NASA CONTRACTOR REPORT

NASA CR-2893

THE BIOSIS DATA BASE: EVALUATION OF ITS INDEXES AND THE STRATBLDR, CHEMFILE, STAIRS AND DIALOG SYSTEMS FOR ON-LINE SEARCHING

Monica Nees and Hannah O. Green

Prepared by
NORTH CAROLINA SCIENCE AND TECHNOLOGY RESEARCH CENTER
Research Triangle Park, N. C. 27709
for Langley Research Center

NATIONAL AERONAUTICS AND SPACE ADMINISTRATION · WASHINGTON, D. C. • SEPTEMBER 1977

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
NASA CR-2893		

4. Title and Subtitle	5. Report Date
The BIOSIS Data Base: Evaluation of Its Indexes and the STRATBLDR, CHEMFILE, STAIRS and DIALOG Systems for On-Line Searching	September 1977
	6. Performing Organization Code
7. Author(s)	8. Performing Organization Report No.
Monica Nees and Hannah O. Green	STR-508
	10. Work Unit No.
9. Performing Organization Name and Address	11. Contract or Grant No.
North Carolina Science & Technology Research Center P. O. Box 12235 Research Triangle Park, NC 27709	NAS1-14208
	13. Type of Report and Period Covered
12. Sponsoring Agency Name and Address	Contractor Report
National Aeronautics & Space Administration Washington, DC 20546	14. Sponsoring Agency Code

15. Supplementary Notes
Langley Technical Monitor: John Samos
Topical Report

16. Abstract

BioSciences Information Service of Biological Abstracts (BIOSIS), publishes Biological Abstracts and BioResearch Index, covering world-wide literature in the life sciences and consisting of more than 240,000 references in 1974. These two secondary sources, jointly also called BIOSIS, are computer-searchable in the batch mode back to 1959.

With the advent of on-line searching in recent years, BIOSIS personnel developed two systems to assist them in interactive querying of their data base. These are STRATBLDR, for building the strategy, and CHEMFILE, a chemical dictionary of compounds and synonyms. STAIRS, an IBM-developed program, was selected for acturally performing the search on the BIOSIS file. Recognizing the need to have these systems evaluated by outside users, BIOSIS asked the North Carolina Science and Technology Research Center to collaborate on this research.

North Carolina Science and Technology Research Center was selected for two principal reasons: long-standing experience in computerized literature searching in general and experience in searching the BIOSIS data base in particular.

This report discusses Nc/STRC evaluation of the hardware and search systems, summarizes the strategies used, analyzes the searches by type of end user, and gives recommendations and conclusions. During the course of the project, the BIOSIS data base became commercially available for on-line searching via Lockheed's DIALOG system. Therefore, throughout this report STAIRS and DIALOG progrsms for searching BIOSIS are compared wherever appropriate.

17. Key Words (Suggested by Author(s))	18. Distribution Statement
Computer searchable Computerized literature	Unlimited - Unclassified Subject Category 60

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price*
Unclassified	Unclassified	67	$4.50

*For sale by the National Technical Information Service, Springfield, Virginia 22161

TABLE OF CONTENTS

I. Introduction

BioSciences Information Service of Biological Abstracts (BIOSIS), publishes Biological Abstracts and BioResearch Index, covering world-wide literature in the life sciences and consisting of more than 240,000 references in 1974. These two secondary sources, jointly also called BIOSIS, are computer-searchable in the batch mode back to 1959.

With the advent of on-line searching in recent years, BIOSIS personnel developed two systems to assist them in interactive querying of their data base. These are STRATBLDR, for building the strategy, and CHEMFILE, a chemical dictionary of compounds and synonyms. STAIRS, an IBM-developed program, was selected for actually performing the search on the BIOSIS file. Recognizing the need to have these systems evaluated by outside users, BIOSIS asked the North Carolina Science and Technology Research Center (NC/STRC) to collaborate on this research, and we were pleased to accept.

NC/STRC was selected for two principal reasons: long-standing expertise in computerized literature searching in general and experience in searching the BIOSIS data base in particular. A section of the Division of Natural and Economic Resources of the state of North Carolina, NC/STRC is also one of six in a network of Industrial Applications Centers (IAC) of the National Aeronautics and Space Administration (NASA). As such, it performs computerized literature searches on almost 60 data bases in response to questions from its clients, who are industrial firms, research institutes, universities and governmental agencies primarily in the southeastern United States.

Organized in 1964 to transfer aerospace technology to the private sector, NC/STRC shortly thereafter developed a computerized search program for the NASA file. Other data bases added to the in-house collection were three textile files, National Technical Information Service, Food Science and Technology Abstracts, and Educational Resources Information Center. NC/STRC also utilizes files at its sister IAC's, other information centers and on-line services available commercially and from the National Library of Medicine, making it one of the largest and most diverse information retrieval groups in the country.

Searches are performed by subject specialists with extensive academic training and industrial experience. Monica Nees, Director of Chemical-Biomedical Services, has a Ph.D. in organic chemistry and many years' experience in scientific information retrieval. Before the start of this project, she had done more than 40 computerized retrospective searches of the BIOSIS data base in conjunction with Mr. William Hoida of BIOSIS. Hannah Green, a Ph.D. in biochemistry, had five years of postdoctoral research experience before becoming an Information Specialist at NC/STRC. The two of us are responsible for all searches in chemistry, biology and medicine, and thus were chosen by BIOSIS to test their systems.

The 1974 BIOSIS data base, 240,000 references, was made available to NC/STRC for on-line searching. In the course of testing the BIOSIS search systems, we did a total of 100 literature searches for our clients from April, 1975 through March, 1976. Because of the experimental nature of the project, the searches were done without charge to the users. This report discusses our evaluation of the hardware and search systems, summarizes the

strategies used, analyzes the searches by type of end user, and gives our recommendations and conclusions. During the course of the project, the BIOSIS data base became commercially available for on-line searching via Lockheed's DIALOG system. Therefore, throughout this report we also compare the STAIRS and DIALOG programs for searching BIOSIS wherever appropriate.

II. Evaluation of the Search Systems

We used the search systems--STRATBLDR, CHEMFILE, STAIRS and the associated hardware--pragmatically as end-users would. The problems described in this section arose spontaneously; we did not seek them out. Because we had such a wide variety of topics, ranging from field biology to chemistry and biomedicine, we were able to study the search systems and indexes much more thoroughly than we could have if the questions had been concentrated in only a few areas.

A. Hardware

The cathode ray terminal used was the IBM model 3275. Overall, we found its operation easy and convenient. The Program Function keys are a particularly attractive, time-saving feature. But the blinking lights on the right front of the terminal were annoying. Because they showed the status of the system, and we had frequent system failures, they were not covered up. For better human engineering the lights should be moved to the side of the terminal.

The terminal was hard-wired from the Research Triangle Park in North Carolina to the BIOSIS computer in Philadelphia. Throughout the experiment, but especially in the first few months, we experienced frequent, prolonged down time, extending several times to many days' duration. The true cause was seldom made known to us but it was usually attributed to our modem (whichever of the many makes and models was then attached).

During the first half of the experiment we did not have an associated printer. Detailed notes of the search strategy were taken by hand. All output had to be printed off-line and mailed. Unfortunately, not much changed after the arrival of the printer, an IBM model 3284. Because signals were being transmitted to the screen at 4800 baud and the printer operates at 400 baud, simultaneous printing could not take place. Thus, a screen at a time had to be copied. The printer was very slow, requiring approximately 65 seconds to cover the entire screen line by line. And the entire screen had to be scanned character by character, even if only the top line contained printing. An end-of-print signal should be incorporated to save time and paper and eliminate the scanning of a blank screen. The printer was used primarily for obtaining a record of the final search strategy or to obtain a few highly relevant references. Hand-taken notes were still necessary. A slower transmission rate with a slave printer would be far preferable to the configuration we used.

It would seem that dial-up access would be preferable to hard-wired, with its high fixed monthly expenses in dedicated equipment and telephone line rentals. However, we would then have been pushed by the clock and would not have felt as free to explore the file or experiment with lengthy, complex strategies. We could not have luxuriated in prolonged browsing, which was the key factor in learning the intricacies of the data base. We

estimate conservatively that the terminal was in use an average of two hours per day for everything from system debugging to actual searching. The commercial dial-up rate for BIOSIS is $75.00 per hour, including line charges. Figuring 200 working days a year, the cost would be $30,000, about three times our hard-wired expenses.

A most valuable tool was the toll-free 800 number which BIOSIS had installed to service their many subscribers about the same time we began the experiment. We utilized it heavily--for assistance on everything from hardware and system crashes to strategy design. An 800 number cannot be urged strongly enough whenever off-site system debugging is undertaken. However, if the system had been thoroughly checked out, both with respect to hardware and software, before we began to use it, the 800 number would not have been used as extensively.

In retrospect, we would have benefitted greatly if BIOSIS personnel had given us a short formal training program at our location once the hardware was installed and functioning. As it was, we plunged into what turned out to be an undebugged search program with only the manuals to guide us. In the early months of the project, we attributed to our inexperience problems which in actuality were those of the hardware and software. Had we been trained on the system, rather than merely being self-taught, we would have been able to troubleshoot more effectively, and with much less frustration.

B. STRATBLDR

STRATBLDR (1), designed to assist in BIOSIS search strategy preparation, was tested in a one year (1974) segment of the BIOSIS file. Defi-

ciencies in STRATBLDR became rapidly apparent, and the use of STAIRS directly was soon substituted in literature search procedures. This section describes the problems with STRATBLDR and outlines alternatives for preparing effective strategies.

Many words and phrases essential for defining search parameters and existing in the BIOSIS file are not a part of STRATBLDR. This results in imprecision and an increased likelihood of omission of useful search terms. Plural forms are ignored. Most problematic are the lack of right truncation and an adjacency operator. For example, (toxin OR poison) AND (fish OR shellfish) produced 23 citations when executed on STAIRS. But (toxin$ OR poison$) AND (fish$ OR shellfish$) retrieved 40 citations.

A search on the effects of aspirin with laxatives could not even be initiated on STRATBLDR. The term aspirin led to the additional terms acetyl and salicylic. The user was then faced with the choice of using salicylic alone or linking acetyl to salicylic with an AND operator. Neither approach is as precise as an adjacency. There was also no way to express the synonym salicylate because it is not a vocabulary word and truncation (at salicyl$) is not possible. Finally, STRATBLDR rejected the terms laxative and cathartic which also were not vocabulary words. In almost all searches begun on STRATBLDR, the lack of truncation and the inability to express adjacencies made it difficult and often impossible to list necessary terms. Vocabulary deficiencies compounded user frustration.

It was not possible to use Cross codes correctly, because code categories were inconsistently and incompletely selected through STRATBLDR. In one search, selection of the phrase "sense organs" in STRATBLDR resulted in Cross code C20001$ when executed in STAIRS. This retrieved only the

citations indexed to General; Methods, in this case the least relevant category. The main interest was Pathology, C20006$. Obtaining all Sense Organs codes (C2000$) would have been preferable. A similar situation arose when selection of Virology-C in STRATBLDR led only to C33502$, the General; Methods subsection of C33500, Virology, General. Further inconsistencies appeared in another search on food preservatives. "Food technology" was selected in STRATBLDR and resulted once more in only one subsection of C13500, Food Technology (non-toxic studies), the General; Methods group C13502$. However, the STRATBLDR phrase "food processing" was specific and did lead to the most appropriate subsection, C13532$, Preparation, Processing and Storage. Twenty-seven relevant codes are listed under food in the printed guide Subject Guide to Cross Index (2). All these possibilities would have to be presented by STRATBLDR for correct Cross code utilization.

Lack of referral to Biosystematic code is very misleading. Selecting the term "human" does not result in a compilation of all citations indexed to Biosystematic code S86215 but only to those few that included the term human in the title or added keywords. Why doesn't STRATBLDR coach with: Human--Use S86215? When the term "algae" is selected, STRATBLDR informs the user to "use also specific names". Here again a list of Biosystematic codes as well as algae names is needed.

Consequently, when transferring the strategy designed on STRATBLDR to STAIRS by the Execute command, only a very incomplete list of documents is generated. The user must laboriously again go through all the manipulations of collecting relevant codes and terms, having gained relatively little from the STRATBLDR experience. It became evident that far better strategies could be developed more rapidly by accessing STAIRS directly

after preliminary preparation with printed guides (2) for vocabulary, Cross and Biosystematic codes.

The mechanics of STRATBLDR searching are inefficient. Terms must be selected one at a time, even when a group of related terms is displayed by the system and the user wishes to use all of them. Here Select and Combine commands would be most desirable. A second cumbersome manipulation is the ordering of search terms followed by the use of commas to place terms in logical groups. Again a Combine command, where terms could be grouped directly either by name or search term number, is preferable and less likely to produce errors. A final STRATBLDR limitation is its maximum capacity of three lines of grouped terms. A somewhat complex strategy or even a simple one utilizing a number of Cross codes easily surpasses this limit, and the search cannot be transferred to STAIRS.

Our conclusion is that it is much more effective to develop search strategies entirely in STAIRS, rather than using STRATBLDR and transferring the incomplete strategies to STAIRS. Initial review of the Subject Guide to Cross Index, Cross Code, Biosystematic Code, and A Guide to the Vocabulary of Biological Literature (2) is essential. This preliminary search preparation is more thorough and requires far less time than an average STRATBLDR session where the search mode selection of terms, one at a time, is slow and tedious.

C. CHEMFILE

A chemical dictionary is an immensely helpful search aid. Even the most experienced chemists rarely know all the synonyms for a given compound. Unfortunately, one defect in CHEMFILE greatly decreases its

effectiveness. All compound synonyms are printed in a line with no punctuation between terms. See Figure 1. As a result it is frequently difficult to find where one name ends and the next begins. Each term list required careful study, because errors resulting from linking the latter part of one name with the first part of the next were a distinct possibility. The insertion of slashes, semicolons or other appropriate delimiters between synonyms is essential.

Figure 1.

CHEMFILE Printout for the Pesticide "Sevin"

```
0001982
ACC NUM0001982
REG NUM000063252
MOL FORNO2C12H11
CAS TYPCARBAMIC ACID, METHYL-, 1-NAPHTHYL ESTER
SYNS    ENT-23,969 CARBARYL CARPOLIN COMPOUND-7744
        EXPERIMENTAL-INSECTICIDE-7744 GAMONIL METHYLCARBAMIC ACID,
        1-NAPHTHYL ESTER N-METHYL-1-NAPHTHYL CARBAMATE
        N-METHYL-ALPHA-NAPHTHYLURETHAN 1-NAPHTHOL, METHYLCARBAMATE
        1-NAPHTHOL N-METHYLCARBAMATE ALPHA-NAPHTHYL-N-METHYLCARBAMATE
        1-NAPHTHYL-N-METHYLCARBAMATE 1-NAPHTHYL-N-METHYLCARBAMATE SEVIN
        UNION CARBIDE-7,744 ARYLAM
WLN     $L66J BOVM1
```

D. STAIRS

In general, STAIRS proved to be a flexible and easy-to-use search system for BIOSIS. Although a few features are slow and cumbersome, most are extraordinarily effective. This section will describe the disadvantanges and advantages of STAIRS for literature searching.

Two improvements would be desirable in the Search mode when truncation is used. If the system could default to the WORDS (title and added

keywords) paragraph, lengthy and consequently slow expansions including authors' names would be avoided. A Select command for choosing appropriate terms from the list generated by truncation would increase relevance as well as decrease the number of items to be processed. We circumvented both of these problems by intersecting a truncated term directly with a previous group whenever possible. That way the expansion is shown only once, not twice. We also quickly learned to avoid certain words. For example, it is faster to search "acid OR acids" than "acid$," which results in three pages of terms.

Though the listing of all words produced by truncation is sometimes helpful, it is not necessary. In cases of doubt about truncation, the Root command can be used. The DIALOG system does not display the variants arising from a truncated term, but optional term expansion can be used to obtain alphabetically related words if necessary.

A special type of truncation for combining the primary and secondary Cross code levels would be extremely valuable. These two code levels represent the topics of major emphasis in BIOSIS references. Grouping them together is common during searching. The DIALOG system has two Cross code options: all levels or first and second levels combined. However, limiting to primary codes alone is sometimes desirable. Because it takes two to three minutes to process approximately 100,000 postings on STAIRS (quite a slow program), terms with high postings should be placed as late as possible in the search strategy sequence. For instance, the Biosystematic code S86215 (for human) has 74,652 postings for 1974. Whenever possible, it was used only in the very last intersection in order to minimize the processing time.

The Save and Execute commands were disappointing and were rarely used. During the Execute phase, every single search statement, including lengthy expansion of terms, is repeated and processed by the computer. This was extremely slow, occasionally requiring as much as 15 minutes. In most cases, re-doing the search was more efficient, because hindsight had improved the search strategy and shortened the execution time.

We found the Change command convenient for shifting to other data bases. It circumvented signing off and then signing on again for the new file.

The Biosystematic and Cross codes are not preceded by S or C respectively in the current BIOSIS files (DUCA or DUCI). This can lead to serious problems when the numbers are the same. For example, 064$ will retrieve papers with Cross code 06400 for Subterranean Biology as well as papers mentioning Beggiatoales, an organism with the bacteria Biosystematic code 06400. There are numerous other Cross and Biosystematic code identities.

The Display command was useful and flexible; one could display all or only some previous search statements. Displaying the number of documents for each term is sufficient for most strategy design. Listing the number of occurrences is redundant, because the difference between occurrence and document postings is usually insignificant. The term canine, for example, has 1052 occurrences in 1047 documents. Simply posting 1047 before canine (1047 canine) would be sufficient. We feel that occurrence data should be eliminated in both the Display and Search modes.

The Purge feature is attractive. Errors or search statements that become irrelevant can easily be eliminated. However, the user must be

careful not to purge statements that will become part of a subsequent operation.

An excellent feature of STAIRS is the rapid Browse. This was immensely helpful for obtaining pertinent synonyms, related terms, Cross and Biosystematic codes, and even genus and species of organisms for which only the common name is known. Retrieving and browsing a known, highly relevant citation and checking its keywords and codes for use in search strategy design is a common, helpful tactic. The Browse format was good for rapid scanning, and the highlighting of untruncated keywords or codes used to retrieve the citation aided greatly in speedy evaluation. Truncated terms should also be highlighted for optimum ease of review.

The fast Browse capability of STAIRS was so useful that we used STAIRS/BIOSIS as a model file for literature searches in other files. After an initial search in BIOSIS, additional terms and appropriate synonyms were found, words causing excessive noise and low relevance were rejected, and the Cross and Biosystematic codes used to index relevant papers sometimes suggested additional fruitful search approaches. Strategies developed on STAIRS were then adapted for Food Science and Technology Abstracts (FSTA), Chemical Abstracts Condensates (CAC) and CAIN.

Several errors and ambiguities were found in the BIOSIS/STAIRS user manual. On page 11, the first three search statements are all numbered 00001; they should be 00001, 00002 and 00003. The Root command, described on page 14, is confusing. The user wonders how the system distinguishes "root smok" for obtaining a list of smok$ forms: smoked, smoker etc. from "root smok" for a selection of the two terms, root or smok. In Appendix I, DUCI is erroneously called DUCB. And in Appendix IV, the best way to

retrieve authors when only the first initial is known is not listed. 'Smith A'$ would retrieve Smith A, Smith A B and Smith A C. This is preferable to 'Smith A '$, suggested in the manual, which retrieves only entries with two initials. Also, in the example for searching on murine blood neoplasms in Appendix IV, there is no need to use Biosystematic code S86375 when the restricting terms mice, mouse and murine are used. In Appendix V, the manual should stress that both the singular and plural of drug (drug$) should be used when searching for drug affiliations. For example, "anti ADJ neoplastic ADJ drug" has 1013 occurrences but "anti ADJ neoplastic ADJ drugs" has 808.

The user manual for BIOSIS/STAIRS would definitely benefit from an increased number of sample searches that illustrate various features of BIOSIS retrieval capabilities. Summary sheets of commands from Sign-On to Sign-Off would also be helpful. The appendices are an especially important and valuable section of the manual, because familiarity with BIOSIS editorial procedures is essential for thorough literature searching.

III. Summary of Search Strategies

Table 1 summarizes the various types of strategies used in the 100 searches which comprised this experiment. A complete list of search titles appears in Appendix A.

Table 1

Summary of Search Strategies

Strategy	No. of Searches
WORDS[a] (alone)	34
WORDS (as one of several parts)	30
WORDS-CROSS[b]	40
WORDS-BSYST[c]	23
WORDS-CROSS-BSYST	3
Not Involving WORDS*	8

[a]WORDS terms from authors' titles and from keywords added by BIOSIS indexers

[b]CROSS CROSS Code index

[c]BSYST Biosystematic index

*Includes CROSS-BSYST, CROSS-CROSS, and CROSS-CROSS-BSYST

The column headed "No. of Searches" totals more than 100 because several approaches were often used on one question. This is not a summary of all the ways tried, but rather of those which yielded results sent to the users. For instance, assume the use of WORDS alone generated a set, and that this set was then intersected with CROSS. If, after this intersection, it was then decided to send only the original WORDS set, this search would be tallied under WORDS, not WORDS-CROSS. In turn, WORDS can represent either a dump of appropriate term or terms, or the intersection of terms with each other.

Of the 100 requests, 34 were answered by WORDS alone. An additional 30 used WORDS alone as part, but not all, of the search. Only eight

searches did not employ the use of WORDS.

Two subsets of the WORDS index, either by themselves or in combination with CROSS or BSYST, were especially useful. Genus-species was employed in seven searches, frequently with as many as a half dozen such entries in each. Geographical location, primarily USA or North Carolina, was used in seventeen searches.

Although WORDS is shown by the above statistics to be a very useful index, these same statistics show that CROSS or BSYST were necessary 74 times. Yet these are the indexes often ignored in manual searching because they are so cumbersome to use manually. Indeed, many individuals who have done manual searches in Biological Abstracts are unaware of their existence. If these same manual searchers can be shown the utility of CROSS and BSYST, they should become eager converts to computerized searching of BIOSIS.

Subsequent sections of this report will concentrate on the types of search questions for which each of the indexes (WORDS, CROSS and BSYST) are most appropriate.

IV. Analysis of Searches by Type of End User

The previous section discussed various types of search strategies used in the project taken as a whole. This one explores in depth the unanticipated observation that the usage of the three major indexes--WORDS, CROSS and BSYST--showed definite patterns which could be related directly to the type of user for whom the search was being performed. Users are grouped into four categories: regular NC/STRC clients, North

Carolina state agencies, MEDLINE operators, and miscellaneous requests. Searches for MEDLINE operators heavily utilized WORDS; those for state agencies both BSYST and genus-species, a subset of WORDS. No major trends were discernible in either the regular client or miscellaneous searches. In this section we will also discuss the strengths and weakness both of the indexes and the STAIRS program used to search them, and will make recommendations for their improvement.

A. Searches for Regular Clients

Of the total 100 searches, 42 were performed for a selected group of our regular clients. These were primarily chemical and pharmaceutical companies who have been using our services regularly for several years. Some research institutes, universities, and a few of our non-chemical clients were also included. Because these regular clients are our major source of income, it was only natural to concentrate on their needs. They were very grateful for this added (and free) service.

1. WORDS only

Approximately one-third of the searches for our regular clients employed only the WORDS index. Other approaches were often tried on these questions, but the output sent to the users was retrieved from the WORDS alone. Table 2 summarizes these "WORDS only" searches.

Table 2

WORDS Only Searches

	Title	Hits
1.	5-Fluorouracil	168
2.	Toxicity of Boron Trifluoride	4
3.	Aphids on Certain Fruits	17
4.	Methods of Increasing the Compatibility of Atrazine with Fertilizers	8
5.	Everything on the Chesapeake Bay	27
6.	Dredging	17
7.	Collagen as a Support for Immobilized Enzymes	8
8.	Changes in Flour Protein during Dough Mixing	24
9.	Single-Cell Protein	6
10.	Odor Control of Tobacco-Related Products	7
11.	Zinc Ricinoleate	4
12.	Functional Properties of Squid	22
13.	Artificial Soils	7
14.	Dehydroacetic Acid	4
15.	Toxicity of Textile Combustion Products	13
16.	Effect of Salts on Natural Vegetation in Freshwater Swamps	86

Because most of these searches were done in the first six months of the project, the "WORDS Only" approach may have been partially caused by a relative unfamiliarity with the data base. More frequently, though, it reflected our constant readiness to take a dump of everything on a topic, in order not to miss anything. This is a very powerful, often overlooked option which should always be considered, even in a large data base. Because we were working with a one-year file, we though of it as small. In reality, its 240,000 documents represent at least a medium-sized data base, in comparison with others available for computerized searching.

In certain cases, the value of a WORDS-only search cannot be overemphasized. It is especially useful in answering requests for "everything" on well-defined topics. For instance, a biomarine institute was building

a document collection on the Chesapeake Bay per se, and was not interested in searching on other aspects such as its tributaries or the animals and plants found in and around it. We merely retrieved everything on "Chesapeake Bay" from BIOSIS and several other files. The users couldn't have been happier. This type of question can be answered very rapidly and economically on-line.

The WORDS index is ordinarily the only one used in retrospective manual searching of Biological Abstracts and BioResearch Index. Indeed, it is the only one most users know about! Utilizing CROSS codes or the Biosystematic indexes manually is almost hopeless. Initially we were skeptical about searching only by augmented titles. But after our experience with STAIRS, we feel much better about WORDS for certain topics. We are very impressed with the quality, consistency and value of the augmented keywords.

After we became more familiar with the file, we sometimes forgot the dump technique and got too exotic in our initial approach to a question. We became ensnarled in complex intersections of various levels of several CROSS codes and then had to back off and return to a more straight-forward strategy.

Our very first search was for everything on 5-fluorouracil, to up-date a multi-file search done about a year earlier. With only a few synonyms for 5-fluorouracil, this topic seemed an ideal starting point. It wasn't! We were, of course, aware of the BIOSIS policy on fragmentation of terms to allow additional access points. Therefore fluorouracil was entered as "fluoro ADJ uracil". For comparison, "fluoro AND uracil" was also used. The unexpectedly large discrepancy in the two postings led to immediate

discovery of a major defect in the search program: an adjacency is not read if its first term is at the end of one line and its second at the beginning of the next. (See Appendix B for details). Several other word-pairs were also tested; we estimate that this defect causes a loss of at least 5 to 10% of the relevant documents. Therefore, in all searches where high recall was especially important, intersections had to be used rather than adjacencies. This often led to a lot of unnecessary noise. We understand that this problem will be corrected the next time the file is loaded on STAIRS. DIALOG does not have this bug--it does read adjacencies "around the corner".

Even aside from the adjacency problem, we found word fragmentation to be much more of a hindrance than a help. Because the guidelines for fragmentation are not always clear, we often had to use both fragmented and unfragmented terms for safety's sake, both on STAIRS and DIALOG. For example, in a search on lithium diiodosalicylate, the term diiodo appeared four times while di ADJ iodo had 32 postings. This dual entry can run up search costs considerably. Consequently we urge that consideration be given to minimizing word fragmentation in the future.

2. (WORDS-CROSS or WORDS-BSYST) and/or WORDS

This section represents a more typical group of questions. Table 3 summarizes the searches.

Table 3

(WORDS-CROSS or WORDS-BSYST) and/or WORDS Searches

	Title	Hits
1.	Formulation of Pesticides	291
2.	Carrageenan Interaction with Proteins	12
3.	All Drugs Administered Rectally	50
4.	pH Treatment of Fish	13
5.	Functional Properties Related to Meat, Fish and Poultry	286
6.	Effect of Urea-Formaldehyde or Formaldehyde on the Olfactory System	14
7.	Dredging	2
8.	Nitrification and Denitrification in Sewage Disposal	61
9.	Toxicology of Coumarins	24
10.	Effects of Chelated Zinc on Wheat and Barley	6
11.	Diazinon in Pest Control for Dogs and Cats	4
12.	Migration and Nesting Patterns in Hawks, Eagles and Storks	169
13.	Birds in North Carolina	13
14.	Effect of Vehicle and Route of Administration on Pesticides and Drugs	76
15.	Histamine and Cotton	3

Terms from WORDS were intersected either with CROSS or BSYST to increase the relevancy of the output. (Several of these searches also had an additional section satisfied by WORDS only.) The intersections were not always a simple one-two process. Several CROSS or BSYST codes were often employed. Frequently, we found that the tertiary level of CROSS introduced far more noise than was tolerable, so that the search then had to be limited to the primary and secondary levels. We strongly recommend that in the future the tertiary levels be much less highly posted.

Unfortunately there is no truncation symbol for primary and secondary levels combined, so each had to be entered separately. This was a laborious process for the larger categories. For instance, to cover all of the categories under C60000, Economic Entomology (frequently needed

for our agrochemicals searches), twenty-two entries are required. These could be condensed into just one, e.g., C600¢, if ¢ were the symbol for primary and secondary levels combined. This highly desirable feature exists in DIALOG via the /MAJ delimiter.

But DIALOG lacks the capacity of searching either by primary or by secondary levels. This is a distinct disadvantage in the few cases where it is necessary to search or negate only a primary level. However, we rarely limited an output to the primary level; too much would be missed if the secondary level were not also included.

The CROSS Code manual and the Subject Guide to Cross Index (2) were the most frequently consulted search tools. Never did we go to the terminal without the CROSS Code manual. Even now, as experienced BIOSIS searchers, we do not depend on DIALOG at all for CROSS codes. The on-line coaching is both inadequate and too expensive. It takes more than two minutes to get an expansion of "CC=600?". And with an expansion you cannot be sure of being led to appropriate CROSS codes. Why, for instance, on an expansion of "econom?" is there no pointer to any of the CROSS codes for Economic Entomology?

Inexperienced users may be lulled into false confidence by just the example given on page R-4 of the DIALOG search manual. An expansion of "CN=Food Tech" gives no indication that there are 15 related terms, though the standard notation "-MORE-" at the bottom of the expansion indicates that additional relevant terms may follow. Perhaps others beside the one selected, "E10 CN=Food Tech--Evalns, Phys, Chem" would be equally or even more appropriate. In contrast, the expansion shown on page R-5 is much more informative. Item E7, "Neoplasms, Neoplastic Agents" is explicitly

shown to have nine related terms, which can then be displayed as illustrated.

Two of our searches, "All Drugs Administered Rectally" and "Effect of Vehicle and Route of Administration on Pesticides and Drugs", employed C22100, the CROSS code for Routes of Immunization, Infection and Therapy. We suggest that immunization be removed and given a separate code. In these searches it caused a very large number of false drops.

3. More Complex Strategies

This was perhaps the most interesting group of questions because several strategies were used on each. Table 4 summarizes the topics.

Table 4

More Complex Search Strategies

	Title	Hits
1.	Health Effects of Dietary Roughage	24
2.	Effect of Vitamin E on Aging and Blood Clotting	57
3.	Pyrethroids, Formamidines and Amidines	112
4.	Effect of Sawdust on Humans	19
5.	Biosynthesis of Alcohols and Related Compounds	117
6.	Systemic Fungicides for Cereal Crops	164
7.	Sarcoptic Mange in Dogs	13
8.	Destruction of Mycobacteria with Chemicals Other than Chemotherapeutic Drugs	23
9.	Carp as Meal for Protein Supplement	0
10.	Economic Impact and Control of Five Insect Pests	207
11.	Anti-Protozoal Vaccines Against 12 Organisms	656

At least two different approaches (in addition to WORDS alone) were employed. The most common were WORDS-CROSS along with WORDS-BSYST, but

other combinations, including CROSS-BSYST and WORDS-CROSS-BSYST, were also productive.

The search on "Anti-Protozoal Vaccines against 12 Organisms" was one of the earliest and also one of the most difficult because of problems with the search program. The output was being segmented by genus and, whenever possible, also by species. During the course of this very long search, a second, very critical defect in the search program was discovered. It is illustrated as follows:

Table 5

Sample of Search Program Error

Search Statement No.	Terms	No. of Documents
1	sodium	4196
2	chloride	2817
3	sodium and chloride	548
4	1 and 2	599
5	4 not 3	51

Thus the total hits resulting from the intersection of the terms A and B are not the same as from the intersection of the search statement number for A with the search statement number for B! All 51 of the "residual" documents did, in fact, contain both sodium and chloride.

This problem was first discovered using the genus-species approach shown below:

Table 6

Additional Sample of Search Program Error

Search Statement No.	Terms	No. of Documents
1	Toxoplasma	132
2	gondii	104
3	Toxoplasma and gondii	104
4	Toxoplasma ADJ gondii	63
5	3 not 4	41
6	(Toxoplasma and gondii) (not Toxoplasma ADJ gondii)	0

The difference of 41 in search statement 5 was suspicious because from previous work on this question, it appeared that all, or nearly all Toxoplasma were, in fact Toxoplasma gondii.

It was first thought that this represented another example of the adjacency problem, but it didn't. All 104 documents in search statement 3 were examined, and all were indeed Toxoplasma gondii. No document had the genus and species terms split between lines. All were written Toxoplasma-gondii (T-g), in either title or keywords, and thus should be searchable as T ADJ g.

Many additional word-pairs were tested; similar discrepancies were found in some cases but not in others (See Appendix B for details.)

No satisfactory explanation for these discrepancies was ever forthcoming. However, the problem "mysteriously" vanished upon an IBM release of a new STAIRS program a few weeks later. From then on, the search program was tested frequently to see if these or any new bugs had crept in. Fortunately, they didn't. We cannot urge strongly enough that similar tests be performed on all systems, not only upon installation, but also throughout their use. It is obvious that the STAIRS program had not been

adequately tested for use in searching before it was released to us, or the adjacency problem and this one would have been discovered and presumably solved. A searcher can take very little for granted, least of all the search program!

B. Searches for State Agencies

NC/STRC is part of the Division of Economic Development of the Department of Natural and Economic Resources (NER) of the state of North Carolina. Therefore it was appropriate to offer free BIOSIS seraches to state agencies, though it was unclear at first exactly how this could best be accomplished.

Afraid of being inundated with requests, we decided to concentrate primarily on other agencies of NER. We are both geographically and organizationally quite isolated from other sections of NER and know very few staff members personally. Consequently, initiating the project was not easy.

1. NER Teasers

After obtaining an NER organization chart, twenty sample searches (teasers) were prepared, based only on the brief descriptions of activities listed on the chart. Table 7 summarizes the teasers.

Table 7

NER Teasers

	Title	Hits
1.	Artificial Reefs	5
2.	Land Use Management	11
3.	Research Vessels	10
4.	Environmental Impact Reports and Studies	26
5.	Reclamation and Mines	4
6.	Fish Hatcheries	10
7.	Wildlife Habitats	14
8.	Terrestrial Wildlife Management	1
9.	Aquatic Wildlife Management	3
10.	Oceanography and Limnology	7
11.	Animal Ecology	8
12.	Plant Ecology	9
13.	Forestry and Forestry Products	16
14.	Pest Control and Economic Entomology	16
15.	Air, Soil and Water Pollution	8
16.	Birds	13
17.	Tobacco	4
18.	Shrimp or Shellfish as Food	10
19.	Fillets as Food Products	6
20.	Water Research and Fishery Biology	25

No attempt was made to give a thorough search, but rather a brief, highly relevant illustrative one. Names and addresses were matched with appropriate output, and the teasers were then sent off with a cover letter (see Appendix C) offering to do free searches.

Perhaps the most important factor in these teasers was the capability of searching by geographical location. More than half of the searches utliized the term North Carolina. Thus it was not merely "Forestry and Forest Products" but "Forestry and Forest Products in North Carolina". This rendered the output much more eye-catching.

The geographical area is often a most important parameter in field biology searches. In recent years BIOSIS, with its augmented terms, has made it possible to search by USA or by the individual states. But

it is difficult to search by broader regions, such as "southeastern United States". How does one search for "eastern European countries" without enumerating all of them? What about "Africa", whose countries are constantly changing? With increasing concern for the environment, geographical parameters will become even more important. We cannot urge strongly enough that this capability be emphasized by additional keywording.

2. NER Requests

Our previous experience with the BIOSIS file had been heavily oriented towards pharmacology, toxicology and agrochemicals. We needed a broader spectrum of topics, and with the teasers hoped to generate questions in areas such as field biology, ecology, and environmental studies. We were not disappointed! The twenty teasers led to eleven search requests summarized in Table 8.

Table 8

NER Requests

	Title	Hits
1.	Fish Ladders	28
2.	Mercury in Fish	90
3.	Viruses in Mollusks	51
4.	Some Aspects of Bluefish, Horseshoe Crab, Red Drum, Snapping Shrimp and Anemones	52
5.	Some Ecological Aspects of North Carolina	86
6.	Symbiosis in Aquatic Organisms	67
7.	Non-Point Source Pollution	181
8.	Vegetative Propagation of Hardwood Tree Species	56
9.	Effect of High Frequency Sound on Fish	53
10.	Containerized Tree Seedlings	14
11.	Stimulating Male and Female Flowering in Conifers	38

The divisions of Marine Fisheries and Forest Resources were the most active users. Some of the searches consisted of multiple, unrelated topics, and therefore the actual number of questions answered was far more than eleven.

Our teasers had done exactly what we hoped for: generated searches in parts of the file we had not used before. By far the most crucial index in answering these questions was the genus-species part of the WORDS file. It was not uncommon to employ more than a half dozen genus-species names in one search. It proved imperative to use both the genus-species as well as the common names. Although we have no hard statistics, it seems that both the genus-species and the common name for a given organism are used in only about half the documents. Thus using only one or the other would seriously affect recall. In many cases the use of the Biosystematic index would have generated too much noise. Sometimes even the genus was too broad. Thus we urge that even more attention be given to indexing as deeply as the species.

At the beginning of the BIOSIS project, we were concerned about possible problems in obtaining a Biosystematic code or genus-species name when only a common name was given. These fears proved groundless. We entered the common name and browsed on STAIRS until we found a document which had only one Biosystematic or genus-species entry. (This approach may not be economically feasible on DIALOG). For genus-species this was often faster than checking reference books or the unabridged dictionary.

One Biosystematic code proved frustrating: S85206, Osteichthyes, is too all-encompassing. It would be most helpful if this huge category of fishes were subdivided.

Of all our users, those from NER were the most faithful about filling out evaluation forms and giving feedback over the telephone. Several expressed real dismay when the project had to be terminated. We have built up a satisfied little poverty-stricken (state agencies) group of repeat users and would like to continue providing them free service. But that would necessitate finding a source from which to recover our out-of-pocket expenses incurred by searching the data base commercially. Is there a possibility of a joint project between us and BIOSIS in this area?

C. Searches Requested or Referred by MEDLINE Operators

For several years we have been doing MEDLINE searches through the Health Sciences Library of the University of North Carolina at Chapel Hill. The MEDLINE operators could not be more cooperative, but we felt that we were always the beneficiaries and could not give much in return. Thus we were delighted to be able to offer them free BIOSIS searches. Table 9 summarizes the MEDLINE searches.

Table 9

MEDLINE Searches

	Title	Hits
1.	Phaeomelanin	4
2.	Affinity Chromatography of DNA and Messenger RNA	24
3.	Neural Crests	18
4.	Gray Lethal Mice	13
5.	Cellulose in Tunicates	0
6.	Chalones	63
7.	Epstein-Barr Virus	205
8.	Isolation of Ribosomes from Rabbit Lymphocytes	7
9.	Cellulase in Termites	5
10.	Various Aspects of Cellulose	147
11.	Autoimmune Reactions in Fish Brain	57
12.	Mucopeptides, and Peptidoglycans in Bone Tissue Culture, Osteocytes or Cartilage	33

The MEDLINE operators came to us primarily when the search topic was "too biological" for their data base (a typical example is "Cellulase in Termites"). On other occasions they wanted to see the type of complementary material BIOSIS had on topics also suitable to MEDLINE ("Affinity Chromatography..."). MEDLINE introduced free-text searching of titles and abstracts in April, 1975, the same time we started this project. Were it not for this, we would have been used much more heavily. Solely with the controlled MeSH vocabulary, it would have been quite difficult to search topics such as "Gray Lethal Mice" and "Phaeomelanin". Here, free-text searching of title and keywords was essential.

In almost all cases we were able to find key references not retrieved in MEDLINE. The end users, primarily graduate students and faculty members, were most appreciative. To quote from the evaluation form on "Phaeomelanin": "Two of these references are exactly what I needed. I could not retrieve them from MEDLINE."

Therefore we have shown the utility of BIOSIS for biomedical questions. However, it is doubtful if BIOSIS will be used as much as it should be in this particular university environment. The MEDLINE operators would be happy to do BIOSIS on-line but feel that most of their users cannot afford it. Their MEDLINE charges are $24 per hour of connect time, including telephone line charges, and 10 cents per page off-line print (4 to 7 citations per page); BIOSIS via DIALOG is $75 per hour, including line charges, and 10 cents per citation. Only direct MEDLINE costs, but not staff time, are charged back to the end user. Many MEDLINE users, including students, pay the modest average cost of $4 to $8 per search out of their own pockets. They cannot afford a much more expensive

search unless it can be charged to a grant. With grant money becoming ever more difficult to obtain, the future of commercial on-line searching in a university environment is clouded. Library budgets are also feeling the pinch, and thus it is unlikely that libraries will be able to offer "free" on-line searching as an overhead item.

D. Miscellaneous Searches

This catchall category includes all searches which did not readily fit into any of the preceeding sections categorized by type of requester. Table 10 summarizes the searches.

Table 10

Miscellaneous Searches

	Title	Hits
1.	5-Fluorocytosine	61
2.	Avian Leukosis Virus	44
3.	Propranolol and Hypertension	37
4.	Anti-thrombins	57
5.	Streptokinase	88
6.	Enzyme and Protein Structure	184
7.	Glass Bead Chromatography	30
8.	Isolation of Histocompatibility Enzymes	52
9.	Magnetobiology and Magnetotherapy	148
10.	Various Aspects of Heparin	586
11.	Extraction of Proteins from Acrylamide Gels	20
12.	Detection of Hyperthyroidism in Humans	44
13.	Phosphoproteins in Viruses	46
14.	Psychiatric Aspects of Aging	187
15.	Effect of Sediment on Fish	41

The searches were performed for a wide variety of reasons, ranging from personal interest to marketing. Unlike the NER questions, for which certain indexes such as the Biosystematic and genus-species were heavily

used, there is no major trend apparent in this section. In fact, the Miscellaneous searches are very similar, both in topics and strategy design, to those done for our regular clients.

The most interesting feedback came from the recipient of "Glass Bead Chromatography", a university researcher "on top of everything" and somewhat skeptical of the value of computerized searching. He was both delighted and embarrassed at the number of highly pertinent references he was previously unaware of. Consequently he became such a convert that he ordered on-line searches of several commercially available data bases, including BIOSIS, even though he had to pay for them personally. We feel that money for computerized searching should be specifically requested in grant or contract applications. The data base suppliers, perhaps in conjunction with the funding agencies, need to do much more missionary work in this area than they have done in the past.

Throughout this report, we have compared STAIRS and DIALOG wherever appropriate. The last search in this Miscellaneous section, "Effect of Sediment on Fish", was specifically designed to test the most important features of BIOSIS searching: truncation of words and codes and limiting CROSS codes to primary and secondary levels. The identical results, 41 hits for the 1974 file, give us confidence in the loading of the BIOSIS file on DIALOG.

The strategy involved an A-B-C logic, where A included appropriate terms from WORDS; B was the truncated Biosystematic code for all Pisces and C the primary and secondary levels of the three CROSS codes for Oceanography and Limnology, Oceanography, and Limnology. The exact strategies used for comparison of the two programs are shown in the two tables below.

Table 11

STAIRS Strategy

Search Statement No.	Terms
1	sediment$ or detritus or bottom ADJ deposit or bottom ADJ deposits
2	S8520$
3	C07510* or C07510- or C07512* or C07512- or C07514* or C07514-
4	1 and 2 and 3

Table 12

DIALOG Strategy

Search Statement No.	Terms
1	sediment?
2	detritus
3	bottom(w)deposit
4	bottom(w)deposits
5	BC=8520?
6	CC=07510
7	6/MAJ
8	CC=07512
9	8/MAJ
10	CC=07514
11	10/MAJ
12	(1 or 2 or 3 or 4) and 5 and (7 or 9 or 11)

On STAIRS, truncation is possible with adjacencies, as in A ADJ B$. This desirable feature is not available on DIALOG. Therefore the concept "bottom deposit(s)", which ordinarily would be written in STAIRS as bottom

ADJ deposit$, was entered as bottom ADJ deposit and bottom ADJ deposits, to correspond to the DIALOG requirements. There is no essential difference in the way Biosystematic codes are handled by the two programs. To get only the primary and secondary levels of CROSS codes in STAIRS, each has to be entered separately, followed by the appropriate asterisks and hyphens. In DIALOG, all levels of the desired category are obtained first; then use of the /MAJ delimiter restricts retrieval to the first two levels. (Later it was learned that all codes could have been combined into one search statement, and then limited as a group by /MAJ, rather than one at a time.)

Based on these and other comparisons, we feel certain that the searches we did on STAIRS could be done just as effectively on DIALOG, and therefore are confident of our ability to search BIOSIS on the commercially available system.

V. Search Evaluation Form

With every one of our searches we sent a detailed cover letter discussing the rationale of the search strategy and commenting on the results. A cover sheet (see Appendix D) explaining abbreviations on the printout was included. A Search Evaluation Form was also enclosed with most output except that for the NER teaser searches. A copy of the Form is shown in Appendix E. Only 24 forms, about a third of the number sent, were returned. The NER requesters, however, sent back every one.

The first question, designed to measure relevancy, had evidentally been poorly constructed. We entered the total number of citations retrieved, and wanted the users to break down that figure (by number or

percentage) into the three categories listed: highly relevant, somewhat relevant, not relevant. Only six did so. Most simply checked one of the three categories.

Replies to the second question were the most useful to us. We often deliberately include even very peripheral material on all searches (not just BIOSIS) and wanted to get a feel for the users' reaction to receiving it. Of the 24 responding, 15 checked "glad to get it", seven "could take it or leave it" and two "would rather not get it". Consequently, we will continue to send peripheral material, but always, as before, clearly labelled as such.

The third question dealt with key references which we may not have retrieved. Eleven users were not aware of any; ten left the question blank; only three said references were missed--but one gave no particulars. The other missed references either were from journals not covered by BIOSIS or covered ones issued too late in 1974 to be included in the data base for that year. Therefore we are quite pleased at the thoroughness of our retrieval.

The overall evaluation of the searches was as follows: ten were "very useful", eleven "somewhat useful" and the remaining three "not very useful". These results were not at all surprising. Whenever we could find little or nothing on the exact topic (which happened more frequently than we had originally anticipated), we always sent peripheral material. For example, on Effect of High Frequency Sound on Fish, the specific interest was only in the use of such sound to herd fish into nets. There was not a thing on this particular aspect. The user's comment was "Thanks for trying--". The same user had earlier asked for a search on Fish Ladders, on which we found nothing. Commenting on the peripheral material sent, he

said: "7 of the (28) citations were of interest, although they had nothing to do with fish ladders. There were concerned with the desired species".

The following is a potpourri of comments from other recipients: on Methods of Increasing the Compatibility of Atrazine with Fertilizers: "This was extremely useful". On Epstein-Barr Virus: "Liked better than MEDLINE. Not clinical." On Viruses in Mollusks: "I think there's one reference missed....In addition there were several I had missed." On Impact and Control of Five Insect Pests: "...needed more information on economics, although this was no fault of the search...(helped) the requestor to know what was available and formulate questions for the next search which is attached". On Mercury in Fish (in North Carolina): "Results ...were not overly beneficial...because there has been very little published in this area". And, finally, on Functional Properties of Squid: "...It leads me to believe that I'm working in virgin area.... Thank you again".

VI. Conclusions

A. With respect to the search systems:

1. We have demonstrated that it is absolutely essential for computer programs and search systems to be tested by end users in a normal environment. Many of the problems we encountered had not been discovered during in-house usage. Therefore we strongly recomment that before any search system is released for general use it first be tested and debugged by knowledgeable off-site end users.

2. STRATBLDR was found to be essentially useless as an aid to searching BIOSIS.
3. CHEMFILE was difficult to use because of the streaming together of chemical synonyms.
4. STAIRS, though slow, was found to be very effective in searching BIOSIS.
5. Identical results can be obtained on both STAIRS and DIALOG.
6. Dial-up access is not necessarily less expensive than hard-wired.

B. With respect to the BIOSIS data base and the indexes:

1. BIOSIS was found to be very responsive to answering a wide variety of search questions.
2. WORDS was used in almost all of our 100 searches. We were impressed with the quality and consistency of the augmented words. Geographical and genus-species terms should be emphasized even more strongly because of their importance for environmental questions.
3. CROSS is by far the most powerful index, yet it is almost never used by manual searchers.
4. BSYST is valuable for limiting output to a specific organism or group of organisms. The code for human was used most frequently.
5. Almost two-thirds of our searches utilized CROSS and/or BSYST. BIOSIS personnel and the on-line vendors must educate users about the utility and strengths of these indexes before truly effective on-line searching can be achieved.

We are indebted to the many BIOSIS personnel who helped and encouraged us, especially Louise Schultz, project supervisor; Patrick Lawrence and John Thomas for assistance with all problems involving the search programs and hardware; Joanne Howard for detailed documentation; and William Hoida for help on strategy design and file structure. We also thank Peter J. Chenery, our director, for his enthusiastic support of this research.

VII. References

1. Schultz, L. "Breaking the Communication Barrier Between Searcher and Literature File: An Interactive Guide", Journal of the American Society for Information Science; (No. 1): 3-9 (1974)

2. Profile Guide; Content Guide; Subject Guide to Cross Index; Cross Code; Biosystematic Code; A Guide to the Vocabulary of Biological Literature. BioSciences Information Service of Biological Abstracts, Philadelphia, Pa.

Appendix A

Summary of Search Titles

and Number of Hits Retrieved

Note: The sequence is identical with that in the main body of the report.

	Title	Hits
1.	5-Fluorouracil	168
2.	Toxicity of Boron Trifluoride	4
3.	Aphids on Certain Fruits	17
4.	Methods of Increasing the Compatibility of Atrazine with Fertilizers	8
5.	Everything on the Chesapeake Bay	27
6.	Dredging	17
7.	Collagen as a Support for Immobilized Enzymes	8
8.	Changes in Flour Protein during Dough Mixing	24
9.	Single-Cell Protein	6
10.	Odor Control of Tobacco-Related Products	7
11.	Zinc Ricinoleate	4
12.	Functional Properties of Squid	22
13.	Artificial Soils	7
14.	Dehydroacetic Acid	4
15.	Toxicity of Textile Combustion Products	13
16.	Effect of Salts on Natural Vegetation in Freshwater Swamps	86
17.	Formulation of Pesticides	291
18.	Carrageenan Interaction with Proteins	12
19.	All Drugs Administered Rectally	50
20.	pH Treatment of Fish	13
21.	Functional Properties Related to Meat, Fish and Poultry	286
22.	Effect of Urea-Formaldehyde or Formaldehyde on the Olfactory System	14
23.	Dredging	2
24.	Nitrification and Denitrification in Sewage Disposal	61
25.	Toxicology of Coumarins	24
26.	Effects of Chelated Zinc on Wheat and Barley	6
27.	Diazinon in Pest Control for Dogs and Cats	4
28.	Migration and Nesting Patterns in Hawks, Eagles and Storks	169
29.	Birds in North Carolina	13
30.	Effect of Vehicle and Route of Administration on Pesticides and Drugs	76
31.	Histamine and Cotton	3

	Title	Hits
32.	Health Effects of Dietary Roughage	24
33.	Effect of Vitamin E on Aging and Blood Clotting	57
34.	Pyrethroids, Formamidines and Amidines	112
35.	Effect of Sawdust on Humans	19
36.	Biosynthesis of Alcohols and Related Compounds	117
37.	Systemic Fungicides for Cereal Crops	164
38.	Sarcoptic Mange in Dogs	13
39.	Destruction of Mycobacteria with Chemicals Other than Chemotherapeutic Drugs	23
40.	Carp as Meal for Protein Supplement	0
41.	Economic Impact and Control of Five Insect Pests	207
42.	Anti-Protozoal Vaccines Against 12 Organisms	656
43.	Artificial Reefs	5
44.	Land Use Management	11
45.	Research Vessels	10
46.	Environmental Impact Reports and Studies	26
47.	Reclamation and Mines	4
48.	Fish Hatcheries	10
49.	Wildlife Habitats	14
50.	Terrestrial Wildlife Management	1
51.	Aquatic Wildlife Management	3
52.	Oceanography and Limnology	7
53.	Animal Ecology	8
54.	Plant Ecology	9
55.	Forestry and Forestry Products	16
56.	Pest Control and Economic Entomology	16
57.	Air, Soil and Water Pollution	8
58.	Birds	13
59.	Tobacco	4
60.	Shrimp or Shellfish as Food	10
61.	Fillets as Food Products	6
62.	Water Research and Fishery Biology	25

	Title	Hits
63.	Fish Ladders	28
64.	Mercury in Fish	90
65.	Viruses in Mollusks	51
66.	Some Aspects of Bluefish, Horseshoe Crab, Red Drum, Snapping Shrimp and Anemones	52
67.	Some Ecological Aspects of North Carolina	86
68.	Symbiosis in Aquatic Organisms	67
69.	Non-Point Source Pollution	181
70.	Vegetative Propagation of Hardwood Tree Species	56
71.	Effect of High Frequency Sound on Fish	53
72.	Containerized Tree Seedlings	14
73.	Stimulating Male and Female Flowering in Conifers	38
74.	Phaeomelanin	4
75.	Affinity Chromatography of DNA and Messenger RNA	24
76.	Neural Crests	18
77.	Gray Lethal Mice	13
78.	Cellulose in Tunicates	0
79.	Chalones	63
80.	Epstein-Barr Virus	205
81.	Isolation of Ribosomes from Rabbit Lymphocytes	7
82.	Cellulase in Ternites	5
83.	Various Aspects of Cellulose	147
84.	Autoimmune Reactions in Fish Brain	57
85.	Mucopeptides and Peptidoglycans in Bone Tissue Culture, Osteocytes or Cartilage	33
86.	5-Fluorocytosine	61
87.	Avian Leukosis Virus	44
88.	Propranolol and Hypertension	37
89.	Anti-thrombins	57
90.	Streptokinase	88
91.	Enzyme and Protein Structure	184
92.	Glass Bead Chromatography	30
93.	Isolation of Histocompatibility Enzymes	52
94.	Magnetobiology and Magnetotherapy	148
95.	Various Aspects of Heparin	586
96.	Extraction of Proteins from Acrylamide Gels	20
97.	Detection of Hyperthyroidism in Humans	44
98.	Phosphoproteins in Viruses	46
99.	Psychiatric Aspects of Aging	187
100.	Effect of Sediment of Fish	41

Appendix B

Interim Report Sent to BIOSIS

Discussing Several Problems with the Search Systems

SPACE AGE ASSISTANCE FOR BUSINESS INDUSTRY

North Carolina Science and Technology Research Center

RESEARCH TRIANGLE PARK, N. C. 27709

P. O. Box 12235

Telephone: (919) 549-8291
TWX Number: 510-927-1804

June 3, 1975

To: Louise Schultz

From: Monica Nees

Subject: Report on On-line Research Project with BIOSIS, April-May, 1975

Introduction

The terminal in the Research Triangle Park was connected to the BIOSIS computer in Philadelphia, with three regular users: Hannah Green, Peter J. Chenery, and me. Since then it has been in use an average of three to five hours a day connect time. During April the three of us used it approximately the same amount of time. During May I was by far the most active user and therefore am the writer of this progress report.

Each section of the report is oriented towards major problem areas encountered in our research. Whenever possible, the following format is used: a statement of the problem; detailed examples; and, if found, the solution to the problem. Also included are questions about problems not yet solved, and subjective as well as objective comments.

Table of Contents

North Carolina Science and Technology Research Center

I. Most Critical Problems: Intersections and Adjacencies

The most critical problem to date was detected very recently, on 5/28/75, and is illustrated as follows:

TABLE I

Search Statement No.	Terms	No. of Documents
1	sodium	4196
2	chloride	2817
3	sodium and chloride	548
4	1 and 2	599
5	4 not 3	51

Thus the total hits resulting from the intersection of the words A and B are not the same as from the intersection of the search statement number for A with the search statement number for B. The difference, 51 documents, was examined in detail and all 51 contained both of the words sodium and chloride.

Other word pairs were tested, and similar discrepancies were found in several cases:

TABLE II

Discrepancies Found

fluoro uracil
Eimeria tenella
polymer fume
Toxoplasma gondii

No Discrepancies Found

potassium phosphate
Trichomonas vaginalis
Eimeria acervulina
Eimeria brunetti
Eimeria necatrix

This problem was originally discovered by the following, which was part of a genus-species search:

TABLE III

Search Statement No.	Terms	No. of Documents
1	Toxoplasma	132
2	gondii	104
3	Toxoplasma and gondii	104
4	Toxoplasma ADJ gondii	63
5	3 not 4	41
6	(Toxoplasma and gondii) not (Toxoplasma ADJ gondii)	0
7	(3) not (4)	41

The difference of 41 in search statement 5 seemed erroneous because from previous searches on this and other files, it appeared that all, or nearly all, Toxoplasma were Toxoplasma gondii. All 104 documents in 3 were examined, and all were, in fact, Toxoplasma gondii. All were written Toxoplasma-gondii (T-g) and thus could be searched T ADJ g. T-g appeared either in the authors' title or in the augmented keywords. None had the genus at the end of one line and the species at the beginning of the following.

We cannot explain the reasons for these discrepancies, nor for the curious effect of the parentheses in 6 and 7: in 7, there is no effect on the number of documents; in 6 -- where words are used -- there is. Joanne Howard was immediately informed of the problem and is working on it. Until these problems are solved, we cannot trust the completeness of any search done on Biosis.

The adjacency function gives erroneous results in other circumstances:

TABLE IV

Terms	No. of Documents
fluoro and uracil	168
fluoro ADJ uracil	158

The difference, 10 documents, was examined. All 10 were fluoro ADJ uracil where the fluoro was at the end of one line and the uracil at the beginning of the following one. We were informed that this is a limitation in STAIRS and cannot be changed. If so, the pitfalls of the adjacency function must be clearly emphasized in the manual.

We are also unable to explain the following discrepancy:

TABLE V

Terms	No. of Documents
(ethyl isopropyl) ADJ alcohol	47
ethyl ADJ alcohol	43
isopropyl ADJ alcohol	0

Is (A or B) ADJ C a permissible operation in STAIRS? If not, this should be stressed in the manual.

The term fume ADJ fever pulled two documents. One had polymer ADJ fume ADJ fever in the title; the other in the augmented words. Yet neither could be pulled by polymer and fume and fever, polymer ADJ fume ADJ fever, or even by polymer and fume or polymer ADJ fume. Why not?

II. Local Printer Essential

A local printer is absolutely essential. Lacking one, we have to take voluminous manual notes, which nevertheless have proved inadequate. It is essential to have automatic documentation of the exact format of the search strategy (i.e., were terms or equivalent statement numbers used in an intersection?), in light of the critical problems discussed in the first section. The COPY command is being used to print from DISPLAY or BROWSE but that is not sufficient. We need complete hard copy documentation as we proceed with the search. On 6/2/75 we initiated a purchase order for the required IBM printer, number 3284 MOD 3.

However, receipt of the actual search output by mail a few days after the strategy was entered has proved quite satisfactory.

III. Incomplete Documentation

Incomplete documentation, especially in the STAIRS manual, caused much wasted time and many frustrations. We have been in almost daily contact with Joanne Howard, who researched our problems, then wrote the necessary documentation and sent it on to us. We cannot understand why this documentation was not in the manual in the first place, especially because we are not the first users of the system.

For instance, page 18 of the STAIRS manual emphasizes the value of the SAVE command: "Even a system failure will not wipe out queries that have been saved." Especially in April we had many "system failures" in that because of line transmission problems we were frequently disconnected from the computer in Philadelphia. Our searches probably average somewhere between 20 and 30 search statements. To protect ourselves from system failure, we routinely saved groups of search statements in small segments. After the then ever-present system failures, we tried to regenerate the searches by recalling each stored segment in sequence. We were never able to recall more than one, but spent weeks in fruitless attempts to do otherwise.

It was not until Joanne Howard's letter of May 23 that what we already had concluded was confirmed: you cannot save more than one named strategy in a STAIRS session. To quote from her letter: "The reason is that the use of the ...SAVE XXXX command causes everything up to the point of entry of the command to be saved under that name". Much time would have been saved had that one sentence, with "everything" underlined, been in the manual.

If this problem related to the SAVE command cannot be circumvented, it is indeed a serious limitation of the STAIRS program. Over a period of time, search questions tend to fall into categories; toxicity, pharamacology, etc. It is advantageous to develop, for instance, a toxicity set, consisting of words, CROSS codes, Bio Systematic codes, all with any suitable truncations, which can be saved and called up again. This set should be useable at any search statement number, not just the first. The capability of using as many stored sets as necessary in the same search is highly desirable.

We realize that two or more strategies developed in STRATBLDR can be executed in STAIRS, if the sets are called up in sequence. However, STRATBLDR-prepared strategies are not complete enough for our requirements, because of their inherent limitations, especially with respect to truncation.

IV. Response Times

For most of May, the response time on certain parts of STAIRS was intolerably slow, reaching three to 10 minutes per operation. This was especially true in recalling from SAVE, EXEC and PURGE of saved statements. The Input Inhibited light was frequently on for several minutes indicating an overloaded computer.

After Pat Lawrence rearranged the overflow file in the last week of May, the EXEC and PURGE commands now take roughly 10 to 30 seconds, a vast improvement over the previous 10 minutes. Response time on normal search statements has also improved, but has not quite reached the desired goal of three seconds or less.

One particular response time is still very slow, and we wonder if there is any way to accelerate it. A search strategy consisting of 35 search statements has been stored under the name ABCD. It is later called up by ..EXEC ABCD. The system laboriously goes through the execution of every single search statement until search statement 35 is reached. The Enter key must be hit after each one. Is there any quicker way to get to the final search statements (in STAIRS, not VM03) or to get a rapid display of the entire strategy?

V. CROSS Codes and Bio Systematics

As expected, the CROSS codes are the most important part of the file. It would be desirable to have a truncation code for only the first and second levels, as well as the existing one for all three levels. The third level often introduces too much noise.

The Bio Systematics are also frequently used, especially the one for human. In some cases in the past, it would have been very helpful to have separate codes for rats and mice. Our requestors frequently ask for one of these species and want to negate the other, or else they want the effect on one separated from that on the other.

In Appendix IV of the User Manual for the BIOSIS/STAIRS System, a confusing example was presented where an apparently unnecessary Bio Systematic code S86375 (muridae) was used: (C24010$ or leukemi$) and S86375 and (mice or mouse or murine). By requiring the mice synonyms, in order to avoid rats, the Bio Systematic code serves no function.

VI. STRATBLDR

STRATBLDR has received rather minimal emphasis in comparison with STAIRS. Although useful for getting related words, it is too tedious. A SELECT command would speed it up tremendously.

Once again, we want to plead for NOT camouflaging the CROSS codes and Bio Systematics. It would be most desirable to have an entry such as Cardiovascular System: C14500, rather than merely Cardiovascular System -C. This should apply to all CROSS codes, not just those of one word in length.

With respect to Bio Systematics, a "hidden" code can be especially dangerous. In a search on 12 different species of protozoa, I wanted to organize the output by genus, and, in some cases, even by species. For the genus Trypanosoma STRATBLDR gave flagellata as a related word. Had I not known that is was a broader term -- a class of protozoa -- I would have used it and gotten references to other flagellata as well, thereby losing the desired specificity. An inexperienced user would be especially susceptible to this pitfall, which could be avoided by an entry such as flagellata: S35200, or, at least, flagellata -S.

"Coaching" by STRATBLDR could be very effective. For instance, if the search term "human" was used, the system could respond: "Use S86215 instead of human".

In STRATBLDR we always write out the term, rather than use its associated 5-digit number. We do not have a printer and would therefore have to write down the number; the term can be more easily remembered. This could be obviated by using the top half of the screen for display by the system, and the bottom half for response by the user.

VII. Searches for Users

Despite all of the problems encountered, we have still managed to do searches for our users. The titles give an indication of the breadth of topics covered:

Calcium in Saliva
Everything on 5-Fluorouracil
Chemotherapy of Coccidiosis
Anti-Protozoal Vaccines
Affinity Chromatography of DNA and Messenger RNA
All Drugs Which Can be Administered Rectally

Chalones
Neural Crests
Propranolol and Hypertension
Nitrification and Dentrification
in Sewage Disposal
Isolation of Histocompatibility Antigens

These topics represent real questions: present searches being done on other files; updates of past BIOSIS searches; solicited topics primarily from our Chapel Hill MEDLINE operators; and a few on our own interests.

The number of hits ranged from eight to almost 700. The latter was on Anti-protozoal Vaccines where the output was segmented by the 12 species requested. It was on this search that most of the bugs in the search system were discovered.

Response from our users has been excellent, and they are all the more delighted because the service is FREE. Because we are still learning the search system, we have not yet attempted any detailed analysis of the output. We're looking forward to when we can.

VIII. Plans for the Immediate Future

Most of our time to date has been spent in debugging the system. Several critical problems have been discovered, and until they are solved it would be fruitless to attempt any detailed comparisons as to the effectiveness of alternate search strategies. We will, of course, continue to do searches for our users, but can make no claim to completeness of the output. We will also continue to notify BIOSIS immediately of any problems we encounter with the system.

In light of the fact that a three-year portion of the BIOSIS file will soon be commercially available on Lockheed's DIALOG system, we wonder if our emphasis on the adequacy of STRATBLDR's entry vocabulary should be re-evaluated. If there are no plans to release STRATBLDR to outside users, either of DIALOG or of any other computerized search system, perhaps this phase of the project should be de-emphasized.

Don't hesitate to contact us with any questions you may have on this report. We're looking forward to your comments and suggestions.

cc: Peter J. Chenery
Hannah Green
Joanne Howard
Pat Lawrence

Appendix C

Example of Cover Letter Sent With Teaser Searches for the Department of Natural and Economic Resources of the State of North Carolina

North Carolina Science and Technology Research Center

RESEARCH TRIANGLE PARK, N. C. 27709

P. O. Box 12235

Telephone: (919) 549-8291
TWX Number: 510-927-1804

TO:

FROM: Monica Nees, Ph.D., Information Specialist

SUBJECT: Free Custom-Tailored Literature Searches on Biological and Biomedical Topics

The North Carolina Science and Technology Research Center (NC/STRC) is participating in a research project with BioSciences Information Service of Biological Abstracts (BIOSIS), publishers of Biological Abstracts and BioResearch Index. These two publications include more than 250,000 references annually on biological and biomedical research. A Fact Sheet describing these data bases in detail is attached.

All references published by BIOSIS in 1974 are available to NC/STRC for on-line computerized searching via a remote terminal at our Research Triangle Park location connected to the BIOSIS computer in Philadelphia. The purpose of our research is to improve the computerized search program, but to do this effectively we need a wide variety of search questions. This is why we are contacting you.

We would be very happy to search the 1974 BIOSIS file on any questions of interest to you or your organization, at no cost to you. The searches can be as complex or as simple as your interests require. They will be custom-tailored to your specific needs by our experienced scientist-searchers. We ask only that you fill out a brief evaluation form after you receive the search output.

The type of printout you will receive is similar to the enclosed sample. A brief description of the search topic appears at the top of the first page. The references in the sample by no means represent a complete search, but merely illustrate some of the information on that topic found in the BIOSIS file.

Please contact us with your search questions as soon as possible. We're looking forward to working with you.

Appendix D

Explanation of Abbreviations Used on Computer Printout of BIOSIS Search Results

North Carolina Science and Technology Research Center

RESEARCH TRIANGLE PARK, N. C. 27709

P. O. Box 12235

Telephone: (919) 549-8291
TWX Number: 510-927-1804

Explanation of Abbreviations Used on Computer

Printout of BIOSIS Search Results

1) ABNUM — e.g. 57068908, = Volume number (57) followed by reference number (068908) in Biological Abstracts or BioResearch Index.

2) CODEN — Coden, a unique five-character abbreviation for source publication.

3) ABBRV — Abbreviated title of source publication.

4) BIBLO — Volume and issue number, year of publication, page numbers of source publication.

5) AUTHS — Author(s).

6) WORDS — Original title of abstract, followed by keywords added by BIOSIS indexers to enrich it.

7) CROSS — CROSS Code numbers referring to subject category indexing. The number before * indicates category in which this reference was published in Biological Abstracts or to which it was assigned in BioResearch Index. The number before - indicates a secondary level of emphasis; absence of * or - indicates a tertiary level.

8) SYST — Numbers referring to codes for BioSystematic or taxonomic classification.

Appendix E

Example of Search Evaluation Form

North Carolina Science and Technology Research Center

RESEARCH TRIANGLE PARK, N. C. 27709

P. O. Box 12235

Telephone: (919) 549-8291
TWX Number: 510-927-1804

SEARCH EVALUATION FORM

These literature search results from 1974 Biological Abstracts and BioResearch Index have been provided at no cost to you as part of a research project being conducted by the North Carolina Science and Technology Research Center (NC/STRC) and BioSciences Information Service of Biological Abstracts (BIOSIS). Please fill out this evaluation form because your response will help us improve the computerized search program.

Name: _____

Address: _____

_____

Telephone No.: _____

Date: _____

Search Title: _____

Number of Citations	
Highly Relevant	_____
Somewhat Relevant	_____
Not Relevant	_____

We often deliberately include peripheral references. In general, what is your reaction to peripheral material?

_____ Glad to get it
_____ Can take it or leave it
_____ Would rather not get it

Are you aware of any key references published during 1973 or 1974 which were not retrieved in this search? Please list.

Overall evaluation of search results:

_____ Very useful
_____ Somewhat useful
_____ Not very useful
_____ Useless

Additional Comments:

Please return form to the searcher at NC/STRC checked below.

_____ Dr. Hannah Green

_____ Dr. Monica Nees

NASA-Langley, 1977