P

OGC 70-0859

26 May 1970

MEMORANDUM FOR: Mr. Houston

SUBJECT: RSM-LITE Comparison Test Results: Application to OGC

1. Attached is the report on the RSM-LITE comparison test, the objective of which was to assess the RSM legal information retrieval capability relative to OGC needs. As a result of the successful completion of this testing and evaluation period, we will soon have the capability to search and retrieve by machine the U. S. Code and all Comptroller General Decisions, both published and unpublished. In total, these two data bases will constitute approximately 50 million words. When these bases are updated by LITE, we will receive copies long enough to update ours.

2. Given the RSM availability in any event, and the assistance of LITE in providing us copies of these particular data bases without charge, we have an in-house capability comparable to the LITE service at minimal cost. Currently, the largest single cost in developing such a system is attributable to converting the source materials to machine readable form. A "rough office survey" would suggest that current and foreseeable OGC research needs would not justify the present costs of converting other legal research materials to machine readable form.

3. The purpose of this "office survey" was to determine what legal research materials were being used by the staff attorneys and, more importantly, the frequency of that use. To accomplish this, each attorney's "signer," which normally should contain all his written research papers and opinions, was reviewed over a 19-month period

A1

from January 1968 through July 1969. This particular 19-month period is probably representative of any other. An analysis of these materials disclosed:

a. The legal research materials cited most often in OGC written opinions and papers, in descending order of frequency were:

1. U. S. Code
2. Code of Federal Regulations
3. Internal Revenue Code (treated separately from U. S. Code)
4. Federal Personnel Manual
5. Comptroller General Decisions
6. Federal Reporter, 2nd Series
7. Federal Supplement
8. U. S. Supreme Court Reports, Law. Ed.
9. Va. Code, D. C. Code & Revenue Rulings (tied)
10. Standardized Regulations, Government Civilians, Foreign Areas, Department of State

b. U. S. Code provisions were cited in a total of 66 opinions over the 19-month period for an average of 3.5 opinions per month. The second most frequently cited research base, the CFR's, appeared in 19 opinions or an average of 1 per month.

c. 82% of the U. S. Code cites were restricted to five titles--5, 18, 28, 31 and 50. Title 5 accounted for 38% of all U. S. Code cites.

4. Doubling these "frequency of use" figures to take into consideration research which does not appear in written opinions suggests the OGC research need is insufficient to justify the cost of converting these materials to machine readable form. It should be noted we currently have a KWIC index of title 50, U. S. Code, and will be receiving a similar index covering title 5, two of the five titles most often used by this office.

A2

5. The foregoing treats with legal research materials such as those comprising our library. A question naturally arises concerning our own internal office papers and opinions. Should these be converted to machine readable form for storage and retrieval purposes? This type material must be readily accessible on a timely basis. It is the opinion of the undersigned that until the technology and more particularly, the economics of that technology, permit all staff attorneys convenient and continuous access to an up-to-date data base, the office should not make any large investment in this direction. This is not to say that interim steps in the direction of computerized storage and retrieval should not be taken. Such interim steps might include converting our office opinions and memoranda to a microfilm or microfiche storage and retrieval system. This would have the effect of solving our current storage problem and yet be a step in the right direction. The economical conversion of microfilm or microfiche to machine readable form should be a reality in the not too distant future.

STAT [redacted]
Office of General Counsel

Attachment

STAT OGC [redacted]
Distribution:
Orig. -✓COMPUTERS
STAT 1 - [redacted]
1 - Chrono

A3

RAPID SEARCH MACHINE v. AIR FORCE LITE ACTIVITY COMPARISON TEST

relative to the needs of

THE OFFICE OF GENERAL COUNSEL
CENTRAL INTELLIGENCE AGENCY

STAT Prepared by: [redacted]
Office of General Counsel
25 May 1970

TABLE OF CONTENTS

INTRODUCTION

In June 1969 the Office of General Counsel (OGC) was notified the Agency was in possession of a specialized machine for searching and retrieving full text materials--the Rapid Search Machine (RSM) STAT prototype I, manufactured by [redacted] A demonstration of its capability was offered using as a data base a segment of the Department of Defense Directives acquired earlier for demonstration purposes from the Air Force LITE (Legal Information Thru Electronics) activity.

Coincidentally, but nevertheless quite timely, a year earlier in the summer of 1968, as the undersigned was preparing to leave for a year's study in Systems Analysis at MIT, the Deputy General Counsel, John S. Warner, suggested that if the opportunity presented itself, an examination of Air Force Project LITE might prove an interesting academic project and be of practical future use to OGC. Mr. Warner's advice was accepted and the undersigned's term project was devoted to an analysis of the state of the art of computerized storage and retrieval of legal information, in particular the system known as Project LITE. The undersigned completed the MIT academic program and returned to OGC the first week in July 1969 in time to view the previously scheduled demonstration on the RSM I.

During this demonstration, we learned the Agency had contracted for the purchase of the RSM prototype II, scheduled for delivery and installation during November 1969. We were advised that RSM II, as compared with RSM I, would function at faster speeds, permit greater flexibility and manipulation of textual materials and provide key-word-in-context (KWIC) output. The fact that the RSM II would be available in any event suggested that for a relatively small additional cost and with the cooperation of the Air Force LITE activity, we possibly had in the offing an in-house capability to search and retrieve full text legal information by machine.

A6

The objective was to assess the RSM capability relative to the particular needs of OGC. Why not compare the RSM capability with that of LITE, an existing operational system? A detailed cost-efficiency analysis comparing the two systems was not intended or needed. If legal data bases of particular interest to OGC could be obtained without charge from LITE, the costs definitely favored RSM since it was available for use in any event. The comparison would be limited to the "search and retrieval" function of the two systems. In response to a particular query, would the RSM produced results equal or exceed those produced by LITE? Whether one system during the actual search and retrieval phase could operate slightly faster or handle greater loads in a given time frame was not considered particularly germane to OGC needs.

On 18 July 1969 the General Counsel, Lawrence R. Houston, in a letter addressed to Lt. Col. Charles R. Minich, USAF, Chief, LITE Branch, wrote the following:

> As you are now aware, we have, located in our Headquarters facility, the prototype Rapid Search Machine (RSM) built by [redacted] I would like to see a practical demonstration of the RSM capability to search and retrieve legal information. In this regard, we have concluded that the most pertinent test and evaluation of the system would be achieved by using two test data bases -- one statutory law and the other, decisional law. The U. S. Code and the Comptroller General Decisions are used perhaps as much as any other body of law in meeting the legal research requirements of my office. Inasmuch as both of these law bases constitute a part of the LITE data bank, it would be appreciated if you could make them available to us on a loan basis.

STAT

* * *

> When all is ready, it would be extremely helpful if one of your staff personnel could come to our Headquarters building, at our expense, and participate in the demonstration. It is our thought that someone intimately associated with the LITE development and its present capabilities could provide us the greatest assistance in evaluating the RSM legal information retrieval capability.

On 28 July we received from LITE ten reels of tape with documentation. Seven reels constituted volumes 19 through 47 of the Published Comptroller General Decisions and the other three reels comprised the U. S. Code through Supplement III.

Due to the need to redo the tape format when changing from RSM I to RSM II use, we decided to wait until the latter was installed before converting the larger part of the LITE tapes (Comptroller General Decisions). During the interim we proceeded with the conversion of the smaller data base (U. S. Code) for use on the RSM I so that we might become somewhat proficient in the operation of the machine and gain some insights into its capabilities and limitations.

The RSM II was installed in November 1969, and the process of preparing the tapes for use was begun. On 3 December John Warner and the undersigned visited the LITE facility in Denver, Colorado for the purpose of seeing firsthand the LITE operation. Mr. Jack Sieburg, Chief, EDP Branch, LITE, took us step by step through the LITE system process.

While in Denver we discussed the parameters of a "rough comparison test" between the LITE system and the RSM II. It was decided LITE would provide us with actual user queries presented to it for search and retrieval purposes. We would receive a total of twenty-four (24) queries-- one-half to be run against the Comptroller General Decisions, volumes 45 through 47, and the other half against the U. S. Code through Supplement III. The undersigned would take two approaches in handling these queries on the RSM. The first approach would be to duplicate as closely as possible the framing of the query actually used by LITE to retrieve material responding to the problem. This is not always possible because of the query constraints available in one system and not the other. The second approach would be to change the framing of the query, where appropriate, based upon the results produced by the first approach and looking to the statement of the problem actually presented for research. This second approach suggests a browsing capability that is only present when the user and the machine interact.

The undersigned would then review and compare the materials retrieved by both RSM and LITE in response to the same problem. Such a review would include a determination of their relevancy to the problem presented. Granted, what is relevant for one user, even when defined carefully, is not necessarily relevant to another researching the same problem. Also, it is quite probable that the framer of the query for LITE had received from the requestor certain additional information surrounding the research problem which did not appear in the simple statement of the problem as made available to the undersigned--information which would make retrieved materials which otherwise seemed irrelevant, appear to be relevant. Taking this into consideration, the undersigned consciously attempted to err on the side of excluding as irrelevant only that which did not appear in the slightest measure associated with the general subject matter. It should be noted, however, that in those instances where only irrelevant materials are retrieved, that fact in itself might be relevant to the point that there are <u>no</u> relevant materials on the subject in the data base. This knowledge of a negative can often times be of great importance in legal research.

The actual test and comparative analysis took place, as time permitted, from January through April 1970. While the test parameters were probably of sufficient size and number to validly support the limited test objective, they are insufficient to draw fast and hard conclusions regarding the value and efficiency of computerized full text legal information retrieval systems in general. Some general <u>observations,</u> however, will be made in this area and properly identified as such.

[redacted] STAT

Office of General Counsel

A9

RSM II v. LITE: DIFFERENCE IN MECHANICS

The following paragraphs treat with the basic differences in the mechanics of the two systems, a knowledge of which is essential to understanding much of what is said about the actual test results. Differences in the hardware, the process of submitting queries and the query logic parameters are presented in that order.

Hardware-Software Comparison:

RSM II is not an all purpose computer, but rather a specialized machine designed for the specific task of searching and retrieving textual type materials. It consists of three basic units--(1) a console which includes a standard keyboard and other operating controls for inputing queries, a cathode ray tube (CRT) used solely for framing queries and an electrostatic printer for output; (2) an equipment and assembly rack; and (3) a tape drive.

The only software requirement is that involved in formating the data on magnetic tape for serial processing of a search. Every character of the full text is searched at the rate of 120,000 characters a second. Printout at a rate of 400 characters a second can be regulated to a designated subrecord(s) of a record, KWIC or an extract of text surrounding the key words searched on, or full text.

Currently there is an undesirable constraint in utilizing the extract (KWIC) mode. Within a record, one cannot restrict KWIC printout to a particular subrecord and print full text of other subrecords. For example, a section of the U. S. Code constitutes one record and is comprised of three subrecords--(1) title and section number, (2) section number and heading, and (3) full text of the section. Regardless of whether full text or only extract printout is desired, one will always want subrecord (1) printed in order to identify the text. Currently a printout of subrecord (1) and an extract of the text can only be obtained by a technique which requires using up query framing space which is at a premium. We are advised the RSM can be modified without great difficulty to overcome this problem.

Currently, the U. S. Code takes up three reels of tape and the Comptroller General Decisions considerably more, requiring the physical changing of many tape reels if a comprehensive search is required. Also, if such a search is required, this process of many reel changes somewhat inhibits the browsing capability--i.e., with each reel change, the user must start the query framing process all over again. We are advised that a Leach tape drive currently being installed will reduce this problem considerably by permitting a much greater density of information per reel of tape.

The LITE system utilizes an all purpose, third generation computer--the RCA Spectra 70/45. Prior to January 1970, LITE had functioned in an emulator mode of operation. The emulator made the RCA Spectra 70/45 computer function as an IBM 1410 computer thereby preventing maximum use of the efficiency and speed of the Spectra 70/45. LITE has no remote devices for input/output. The operating system provides for batch processing of queries. Key punch cards are used for query input and output is by high speed printer.

LITE has developed over the years a number of software programs which permit manipulation of their data bases to provide special products in addition to the general search and retrieval function. These products constitute indices, collations, frequency listings, etc., which are in themselves research tools that have a continuing benefit over a period of time involving many research problems.

LITE utilizes Random Access Card Equipment Units (RACE) to eliminate serial processing of text files. The system searches against a "vocabulary file," consisting of all the words used in the full text except for common words such as "of," "and," "or," etc. LITE has a standard common word list of 120-plus words. The elimination of these words reduces the total volume of text words searched against by approximately 40 to 50 percent. The "vocabulary file" contains the key word along with the associated string of address locations in the "text file" which contains the full text information for retrieval. LITE also has three modes of printout--(1) citation, (2) KWIC and (3) full text. The key words used in the search are easily located in the printout by appearing in the right margin next to the line of text wherein the word appears. This saves considerable time in browsing an extract of text for purposes of determining relevancy to the question asked. The RSM does not have a similar feature.

Comparison of Process of Submitting Queries:

It should be recalled at this time that the objective of the comparison test was to assess the RSM capability relative to the particular needs of OGC. How would a staff member of OGC submit a query to either system?

Looking first to RSM, he would call a number and reserve time on the machine. Currently there is no difficulty in obtaining sufficient timely access to the RSM to meet OGC needs. He would then pick up from the tape library the data base desired, and mount and dismount the tape(s) on the tape drive.

At this point he sits at the console, and the interplay between researcher and the retrieval device begins. The researcher frames queries, narrowing, broadening or terminating those queries as suggested by an on the spot review of the extract printout. Much the same way traditional manual research is done, the researcher can, by process of trial and error, consult one term or phrase, then another, until he pinpoints the materials needed.

It should be emphasized that the mechanics of operating the RSM are not in any sense overly difficult or complicated. An hour or two of instruction at the console would generally be sufficient to permit one to operate the machine on his own.

Looking next to LITE, the OGC staffer could relay his research problem to the Denver facility via telephone or letter. LITE has over the years conducted one and two-day training courses in the art of framing queries for submission to the system. A user so trained would then frame his own query on a special form and submit it to the Denver facility. Evidently this process has not been too successful.

LITE now suggests that the best results are obtained by calling one of three LITE staff attorneys who is an expert in the LITE framing process. The research problem would be presented and discussed over the phone, framed by the expert, and the results mailed out that same day or the next. Prior to mailing, the framing attorney reviews the results. His review might consist of, among other things, an informal notation (to the Requestor) of any peculiarities in the output, ideas for possible reframing or follow-up searches, etc. No attempt is made to determine whether the search gives "the answer" to the problem presented, as this is considered the Requestor's decision.

What should be noted here is that the Requestor may, indeed, need to reframe, given the first set of results. Once the question is put to the computer,that is it. To modify the request, it must be resubmitted. Each time a search request is submitted by our OGC staffer the charge is fifty ($50) dollars. Acting through an intermediary can be costly, not only in monetary terms, but also the time required to submit and resubmit queries.

Query Framing and Logic Parameters Comparison:

The CRT screen on the RSM enables the formulation of two separate search queries via the CRT keyboard. Data typed into the query areas of the CRT screen can be edited or modified (type, insert and delete operations) prior to the transmission of the data into the RSM.

Each search query on the RSM can be typed as either a single field containing 80 characters or less (single field request mode) or divided into four 20-character fields (multiple field request mode) depending upon the data search being conducted. The query data entered is formulated as a series of terms linked together in a specified manner using special term operators as required. Logic operations (AND/OR) can be performed between terms in a field by using an "at least" or "M out of N" number of term matches required to satisfy a field. When formulating multiple field requests, interfield AND, OR, AND NOT, and OR NOT logic operations are performed using special interfield symbol connectives.

These limited framing parameters put certain constraints on the number of synonym/equivalence (S/E) terms that can be included in a given search query. For example, in the "single field request mode" the framer can require that any number of terms, from among the total number of terms that can be typed into the 80 character space, be in a document to satisfy retrieval. He cannot, however, require that term "A" AND any one of a following list of S/E terms be present. Using the "multiple field request mode" the number of S/E terms can sometimes be stretched to three within the 20 character field limitation, but more often this limitation permits only two such terms. For example, the framer can require that terms "A" AND "B" AND either "C" or "D" be present to satisfy retrieval. Terms "C" and "D" are S/E's and the field character limitation might possibly permit one additional S/E.

LITE, on the other hand, is not so limited. While the logic operations (AND, OR, NOT) are basically the same, there is no 80 character limitation per search query. A single search query can be a complex of several pages of terms. Without the restricted query character limitation, the single query can contain as many S/E terms as desired. In fact, LITE batches up to 28 individual queries in a group. What all this means is that to duplicate a single LITE query may take several queries on the RSM.

The query character limitation on the RSM is somewhat mitigated by the use of a "universal multi-character match" symbol, "?", which can be used as a prefix, suffix, or within a word or term to match an indefinite number of character positions. This operator is especially useful as a plural absorbing suffix or as an abbreviation device when only the first few letters of a word are required.

It is our understanding that LITE, as of the first of this year, has a comparable "universal character match" capability. When framing the queries used for the comparison test, however, each key word, including the S/E of a "root" word had to be spelled out.

On the RSM one can restrict retrieval to documents which contain the searched upon key words in a contiguous order such as a phrase, or to those documents which contain the key words anywhere within the document, not necessarily contiguous. There is currently no capability to restrict retrieval to documents containing two or more key words in the same sentence but not necessarily contiguous. For example, in a particular U. S. Code section the Central Intelligence Agency Act might be referred to solely as 50 U.S.C. 403, or as 403, et seq. of title 50, or other variations of basically the same content. On the RSM one must frame all the possible ways in which this can be stated. On the other hand, if one had flexibility within a sentence structure, the query would simply be to retrieve all sections wherein 50 and 403 appear in the same sentence. The particular order within the sentence would make no difference. Clearly, if one has to specify all possible alternatives in such a situation, then the 80 character query limitation becomes somewhat restrictive.

LITE not only has this "within a sentence" restriction capability (hereinafter referred to as "sentence constraint"), but an even more restrictive "within so many words" (hereinafter referred to as "word constraint") capability. For example, the query could be to restrict retrieval to those documents, wherein key word "A" is anywhere within five words on either side of key word "B." In addition, LITE has a "within so many sentences" constraint (hereinafter referred to as "paragraph constraint").

COMPARISON TEST RESULTS

The general parameters of the comparison test and the basic differences in the mechanics of the two systems have been outlined, setting the stage for the results detailed in the following paragraphs. All that remains is to define two terms that will repeatedly appear in the presentation. Each of the twenty-four research queries used in the test base is classified either as a "simple" or "complex" query.

A "simple" query is one where knowing the location of a given word or phrase is the only objective. To retrieve every section of the U. S. Code containing the term "self incrimination" is an example of a "simple" query. To retrieve all references to the "Classification Act of 1949" or to "reimbursement of funds or appropriations" are still other examples.

A "complex" query is one where the objective is to retrieve those materials which will assist in resolving a particular problem. For example, retrieve all U. S. Code sections or Comptroller General Decisions which speak to the problem of "liability for the negligent act of a service employed policeman directing traffic on a road subject to concurrent State and Federal jurisdiction."

It so happened that one-half the test queries (12) constituted "simple" queries - 8 against the U. S. Code (U.S.C.) and 4 against the Comptroller General Decisions (C.G.). In 9 of these 12 the LITE query framing could be duplicated on the RSM - 6 against U.S.C. and 3 against C.G. In these 9 searches the same documents, no more nor less, were retrieved by both RSM and LITE.

The LITE framing of the remaining 3 simple queries (2 against U.S.C. and 1 against C.G.) could not be duplicated on the RSM. LITE's framing used "sentence constraints" in the two U.S.C. queries and a "word constraint" in the C.G. query. Utilizing the same key words, but with the broader "document constraint," RSM output nevertheless duplicated that of LITE in two of the queries. In the other, RSM pulled 513 U.S.C. sections as compared with 431 pulled by LITE. Those sections pulled by LITE were included in the greater number pulled by RSM. When the key words in this particular query were framed to be contiguous, the RSM pulled only 19 sections. This query required a sentence constraint, and the best RSM could do was require a document constraint causing the user to review those 513 sections retrieved to find the 431 that met the objective.

Given the query character limitation on the RSM, it should be noted with regard to these 12 simple queries that on an average it required 1.25 searches on the RSM for each search framed by LITE.

Turning now to the twelve "complex" queries (4 against U.S.C. and 8 against C.G.), not a single LITE query frame could be duplicated on the RSM. Each LITE frame used one or more of the various word, sentence or paragraph constraints. In each case RSM, using the broader document constraint, naturally retrieved more documents than LITE, always including those retrieved by the latter. The LITE/RSM document retrieval comparison rate looked as follows - 13/20, 9/12, 87/134, and 36/62 as against U.S.C. and 2/11, 6/16, 8/18, 3/9, 1/7, 2/19, 8/10 and 1/31 as against C.G.

In 6 searches (3 against U.S.C. with LITE/RSM retrievals of 13/20, 9/12, and 36/62; and 3 against C.G. with retrieval comparisons of 3/9, 1/7 and 2/19), while there were relevant retrievals among the jointly pulled documents, the additional documents pulled only by RSM were irrelevant in responding to the problem presented. Reframing and browsing failed to produce additional relevant documents.

In 3 other searches (all against C.G. with LITE/RSM retrieval comparisons of 2/11, 6/16, and 8/18), all documents retrieved by both systems were irrelevant. In 2 of these searches, reframing and browsing on the RSM failed to find relevant documents - a fact which is relevant in itself. There probably are no relevant materials on the subject in the data base. As to the third search, reframing and browsing narrowed retrieval to 3 documents, all of which were relevant. This particular search is discussed in more detail further on.

In the remaining 3 searches (1 against U.S.C. with a retrieval comparison of 87/134 and 2 against C.G. with retrieval comparisons of 8/10 and 1/31), the additional documents retrieved by RSM included those that were relevant. In all 3 cases, LITE's use of a word constraint was too restrictive. If LITE had used the less restrictive sentence constraint, all the relevant documents missed would have been retrieved.

With regard to these 12 complex queries, it required an average of 2.66 searches on the RSM to duplicate as nearly as possible each search framed by LITE.

The foregoing discussion has described those results which followed from duplicating as nearly as possible the query as framed by LITE. Let us now look to some results of specific searches wherein man and machine interplay.

In one of the 6 simple queries where RSM could duplicate the LITE frame, the objective was to retrieve all sections of the U.S.C. containing the term "research grant(s)." Both systems retrieved the same 5 sections. The query was then reframed on the RSM ("gran? ? Resea?") and 15 additional sections were retrieved containing either the term "grant(s) for research" or "grantsinaid for research."

Another simple query sought the retrieval of Comptroller General Decisions referring to "equitable adjustment(s) arising from dispute(s) clause(s) in contract(s)." The LITE frame (underlined phrases and term) was duplicated on the RSM, and both systems retrieved the same 3 decisions. The query was then framed to broaden the retrieval by eliminating the contiguous word constraint between "equitable" and "adjustment(s)" and also "dispute(s)" and "clause(s)." In effect, the only constraint was a document constraint, and the RSM retrieved only 3 decisions in addition to those pulled with the more restrictive query. Two of these newly retrieved decisions satisfied the query objective.

In one complex query the objective was to retrieve all Comptroller General Decisions relevant to solving the following problem: "Are state license plates required for government-owned vehicles operated by contractors on public roads, outside the confines of contractor operated installations?" LITE's frame included sentence constraints and retrieved 2 decisions. RSM with its broader document constraint retrieved 11 decisions, including the 2 pulled by LITE. All these decisions were irrelevant. In broadening the RSM query by eliminating one key word, 14 decisions were retrieved including 7 previously pulled. Three of the 7 newly retrieved decisions were relevant - 2 of these 3 were considered highly relevant.

Finally, in still another complex query, the results evidenced the classic problem in framing queries - the difficulty in second guessing the terms used by the particular authors of the materials to be retrieved. The objective was to retrieve from title 26 of the U. S. Code any sections relevant to the question of whether "the statute of limitations is waived for filing tax refund claims for prior years in connection with death in a combat zone."

LITE, using a sentence constraint, retrieved 13 sections and RSM, with the broader document constraint, retrieved 21 sections, including the 13 pulled by LITE. Only a couple of these sections had even marginal relevancy. Reframing and browsing on the RSM retrieved nothing more relevant. An OGC staff attorney, who continually works with title 26, was given the above quoted statement of the problem and requested to see what he could come up with by manual research. After 30 minutes of research he found a section more on point than anything retrieved by machine. An examination of this manually retrieved section disclosed that had one term used in the framed query been deleted, the section would have been pulled by the RSM. Armed with this fact, the query was appropriately reframed and 9 sections were retrieved, only one of which had previously been pulled. Included among the 8 newly retrieved sections was the section found by manual research and more significantly, a section even more on point and probably dispositive of the problem.

It is worthy to note, that the term deleted from the query frame was "waive" and its synonym/equivalence, "toll." Given the nature of the problem, this term "waive or toll" would rationally be considered most germane and probably would have been among the last terms to be deleted from the query.

CONCLUSIONS

In response to a particular query, did the results produced by RSM equal or exceed those produced by LITE? LITE results would be exceeded only if additional pertinent or relevant documents were retrieved by RSM.

In 17 of the 24 test queries (11 simple and 6 complex), the RSM results either equaled or exceeded those of LITE. In the 7 searches where those results were exceeded, 4 were the direct result of the browsing capability allowed by the interplay between user and machine. The other 3 resulted because LITE's framing of the query was too restrictive, excluding relevant documents.

In the remaining 7 test queries (1 simple and 6 complex), while the RSM retrieved documents included those pulled by LITE, the additional documents were of no value and the user was required to sort through a greater number of documents to separate the relevant from irrelevant. The additional documents ran from a low of 3 in one search to a high of 82 in another.

Just as there is a danger in using too freely the "universal character match" resulting in a large irrelevant retrieval, there is also the danger of being too restrictive by the use of word and sentence constraints. It seems quite evident, however, that a "sentence constraint" would greatly enhance the search and retrieval capability of the RSM. The need for the more restrictive "word constraint" is not so evident. With a sentence constraint, the RSM would have equaled or exceeded LITE results in each of the 24 test queries. As can be seen, however, from the test results, the RSM browsing capability mitigates to a considerable extent this current lack of a sentence constraint, and in fact, with regard to many queries, will more than compensate.

At this point, it should be reiterated that the purpose of the comparison test was to assess the RSM capability relative to OGC research needs. What are these needs? Suffice it to say, that the

U. S. Code and Comptroller General Decisions are among the legal research materials used most often by OGC. Even so, a "rough office survey" suggests that the need to utilize the RSM search and retrieval system probably would not exceed, in any event, an average of six searches a month. Given such a limited need, it should be readily understandable why the quantitative factors of speed and volume of searches are not as germane to OGC as the qualitative factors of accuracy and thoroughness provided by a browsing capability. Taking all the above into consideration, it seems safe to conclude that the RSM search and retrieval capability is comparable to that of LITE and more adequately meets the current and foreseeable needs of OGC.

POSTSCRIPT

On 8 and 9 April 1970 Mr. Jack Sieburg, Chief, EDP Branch, LITE, participated in discussions and a demonstration of the RSM capability at Agency Headquarters. As previously indicated, the purpose of his visit was to assist OGC in evaluating the RSM capability.

During the two days, Mr. Sieburg was familiarized with the RSM operation and query techniques and reviewed some preliminary results of the comparison test. He took an active part, based upon the previous day's RSM familiarization, in the presentation of a three hour general demonstration attended by Messrs. Houston, Warner and [redacted] of OGC and Mr. Maurice H. Lanman, Assistant General Counsel (Fiscal Matters), Department of Defense.

In a post-demonstration meeting in Mr. Houston's office, Mr. Sieburg offered his opinion that the RSM search and retrieval capability was comparable to that which LITE could provide. He made special note of the RSM browsing capability and its apparent value. There was agreement among all present that OGC, by periodically receiving from LITE updated tapes on the two data bases, had an in-house search and retrieval capability comparable to and in some respects surpassing that which LITE could provide.

Before concluding this report, some general observations should be noted. These comments treat with the quality of output and will in most part be equally applicable to all full text machine systems.

The commentary in this field suggests that there will be a marked difference in efficiency between searches against statutory data bases and searches against decisional data bases. The rationale is that the language of the former is more precise and, therefore, more accurately predictable by the query framer. Let it be said simply that the results from the limited comparison test offered no evidence to support this oft stated conclusion. Needless to say, it seems this system of retrieval works best against "simple" queries, whether they are directed against statutory or decisional data bases. In this regard, it might be said that most queries of the simple category will be directed against statutory bases and thus give credence to the commentary.

All full text systems retrieve on the basis of "word matching." In other words, what comes out as a final product is only as good as what goes in by way of query. The onus is placed upon the human element, not the machine. The system is only as good as the ability of the individual to frame queries. If the query is framed too broadly, then one is confronted with the same problem he started with, that is, he has retrieved too much material, requiring further sorting to separate the relevant from the irrelevant. If framed too narrowly, then relevant materials are excluded. To achieve the ideal a majority of the time takes considerable experience in framing queries against any particular data base. Assumptions as to what terms a particular data base will or will not contain usually are proved incorrect.

This suggests the researcher is going to have to be "attached" to the machine during the searching process. The ability of the researcher to remain in his office and through a remote terminal call forth, search and retrieve from, a particular data base, suggests the ideal system. There would be no need to reserve and manually handle the data base. The technology is here--when time sharing/real time equipment is economical, the ideal will be reality.