ED 073 782 LI 004 192

AUTHOR Buckland, Lawrence F.; Madden, Mary
TITLE Investigation of the Searching Efficiency and Cost of Creating a Remote Access Catalog for the New York State Library. Final Report.
INSTITUTION Inforonics, Inc., Maynard, Mass.
SPONS AGENCY New York State Library, Albany.
PUB DATE 22 Dec 72
NOTE 77p.;(0 References)

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS Catalogs; Cost Effectiveness; *Library Automation; On Line Systems; Relevance (Information Retrieval); Search Strategies; *State Libraries
IDENTIFIERS *New York State Library

ABSTRACT

From experimental work performed, and reported upon in this document, it is concluded that converting the New York State Library (NYSL) shelf list sample to machine readable form, and searching this shelf list using a remote access catalog are technically sound concepts though the capital costs of data conversion and system installation will be substantial. The two primary areas of investigation covered in this report are: (1) pilot conversion to machine readable form of a portion of the NYSL shelf list; the purpose of this conversion process itself being the creation of a file of machine readable records which can be searched by a computer under the control of a telecommunication computer terminal. The purpose of the pilot conversion test is to determine costs of conversion, and any unusual technical problems; and (2) experimentation with, and use of, the initial product of the pilot conversion in catalog searching. The purpose of the search test is to determine technical feasibility of the search process where a user must formulate a query as a logical combination of alphabetic search words, a process far different than the mental eye-brain scanning of entries on catalog cards. (Author/SJ)

onics

Final Report

INVESTIGATION OF THE SEARCHING EFFICIENCY AND COST OF CREATING A REMOTE ACCESS CATALOG FOR THE NEW YORK STATE LIBRARY

Prepared by:

Lawrence F. Buckland
Mary Madden

Inforonics, Inc.
December 22, 1972

TABLE OF CONTENTS

1. INTRODUCTION

This report describes the results achieved so far in an experimental program in machine-assisted bibliographic control undertaken by the New York State Library . This program was undertaken for several reasons. First, the new library building on the South Mall will present catalog access problems, with Reader and Technical Services separated at such distances that the one public card catalog will be inadequate. Second, it seems likely, but has not yet been proven, that a single computer record may substitute for a great number of present files; efforts now being spent on file maintenance and searches may be reduced with automation Third, the long-range purpose of statewide information network development will be aided by the construction of a NYSL bibliographic data bank available for remote external access.

Investigations were undertaken to provide a basis for well-informed decisions on the advisability of a computer-assisted catalog and on the most desirable forms of storage, access, and display of information.

The two primary areas of investigation covered in this report are:

1. A pilot conversion to machine-readable form of a portion of the NYSL shelf list; the purpose of this conversion process itself being the creation of a file of machine readable records which can be searched by a computer under

the control of a telecommunication computer terminal. The purpose of the pilot conversion test is to determine costs of conversion, and any unusual technical problems.

2. Experimentation with, and use of, the initial product of the pilot conversion in catalog searching. The purpose of the search test is to determine technical feasibility of the search process where a user must formulate a query as a logical combination of alphabetic search words, a process far different than the mental eye-brain scanning of entries on catalog cards.

The use of actual shelf list data in the experiment provided a real environment for the conversion and therefore experimental costs of conversion can be projected so that the cost of converting the entire job can be computed.

The age of the shelf list results in several methods having been used in its construction so about 35% of the shelflist records are incomplete. In addition they are so messy in appearance (fields crossed out or written-in what seems to be an indiscriminant fashion) as to be confusing to the personnel involved in the tagging, editing and keyboarding.* The only attributes the shelf-list file might be considered to have with respect to a massive retrospective conversion effort are: (a) it is organized logically by subject area and, (b) the records needed for conversion are easily separated from the file.

* The effort to complete the shelf-list records was deemed too expensive.

Assuming the experimental sample is representative of the entire list, the shelflist consists of the following types of records: L.C. cards (46%), Order slips (19%), NYSL original cataloging (9%), Serial Line Cards (10%), and various other types, i.e., Short Original Cataloging on hand written or typed cards (16%).

During the course of the experimentation, a consulting group from Inforonics, Inc., a firm with a substantial history in library automation, was engaged to review progress.

1.1 Study Approach

The approach taken by Inforonics in this review was to examine the ongoing NYSL experiments directed at key questions, namely cost of input and search utility. For the conversion study the results of the pilot test were analyzed and costs calculated for each function of conversion. In the search investigation, Inforonics helped design the search experiments and the procedure for the documentation as well as analyze the data obtained in the experiment searches run by the NYSL staff.

The use of microfilm to duplicate and distribute copies of the NYSL catalog as a Remote Access Catalog was not included in the experimentation. In any future plans however, it remains an alternative which should be considered.

The primary end product of the investigation so far has been an estimation of (1) the costs of conversion, and (2) the accuracy of searching when compared to manual card searching. However observation of the experimental apparatus and procedures yielded a great amount of qualitative information, which should be useful in the longer range planning of NYSL automative activities.

This report, in addition to describing Inforonics' work, also contains the results of the NYSL in-house project staff work, and should be considered a final report of the entire conversion and searching project.

2. CONVERSION EXPERIMENT

2.1 Background

Estimating the costs and technical problems involved in converting the NYSL shelf list to machine readable form for use in a Remote Access Catalog is so complex that the observations and measurements of an actual production test environment is needed. To satisfy this need, the NYSL staff carried out a pilot project to convert a segment of the NYSL shelf list. The project had three main components: an in-house data preparation effort, a contracted data tagging and keyboarding effort, and an in-house EDP file validation and conversion effort. The project was begun in the fall of 1969 and had progressed to a stage where proposals for tagging and keyboarding could be solicited in November. A contract was awarded shortly afterward. The bulk of the work on this project was carried on during the year 1970. Its end product was a file which was to be used as a test file for subsequent experiments in searching a Remote Access Catalog.

2.2 File Conversion Experiment Design

The procedure for converting the shelf list to machine readable form was developed to adhere to several basic experimental design policies.

1. The use of Dewey Class 550-599 as an experimental sample. This sample was considered representative of the total shelf list, and was common enough in topic to allow good search experimentation.[*]

2. The use of two types of files of machine readable data elements in the converted file -- one a fully coded MARC II and the second a modified MARC II containing an abbreviated list of elements.

   These two types of files would allow experiments yielding possible cost reduction of coding a modified MARC II record. If so, then experiments were needed to see how much its search capability would be curtailed when compared to a full MARC II record.

3. The use of three types of staff for manuscript preparation, tagging, and editing tasks: clerical, semi-professional, and professional. The skills required in converting were relatively unknown and allowing for the use of all types of personnel would yield data useful in matching conversion tasks to the skill levels of different library personnel.

[*] In addition the subject areas chosen for this conversion were to be of use to the Science and Technology Section which did not have a catalog of its own particular collection.

2.3 Experiment Design Task

The following design tasks were carried out in preparation for the pilot test conversion:

2.3.1 Encoding worksheet design

A worksheet was designed on which a copy of the shelf list card could be affixed. The worksheet contained spaces for MARC tags and other cataloging information needed in the record. A copy of a completed worksheet is shown in Figure 1.

2.3.2 Microfilming shelf list sample

The conventional method of duplicating library card files by microfilming and Xerox Copyflo enlarging was found to be the lowest cost and least disruptive method to the library procedures. The Xerox enlargements were to be subsequently affixed to the worksheets.

2.3.3 Tagging manual preparation

A tagging manual was developed by extracting pages from the L.C. MARC Manual which contained the information which seemed to be most usable.

There were two types of shelf list data entries: complete category consisting of LC or N cards, called Category 1 and incomplete cataloging consisting of serials cards, order cards, and miscellaneous incomplete cards called Category 2. Each of these samples (Category 1 and Category 2) were encoded in two ways: Full MARC encoding* called Task I (to be done by the vendor) and Modified MARC called Task II to be done by the NYSL tagging staff.

*Full MARC is a slight misnomer because Full MARC can only be done with book in hand, not from catalog cards. Full MARC encoded from catalog cards misses only an occasional fixed field however, which we consider minor in this experiment or in any planned conversion.

Thus there were four types of records possible in the experiment, each with its own range of tags depending on the extent of the cataloging available.

2.3.4 Tagging staff training

A tagging and editing staff was formed, to handle the coding of the source documents (3" x 5" paper slip copies of the shelflist records stapled to the coding sheets), of six part-time library science students from the State University of New York at Albany and four full-time clerical employees. Each person received approximately 1-1/2 days of training prior to their tagging the source documents. This training period consisted of practice tagging of sample L.C. cataloging records which were specially chosen to illustrate most MARC variable and fixed fields and as many variations of these fields that might possibly occur.

2.4 Conversion Procedure

The following steps were used in the conversion process:

1. The xerox copies of the shelf list were affixed to the work sheets.

2. The worksheets were separated into groups by the NYSL project staff. Task I documents contained those worksheets to be both tagged and encoded by the vendor. Task II documents contained the worksheets to be tagged by the NYSL experimental project staff.

3. The Task I worksheets were sent to the vendor.

4. The Task II worksheets were sent to the experimental project staff (library school students, NYSL clerical, and professional staff). The worksheets were tagged, edited, and sent to the vendor for keyboarding.

5. The vendor tagged the Task I worksheets and then encoded both task I and II worksheets by the following procedure:

   a. The tagged worksheets were transcribed by typing on an OCR typewriter.

   b. The typed worksheets were read on an OCR scanner creating a magnetic tape of typed line images.

   c. The OCR output was run thru a print out program, which contained a simple validator, producing a listing with error messages.

   d. The listing was proofread and marked for editing.

   e. Typed lines containing errors were retyped and merged with the original file, replacing the incorrect lines.

   f. The corrected file was processed to arrange it in class order and to convert it to (1) a MARC II input format and (2) a BCD listing tape.

   g. The BCD tape was listed, and the list and the MARC tape were delivered to N.Y.S.L.

6. The tape delivered to NYSL was translated to the Control Data character set and an NYSL internal format.

7. The translated tape was verified to assure that the file conformed to the NYSL version of MARC. Invalid records were deleted to be re-input by the vendor, reprocessed through the NYSL validation system.

8. The NYSL MARC tape was converted to the form which SUNY Biocommunication Network computer staff could enter into its system.

2.5 Problems with the Experimental Operation

Many problems occurred in the conversion process which had their root in the (1) lack of time to properly plan experiments, caused because of pressure to make fiscal expenditure committments and (2) Misrepresentation of capability on the part of the tagging and keyboarding vendor. The net result of these problems was (1) a delay in schedule and (2) an obscuring of the measurements of parameters from which cost and production estimates could be made.

2.5.1. N.Y.S.L. Project Control Problems

Delays and reprocessing were caused by inadequate document control procedures. Although the pilot operation was experimental and covered only a small fraction of the NYSL shelf list, the actual numbers of documents, (approx. 20,000), batches (800), and number of processing steps (approx. 10) were large enough to require strict controls. Microfilm was not inspected properly, supplies (worksheets) ran out, inadequate backlog of work (due to delays in microfilming deliveries) and lost batches of documents, all contributed to excessive time spent in expediting, reprocessing, and rescheduling.

As well as possible, time spent in these activities was removed from the production time measured, but it is possible that some nonproductive time was not accounted for which would affect the accuracy of the data collected on labor time required to tag, edit, and correct entries.

2.5.2 Vendor Processing Control Problems

The vendor had inadequate file control, error detection, and manuscript control procedures. Detecting errors, finding original manuscript to be resubmitted, and general follow-up was left to the NYSL staff, contributing much to their administrative workload. This effort was considerable because the Dewey vendor did not supply the specified printout in sequence so locating records for checking was exceedingly difficult.

Finally the errors found by computer validation were found so late in the project that the use of the errors to correct tagging and editing procedures was impossible. The taggers themselves did not have the benefit of learning from these mistakes.

2.5.3 Problems with non-LC cataloging procedures

Some entries in the shelf list had items which were difficult to fit into the MARC II data item set. This is a real problem however, and would occur even in a properly designed production system. Further study is required to determine whether these entries will require being revised or recataloged.

2.5.4 Inadequate data base analysis

The lack of time for planning caused several hasty decisions on the specifications of the Task I and Task II data bases, causing problems in the resultant encoded data. The assignment of the 082 (LC DDC's number) tag to the NYSL Dewey number caused ambiguity because its structure is different from the LC suggested Dewey number. The 490 (MARC) tag was used without its indicator, which caused the field to be meaningless. Finally the holdings statement field 901 (local data) was improperly designed so that computer analysis of its contents would be exceedingly difficult. This last problem did not affect the production process, however, because its implications were in future use of the data for circulation control.

2.6 Experimental Results

The results obtained in the Conversion experiment consist of a determination of a conversion cost per record, some qualitative judgment about possible cost reduction, and an error analysis of the encoded records.

2.6.1 Conversion cost per record of Task I records full MARC

The cost per record of the full MARC record was $1.65/record. This value was computed from the total Task I vendor quotation of $3844.00 eliminating the setup and programming costs of $2194.00.

2.6.2 Conversion cost per record for Task II records modified MARC

The cost required by the conversion process is estimated to be $1.74/record, broken down into labor ($1.49), and computer material and services costs ($.27). A breakdown of these costs is shown in Table I.

In calculations to make the vendor costs and NYSL costs comparable, NYSL direct labor costs have been burdened with 100% overhead for supervision, payroll benefits, facilities, and technical support.*

* The president of the vendor company told us for a job the size of the NYSL experiment (20,000 records) that the cost would be apportioned 25% keyboarding, 12% verification, 15% scanning, computer editing and conversion, and 48% overhead and fees. The overhead and fee of 48% of the total is approximately 130% of the direct labor costs. We estimate the non-fee cost (overhead) to be approximately 100%. This assumption places the fee at 11%, which is reasonable.

2.6.3 Cost of conversion of the NYSL Shelf List

The conversion of the entire shelf list would require 128,000 man hours and $1,113,600 in total funds. A project of this magnitude carried out over a period of 4 years would require 18 full time staff. Carried out over 2 years it would require 36 full time staff.

2.6.4 Possibilities for cost reduction

The cost of conversion estimated is a realistic practical figure and will not decrease with increase in volume of titles processed or with minor technical improvement of the system. The possibilities for further reduction lie in three areas; the use of format recognition, the availability of additional RECON records from the Library of Congress, and the use of MARC records produced by other libraries.

2.6.4.1 Format recognition

The use of format recognition will probably reduce the cost of conversion slightly. Typing costs are higher because it is a more difficult task. Some of the costs saved in tagging are expended in additional editing. Potential cost savings are not available from any published source, and in our opinion will not exceed 10%. Data will be forthcoming from the Library of Congress shortly comparing their costs of tagging vs. format recognition.

Table I

Task I

Vendor Cost $3.80/record

Task II

Function		Direct Cost	Overhead	Total
NYSL tagging labor		.26	.26	.52
NYSL revision labor		.087	.087	.17
Vendor typing labor		.27	.27	.54
Vendor verifying labor (proofreading)		.13	.13	.26
				$1.49
Vendor Computer				.16
NYSL material & services				
Manuscript	.01			
Microfilm & enlargement	.08			
Xerox copy	.04			
	.13			.13
Total Cost				$1.78

2.6.4.2 Library of Congress REprosective CONversion (RECON)

The use of RECON tapes from the Library of Congress will involve only computer expense which is 10% of the total, saving 90% of the total cost (all labor). Records can be converted at a cost of $.16. The RECON tapes presently cover English language imprints back to 1968, so it is reasonable to assume that 100,000 titles are already available in encoded form.

2.6.4.3 Other libraries machine records

A promising area of cost reduction is the use of encoded records of other libraries. These records are being encoded by several groups in large quantities, and as time progresses, the encoding formats are progressively closer and closer to being MARC identical. At present there are 2.5 million records to be encoded which probably would be useful. There would be additional computer programming and operating costs associated with their conversion, which we estimate to be $.10/record based on the use of 500,000 records. The computer conversion cost of such records would be approximately $.26. In addition no data is available which allows estimation of the percentage of the NYSL shelf list contained in these available encoded MARC files.

2.6.5 Conversion costs at other libraries

A telephone survey of other conversion projects was made, and the following costs were obtained. These costs are not exactly comparable to the experimental costs because (1) different methods of accounting are used for overhead and computer costs and (2) there are variations in the accuracy of the final product.

NYSL Full MARC Task I	$1.65/record
NYSL Modified MARC Task	1.78/record
Library of Congress	2.96/record (exclusive of computer cost.
2nd Commercial Vendor	2.60/record

2.6.6 Analysis of errors

One of the significant results of comparing the Remote Access Catalog Conversion Project with other conversion efforts is the relationship of encoding cost to percentage errors in the final product. The data obtained were not accurate enough to compute quantitative relationships of cost to error percentages however, it was possible to compare the NYSL experiment with error data from a second commercial vendor.

The tagging and typographic errors contained in the keyboarded copy we have separated into two types, defined as follows:

Logical errors - those errors which can be detected but not corrected by a computer program of moderate complexity, but without extensive dictionaries.

Spelling errors - errors in spelling of any string of characters in an item including spacing and punctuation.

The errors at successive stages of the two input processes are compared, expressed as a percentage of MARC II records in error. Some data are not available.

NYSL Experiment ($1.78/record)

	logical	Spelling
at keyboarding	unknown	unknown
after 1st edit	14.1%	14%
after NYSL error analysis and vendor re-edit	1.1%	14%

2nd Commercial Vendor ($2.60/record

	logical	Spelling
at keyboarding	8%	30%
after 1st proofreading & edit	.1%	1%
after 2nd proofreading, checking and edit	.0%	.02%

We think this comparison is a useful one for it points out that the difference between a very good file and a file with considerable typographic errors is one additional high quality proofreading and editing pass which contributes approximately 50% increase in cost.

2.6.7 Effect of error on file usage

Although the error rate in the NYSL final product is quite high, spelling errors occurring in 14% of the records encoded, only about 2% of these errors could cause errors in the remote access catalog searching experiments*. These serious errors were contained in words in the elements potentially useful as search elements, such as short title. The bulk of the nonserious errors were punctuation, spacing or errors in fields not likely to be searched.

Although the file error rate might be acceptable for machine search purposes, in the use of a file in technical processing or the production of printed catalogs, 14% error would be above that acceptable by cataloging tradition. The only severe shortcoming is really an esthetic one.

* This will be discussed in greater detail in Chapter 3 on Searching.

A powerful concept which can be easily applied to a machine system is the correction of a catalog file based on reports of errors from users. In a manner similar to the way in which L.C. corrects and reprints its cards upon notification from users, NYSL could correct its machine file. As long as the file is accurate enough to be acceptable for use (which the experimental NYSL file is, we think) it can be edited every time a user spots an error.

One can carry this line of reasoning through to the conclusion that all systems no matter how accurate accept user input for error correction. Speaking loosely in a mathematical sense, it probably costs an infinite amount to create a large file with zero error. As a corollary, each succeeding error found costs more to find than its predecessor. It seems practical therefore to let the users find them at some point, for their effort costs nothing as it is a byproduct of their searching activities.

3. ON LINE SEARCH EXPERIMENTS

The two concepts underlying the Remote Access Catalog are "Remote" which means that access can be done at places remote in and outside the library and "access" which implies that a searching capability is available in the system. The problem of "remoteness" is not a difficult one and many successful computer systems exist which operate at a distance from their users connected by tele-communications.

The problem of search is not simple however and there are few systems in operation on any large scale and none which can perform in any demonstrable way what the proposed NYSL Remote Access Catalog is supposed to do.

To investigate the unknowns in the proposed concept an experimental program was carried out which allowed project personnel to search the data base converted from the 550-599 sections of the NYSL Catalog by a variety of access points. This search experimentation could be evaluated qualitatively from a users point of view and also could be compared to manual card catalog searching.

3.1 Experimental Design Policy

The basic policy decision in the design of the search experiments was to use the Upstate Medical Center BioMedical Communication Network (BCN) search system. This was a laudable decision because much can be learned from this system without any programming cost. Any possible disadvantages of not being able to do exactly what one would like to are far outweighed by the cost saving estimated to be in the hundreds of thousands of dollars. Any answers to questions one can't develop experimentally may be arrived at analytically given measured values of experiments performed in those areas where one can.

The Biomedical Communication Network is a group of libraries connected to a search center by telecommunication lines. The search center located (at the time of the experiment) at the Upstate Medical Center in Syracuse, New York, accepts all search requests and displays results via IBM 2740 typewriter terminals. The center's computer can also serve as a communication switching mode which allows one library to communicate with another.

The use of the Upstate Medical Center search system, which is an on-line version of the IBM DOC PROC System, as an experimental tool gives one the following capability. A machine file is created where words representing authors (or more generally personal and corporate names), titles, subjects, and the MARC fixed fields are stored in a computer memory. These data elements can be searched in isolation or combined in a logical and, and/not, or, and/or combinations. Additionally a list of stop words is provided so that one does not need to concern himself with problems of initial articles or non-significant words.

The second experimental tool available to the project is the NYSL manual card catalog containing the identical data. In this form the titles can be searched by a set of access points consisting of the initial words of the author, title or subject heading. Given these experimental tools and the technical requirements stated above the next step is to design an experimental search procedure which yields data on the utility of the remote access catalog.

The general procedure followed in the selection and setup of the data sample and documentation of searching was as follows.

3.2 Preparation for Experiment

The preparation for the experiment included five tasks to prepare data and personnel for the actual test searching. These tasks were:

1. Train New York State Library and Inforonics personnel to be able to understand the capabilities of the Biomedical Communications System.
2. Convert 18,000 records, previously encoded from monographs and serials of the New York State Library shelf list into the MARC format, into the internal operating format of the BCN system.
3. Develop experimental procedures for performing searches and collecting and tabulating data.
4. Select a group of test search requests from requests originated by libraries in the New York State Interlibrary loan (NYSILL) Network.
5. Assign responsibilities for various segmented analytical and evaluation tasks to Inforonics personnel and the New York State Library personnel.
6. Develop tools for evaluating the effectiveness of the search results.

3.2.1 Training of project personnel

New York State Library and Inforonics staff studied the capabilities of the Biomedical Communications DPS system from system manuals and by actual visitation to the Biomedical Communications Center at Upstate Medical in Syracuse. A small subset of the 18,000 encoded shelf list records had been converted to the Biomedical internal format and by actual operation on this small sample the project personnel learned the command language of the BCN system. Additional NYSL staff were taught to use the BCN system by project personnel who had acquired knowledge first hand by visits to the Syracuse BCN facility.

There were two difficulties with learning to use the BCN query language. First, it is a general search system so the data base to be searched and printing options for successful matches had to be specified with every query. Thuswith every query there was more than just the search words to remember and to key. Secondly, the search syntax was very rigid. Some words and symbols had to be used in a spe^ified way. "option" began each search. A semicolon ended each line, "end" ended each search. Furthermore, each word to be searched had to be identified by the searcher as personal name, corporate and conference name, or subject and title word.

Neither of these problems were insurmountable, but they did make learning the BCN query language difficult.

3.2.2 Data base conversion

There were two steps to converting the data base to the internal format used by the BCN-DPS system. The first step was to determine which data fields and/or MARC subfields would make the keyword indices for searching. In order to keep costs down, NYSL had agreed to pay the rental for only one disk pack. Unfortunately, although the amount of disk system needed for the 18,000 record file could be predicted, there was no means of predicting how many unique keywords the various data fields would generate, nor how much disk storage the indices would need. Therefore, it was decided to index a limited number of fields, and as a result the disk pack was no where near capacity. With better storage prejections more fields could have been indexed.

The second step was the programming to convert the MARC formatted data base. This was done by BCN personnel. The conversion of the data base was checked by displaying records in response to test queries.

When the conversion was deemed satisfactory the data base was converted. The indexes were created, and were used to produce keyword frequency lists. The indexed word had one of three prefixes depending on the data field in which the word was found, 'Ø' for personal name used as author, added entry, or subject added entry; '1' for corporate or conference name used as author, series, or added entry; and 'blank' for title and subject words. The word frequency lists were in three sections according to prefix. Each section was arranged alphabetically with the frequency of occurrence of the word in the data base and with the count of documents in which the word appeared.

These printed indexes were necessary when formulating a search. The user could determine if a word was indexed as he guessed, for example as subject or title; and if indexed, in how many documents. This latter would help the user decide if he needed more ind x terms to narrow the search, i.e., produce fewer "hits". As helpful as these indexes were however, they implied an extra look-up and extra time before the computer search itself. It would have been more helpful to have the computer do this lookup, and report if the word or combination of words was indeed in the indexes and the frequency count.

A master list of the complete data base would be needed for evaluation so the 18,000 MARC records were sorted by Dewey decimal number and printed.

3.2.3 Assignment of analysis and evaluation tasks

The performance of the necessary searching, data collecting and evaluation tasks during the experiment posed some difficult scheduling and personnel assignment problems, which were further complicated by system failures and the early termination of text searching. The plan involved the assignment of personnel for the various tasks and the specification of procedure for an experimental work flow.

				Eliminate Data Base Errors Line 4 & 5		Eliminate Operator Errors		Eliminate Equipment Failures	
1. Total Searches		640	100%	485	100%	379	100%	325	100%
2. Successful		303	47.4%	303	62.5%	303	60%	303	93%
3. Unsuccessful		337	52.6%	182	37.5%	76	20%	23	7%
4. Failures Untraceable	86								
5. Request record not on data base	69								
		155	24.2%						
6. Query typing error	37		5.8%						
7. Search syntax error	37		5.8%						
8. Search strategy error	34		5.3%						
		108	16.9%						
9. Terminal failure	17		2.7%						
10. System failure	36		5.6%						
		53	8.3%						
11. Machine record incorrect		45	.7%						
12. Request information misleading									
13. Unexplainable									

Successful - Unsuccessful Searches & Why

	I	II	IIA	III	IIIA	IV	Total
Total searches	262	167	48	72	41	50	640
Total successful searches	149	80	1	33	2	38	303
Total unsuccessful searches	113	87	47	39	39	12	337
Reason for failure:							
Query typing error	15	21				1	37
Request record not on data base	29	33	12	7	17		69
Search strategy	7	4		13		10	34
Search operator error	13	12		11		1	37
Terminal malfunction	7	8		2			17
System malfunction	24	3			9		36
Request information misleading	17	4	1	5		2	29
Machine record incorrect	1	1		1	1		4
Untracable			47		39		86
Should have been retrieved		1					1

Successful - Unsuccessful Searches & Why - Percentages

	I	II	IIA	III	IIIA	IV	Total	
Total searches	262	167	48	72	41	50	640	
Percent successful	56.8%	47.9%	2.1%	45.8%	4.9%	76%	47.4	
Pe-cent unsuccessful	43.2%	52.1%	97.9%	54.2%	95.1%	24%	52.5%	
Reason for failure							% of Total Errors	% of Total Searches
Query typing error	13.27%	22.9%				8.33%	10.97%	5.77%
Request record not on data base	25.6%	37.9%		17.9%			29%	15.25%
Search strategy	6.19%	4.54%		33.3%		71.2%	10.08%	5.312%
Searcher error	11.5%	13.7%		28.2%		7.1%	10.97%	5.78%
Terminal malfunction	6.19%	9.2%		5.1%			5.04%	2.65%
System malfunction	21.23%	13.5%			23.07%		10.68%	5.61%
Request information misleading	15%	4.02%	2.1%	12.8%		16.6%	8.45%	4.44%
Machine record incorrect	.87%	1.7%		2.55%	2.55%		1.33%	6.99%
Untraceable			97.9%		79.4%		13.64%	7.17%
Should have been retrieved		1.68%					.29%	.153%

3.2.3.1 Personnel:

There were three principal participants involved in the on-line search experiment:

1. A search coordinator who selected the requests, grouped the requests, evaluated the computer searches, and when necessary re-searched the computer searches. Robert Vines was the search coordinator.
2. A search assistant who was under the direction of the search coordinator, and who did most computer searches and some comparative searches in the card catalog.
3. A tabulator, who kept tally sheets of all requests and computer searches, and kept the search coordinator informed of what types of searches had been done and which had not. Mary Madden, of Inforonics, was the tabulator. Originally, the search coordinator would determine why searches failed and how these searches should be re-searched successfully. When it became apparent that there was not enough time to re-search requests, the coordinator sent them to the tabulator, whose task it became to determine why searches failed.

3.2.4 Selection of search requests

The searches were grouped into five types or categories. In all cases 'A' group were those not found in the New York State library card catalog.

a. Type I were presumed to be personal author and title requests.
b. Type II and II-A were presumed to be Title Main Entry Requests.
c. Type III and III-A were presumed to be corporate and/or Series Entry Requests.

Percentage of 'Hits' per Successful Search

	I	II	IIA	III	IIIA	IV	Total
Total document searches	262	167	48	72	41	50	640
Percentage successful searches	56.8%	47.9%	2.1%	45.8%	4.9%	76%	47.4%
Number of hits							
1	73%	67%	100%	75%	50%	87%	74%
2	20%	13%		15%	50%	13%	17.1%
3	4%	9%		6%		3%	5.3%
4	1%	5%					2%
5	.6%						.4%
6		2.5%		3%			1%
16	.6%						.4%
42		1.3%					.4%
45		1.3%					.4%

Number of 'Hits' per Successful Search

	I	II	IIA	III	IIIA	IV	Total
Total document searches	262	167	48	72	41	50	640
Total successful searches	149	80	1	33	2	38	303
Number of 'hits'							
1	109	54	1	25	1	33	223
2	30	11		5	1	5	51
3	6	7		2		1	16
4	2	4					6
5	1						1
6		2		1			3
16	1						1
42		1					1
45		1					1

d. Type IV and IV-A are synthetic requests which were derived from existing file records on the data base.

e. Type V and V-A were presumed to be subject searches.

3.2.5 Procedure

The procedure and experimental work plan consisted of the following steps:

1. The search coordinator grouped all requests into one of four categories (Type I, II, III, IV).
2. The search coordinator numbered all requests.
3. The search coordinator gave a group of requests to the search assistant. In the beginning requests were author-title only (Type I).
4. The search assistant searched each request on the BCN-DPS system. He could re-search any request up to three times, assuming each time the search is a syntactically correct BCN-DPS search with no spelling errors. This was the original plan, time being of the utmost importance forced the abandon of this, so that most searches had only one try.
5. The search coordinator evaluated each search, and recorded his findings on the "search evaluation" sheet.
6. Searches were divided into three groups: successfully completed searches, successful searches with too many hits, and unsuccessful searches. The search coordinator was to record in a log book the status of each search. This was also not done, because of the time element.

Request Record vs. Machine Record - Percentages

	I	II	IIA	III	IIIA	IV	Total
Request record not different from machine record	19.12%	40.11%	2.1%	31.94%		10.%	22.97%
Request record differed from machine record	40.45%	13.77%		36.12%	14.63%	88.%	47.31%
Not determined if request record differed from machine record	40.43%	46.12%	97.9%	31.94%	85.37%	2.%	29.22%
Type of differences:							
Completeness of bibliographic record	61.32%	26.08%		19.23%	16.66%	61.35%	34.31%
Word order	.94%				16.66%		.65%
Spelling	14.15%	13.04%		11.53%	16.66%		7.51%
Main entry and/or title entry different	17.92%	34.78%		53.84%	50%	38.63%	19.93%
Other	5.66%	26.08%		15.38%			5.22%

Request Record vs. Machine Record

	I	II	IIA	III	IIIA	IV	Total
Request record not different from machine record	51	67	1	23		5	147
Request record different from machine record	106	23		26	6	44	306
Not determined if request record differed from machine record	105	77	47	23	35	1	288
Types of differences:							
Completeness of bibliographic record	65	6		5	1	27	105
Word order	1				1		2
Spelling	15	3		3	1		22
Main entry and/or title entry different	19	8		14	3	17	61
Other	6	6		4			16

Percentage of Words Used in Each Search

	I	II	IIA	III	IIIA	IV	Total
Ø search words	1%						.5%
1 search word	16%	28%	33%	18%	31%	38%	23%
2 search words	38%	35%	33%	22%	40%	16%	33%
3 search words	35%	26%	27%	42%	24%	28%	32%
4 search words	9%	11%	4%	17%	2.4%	10%	9.5%
5 search words	1%	1%	2%			2%	1%
6 search words						4%	.3%
7 search words							
8 search words						2%	.2%
9 search words							

Number of Words Used per Search

	I	II	IIA	III	IIIA	IV	Total
Ø search word	3						3
1 search word	41	46	16	14	13	19	149
2 search words	99	58	16	16	17	8	214
3 search words	92	43	13	30	10	14	202
4 search words	24	18	2	12	1	5	62
5 search words	3	2	1			1	7
6 search words						2	2
7 search words							
8 search words						1	1
9 search words							

Data Elements Used in Searches

	I	II	IIA	III	IIIA	IV	Total
Main entry	41		7	25	35	16	125
Title	15	119	36	11	4	8	193
Main entry + Title	201		3	35		4	243
Main entry serial indicator		1					1
Title and serial indicator		42					42
Truncated main entry			1	1	1		3
Truncated Title	2	4	1				7
Truncated title and serial indicator		1					1
Main entry with series						1	1
Series						11	11
Subject						3	3
Scan general note						4	4
Scan dissertation note						3	3
Null search	3						3

Percentages of Each Data Elements Used in Searches

	I	II	IIA	III	IIIA	IV	Total
Main entry	15.64		14.58%	34.71%	87.8%	32%	19.52%
Title	5.72%	71.25%	75%	15.29%	9.75%	16%	30.15%
Main entry plus title	76.71%		6.25%	48.61%		8%	36.09%
Main entry and serial indicator		.59%					.15%
Title and serial indicator		25.14%					6.56%
Truncated main entry			2.08%	1.38%	2.43%		.46%
Truncated Title	.76%	2.39%	2.08%				1.09%
Truncated title and serial indicator		.59%					.15%
Main entry and series entry						2%	.15%
Series						22%	1.71%
Subject						6%	.46%
Scan general note						8%	.62%
Scan dissertation note						6%	.46%
Null search							.46%

7. Nothing further was done to successfully completed searches.

8. Successful searches with too many hits were to be researched on the computer by the search coordinator to determine appropriate means of reducing the number of hits. This did not really prove a problem.

9. Unsuccessful searches were sent to the tabulator, to determine why they failed.

10. "Search Evaluation" sheets, hard-copy (from the terminal), and copies of NYSILL requests were sent to the tabulator.

11. i. Unsuccessful searches, if there was a Dewey number on the request which was within the required range, were to be searched on the Dewey listing of the computer data base to determine if the title was in the data base. If the title was in the data base, the manual search was to continue to ascertain why the title was not retrieved. In all but one case, the information from the Dewey listing, and the computer search were sufficient to determine why the search had failed. If the title was not in the data base, the manual search was to continue to determine why the title was not in the data base.

ii. Unsuccessful searches with no Dewey number on the request were to be searched in the public catalog to determine if the title was indeed in the library, and if so, what the Dewey number was. If the title was not in the library the search ended. If the title was in the library but should have been in the data base, the search was to continue to ascertain if the titl· was in the data base, and why was it not retrieved; or if it was not in the data base, why not. The results of these searches were to be recorded on the "search evaluation" sheet. This is fact was not done, but might be a good study to do.

12. Once the routines have been established, synthetic requests (Type IV) were given to the search assistant. These, time permitting, were searched in the Public Catalog first, and then on the computer.

13. Searches for retrieval item experimentation were to be done by the search coordinator. These were also to be done after the basic routines were established. These were not done due to the pressures of time.

14. Subject searches were also to be done, after the bulk of the other searches were done. There were however no NYSL subject requests to be found.

3.2.6 Development of evaluation methods

Search effectiveness is defined as the percentage of requests matched by entries in the data base which represents titles actually desired. The numerical value of search effectiveness varies with difficulty of search which in turn depends on type of request, search logic and vocabulary used, etc.

3.2.6.1 Search evaluation sheet

The test search data were recorded on a Search Evaluation Sheet (Figure 1) in order to make subsequent analysis and tabulation more convenient and to determine if the search was successful. A description of the Search Evaluation Sheet follows along with experience gained in using it. In its mode of use the tabulator (Madden) would be able to evaluate the searches without reference to the majority of NYSILL requests on the machine file. Due to the pressure of time, the forms were not always filled out completely; however the designated data fields still determined which items to evaluate a search by.

Item 1* - "Type"

The searches were divided into four categories or types so that in addition to measuring the effectiveness of the total sample of searches, sub totals could be determined by'

1. author title requests from NYSILL "I"
2. title main entry requests from NYSILL "II"
3. corporate main entry requests from NYSILL "III"
4. searches of all types not from NYSILL but simulated from entries known to be in the data base "IV"

* Item numbers refer to figure 1.

1 TYPE | II | 4 |

1,2,3,9,10,12,19,26,42,45,58   2:26—2:38

2 SEARCH# | | 3 DATE | 7 | / | 71 |

4 MONOGRAPH? ______ SERIAL? ______ KNOWN ITEM? ______ SUBJECT? ______

5 NUMBER OF SUCCESSFUL MACHINE SEARCHES ______

6 NUMBER OF UNSUCCESSFUL MACHINE SEARCHES ______

7 RECORD IN MACHINE FILE YES? ______ NO? ______

8 RECORD IN CARD FILE (P.C.) YES? ______ NO? ______

9 INFORMATION PRESENT ON REQUEST

10 FIELDS PRESENT ON MACHINE RECORD

11 FIELDS PRESENT ON SHELF LIST RECORD

12 REQUEST ACCESS POINTS EXACTLY LIKE MACHINE RECORD. YES? ______ NO? ______

if no, how do they differ?

SPELLING: ______
COMPLETENESS: ______
WORD POSITION: ______
OTHER DIFFERENCES: ______

13 EXPLANATION:

Figure 1

A fifth type "Subject" was provided for but not used because of lack of time.**

In addition each of the searches was catagorized according to whether or not it was found in the public catalog, an "A" being used to designate such if it was.

Item 2-Search Number

This item was devised to aid correlating NYSILL requests and computer searches. At first one computer search represented one NYSILL request, later several requests were included in one computer search. As it turned out the Search Evaluation Sheet became a cover sheet for a set of accompanying console printout sheets.

Item 3-Date

Self explanatory, however it was found useful to include the time of day of log on-log off of the search.

Item 4-Search Category

Monograph, Serial, Known item, and Search categories were used to record other attributes of a request in addition to Item 1.

** The BCN system was shut down during the course of the experiment and summer vacations limited the labor which could be spent by the searcher. The labor was deemed better spent in getting more searches done rather than transcribing data about them on the sheet. This of course shifted an unanticipated burden on the evaluator (contractor) because most data had to be gathered from the original documents.

Item 5-Number of successful machine searches

Item 6-Number of unsuccessful machine searches

These items recorded the number of successful and unsuccessful distinct searches accumulated for the requests. These data give an indication, before careful review of the runsheet, as to how well the computer search fared.

Item 7-Record in machine file

Item 8-Entry record in card file

This data was collected only for unsuccessful searches because if the search succeeded the record was in both files. These two questions provided data for the evaluation and follow up of unsuccessful searches.

Item 9-Information present on request.

The entry on the NYSILL request was to be transcribed onto this form. In most instances the NYSILL request itself was attached to the evaluation sheet.

Item 10-Marc Field present on machine record.

The MARC elements on each machine record matched was also to be filled in by the search coordinator when he verified the search results.

Item 11-MARC fields present on shelf list record.

This data was needed for verification purposes. Also, it revealed which records that should have been in the data file were not. A public catalog search would help determine why the record was not in the data base. Was it: not in 550-599 range, new acquisition, member of lost batch, etc.

Item 12-Request access points exactly like machine record.

This question was designed to record the variation that was found to exist between the request words and the machine record words. It was thought to be one of the most important items of experimental data to be collected. The design character of operational systems anticipates that matching must take place with incomplete or inaccurate data. The nature and possible kinds of inaccuracies which might occur must be known.

Item 13-Explanation.

This item was included for noting any problems or special conditions and was mostly used for elaborating differences in search words and machine search words.

3.2.6.2 Tally Sheet

The Tally Sheet was used to summarize the data recorded on the search evaluation sheets, and to record additional information about each request. The columns on the Tally Sheet included the following items:

1. Request search identification.

The request searches were identified by both the request number from the NYSILL request, and the log in-log out time for the search. If several requests were included in one computer search the total time for the entire search is shown divided by the number of searches to indicate an average time for each search.

2. Date of computer search.

3. Total searches for this request.

This column recorded the number of times a search was made for a single request. The experiment plan was to search each a maximum of three times if it was unsuccessful the first and second time. However of the 645 searches made, only 60 were second searches, and only 16 were third searches.

4. Monograph, Serial, Known Item, Subject.

This field was used to tally the request categories recorded on the search evaluation sheet.

5. Request.

This column was used to tally the type of request data appearing on the search evaluation sheet.

6. Record in Machine file, Record in Public Catalog.

This data was needed to determine whether an unsuccessful search was due to the fact that the title was not in the libraary (Public Catalog) or had not been encoded into the machine files. In many cases the fact that a search was successful, was sufficient evidence to indicate 'yes' for both questions regardless of what the search evaluation sheet said. Type III-A and type II-A requests, which did not have Dewey numbers listed were unverifiable without access to the Public Catalog and therefore on the NYSILL request were not tallied on these questions.

7. Information present on request.

These columns record which data elements were on the NYSILL Request.

8. Fields present on the machine record.

These columns record the data fields present on the bibliographic record. The Dewey list was the source of this information.

9. Request access points = machine access points

If no, how do they differ? This column summarizes the data on the separate search evaluation sheets.

10. "Request access points searched"

The field recorded which data fields and how many words of these fields were used in the computer search. Also, noted here was the use of the special search features of the BCN system. This information was taken from the computer hard copy.

11. Total number of hits.

The number of documents which matched this search are recorded in this column. In a multiple request search, the total hits for all requests were not recorded in this column. Only the number of hits that matched the search terms for the single request were tallied.

12. Success

Success means that the desired item was easily identified among the total number of items matched. No analysis of the ratio of relevant to non-relevant matches was made.

13. Reason for failure.

In the event that the search was unsuccessful, the tabulator had to determine why. The possible reasons for failure are:

a. typing error (on the computer)

b. not on the data base

c. search strategy

d. operator fumble (other than typing), e.g., incorrectly formulated search, omission of 'list' statement

e. terminal malfunction

f. system malfunction

g. request information misleading

h. machine record incorrect

i. untracable (type III-A add II-A)

j. should have been retrieved? (The search was systactically correct and the record was on the data base, but the search failed.)

3.3 Test Search and Evaluation

The search experimentation was carried out using the personnel, procedures, and forms just described and yielded good experimental results. The BCN system was understood to be an experimental tool only, and its shortcomings while they aggravated the experimenter somewhat, did not seriously affect the data collected.

3.3.1 Conclusions about search effectiveness

Table 3.2.6 is a display of the total results of successful and unsuccessful searches, and totals of errors which caused the unsuccessful searches. These figures represent the search effectiveness which could be expected if the BCN System were put into operation searching NYSILL requests using the current experimental procedure.

At first glance it might seem that the use of the on-line BCN system was a failure since the majority (52.5%) of the searches were unsuccessful. See lines 2 and 3. However this is not really the case because many unsuccessful searches were due to failures which would be corrected in an operational system. The following analysis of the data, points out such errors and describes briefly what can be done about them along with specific tasks for further experimentation.

3.3.1.1 Untraceable searches

The NYSILL requests, which were not found in the public catalog, became "A" requests for the on-line experiment. None of these "A" requests had a Dewey Decimal number on the NYSILL requests, and therefore it was impossible to trace them accurately in the Dewey ordered listing of the data base to see why they failed. The search tabulator using Dewey Decimal Classification classified the requests in an attempt to trace these records to determine why the search failed. Twenty-nine of the 86 requests could be classified nearly unambiguously but even so they were not found in the data base listing even through a scan of several Dewey classes bounding the classified NYSILL request.

We conclude therefore these 86 untraceable searches are due almost entirely to the machine record not being on the data base, and have assigned them to that error category.

3.3.1.2 Request record not on the data base

If the request record was not successfully retrieved, and if the Dewey number on the NYSILL request was not found on the Dewey listing of the data base, then the record was assumed not to be on the data base. Instances did arise when the Dewey number on the NYSILL request was incorrect, but these records were retrieved, or another record with the given Dewey number was listed in the Dewey Decimal listing of the data base. In this latter case, two records with the same Dewey Decimal number, the search tabulator had no means of determining which record had the incorrect Dewey number, so it was assumed the NYSILL request was correct.

The number of searches for which records are not in the file is 155 combining the untraceable and those known not to be on the file. In an operational system all the library's holdings will be on the data base. The computer will not be asked to retrieve what is not in the data base, unless for an acquisition support system. Thus the percentage of unsuccessful searches due to records not on the data base will not be as high as in this experiment, unless the data base is not kept up-to-date.

The new measure of effectiveness excluding these 155 searches is 62.5% successful and 37.5% unsuccessful.

3.3.1.3 Operator errors

There were three types of operator errors - query typing errors, search syntax errors, and search strategy errors.

3.3.1.3.1 Query typing errors

As the search assistant types a search in the BCN Doc. Proc. System command language, there is a chance he will make a typing error. This was more likely to happen at the beginning of the actual test searching and continued until the operator became more skilled. The number of failures due to operator typing was 37 or 5.8%.

3.3.1.3.2 Search syntax errors

As the search assistant formulates and types the search, there is a chance he would make an error in the required syntax of the search. For example, he might forget to ask that the search results be listed, or he might forget to specify the data base to be searched. The number of failures due to improper search syntax was 36 or 5.6%.

3.3.1.3.3 Search strategy errors

The search assistant was not instructed as to the meaning of type I, II, II-A, etc., designations. Given a group of requests to search, he made his own inference as to what type of citation each request was, and subsequently based his search strategy this inference. In some instances when the search assistant's inference was incorrect, the subsequent search strategy could not possibly result in a successful search. In other instances, incorrect inferences made no difference. The number of such failures was 33 or 5.2%.

3.3.1.3.4 Effect of operator errors on search accuracy

The total number of operator errors were 106 which caused a total of 16.6% of the searches to be in error. In an operational system several factors would exist which would prevent such errors and would correct those which were initially present so that they would be eliminated by the time the search was completed.

First in an operational system the terminal query language would be tailored to catalog searching so that it would be a simple console process not requiring the memorizing of a series of complex commands nor the typing of such search constants as database name. This simpler language would eliminate a large number of typing and search specification errors.

Secondly an operational system would be staffed by library personnel skilled in the reference and search function, who would become, after training, thoroughly familiar with the basic operation, and would make fewer conceptual errors.

Thirdly, a system designed specifically for search would have terminal language diagnostics and feedback error messages and comments so that logical errors which enter the system could be presented to the searcher for correction. Also in an operational system the correct portions of a modified query would not have to be retyped, rather they would be modified by an on-line editor

Finally the search speed of an operational system would be faster, so that searches which contained errors and produced ineffective results could be searched again with minimum expenditure of time and effort.

3.3.1.4 Equipment errors

The equipment failed during the testing due to both software and hardware. For purposes of the experiment the errors were categorized into terminal failures and system failures.

3.3.1.4.1 Terminal failures

These failures were primarily due to the fact that the terminal used was a light duty one, and appeared never to have been heavily enough used to work out its mechanical difficulties. Under the initial spurt of heavy use in the experimentation, failures occurred which were then fixed. There were 17 failures for a percentage of 2.7%.

3.3.1.4.2 System failures

The software and hardware of the BCN system failed 36 times for a rate of 5.6%. The combined equipment failure rate was 8.3% which in an operation system could be overcome by duplication of consoles, and repetitive searching. In all commercial time sharing systems, the downtime is less than 1% and the mean downtime interval is approximately ten minutes. Also most systems have a "fail-safe" feature so that most search queries and partial results are saved and a search can continue from where it left off.

If such procedures can be established to eliminate these failures, then the search effectiveness is 93% successful and 7% unsuccessful.

3.3.1.5

The remaining three reasons for unsuccessful searches can not be solved by present known developments. In an operational system, these are basic limitations and will contribute to error.

3.3.1.5.1. Machine record incorrect

These failures are a result of the input system used. They should be analyzed further to determine if there is any pattern amongst machine record errors. For example, are they tagging errors, or keying errors. Are keying errors most likely when inputing numbers? At this point in the experiment these errors do effect the success rate, but in any future system for the New York State Library machine records errors can be held to a minimum.

3.3.1.5.2 Requests information misleading

This was a particularly perplexing problem amongst requests that appeared to be author title requests, but which turned out to be requests for a member of a monograph series. Further work should be done to design a search strategy which will resolve many of these ambiguities automatically. Users should also be encouraged to re-search requests, trying as many different access points as possible.

3.3.1.5.3 Should have been retrieved.

It was impossible to determine why one particular search failed. The record was on the data base, the search was typed correctly, the search was formulated correctly, and the terminal seemed to be working properly. This we would label unassignable error which it would appear would occur in an operational system as it did in the experiment.

3.3.2 Characteristics of requests and matches in the on-line search experiment

During the experiment several data items were recorded to discover how particular classes of requests matched. These data are compiled and tabulated in the following paragraphs under the following topics:

a. The distribution of matches per successful request search

b. Differences between request record and machine catalog record

c. The distribution of the number of words used per search

d. Data element usage

3.3.2.1 The Distribution of matches per successful search

Table 3.2.7.1 is a display of the number of documents retrieved per successful search. The majority (74%) of the searches retrieved only the desired document, 17% of the cases retrieved the desired document and one additional document, and 5.3% of the searching cases retrieved the desired document and two others.

In preparation for searching, it was decided five was a tolerable number of hits, although only one of the five was the desired item. These results show if the search was successful, then it was very successful in terms of retrieving less than the allowed five documents.

One must remember all cases were not successful, and the above results may be only a portion of the picture. The unsuccessful searches which retrieved documents could also be tallied, to ascertain if the majority of all searches retrieve only one document, and if there is not a greater range in the number of documents retrieved. Also, this number of matches is dependent on file size so that if additional records were included, additional non-relevant matches would be made.

3.3.2.2 Request record v. machine record

The statistics on Table 3.2.7.2 show that differences did exist between request records and machine records. There were differences in 47.81% of all searches and there were no differences in 22.97% of all searches. This finding corroborates what we already know, nwmely that few library patrons will know the exact catalog entry. Any query system designed for on-line searching without the book in hand will need to bear this in mind.

It was not determined if the request record differed from the machine record in 29.22% of the searches. It was assumed that these cases would later be re-searched. These requests should be re-searched until it can be determined whether or not the request record is the same as the machine record.

The type IV requests had a high percentage of different entr'es (38.63%) and a high percentage of incompleteness (61.36%), because they were constructed to be misleading and often leading words were omitted from the request.

Type I records had a high percentage of incomplete records (or entries) 61.32%, because NYSILL requests often include only the author's initials, instead of the full name. This did not hinder retrieval. The surname and initials, in most cases, uniquely identify the author.

3.3.2.3 The Distribution of the Words used per search

The majority (88%) of all searches used 1, 2, or 3 words in a search. 'Ø' search words was an error in search syntax on the part of the searcher. 23% used 1 word, 33% (the largest group) used 2 words, and 32% used 3 words.

The range of words per search was 4 for types III and IIIA, 5 for types I, II, IIA, and 8 for type IV. Type IV, the synthetic searches, tended to use more words probably because the searcher was not sure what he was searching.

No particular number of words per search was used in a majority of cases.

In conclusion, no matter what type of request record is being searched, the search system must allow for multiple word searches. Search algorithms should be based on from one to three words of the search request.

3.3.3 Percentage Data Elements

Table 3.2.7.4 entitled "Percentage of Each Data Elements Use in the On-Line Searches" is a tabulation of the MARC elements used in the searches along with a tabulation of the use of the BCN-DPS special search features. The MARC elements that were used either alone or in combination with other MARC elements are: main entry, title, serial indicator, series, subject, general note, and dissertation note. The BCN-DPS transaction search allows one to spell the beginning of a word and punctuated with a special symbol ($). This is useful in searching for inflected forms. For example, one can search alternate, alteration, alternatives by typing alter($). A second Biomedical search feature that was used is the scan feature whereby one may search data fields that were not automatically indexed by the BCN.DPS system. Thus one may search the general note field for a report number, or the dissertation note fields for indication that the document is indeed a dissertation.

As the table indicates the frequency of the use of MARC elements depended very much on the type of request being searched. The search assistant would infer from the NYSIL request that he had a personal author citation, corporate author citation, etc. Personal author request (Type 1) was searched under main entry and title fields in 76.71% of the searches. Title main entry (Type II) was searched under title in 71.25% of the searches. Synthetic requests (Type IV) required the greatest diversity of

MARC data element categories. This is largely because the synthetic requests were assimilated from a wider range of data fields than are liable to be in the NYSIL requests.

In general fixed field information was not used in any of the searches. The serial indicator is the exception to the rule and it was used in 6.86 of all searches. It should be kept in mind that there were not many serial entries included in the data base, 1000 of the 18,000 records were serials. Therefore successful searches of serial records were not as numerous as non-serial searches.

The BCN truncation feature was used in 1.7% of all searches. The scan feature was used in 1.00% of all searches. It should be noted that the scan feature was only used on data fields non indexed automatically by BCN-DPS. It is costly to use the scan feature, as it searches character by character. Therefore, it is wise to avoid the scan feature whenever there are other possible search access points available. However, the utility of such a feature and the truncation feature should be kept in mind.

It should be noted that many of the MARC data fields were not used in the searches. Some of these data fields are: imprint, collation, edition, government printing office number, SBN number, etc. Many of these fields are either incomplete or omitted in the NYSIL request so that the search assistant might not have had them in front of him when he made his request and therefore did not use them.

Section 3.3.4 Synthetic searches

Synthetic search requests were so named for their method of development. Catalog citations within the given Dewey classification range were rewritten as citations without the required NYSLL information. In other words they were rewritten as vague reference questions first encountered at the reference desk before the search refinement given NYSLL requests. A total of 74 synthetic search requests were searched by the Science and Technology Division in the catalog, but only 50 were searched by the search assistant in the machine data base so all statistics refer to the 50 searched in both files.

Synthetic Search Results

Found in both	19	38%
Found only in card cat.	3	6%
Found only in machine file	19	38%
Neither	9	18%

Overall the computer search fared much better than the human search. This is true despite the fact that these vague synthetic searches were searched by personnel from the Science and Technology divisions, who would be most familiar with the subject area. No record was kept of how long each catalog card search lasted, or of how many false starts were made before some answer, the correct title or no title, discovered. This information is available for the machine searches; only two of which were searched more than once, and one of these was not necessary. It is very likely that the card catalog searched were tried more than once.

It is a very rare and valuable searcher who can solve reference questions with one try. On the other hand, the machine techniques which although not well suited to library problems still had a success rate of only one search of over three times that of the (experienced) human searcher.

The card catalog was successful in 3 searches (6%) in which the machine failed. It is interesting to note why the machine search failed. In the first case, the search assistant "and-ed" three title words. Unfortunately one of these words was in fact not in the title, so the search failed due to an operator error. The human searcher probably knew the anacronym included in the title, and so had little trouble in the card catalog. In the second case, the search assistant searched on more of the series entry than was indexed. This points up an improvement for future systems: all series entries should be fully indexed. In the third and last case, the author's surname was searched in the general notes and not in the contents notes, as it should have been. This is an operator error. If he knew the surname was not the main entry, but a note then he should have guessed that it was a content note not a general note. Thus there were two operator errors and one data base limitation failure. Is is very likely both operator errors would have been caught on a second search. Improved indexing could be obtained through the use of a larger file so that the entries could be found by query using any access word.

In sum these three cases do not really prove the card catalog superior to the machine search strategies or files, but rather show problems that must be provided for in an on-going library search system. The machine searches were far more successful on only one search than the card catalog searches.

4. THE USE OF MODIFIED MARC RECORDS IN THE NYSL EXPERIMENTAL DATA BASE

The design of the conversion experiment provided for two types of encoded records. One a complete catalog record encoded using a complete set of MARC II tags, and a second complete catalog record encoded using an abbreviated set of tags. The intent of experiment with an encoded file using abbreviated tagging was to determine whether costs would be saved converting to such a file, and if so whether searching effectiveness would be impaired using abbreviated tagging. In the conversion experiment, two tasks were carried out to create these files. Task I created fully tagged records and Task II created fully tagged records and Task II created records with abbreviated tagging.

4.1 Specification of Task II Tagging Elements

The file to be created in Task II differed from fully tagged MARC II records in two general ways, one the record format and character set were different, and two the data elements included in the record were different.

4.1.1 Differences in record format and character usage

The following list describes differences of the NYSL Task II record from the standard L.C. MARC record (Task I record)

1. Task II records use the character set specified in Table 1. No lower case alphabetic characters are used.
2. The character "$" is used as the subfield delimiter, "≡" is used as the end of field mark, the two characters "≡≡" are used as the end of record mark.
3. Accent and diacritical marks are not used.

4.1.2 <u>Differences in data elements in the Task II records (abbreviated tagging)</u>.

1. All indicators are blank except in fields

Field	1st Indicator	2nd Indicator
041	0 or 1	b̸
100	b̸	0 or 1
110	b̸	0 or 1
111	b̸	0 or 1
260	0 or 1	b̸
400	b̸	0 or 1
410	b̸	0 or 1
411	b̸	0 or 1
505	0, 1 or 2	b̸

2. The following fields are never present:

040
051
060
070
071
241
350

3. The following subfields never occur:

050 subfield "B"
300 subfield "B" and "C".

4. In field 008

Character 22, Intellectual code must always be b̸.
Character 30, Festchrift indicator must always be 0.
Character 31, Index indicator must always be 0.
Character 32, Main Entry indicator must always be 0.
Character 33, Fiction indicator must always be 0.
Character 38, Modified record indicator must always be 0.
Character 39, Cataloging code must always be b̸.

5. Character 22 of the record leader is the completeness indicator. LC does not use this indicator.

I-record does not contain all bibliographic information.
C-record has complete bibliographic information.

6. Character 23 of the record leader is Task indicator. LC does not use this indicator.

1-Task 1
2-Task 2

7. Data field 001 contains the NYSL Accession number, not the LC-Card Number. This field is always 13 characters (including end of field mark) right justified and blank filled.

8. Data field 010 contains the LC-card number, LC does not use this field. $a is the only subfield code. The indicators are not used.

9. Data field 082 contains NYSL-Call Number, not the LC-Dewey Decimal number. $a is the only subfield code and can occur only once. The indicators are not used.

10. Data field 901 is the NYSL-Holdings, LC does not use this field.
    $a - delimits the start of the data.
    $b̸ - separates each accession number from the holding information.
    The indicators are not used, for example:
    b̸b̸$A1406780$b̸c.1,b̸1406781$b̸C.2,b̸1406782$b̸C.3≡

11. Data field 902 is the NYSL-Batch control field, and is always present. LC does not use this field.
    $a - is the only subfield code.
    The indicators are blank. The data is a 2 digit Batch number, a 1 character Cataloging code ("A", or "C") and an optional inventory code (if present it is "M", i.e., missing in last inventory).

12. Field 008, character 28, the Government Publication Indicator, is not like LC MARC. It contains:

b̸	not a government document
F	federal
L	local
S	state
A	foreign
U	internation

13. Character 7 of the leader, the Bibliographic level is not like LC MARC. It contains:

M	monograph
S	serial, with complete holdings listed
P	serial, with incomplete holdings listed

4.2 Experimental Results

The results of the conversion experiment showed that there was no discernable difference in the cost of encoding either the Task I or Task II records. Although it would seem that there would be savings using the abbreviated tagging, the following observations were made which support the finding:

4.2.2 Number of characters per record nearly equal for both records

The number of characters per record saved by abbreviated tagging was insignificant. Therefore all keyboarding, proofreading and data file correction costs were equivalent.

4.2.3 Editing and tagging difficulties not related to abbreviated tagging

The last remaining cost element, editing and tagging as it turned out, is not reduced by an abbreviated set of tags. The difficult tagging decisions (those which contribute excessively to cost) are of a cataloging nature and exist whether or not an abbreviated set of tags are used.

NYSL MARC TAPE CHARACTER SET

Tape: 7 tract odd parity 800 bpi

Octal	Character	Octal	Character	Octal	Character	Octal	Character
00	0	20	+	40	-	60	ƀ blank
01	1	21	A	41	J	61	/
02	2	22	B	42	K	62	S
03	3	23	C	43	L	63	T
04	4	24	D	44	M	64	U
05	5	25	E	45	N	65	V
06	6	26	F	46	Ø	66	W
07	7	27	G	47	P	67	X
10	8	30	H	50	Q	70	Y
11	9	31	I	51	R	71	Ƶ
12	:	32	<	52	!	72	]
13	=	33	.	53	$	73	,
14	≠	34	)	54	*	74	(
15	'	35	"	55	#	75	_
16	%	36	@	56	&	76	≡
17	[	37	;	57	>	77	?

Note:

15	apostrophe
75	underscore
76	triple hyphen

5. CONCLUSIONS AND SUGGESTIONS FOR FUTURE RESEARCH

From the experimental work performed, we conclude that converting the NYSL shelf list sample to machine readable form, and searching this shelf list using a Remote Access Catalog are technically sound concepts. The conversion procedure used is workable and would be practical in a production environment with a few technical modifications. A second finding is that the search methods used in the experiment are satisfactory. The search "power" was found adequate using the experimental system and although slow, by analysis one could design a system which can search faster and more conveniently than manual searching of card files.

5.1 Cost of a Remote Access Catalog

Although the experiments showed that the projected conversion and search concepts supporting a Remote Access Catalog to be technically satisfactory, the capital costs of data conversion and system installation will be substantial.

Data conversion costs were calculated to be $1.74 per entry so that $1,113,600 is required to convert the 640,000 entries in the shelf list. By comparison with other systems, the conversion costs were shown to increase with an increase in quality. In order to reduce costs, any projected NYSL conversion should make the most of retrospective conversion performed by L.C. and others. Per record costs using available records are estimated to cost $.26. Although no data are available from which one could calculate a percentage of one's holdings available in other libraries, there is a potential cost reduction of 85%.

No data are available at this time from which one could project savings using format recognition. From the experimental work performed on encoding records with abbreviated tagging system, we conclude that the cost savings are insignificant. Within the accuracy of the experiment discernible cost savings could be verified. Further, if there were minimal savings they have to be weighed against the disadvantage of not creating standard records which could be shared with other libraries, a concept that appears to offer greater data base encoding savings.

5.2 Future Research

During the course of the experimentation, problems occurred, questions arose, and new ideas occurred to the investigators which could form the basis for further research. We have divided these suggestions for future research into three categories:

a. Investigation of new or modified concepts in conversion and search.

b. Specification of improvements needed for a production system.

c. The specification of new uses for a remote access catalog.

5.2.1 Investigation of new conversion and search concepts

5.2.1.1 The use of on-line prompting terminals to convert data

An alternative to complete tagging of MARC records is to have the typist encode data by following the queries of a computer program connected to the terminal. Such a procedure holds the prospect of substantially reducing the amount of tagging required.

5.2.1.2 Search of NYSL sample in other data bases

It would be useful to search a sample of NYSL entries in the libraries which have encoded their retrospective collection in MARC II format. This overlap analysis would provide information needed in the calculation of the costs saved by using the MARC records of other libraries.

5.2.1.3 Investigation of the loss of utility of a catalog containing a certain percentage of error.

There is an indication that errors in a catalog do not impair its usefulness as much as its esthetics. A study of this problem with the objective of determining an acceptable error rate would be useful, and perhaps point the way to lower cost conversion of an acceptable quality.

5.2.1.4 Investigation of additional search logic capabilities

There are problems of searching where the truncation feature is not powerful enough, such as text book and textbook. In addition it is thought that positional description of terms (A followed by B) would be useful. An analytical investigation of these and other complex search problems such as corporate author would be useful and necessary before any large committment was made to a future system.

5.2.1.5 Study of search system response time characteristics

In the experiment, waiting for searches, although it was understood as a limitation of the experiment at the outset, was bothersome. However little or no data was collected which would tell one what an optimum response time was. Also it is not known whether a complete search is needed rapidly or merely the search to the first item which can be displayed. In the latter approach, the computer system would use the rcading time taken by the user to perform additional searches, and would relieve peak loads. An analysis of this search response time requirement would be useful in determining equipment configuration and cost characteristics of any Remote Access Catalog Configuration.

5.2.1.6 Search system respor e time

One result of the search experiment pointed out that there may be two classes of searches, one of which is bothersome to wait for, the other not. When the result of a search is unpredictable or where feedback is necessary to reformat the search then rapid system response is necessary. Searches which are predictable such as interlibrary loan request for monographs searched by personal author main entry do not need a rapid response because the outcome is definite and depending on match or no match the next step is known. We think this is an area worthy of further study.

5.2.2 Specifications of improvements needed in a production system

The present experiment experienced a great deal of difficulty in document control. It would be useful to specify a control system consisting of both manual procedures and computer tabulation which would insure that production flowed smoothly with no lost documents or duplicated conversions.

5.2.3 Study of additional use of the remote access catalog

During the course of the experiment several questions arose concerning the use of the data base for other than the Remote Access Catalog and the effect of those uses on the requirements of the data base. In addition, the large cost of the data base suggests that it be used for other purposes. It seems appropriate therefore to study the possible uses of such a base and document them into a set of general system requirements.

APPENDIX

TERMINAL
PRINTOUT
SAMPLE

```
s#
      ENTER SEARCH NOW.
option geobook1,stat;
l1 @scoot;
l2 plant & symbiosis;
l3 @goetsch;
l4 ants;
l5 @rogers;
l6 techniques & autoradiography;
l7 l1 & l2;
l8 l3 & l4;
l9 l5 & l6;
l10 l7,l8,l9;
list ddc,mea,til,imp;
end;
SEARCH INPUT ACKNOWLEDGED.
OPTION GEOBOOK1,STAT;
L1 @SCOOT;
L2 PLANT & SYMBIOSIS;
@SCOOT
DERR012 (WORD) KEYWORD IS NOT IN DICTIONARY
L3 @GOETSCH;
L2       =========000000==========
L4 ANTS;
L3       =========000001==========
L5 @ROGERS;
L4       =========000033==========
L6 TECHNIQUES & AUTORADIOGRAPHY;
L5       =========000028==========
L7 L1 & L2;
L6       =========000002==========
L8 L3 & L4;
L7
DERR012 (WORD) KEYWORD IS NOT IN DICTIONARY
L9 L5 & L6;
L8       =========000001==========
L10 L7,L8,L9;
L9       =========000002==========
LIST DDC,MEA,TIL,IMP;
L7
DERR012 (WORD) KEYWORD IS NOT IN DICTIONARY
L10      =========000003==========
END;
KYWD01            0000000001
KYWD02            0000000001
KYWD03            0000000033
KYWD04            0000000028
KYWD05            0000000001
KYWD06            0000000001
KYWD07            0000000001
KYWD08            0000000002
RESULT            0000000002

0002         L10
DDC:         578.4/R724
MEA:         ROGERS ANDREW W.
TIL:         TECHNIQUES OF AUTORADIOGRAPH , BY ANDREW W. ROGERS.   #233
IMP:         AMSTERDAM, NEW YORK, ELSEVIER PUB. CO., 1967.

1 010    0   L10
DDC:         595.796/G599
MEA:         GOETSCH WILHELM, 1887-
TIL:         THE ANTS. TRANSLATED BY RALPH MANHEIM        #20
IMP:         ANN ARBOR, UNIVERSITY OF MICHIGAN PRESS 1957
```

Name Clure

Date 2/16

Language LAN□ ENG	
Fixed Fields	
FFD Govt Pub 1. F	Conf/Meeting 2.
Publisher is M E 4.	Biography 12.
Subject is M E 13.	Pub Date Key 20. S
Date 1 21. 1951	Date 2 22.
Country of Pub 23. DCU	Repro Form 25.
Contents Forms 26.	Bib Level 27. M
Modified Record 28.	902. A

SAMPLE MANUSCRIPT

New York State Library

664

NOT

author entry originally:
‡ Muesebeck, Carl Frederick William ‡ 1894–

DDC
t1C‡bc

CRP
IMP‡bc
COL
SEC‡‡v
NO6
Su‡‡z
AEP‡d

CAL
LCN

595.79
M948

755481

‡
~~595.79~~
~~M948b~~
~~v.1 2/14/59~~
~~970637~~
~~v.2 5/3/67~~
~~1240940~~
5/3/67

Hymenoptera of America north of Mexico; synoptic catalog. ‡ Prepared cooperatively by specialists on the various groups of *Hymenoptera* under the direction of C. F. W. Muesebeck, Karl V. Krombein and Henry K. Townes. Washington, ‡ U. S. Govt. Print. Off. ‡ 1951.

1420 p. fold. map. ‡ 24 cm. (U. S. Dept. of Agriculture. ‡ Agriculture monograph ‡ no. 2)

——— Supplement. 1st– Sept. 1958–
Washington, U. S. Govt. Print. Off.
no. 24 cm.

~~HD1751.A918 no. 2 Suppl.~~

1. Hymenoptera ‡ North America. I. Muesebeck, Carl Frederick William ‡ ~~1894–~~ ~~(Series)~~

HD1751.A918 no. 2 ~~595.79097~~ Agr 51–324 rev

~~U. S. Dept. of Agr. Libr. for Library of Congress~~ ~~[illegible] no. 2~~ ~~made T~~

TAG	DATA
901	970637 ‡v.1, 1240940 ‡v.2

12/