHECKING TERM IN CONTROLLED VOCABULARY	Valuable.

TABLE 6., CONTINUED

3. SEARCH SPECIFICATION	
a. MUST NAME FIELDS	Not needed.
b. MAY LIMIT FIELDS	Yes
c. BOOLEAN AND	Yes
d. BOOLEAN OR	
e. BOOLEAN NOT	Yes
f. NESTING OF BOOLEAN EXPRESSIONS	Desirable.
g. PHRASE AND DISTANCE SEARCHING	Desirable.
h. NATURAL ENGLISH	Valuable.
i. RANGE SEARCHING	Slightly desirable.
4. TUTORIAL FEATURES	
a. "HELP" COMMAND	Yes
b. DOCUMENTATION ON-LINE	Yes
5. SEARCH STRATEGY	
a. RANDOM	Consecutive is unacceptable, typical use is random.
b. SEQUENTIAL	
c. INDEXED SEQUENTIAL	
7. REPORT GENERATION	
1. LANGUAGE TYPE	
a. STANDARD OUTPUT	
b. SELECT AMONG OPTIONS	Desirable.
c. OWN CODE	
2. MEDIA FLEXIBILITY	
a. OFF-LINE PRINTING	
b. DISPLAY OF GRAPHS	
c. SPECIAL OUTPUTS	Slightly desirable.
3. SPECIAL CAPABILITIES	
a. SORTING	Valuable.
b. SEQUENCING	Slightly desirable.
c. COUNTING	Slightly desirable.
d. ARITHMETIC	Slightly desirable.
G. SECURITY PROTECTION	
1. TERMINAL	Users must be charged.
2. DATA BASE	Data base will be public domain.
3. RECORD	
4. FIELD	

This is an important question largely because of the research which tends to cast doubts on the consistency of manually prepared document analysis. Salton reports in a number of research studies that automated language analysis procedures can provide benefits. One of the major results of a recent study was that although simple word extraction followed by boolean search does not produce retrieval results equivalent in effectiveness to standard manual indexing techniques, a variety of techniques can be added to obtain retrieval whose effectiveness exceeds conventional manual methodologies. When these factors are added to the expense of preparing the index and thesauri by hand, the argument to stay with automated techniques becomes stronger.

One wonders openly whether the choice to implement the HDB in the way it now appears was made with full understanding of the implications of this ever-expanding body of research or merely as a result of the convenience, or even the personal bias of one of the early workers on the project. Certainly the CERL HDB managers have a choice between two courses of action. One is to use a manual indexing technique coupled with a manually prepared thesaurus or set of key words. Given this choice several of the commercially available and tested retrieval systems could be adapted and the HDB would be searched with techniques similar to those now successfully being used to search the major document data bases in use today (NTIS, ERIC, Chem Abstracts, etc). The

other choice is to continue with the more forward-looking but less proven techniques of automatic content analysis with natural language queries as represented by the SMART system. The system in use today by the HDB lies somewhere in between the two extremes, since laborious indexing is done as the statements are prepared, and boolean searches of a sort are done on the basis of these indexes. But this is complemented by running SMART, which does not make use of the indexes at all.

6.2 Suitability of SMART

There are some questions and reservations about the use of SMART as a major tool for the HDB.

First of all, the SMART implementation requires a very large core region on a 360 system. The current implementation was intended as a vechicle for experimental work in information retrieval techniques. As a result, much of the size of the code is concerned with measuring retrieval performance. Considerable portions of the code which is loaded from the SMART library is never actually executed. A production version could conceivably be produced which would not include as many measuring tools and thus could be somewhat smaller. One current goal of the Cornell group is a modular implementation so that one could load only necessary modules for a production environment implementation. An alternate solution would be to implement the code (which is primarily Fortran with some assembly language subroutines) on a virtual memory operating system.

Secondly, the current implementation is strictly a batch system. An attempt has been made at Cornell to implement SMART under TSO, but that work now appears to have fallen by the wayside. It seems apparent that an interactive system would make it easier for the user to modify his searching strategy based upon what he is finding, rather than submitting a number of batch jobs, all of which must do the complete search. As an information system to be used by information specialists this major inconvenience might be overcome, but in our view it is unlikely to ever be regularly used by customers directly in this mode.

Some thought should be given to why commercially available systems are not offering automatic content analysis and natural language queries in quite the same way that SMART attempts to do. The systems which are available commercially seem to be universally built on some variation of key word searching and boolean expressions for search requests. The commercial systems have a long (up to 10 years) period of development behind them. When these efforts started natural language processing was not sufficiently developed to make it worth the commercial risk. Some would argue that it is still not worth the risk. The fact that so many commercial systems use key words tends to suggest that the technology is accepted and a long term period of support can be envisioned. The implication of all of this to the HDB is that if what is needed must feature natural language queries with

no manual preparation of key words, a non-commercial, semi-experimental system is the only choice. However, if the current HDB can be expanded (either by hand or with programs) to include key words, one of the commercially available systems will provide reliable long term service of a less sophisticated nature. It may even be that simple full text searching of the statements themselves using a controlled thesaurus, could be used on one of the commercial systems.

These conclusions should not preclude pursuing the goal of interactive language query systems. To whatever extent this capability is crucial to the long term goals of the COE, it should be pursued as an adjunct to systems like the HDB. However, a completely adequate job of data storage and retrieval in support of an Information Access Center (IAC) for the HDB can be done with commercially available systems. Unfortunately, some backtracking will be necessary to associate appropriate keywords with each of the habitability statements if that course is taken.

6.3 The Choice of Continuing With SMART

One possible course of action is to continue using the SMART system in its present form. This can be done with or without the concurrent use of the package of programs loosely associated with AND/OR. Options that are directly available to CERL at the present time include continuing with the CSO installation and running the software that is now available, transferring the SMART system to the Amdahl installation at the University of

Michigan, or transferring it to the IBM 360/91 at UCLA. Of course, it is always possible to put the system on some nationally available time-sharing service, but that would cause some (solvable) difficulty with the local availability of printouts.

The DEC-10 and 360/75 installation at the Computing Services Office (CSO) of the University of Illinois is expected to stay available in its present form only through the middle of 1977. At that time the present indication is that the DEC-10 system will be taken out of service. The general expectation is that the 360 will stay in service through the middle of 1978, because of the demand by university users for whom conversion will be impossible before that time. Thus, there need be no rush to bring up a different system if one is willing to tolerate the long turn-around time for SMART jobs. Some of this turn-around time is a result of having to ask that the disk with the HDB be mounted each time it is needed. Requesting that the disk be permanently mounted would reduce that delay but produce some small increase in costs.

Another course of action, if the choice is to stay with the SMART system, is to move the library to the installation at the University of Michigan at Ann Arbor. At least two projects at CERL are currently using the University of Michigan system with apparently good results. If the choice is the Michigan system, then the next choice is what to do about the AND and OR programs.

However, the very nature of the Michigan Time Sharing (MTS) system provides a useful solution. MTS is a superior time sharing system which allows interactive access from user programs in a rather general way. The system has an impressive collection of interactive services, including a good editor, document preparation systems, and convenient handling of large disk files. It would be possible to recode the AND and OR programs, as well as the programs which allow one to see the documents (statements) and bibliographic entries. The programming could all be done from CERL with interactive terminals. Terminal access to MTS can be directly by FTS line, or can be arranged in the same way that some other projects at CERL employ. They dial to a phone port at CAC which is attached to a multiplexer, the other end of which is a port on the MTS system.

The multiplexer equipment now in use is available as excess capacity on a system installed by CAC for another project. That project is currently expected to continue at least through January of 1977. The excess capacity of this line is expected to be available so long as that project continues to be funded, which is expected to be for more than another year. In the worst case, that in which the CAC project no longer needs the access to MTS, it would only require four regular users at CERL to justify pooling costs to put in this equipment themselves, pay the same rate each that is currently being paid by CERL users for this service, and have the multiplexer strictly for CERL use. The

total budget for the multiplexer connection is on the order of $800 per month. As few as four projects could use such a communications system to keep their total costs below long distance access costs.

Remote job entry from the Unix system at CAC is now available. Files of card images are transmitted to MTS from Unix disk files in a manner similar to Hasp work stations. Printed output from UM is available on the equipment at CAC. This service is expected to continue so long as the line to Michigan is needed and there are funds to support it. Charges for use of the local system come as a separate bill from the computer charges assessed at Michigan. The communications cost is currently billed as a fixed monthly cost for the use of the multiplexer and associated phones.

The costs at UM are said to be reasonable according to the CAC users of MTS. The only noticeable startup costs for going to this solution would be costs associated with sending the HDB files to Michigan, and the costs of reprogramming the programs other than SMART which are needed to continue the present mode of operation. However, our experience with MTS indicates that the level and reliability of the service at Michigan warrant its serious consideration.

Still another avenue is open to facilitate staying with the SMART system without being concerned with the continuing availability of the CSO 360. The Campus Computing Network of UCLA is available on the ARPANET and can be accessed as easily from any arpanet node as it can from the CAC. The time sharing system there is TSO, which certainly does not compare to MTS in terms of its friendliness to the user. However, large batch jobs can be run at CCN, and printouts can be returned to the printer at CAC. The charges for this connection could probably be kept to on the order of $5 per hour connected to the network, plus the normal user fees at CCN. The 360/91 installation at CCN is one of the more reliable places we have come in contact with over the last two years.

The interactive portions of the HDB tasks would have to be recoded under the TSO system at CCN. However, since they are now coded in Fortran, a mere conversion would suffice to make the system as useable in that environment as it is in its present environment. Costs there are comparable to costs at the University of Illinois, except that in our experience jobs which require a large region size (as SMART does) generally are cheaper to run at CCN. Also, because of more core on the CCN system, large jobs can be run at any time of day and the turn-around time is generally better than for a comparable large job on the 360/75 at CSO. Also, the processor out there is much faster and as a result, the wait for results of a query should be much shorter.

In either the University of Michigan or the UCLA situations, the disadvantage of SMART being a strictly batch system would still apply. However, with the appropriate cooperation of the originators of SMART at Cornell, either of these systems would be suitable for converting SMART into an interactive system. This would be no small undertaking. It could not (or should not) be done without the active cooperation of the group at Cornell who are intimately familiar with the inner workings of SMART. The software development for such a task would conservatively take about a year for about a two to two and one-half man-years of programming.

As a purely batch system, SMART could be installed on one of the nationally available time-sharing systems which offers IBM equipment. If the remote job entry equipment at CERL is sometime attached to such a service, it would be easy to move a copy of the SMART library to such a service and run just the SMART system as pure batch jobs. If the system also supports time-sharing service, the AND and OR programs could be recoded just as they would have to be with any of the other choices.

7. Developing a Replacement for SMART

While the SMART system in its present implementation is not quite satisfactory for the production stages of the HDB effort, careful consideration must be given to any proposals to change systems at this stage of development. Certainly a change from the implementation on the IBM 360 and DEC-10 system at CSO is

going to be necessary, because those systems are scheduled to be phased out of service over the next two years. Section 6 discussed some of the issues which must be addressed in an information retrieval system suitable for the HDB, but outlined the options available to CERL if their decision were to stay with the SMART program.

The assumption in this section is that a decision has been made to abandon the SMART programs and develop or find something else. Given that assumption, two avenues of investigation are open. One is to develop the CELDS system which is performing a similar function for environmental data bases in conjunction with another group at CERL. The other is to make the necessary modifications to the HDB to make the information retrievable using one of the nationally available information retrieval. The DIALOG system at Lockheed is given as an example because it comes the closest to meeting the criteria outlined in Section 5.

7.1 Revision of CELDS for the HDB

Any initial implementation of HDB on a CELDS-copy retriever would have to include at least the capabilities that ANDOR, BIBAX and DOCAX already provide to HDB users. CELDS provides these options now, and in addition provides:

1) all functions are combined into one retrieval language.
2) SAVE interesting and often used output sets

3) HELP

4) partial search tells user how many statements satisfy sub-expressions

5) parentheses and full expression nesting

6) OOPS to return to previous statement-set

7) allows multiple values per field

8) off-line printing

9) simple logon-logoff

10) retains fast response time even for very large databases

To convert to a field-oriented system (like CELDS) the HDB could be broken into the following fields:

ACC - accession number
DOC - document number
STMT- statement number
DATE- date published/ researched/ input
[unknown for current database]
BIB - bibliographic data
AUTH- name of author(s) [unknown for current DB]
FUNC- functional area code
TRFC- training facility code
PHYS- physical settings
ENV - environmental descriptors (however many apply)
OCCU- occupants
PSTR- posture
INVM- involvement
ORGF- organizational functions
SFCN- function of statement
TEXT- the text of the statement
KEY - keywords [unknown for the current DB]

Several new values would have to be added to the SFCN field including "objectives", "data", and "procedure". Several of the fields (such as PSTR and INVM) could be dropped and their values used as KEYWORDS. It would help streamline the list of fields without loss of generality. The DATE field is a useful field to

include, but not strictly necessary. The only non-searchable fields would be BIB, TEXT, and DATE.

CELDS-like format includes one line per field and each line is prefixed by accession number and field number. The current HDB lines are suffixed by statement number, card number, and document number, and separated (unnecessarily) by 'NEXT TEXT' cards. Converting data formats would be fairly simple, except that a few desirable fields would be missing [e.g. keywords] and the current HDB uses digit strings for the indexes. Names would be much easier for novice users to read. These could be converted automatically.

Two CELDS input programs would have to be modified slightly (made more general) to accomodate the different field names. The CELDS retriever program would have to be modified to use the new fields also. The inversion program would have to be run on the newly created Habitability Data Base.

The next obvious improvements would include adding keywords to the database, and adding an on-line thesaurus to the retriever. Then the combined retriever could be modified to use the thesaurus to recognize concepts in a very SMART-like environment. Concept numbers and weighting are not currently practical for interactive searching, but this could make a fascinating research project.

7.2 Using a Commercial Information Retrieval System

One of the more popular and widely used of the commercially

available information retrieval systems is the DIALOG system operated by Lockheed in Palo Alto, California. This system was included in the survey discussed in Section 5. If a commercially available system is considered as a home for the HDB, certainly DIALOG should be considered a prime candidate.

The decision to move to a commercial retrieval system presents questions both of a technical nature and of a purely operational nature. We address both kinds of questions, but from the very limited basis of the specific information which is available to us in the course of this investigation. We consider first the technical questions of what would be required to put the HDB into the DIALOG system.

Putting the HDB into DIALOG would require almost exactly the same amount of effort as putting it into CELDS-format. DIALOG is a field-oriented system with full text searching ability, but not natural language query. The HDB would almost certainly have to be converted to a DIALOG format, and keywords should be added. DIALOG would require a very complete thesaurus, which would then be available on-line. Full text searching in DIALOG requires an exact match to the words in the statement.

The DIALOG system works primarily with searches on predefined fields. Although the system is designed for bibliographic retrieval, the similarity to the information in the HDB suggests that only a small perturbation of the HDB would be required for conversion to DIALOB. The fields in the index used

with the HDB statements could be made into fields in the DIALOG sense. The statements in the HDB are similar to abstracts and thus could be treated by DIALOG in the same way abstracts are treated. The task of converting what now exists in HDB to a form suitable for DIALOG could be assisted by some of the text processing capability in the UNIX system at CAC.

One conceptual dissimilarity between the two systems is that in DIALOG all information of one record (or set of records) concerns a single document, and there is no field to refer to a parent document. In HDB, on the other hand, the basic information is a statement, several of which come from the same parent document. It would be possible to think of each HDB statement as a document in the DIALOG sense, providing that an extra field is added to give reference to the parent document. Also, in this context, it would probably be advisable to encode keywords for each of the HDB statements. Other information would be based on the parent document. This would probably need to include the author or some other reference to the source, the date if that applies, the corporate author if one exists, and inevitably, the key words for the parent document.

Another pressing need, in the event of this choice as well as several others, is for a completed thesaurus for the HDB. The approach taken in the thesaurus for the early portions of the HDB is a step in the right direction, but it needs to be expanded to include terms peculiar to the whole range of habitability

statements, not just the limited subset available to CERL now. By standardizing the HDB vocabulary, and by carefully keywording, DIALOG could be a fast, easy to use system for retrieving from the HDB.

It is impossible to determine the cost of putting the HDB on DIALOG nor estimating what it would cost to run, except by comparing the complexity of HDB to some of the other available databases for which at least representative user costs are available. The cost for accessing the NTIS data base, for example, is $25 per connect hour. (The system is purely interactive.) In addition to this a communication cost is added dependent upon the mode of access. For access via Telenet this charge is $8 per hour. Since the HDB is considerably smaller than NTIS, one would expect the charge to be less, except for the fact that fewer customers might mean higher prices.

The documentation of DIALOG makes it very clear that they will not be able to predict the cost for a new database and the accompanying service to access it. Such an estimate could be nothing but a raw guess without an extremely detailed proposal from Lockheed. One immediate suggestion is that Lockheed should be contacted, given as much information as possible about the HDB, including this report, and then asked to submit a cost proposal. Sales brochures for DIALOG indicate the price for out of the ordinary services as "negotiable."

Although we have no idea what it costs to put up the NTIS information on DIALOG, it seems clear that one of the justifications is the wide interest in accessing the NTIS database, and thus the customer base with which to recover the installation costs. For a special purpose client like CERL the cost cannot be spread over so many customers and thus the apparent cost will seem higher. In order to operate an IAC which includes the capability to search the HDB for clients, CERL would thus have to pass on fairly high operational costs to the client or operate the service at a loss until the number of clients spreads the cost out over a wider base of users.

There is another whole question which is still unanswered as to whether Lockheed would even be interested in putting HDB on their system. Certainly the customer base at the present time would not warrant their covering the cost of transforming the HDB into a form suitable for DIALOG. CERL would either have to do that themselves or pay Lockheed to do it. Now it is certainly true that Lockheed is intended to be a profit making venture, and thus they may be willing to put whatever someone wants onto their system for an appropriately large sum of money. However, it may be that their growth plans do not allow for yet another potentially large data base to come on the scene in the near future. If this is true they will not be able to put the HDB database on DIALOG, regardless of whether or not they could recover their costs for doing so. We were able to contact DIALOG

users, and use DIALOG on-line. The DIALOG users we sampled were largely pleased with Lockheed service.

LIST OF REFERENCES

[1] T. A. Davis, "Conceptualization of Habitability Expressions for the Habitability," Data Base U.S. Army Construction Engineering Research Laboratory Interim Report D-68, U.S. Army Construction Engineering Research Laboratory, Champaign, Illinois, August 1976

[2] Dennis W. Fife, Kirk Rankin, Elizabeth Fong, Justin C. Walker and Beatrice A. Marron, "The Technical Index of Interactive Information Systems," U.S. Department of Commerce National Bureau of Standards Technical Note No. 819, U.S. Government Printing Office, Washington, 1974

[3] Wayne Hamilton, "General Overview of the HDB File System," Portion of an unpublished report, 1976

[4] U.S. Army Construction Engineering Research Laboratory, "Demonstration Instructions for the Prototype Habitability Data Base," U.S. Army Construction Engineering Research Laboratory, Champaign, Illinois, (Mimeographed)

[5] U.S. Army Construction Engineering Research Laboratory, "User's Instruction Manual for the Prototype Habitability Data Base," U.S. Army Construction Engineering Research Laboratory, Champaign, Illinois, (Mimeographed)

[6] Gerard Salton, ed., The SMART Retrieval System, Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1971

BIBLIOGRAPHY

[1] G. J. Baker, "Database World," Database Journal, Vol. 6, No. 11, 1976, pp.18-23

[2] G. J. Baker, "The Correct Use of Codasyl DBTG Sets," Database, Vol. 6, No. 2, pp. 19-21

[3] Barnett and Lightfoot, "Information Management Systems (IMS), a User's Experience with Evolutionary Development," in Data Base Management Systems, Proceedings of the SHARE Working Conference on Data Base Management Systems, Montreal, Canada, July 23-27, 1973, Donald A. Jardine, ed., North Holland Publishing Company, 1974

[4] M. Bibby, "User Experience with IDS at STC," Database Journal, Vol. 6, No. 5, 1976, pp. 7-13

[5] Vaclav Chvalovsky, "Anything New in Data Base Technology?," Datamation, Vol. 22, No. 4, April 1976, pp. 54-55

[6] L. J. Cohen, "Is Database the Way of the Future?," Database Journal, Vol. 6, No. 11, 1976, pp. 15-18

[7] Leo J. Cohen, Data Base Management Systems, Performance Development Corporation and Q. E. D. Information Sciences, Inc., 1973

[8] Robert M. Curtice, "The Outlook for Data Base Management," Datamation, Vol. 22, No. 4, April 1976, pp. 46-49

[9] C. J. Date, "Relational Data Base Concepts," Datamation, Vol. 22, No. 4, April 1976, pp. 50-53

[10] T. A. Davis, "Conceptualization of Habitability Expressions for the Habitability," Data Base U.S. Army Construction Engineering Research Laboratory Interim Report D-68, U.S. Army Construction Engineering Research Laboratory, Champaign, Illinois, August 1976

[11] E. J. Emerson, "DMS 11,000 User Experience," in Data Base Management Systems, Proceedings of the SHARE Working Conference on Data Base Management Systems, Montreal, Canada, July 23-27, 1973, Donald A. Jardine, ed., North Holland Publishing Company, 1974

[12] Mrs. S. Fenlon, "The On-Line Patient Index at Addentrooke's Hospital, Cambridge," Database Journal, Vol. 6, No. 7, 1976, pp. 22-26

[13] Dennis W. Fife, Kirk Rankin, Elizabeth Fong, Justin C. Walker and Beatrice A. Marron, "The Technical Index of Interactive Information Systems," U.S. Department of Commerce National Bureau of Standards Technical Note No. 819, U.S. Government Printing Office, Washington, 1974

[14] Wayne Hamilton, "General Overview of the HDB File System," Portion of an unpublished report, 1976

[15] D. Hannaford, "TOTAL--A Detailed Analysis," Database Journal, Vol. 6, No. 7, 1976, pp. 9-14

[16] L. C. Hobbs, "Future Trends in Hardware," in Data Base Management Systems, Proceedings of the SHARE Working Conference on Data Base Management Systems, Montreal, Canada, July 23-27, 1973, Donald A. Jardine, ed., North Holland Publishing Company, 1974

[17] Susanne M. Humphrey, "Searching the MEDLARS Citation File On-Line Using ELHILL and STAIRS: An Updated Comparison," Information Processing & Management, Vol. 12, Pergamon Press, Great Britain, 1976, pp. 63-70

[18] F. E. Johnson, "IDS--A Brick in the Database Tower of Babel," Database Journal, Vol. 6, No. 5, 1976, pp. 2-6

[19] Database Journal, "The Interim Report of the ANSIX3SPARC Study Group on Database Management Systems," Database Journal, Vol. 6, No. 11, 1976, pp. 10-14

[20] Database Journal, "The IBM Database Range--VANDLI, DLI

ENTRY, DLI and IMS," Database Journal, Vol.6, No. 10, 1976, pp. 2-9

[21] Database Journal, "Database World," Database Journal, Vol. 6, No. 10, 1976, pp. 19-23

[22] Database Journal, Vol. 6, No. 5, "IDMS and the 2900 Series," Database Journal, Vol. 6, No. 5, 1976, p. 19

[23] Donald W. King and Edward C. Bryant, The Evaluation of Information Services and Products, Information Resources Press, Washington, 1971

[24] U.S. Army Construction Engineering Research Laboratory, "Demonstration Instructions for the Prototype Habitability Data Base," U.S. Army Construction Engineering Research Laboratory, Champaign, Illinois, (Mimeographed)

[25] U.S. Army Construction Engineering Research Laboratory, "User's Instruction Manual for the Prototype Habitability Data Base," U.S. Army Construction Engineering Research Laboratory, Champaign, Illinois, (Mimeographed)

[26] P. A. Lavallee and S. Ohayon, "DMS Applications and Experience," in Data Base Management Systems, Proceedings of the SHARE Working Conference on Data Base Management Systems, Montreal, Canada, July 23-27, 1973, Donald A. Jardine, ed., North Holland Publishing Company, 1974

[27] E. T. Lee, "Shape-Oriented Storage and Retrieval of Geometric Figures and Chromosome Images," Information Processing & Management, Vol. 12, Pergamon Press, Great Britain, 1976, pp. 35-41

[28] S. Lie-Nielsen and J. Refsnes, "SIBAS--The Portable Database Management System," Database Journal, Vol. 6, No. 11, 1976, pp. 2-9

[29] J. Lockeretz, "User Experience with IMS at Esso Petroleum Company," Database Journal, Vol. 6, No. 10, 1976, pp. 10-18

[30] I. G. MacDonald, "Univac's Interpretation of the CODASYL DBTG Proposals," Database, Vol. 6, No. 2, pp. 3-7

[31] R. Maskell, "LEXICON--An Established Data Dictionary System," Database Journal, Vol. 6, No. 7, 1976, pp. 15-21

[32] W. E. Mercer, "User Experience 'TOTAL'," in Data Base Management Systems, Proceedings of the SHARE Working Conference on Data Base Management Systems, Montreal, Canada, July 23-27, 1973, Donald A. Jardine, ed., North Holland Publishing Company, 1974

[33] Alice Ray, "Habitability Information System, Final Report, Pt. 1," Champaign, Illinois, June 1975, (Mimeographed)

[34] K. A. Robinson, "DMS-1100, An In-Depth Evaluation," Database, Vol. 6, No. 2, pp. 8-14

[35] Gerard Salton, Dynamic Information and Library Processing, Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1975

[36] Gerard Salton, ed., The SMART Retrieval System, Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1971

[37] G. Michael Schneider and Edouard J. Desautels, "Creation of a File Translation Language for Networks," Information Systems, Vol. 1, No. 1, Pergamon Press, Great Britain, 1975, pp. 23-31

[38] Michael E. Senko, "Information Systems: Records, Relations, Sets, Entities, and Things," Information Systems, Vol. 1, No. 1, Pergamon Press, Great Britain, 1975, pp. 3-13

[39] William Howard Stellhorn, "An Experimental Information Retrieval System," Department of Computer Science Report No. 657, University of Illinois at Urbana-Champaign, Urbana, Illinois, July 1974

[40] Michael Stonebraker, "Getting Started in INGRES, a Tutorial," 1975

[41] Michael Stonebraker, Eugene Wong, Peter Kreps and Gerald Held, "The Design and Implementation of INGRES," Electronics Research Laboratory Memorandum No. ERL-M577, University of California, Berkeley, Berkeley, California, January 1976

[42] Roger K. Summit and Oscar Firschein, "Document Retrieval Systems and Techniques," in Annual Review of Information Science, Carlos A. Cuadra, ed., American Society for Information Science, Washington, 1974, pp. 286-331

[43] Daniel J. Tanner, "User Ratings of Software Packages," Datamation, Vol. 21, No. 12, December 1975, pp. 139-154

[44] Massachusetts Institute of Technology, "Janus Beginner's Manual," Massachusetts Institute of Technology, Cambridge, Massachusetts, 1975, (Draft)

[45] Massachusetts Institute of Technology, "Janus User's Manual," Massachusetts Institute of Technology, Cambridge, Massachusetts, 1975, (Draft)

[46] P. Thorpe and M. Cocks, "TOTAL--Market Leader in Independent Database Management Systems," Database Journal, Vol. 6, No. 7, 1976, pp. 2-8

[47] R. L. Welsh, "User Manual for the Computer-Aided Environmental Legislative Data System," U.S. Army Construction Engineering Research Laboratory Interim Report E-78, U.S. Army Construction Engineering Research Laboratory, Champaign, Illinois, November 1975

[48] Martha E. Williams, "Use of Machine-Readable Data Bases," in Annual Review of Information Science, Carlos A. Cuadra, ed., American Society for Information Science, Washington, 1974, pp. 221-284


BOUND BY