A Forum on Centralization and Documentation

In March of 1965 the editors received the letter entitled "Centralization of Document Searching Facilities" which follows. A violent controversy, described by one commentator as a barroom brawl, has been raging within the Federal government on the subject of that letter. Printing the letter without hearing all sides would therefore have been clearly unfair. Opening the pages of a technical journal to the echoes of a barroom brawl will be criticized by some. Nevertheless, in the era of "Big Science," the technical and the political unavoidably overlap, and pretending that the overlap does not exist is fit only for an ostrich.

The focus of the controversy is a report sponsored by the National Science Foundation entitled "Centralization and Documentation," available through the Clearinghouse for Federal Scientific and Technical Information, United States Department of Commerce.

Several knowledgeable participants or observers of the technical and political facets of the controversy were invited to comment on the letter in 1500 words or less and the authors of the letter were allowed a closing comment of 500 words or less. The responses of the Honorable Roman C. Pucinski, Congressman from Illinois, Dr. Allen Kent, Director, Knowledge Availability Systems Center, University of Pittsburgh, Dr. Mortimer Taube, Chairman of the Board, Documentation, Inc., Dr. Harold Wooster, Director of Information Sciences, Air Force Office of Scientific Research and Dr. Gerard Salton, Harvard University, follow.

—Anthony G. Oettinger, *Chairman,*
ACM Committee on United States
Government Relations

"Centralization of Document Searching Facilities"
A Letter by M. L. Ernst, V. E. Giuliano, and P. E. Jones
Arthur D. Little, Inc., Cambridge, Massachusetts

The course of future federal action for coping with the "information explosion" is the subject of continuing debate among concerned individuals in government, the scientific community and various commercial interests. Numerous alternatives ranging from laissez-faire to massive federal intervention have been proposed and considered. A frequently heard proposal has been the implementation of large regional (or discipline-wide) federally-sponsored information centers in which some form of mechanized literature searching system is to be used. Unfortunately, there has been very little available evidence to indicate the probable effectiveness or even to support the feasibility of this alternative.

About eighteen months ago Arthur D. Little, Inc., completed a study of the feasibility of implementing very large centralized facilities for the exhaustive searching of large collections of documents. The study was done for the National Science Foundation, and the results were summarized in a report entitled "Centralization and Documentation." The report is one of very few existing documents which make clearcut statements about the literature centralization issue, at a time when this issue is of great public concern. Although the study was quite limited in scope and the conclusions drawn were clearly limited by the scarcity of the data available at the time, the report met a quite unexpected demand and has been widely circulated, discussed and reviewed—over 5,000 copies have been made in three printings and two editions. The report is currently available through the Clearinghouse for Federal Scientific and Technical Information, U.S. Department of Commerce.

Recently this report has tended to become the focus of a part of the controversy over appropriate federal action. In the vacuum of available evidence, there has emerged a growing tendency to regard the scope of the conclusions and recommendations of the report as being far broader than they really are. In the volatile context of a heated debate, such extensions have already led to misunderstandings and might well, if unattended, lead to errors in judgment.

The purpose of this letter—by the writers of that report—is to recapitulate and discuss our findings as they were reported, in an attempt to dispel a few of the misunderstandings that have been artificially created by wishful or fearful interpretations of our conclusions and recommendations.

The recommendations resulting from the study were:

1. Do not support large-scale centralization of mechanized-document retrieval facilities at this time. A large centralized facility drawing upon the current state of the art of document retrieval techniques could probably not achieve the main objective for which it was designed—provision of an effective, exhaustive, document retrieval capability to supplement efforts to prevent duplicate research or development investments. Responsibility for showing that a proposed centralized facility would be feasible and would satisfy this objective must be borne by the proponents of centralization, employing quantitative evaluation techniques such as those we have developed.

2. Support the undertaking of a comprehensive program to determine the real informational needs of scientists and engineers. Such a survey is a prerequisite to the possible support of centralized document searching facilities in the future, to insure that such facilities will serve real functions, and that they will in fact be used. To be meaningful, the survey must be conducted with considerable imagination and insight.

3. Before undertaking extensive efforts to develop aids such as elaborate word thesauri for existing, partially centralized, mechanized information retrieval systems, investigate and develop further the use of statistical techniques—both for the automatic generation of thesaurus lists and for the automation of some of the functions currently performed by human intermediaries.

4. To support such a program, test-operate one of the existing medium-sized coordinate retrieval systems on a statistical associative basis. The current state of the art of these associative techniques will permit such an undertaking, and a great deal could be learned from it; moreover, the users of the system might realize substantial benefits.

5. For activities which are not concerned with exhaustive literature search operations, support centralization on an individual project basis, after cost/effectiveness analyses have demonstrated—quantitatively—that adequate service levels and over-all benefits will accrue.

To place these findings in perspective, it is important to recognize that well over two years have elapsed since most of the work was done. We have not been actively concerned with the topic of centralization since the time of the report. Although we are un-

aware of any subsequent study which has brought significant new facts to light, the perimeters of the problem have clearly changed since the work was reported.

We therefore find it necessary to stress that the study and report were limited in scope to the analysis of a very particular type of centralization based on use of very specific techniques for very specific purposes. Namely, the report is addressed to the feasibility of centralized *facilities* for *relatively exhaustive search* according to *subject content* of *very large collections* of *term-indexed* documents. The report is primarily focused on the feasibility of exhaustive searching of a very large collection, perhaps several million documents, although conclusions relating to exhaustive searching are valid for smaller collections of 200,000 or more documents. Moreover, the study assumed that a main purpose of such search would be to identify exhaustively all or nearly all the information on a given complex topic, to help avoid potential duplication of scientific or engineering effort. This reflects the attitude toward the "information explosion" and centralization issues prevalent two years ago; there are many other possible modes of centralizing document services, but these are not treated in our report.

The emphasis on exhaustive searching arose as a consequence of a commonly-heard argument favoring federal involvement in centralization of document facilities: if the literature could be searched effectively before research was begun, duplication of research and engineering effort could be avoided, and valuable economies achieved. Although this argument is itself subject to challenge on a number of different grounds, it still enjoys a certain degree of currency. According to this argument, however, it follows that to prevent the duplication of previous work, a search through the prior art would have to be thorough (i.e., relatively exhaustive) to provide confidence that the matters to be studied had not been treated before. Hence the argument favoring centralization implied that the mechanized search component should be capable of good performance (without excessive labor) on relatively exhaustive retrospective searches for similar prior work. Thus we devoted our attention to the prospects of attaining such exhaustive search capabilities.

It is argued in the report that if a centralized facility of this type were to be implemented immediately, it would have to use some form of mechanized "coordinate" searching system for identifying document references. This continues to be true today. Our results are based on the analysis of a mathematical model of coordinate searching systems; this model was developed in order to enable evaluation of such performance aspects as: the expected number of documents to be retrieved as a consequence of different types of searches, the expected recall and precision ratios (which are measures of system performance), and the expected search effort required. The modeling procedure had provisions to reflect the effects of human intermediaries in the search process, to the extent that these effects were known based on the data available. We also studied three classes of techniques which can be brought to bear for improvement of the document retrieval process: word thesauri, citation indexes, and statistical methods of index term association.

The data we had to draw upon was quite limited, and the following statements were included in the discussion of our results:

> As the data available for this application are admittedly sparse, we must be cautious in drawing conclusions from our results. . . . Since the deficiencies present in our evaluation techniques could largely be remedied through the collection and application of routine operating data, it seems reasonable to accept tentatively the results of our evaluation, placing the burden of proof of feasibility on those who would seek radical expansion in documentation center scope and capacity. The methodology for evaluation is now available, and its application should be considered a fundamental requirement before the investment necessary to develop a large centralized system is made.

Much of the current controversy over the report apparently stems from fear that the first recommendation tends directly to deny the usefulness of some of the large existing mechanized literature searching services such, for example, as are currently operated by the National Aeronautics and Space Administration, the Atomic Energy Commission, or the National Library of Medicine. The document collection sizes of several of the agencies have grown considerably since the report was published, approaching or exceeding 200,000 documents in size. They have not been entirely free from the difficulties predicted by the model, nor have the difficulties become critical, to our knowledge.

We want to make it clear that no inference as to the overall value of these existing government systems can or should be drawn from our study alone since:

(a) These existing collections have come into existence in machine searchable form largely as a by-product of announcement journal publishing activities (such as NASA's "Star" or the "Index Medicus" of the National Library of Medicine). Evaluation of the usefulness of announcement publications was not within the limited scope of our study.

(b) It is not at all clear that a main use of the existing retrieval systems is for exhaustive searching. To the extent that exhaustive literature search operations are not the principal function, recommendation 5 rather than recommendation 1 is most applicable to these existing systems.

However, the report does express skepticism about the usefulness, effectiveness, and service levels of the mechanized *search* component of existing systems. We continue to wonder whether the systems are really used extensively for *searches*. If not, it is enormously important to ascertain the reasons. These questions remain open, for the "adequate cost/effectiveness analyses based on the operating data of the existing systems" recommended in item 5 have not been announced. Were they to be conducted, it is entirely possible that actual system operating data could be used to justify the feasibility or even the desirability of expanding existing systems for nonexhaustive searching purposes. Thus our position that the burden of proof of feasibility should rest with those who seek major expansion in any present system is coupled with an opportunity for them to prove their case; we urge the proponents of such action to gather the much needed evidence. The fact that the report articulates detailed criticism, however, does not provide grounds for treating it as a pronouncement of uselessness upon existing mechanized documentation systems; it is more accurate to treat the report as an indicator that quantitative evaluation of the performance of mechanized coordinate searching systems is both possible and highly desirable before they are radically expanded.

Another area of partially artificial (and partially real) controversy over the report relates to recommendations 3 and 4, which treat associative retrieval methods as possible next steps in developing machine searching capabilities.

Most existing collections are today searched with questions which logically combine terms from a vocabulary by means of "and" or "or" operations, and retrieve documents which precisely satisfy a logical formula consisting of such combinations. Associative searching is based on the exploitation of index term usage statistics to derive numerical measures of association among index terms. Potentially, certain highly desirable capabilities are thereby conferred upon the machine searching system which are not available in the existing coordinate searching systems, namely, a potential capability for automatically generalizing a user's request to make it more compatible with the vocabulary of the retrieval system, and a potential capability for automatically matching the user's depth-of-search requirement to system parameters, by ranking the documents presented to users in decreasing order of probable relevance.

While we have recommended bringing the associative techniques to a point where they can be tested, we do not recommend

the associative techniques as a panacea for alleviating all the ills of machine searching based on coordinate indexing. Indeed, such a position would be self-contradictory since associative searching is simply a different method of using what is already basically a coordinate-indexed document collection. It is more to the point to reflect that these techniques are still experimental and remain unevaluated from a use viewpoint. Although work has continued in numerous research groups there have been no dramatic breakthroughs. Accordingly, the statement directly preceding the recommendations in our report continues to hold:

> Although preliminary experimental results appear promising, larger-scale tests must be conducted, and a number of technical problems must be solved before a system incorporating these concepts can be developed which will be of sufficient power to enable high-performance searching of very large centralized collections such as those considered here.

In other words, while the promise of the associative searching methods has in no way diminished, there is no reason today, any more than there was two years ago, to plan on using associative methods, or any other machine technique, as a means for obtaining high quality exhaustive searching of very large collections.

Nonetheless, long range planning does dictate that when the statistical associative techniques become available for pilot operation, as will be the case in at least one major government installation shortly, the associative system should also be subjected to rigorous quantitative evaluation of the type indicated in recommendation 5 of our study. This will at least clarify some of the scientific questions under dispute.

In conclusion, we must emphasize the fact that the foregoing remarks, the report, and the whole issue of machine literature searching represent but one topic in the debate over the nature of federal involvement. The broader context of centralization of documentation resources contains many important topics, less glamorous perhaps than the machine searching methods so far discussed, but each worthy of investigation and study in its own right.

(a) The desirability, feasibility and economics of centralized versus decentralized physical storage of documents, taking into account costs of telecommunication and reproduction techniques. Such centralization might be particularly advantageous in the case of relatively little-used foreign publications, for example.

(b) The desirability, feasibility and economics of centralizing the cataloging and catalog card preparation of the journal and technical report literature.

(c) The desirability, feasibility and economics of centralized machine encoding of texts of journals and documents, where the tapes are to be disseminated to user organizations for decentralized use.

(d) The desirability, feasibility and economics of further expansion of document announcement publications, and questions relating to the number, size, and depth of coverage of such publications.

Much of the controversy about machine searching methods in information centers has had the unfortunate consequence of distracting attention from the full range of issues that bear upon the question of centralization and federal action. Our study of centralization has helped us appreciate the significance of the decisions to be made. We hope this letter will help to restore a technical perspective upon some of the issues involved.

Invited Comments on "Centralization of Document Searching Facilities"

By the Honorable Roman C. Pucinski

The admitted shortcomings of the ADL report constitute the best rebuttal for some of its conclusions. The statement about the recommendation of burden of proof begs the question of Lord Coke, "If I am wrong, what makes you right?"

I believe, however, that the basic issue in the ADL report seems to have been lost in a cloud of argument about coordinate indexing and associative indexing, etc. In the content of what follows, the above argument is akin to an argument about the choice of paint for a new building.

In order to bring the subject in its proper perspective, I should like to state what I believe the real basic issues are.

We have presently hundreds of institutions of specialized information centers supporting various research activities in government, private corporations and universities. Each of these specialized information centers is a treasure house of long cultivated specialized skills for processing technical information. Each of these centers has acquired expertise in the subject matter it covers, and knows the best way to handle some of the specific problems the center may have. The existence of each one of these centers is a precious possession which we must encourage, support and augment, and where necessary add new ones.

However, we must realize that science no longer recognizes the boundaries of scientific discipline.

The language of nature is interdisciplinary and spreads across the invisible borders of all scientific areas of inquiry. Concepts in thermodynamics are related to problems in biology, psychology information theory, economics and ethics! The problem we must address ourselves is: how can we harvest from this cross-fertilization of ideas; how can scientists best benefit from this wealth of information, in the most efficient possible way; how can we best prepare ourselves to take advantage of the rapid pace of advancing technology which can be used to improve information handling; how can the existing specialized information centers be helped to increase their effectiveness?

To paraphrase President Johnson, this is not a Defense Department problem, NASA problem, NIH problem, or Patent Office problem. This is a *National* problem.

From the testimony before my committee and others, there seems to be little doubt that what we really need is a national nexus, a switching network to harness information for our scientists and make it available wherever needed.

We must get ready for the day when each scientist will have available a pocket-size, portable TV screen tied in with the National Information System, which in turn will be tied in with all information sources throughout the world. In a matter of seconds, a scientist will be able to communicate and interrogate the world's storehouse of information and reproduce instantly any article or portion he may need.

What is even more important is the need for official recognition and the awareness that scientific technical information is our greatest national asset and as such must be treated with equal support and status usually accorded steel, oil and other major industries.

There is a crying need for coordinating the far-flung activities in DPIR on a "real-time" basis. This includes standardization, cooperative efforts in basic and applied research, here and with countries abroad, training of information scientists, etc.

Our scientists can become more productive if we remove their lingering doubts about the originality of their work. Indeed with advanced computer technology computers with a self-purging system of duplicative efforts would provide a significant spur for greater initiative in assuring nonduplication of work, with an overall effect for greater diligence similar to the effect computers now have in individual tax reporting.

Let me give you just one example. An alleged novel method for detecting peptides was reported in 1962 in the *Journal of Biological Chemistry*. The same method was reported four years earlier, in 1958, in *Analytical Chemistry*. The author of the later publication acknowledged his needless duplication, which we hope will make him more diligent next time.

Clearly a National Information System could provide the leadership for charting this nation's overall views and provide guidelines for not only improving information handling in the U.S.A., but also to join hands with our allies and friends abroad in a cooperative effort which needless to say would be invaluable in terms of saving time, money and manpower.

These and many other programs can assume reality and meaning only through the establishment of a National Information System built on the foundation of a clearcut national mandate operated under the prestige and wisdom of our scientists and professional societies, and with the gratitude of a grateful nation of people.

By Allen Kent

My remarks on the letter by Messrs. Ernst, Giuliano and Jones are rather negative in nature, since the recommendations seem insufficiently supported. Excerpts to suggest this are given liberally in the following.

The title of the letter was "Centralization of Documentation Searching Facilities." It referred to a report entitled "Centralization and Documentation," submitted by Arthur D. Little, Inc. to the National Science Foundation. Wishing to comment on the original report as well as the letter, I searched my files and found a report, as issued by the Office of Technical Services (PB 181548), with only a corporate author (Arthur D. Little, Inc.) given. The recommendations listed in the subject letter are identical with those given in PB 181548 and the latter was therefore assumed to be the report in question.

Furthermore, in order to place the report and letter in the context of the size of project from which they emerged, I checked *Science Information Notes* and found the following announcements under "Grants and Contracts" from the National Science Foundation:

April–May 1962

Arthur D. Little, Inc., $60,000 contract for study of the degrees of centralization of facilities desirable for the storage and dissemination of scientific documents.

October–November 1962

Arthur D. Little, Inc., $148,000 contract for study of the degree of centralization of facilities desirable for the storage and dissemination of scientific documents.

I assumed that the contract(s) referred to above was the one which resulted in the report in question. It was not clear whether the second grant was a corrected figure or an additional payment, making for a contract total of $208,000. In either case, the amount of the contract was rather substantial, and caused some wonder about the statement in the letter that "the study was quite limited in scope."

As a matter of fact, the report issued by Arthur D. Little, Inc. was not so modest in its stated objectives. In the first paragraph of page one, it is said that:

> This report considers the feasibility of centralizing facilities for the storage and retrieval of scientific documents. Our main objective has been to furnish operational analyses which can provide a basis for formulating government policy on centralization of such facilities.

Perhaps, with this main objective, it should not be surprising to Messrs. Ernst, Giuliano and Jones, as expressed in their letter, that: "Recently this report has tended to become the focus of a part of the controversy over appropriate federal action."

Now let us get to the matter of the main conclusion of the report, as follows:

> . . . development of large centralized searching systems of the type studied is not a desirable course of action to be pursued at this time. Precision of the order of 5% and recall of the order of 50%, as typically obtained from present large systems, would be vastly unsatisfactory in systems ten times as large, even if everything else except the number of documents retrieved were held constant.

This conclusion is reached despite the fact that the report suggests that performance data on large systems is rare, as evidenced by the statement: "Performance and cost data on existing large documentation systems are suprisingly sparse . . ." Furthermore, the lack of data has resulted in a number of assumptions being made in the report; one of which is given as follows:

> Because of almost complete lack of data, we have been forced to make what appears to be reasonable assumptions; the assumption that term usage frequency in searches is nearly proportional to usage frequency in indexing.

This assumption is not, in my opinion, defended adequately in the report. Rather, in leading up to the main conclusion the authors state that they have been primarily influenced by the work of Cleverdon:

> The practical aspects of evaluating retrieval system performance, particularly in terms of recall and precision ratios, have been studied in depth by Cleverdon. We have been primarily influenced by this work and that of Bourne, et al. . .

The work of Cleverdon has been reviewed recently,[1] and criticism has been directed at the inaccurate interpretations and generalizations of the data gathered during this work. Since Arthur D. Little was "primarily influenced" by this work, it is perhaps wise to re-examine the conclusions and recommendations which were so influenced.

Apparently, the authors did not develop their own data to support their conclusions. Although one of their recommendations contains the statement:

> Responsibility for showing that a proposed centralized facility would be feasible and would satisfy this objective provision of an effective, exhaustive, document retrieval capability must be borne by the proponents of centralization, employing quantitative evaluation techniques such as those we have developed.

Nevertheless, the report cautions us as follows:

> As the data available for this application are admittedly sparse, we must be cautious in drawing conclusions from our results. . ..

And the report goes on as follows:

> Since the deficiencies present in our evaluation techniques could largely be remedied through the collection and application of routine operating data, it seems reasonable to accept tentatively the results of our evaluation, placing the burden of proof of feasibility on those who would seek radical expansion in documentation scope and capacity.

Nevertheless, the report seems to seek to close the door on the collection of routine operating data that would be considered valid, since it is stated that a system becomes "centralized" and therefore of concern in the study reported somewhere above the 200,000 document level. Nevertheless, it is reported that collections with automated search systems are limited to sizes of 150,000 documents or less.

Thus, conclusions derived from routine operating data would

[1] SWANSON, D. R. The Evidence Underlying the Cranfield Results. *Library Quarterly, 35*, 1, (1965), 1–20.

have to be extrapolated from the smaller systems to the large—making assumptions with regard to mass effects that would occur —the very matter that is to be validated.

> It is also quite disturbing to find a statement in the letter: Although we are unaware of any subsequent study which has brought significant new facts to light, the perimeters of the problem have clearly changed since the work was reported.

This statement is surprising, especially since I could find nothing in the letter to indicate how the authors perceived changes in the perimeters of the problem.

My comments to the letter and report have been superficial and necessarily brief. However, it must always be so when one is asked to comment briefly on a study which has been financed in a substantial manner. One can only question the conclusions and recommendations that seem to be presented without adequate supporting information.

Perhaps my most serious question then is with regard to the lack of data to support the recommendations. It behooves the report writers not to fall into the same trap that they wish others to avoid—and this it seems to me they have not done.

By Mortimer Taube*

Swanson's paper, "Evidence Underlying the Cranfield Results" [*Library Quarterly* (Jan. 1965)], and my paper, "A Note on the Pseudo-Mathematics of Relevance" [*Amer. Doc.* (Apr. 1965)], make it quite clear that the relevance-recall mathematics in the Cleverdon Studies and the Little Report, is not valid and cannot supply a rational basis for any conclusion except in the material sense that a false proposition implies any proposition.

The authors of the report entitled "Centralization and Documentation:"

> find it necessary to stress that the study and report were limited in scope to the analysis of a very particular type of centralization based on use of very specific techniques for very specific purposes. . . . Moreover, the study assumed that the main purpose of . . . search would be to identify exhaustively all or nearly all the information on a given complex topic, to help avoid potential duplication of scientific or engineering effort.

Since no serious information people have proposed the creation of a centralized system for such a purpose, either exclusively or primarily, the recommendation that such centralization not be supported is supererogatory. The very concept of exhaustive searching of large collections to avoid duplication of research is confused. The fact is that the requirement for exhaustive search to avoid duplication of research varies inversely with the size of a collection. It is a well known fact of library practice that smaller libraries with modest resources must catalog and search more exhaustively than large libraries; and specialized information centers having relatively small collections must index and search more exhaustively than large information services. On the matter of the avoidance of duplication, if no "relevant" information is found, then one may require assurances that the search has been exhaustive; but if an "ordinary" search discloses duplication (as it most likely would in large collections), exhaustive searches are unnecessary, and for this purpose constitute a false requirement. As opposed to the discovery of duplication, a "state-of-the-art" search should be exhaustive; but state-of-the-art searches involve much more than the formal interrogation of a mechanized store, however large or well indexed.

The authors admit, however disguised this admission is, that it is contradictory to suppose that coordinate indexing with association will increase the "relevance-recall" rating of a coordinate indexing system without association. Nevertheless, they wish to continue to spend government funds for a research activity to achieve this contradictory purpose. The government would be as well served if it supported research to square the circle or build perpetual motion machines.

There is the question of burden of proof. Why does it seem reasonable, as the authors say it is, to accept bad mathematics based on "sparse data" as the basis for subsequent evaluation? There exists a total literature on cost and systems evaluation that the authors have disregarded. The burden of proof is on them to show that the kind of models they have built and tested have any relevance to the design and improvement of real systems. Without such proof, the authors' models of information systems can be taken as equivalent to the "models of the brain" which disgrace the literature in the computing field and are intended to establish results already accepted before the model is contrived.

There is a background to both the original ADL Report and the communication to the ACM which should be made explicit. After a careful analysis of the Report some months ago, I proposed to both the National Science Foundation and the authors that the Report be withdrawn to avoid public controversy that might be more political than scientific. Even the National Science Foundation, which sponsored the original Report, classed it among "hasty and preliminary efforts"; but the Foundation thought its publication justified on the grounds that it would lead to healthy controversy concerning a key issue. Further, one of the authors justified its conclusions based on "sparse data" on the grounds that it was necessary to stop the Stafford Warren scheme. We may accept the hard necessity that science is now a political weapon, but this necessity should not excuse interlarding science itself with political argument.

By Harold Wooster

Back in 1954 the Powers that Be decided that a US Scientific Satellite would be a Good Thing. Redstone Arsenal, which had certain hulking pyrotechnic devices with a proven capability for putting the then Secretary of Defense into orbit, was ruled out on the grounds that these might be useful someday. The job was given to the Navy if they would promise faithfully not to use any military hardware in building Vanguard.

Then came 4 October 1957. Sputnik flew. It seems obvious now that if we had quit searching von Braun for matches, and scattered a few subscriptions to "Astounding Science Fiction" in the right places, we could have been first into space.

As always in times of crisis, patriotic citizens volunteered to build a Greater Galactic Kluge. The first audible indication of the information, or kluge, explosion came when Allen Kent and Merritt Kastens collided mid-stage in February of 1958. Since then there have been a Taube plan (the job is obviously too big for the Federal Government; let Documentation, Inc. handle it), a National Federation of Science Abstracting and Indexing Services plan, a Kelsey (F. Ellis) plan, a Kelly (J. Hilary) plan, a Pucinski plan (let NSF do it, only in Chicago) and, last but not least and probably not last either, a Warren plan.

There has been little opportunity to point out that the several Galactic Kluges share in common a certain glorious ignorance of economics, of user requirements, of what technology can and cannot do, of the complicated interplay of biogeopolitics and, for that matter, of the relative unimportance of subject indexing and information retrieval to the working scientist.

The rules of formal debate do not apply in barroom fights, where one lashes out with the nearest handy object. The report under discussion has proved to be at least as useful as a broken beer bottle for this purpose. Not to read, Heaven forbid, but to roll up and hit people over the head with. The battle cry "Do you know that the Science Foundation sponsored a study by A. D. Little which proved that centralized information services won't

*Deceased September, 1965.

work?" has been proved in combat to be unanswerable by people who haven't read the report either.

I regret now that I have had to read the report in detail. My scholar's conscience compels me to point out that it, or at least the edition of July, 1963, C-64469, is a bibliographic bastard. It has a corporate mother, a Federal (or Fairy?) god-mother, but no acknowledged personal father(s). The three signers of the Manifesto under discussion have presumably had some connection with the report—one of them does tie with Myron Kessler for the number of citations in the "References"—but I can not prove from the text of either the report or the Manifesto that they are its natural father(s). Is there now a public confession of paternity?

Unfortunately for any honest use, the arguments of the report rest upon a most unsound premise indeed, one iterated in the Manifesto, that: "If a centralized facility of this type were to be implemented immediately (Right now, or X years from the date the first piece of paper is signed?) it *would have to use* (my italics) some form of mechanized 'coordinate' searching system for identifying document references. This continues to be true today."

This statement is not true today, and was not true then. It is a flagrant example of myopic, xenophobic, Cantabrigian, computer-intoxicated commercial and intellectual parochialism at its chauvinistic worst. There may well be instances when it is desirable, for reasons of salesmanship, gamesmanship, prestige, novelty, access time, a shortage of skilled inexpensive labor, or the simple brutal fact that it is often easier to buy a computer, or hire a contractor, than it is to get manpower spaces, to go down the mechanized coordinate searching route. But you don't *have* to.

In the real world, outside of the swirling miasmas of the Charles River, there are document systems which handle several times 200,000 documents very nicely, thank you, without computers or coordinates. Terzi at IDAMI, the Italian Automatic Documentation Institute in Milan, for instance, uses Universal Decimal Classifications as multiple subject headings; Ember at the University of Montreal does it with the elegant Symbolic Shorthand Notation; A. D. Little's competitors at Batelle have their own home-brewed method of extracts and multiple prefilings which works—to date without any visible limitation on subject and size. There are even unverified rumors floating around Washington that on occasion successful subject searches have been conducted, *sans* benefit of computer, at the highly unmechanized US Patent Office and the Library of Congress.

One of the best lines of reasoning in the report has been omitted from the Manifesto, probably for the same reasons that have made me quit using it—it turns in your hand and bites you. This is the simple effect of size. The argument runs thusly. Let us assume that indexing efficiency does not fall off with the size of the collection (although, of course, it will fall off—Du Pont does a better job of indexing chemical literature than does the Library of Congress). "Precision of the order of 5% and recall of the order of 50%, as typically obtained from present large systems, would be vastly unsatisfactory in systems 10 times as large, even if everything else except the number of documents retrieved were held constant. . . . If 1,000 documents were retrieved as a result of an exhaustive search of a large collection, readers would simply not be willing to wade through 1,000 documents to find the 50 that were relevant."

This is probably still one of the best arguments against the Galactic Kluge and the central search computer. Librarians laugh at it though—the flaw is too obvious. You don't make one big pile of books. You put your books on Medicine into a library of medicine, on Agriculture into a library of agriculture. This is called classification. If a book on agricultural medicine creates problems, buy two copies. You end up with lots of little piles rather than one big pile, and when you have to make a search you look through some of the little piles rather than all of the big pile.

Scaling factors are touched upon in another place in the report, neglected in the Manifesto. "Existing computers, operating serially, do not appear to be capable of handling the problem economically for collections with 9000 or more terms and over 200,000 documents, even if the simplest associative techniques are employed." There are subtle hints that for a small additional fee the anonymous authors would be glad to develop an analog network that might be able to handle large collections, but that right now 100,000 documents is probably the practical upper limit.

I have no quarrel, mind you, with the recommendations of the study. My only concern is with the way the Putative Progenitors got there. I detect subtle whiffs of heated emery and axes being ground in their numbers 3 and 4, but why not? It just might work. Far stranger things have happened in the last 8 years.

The suggestions for future studies are far less glamorous than Kluge building, but these are things that the Kluge builders should know before they start mooring their aerial castles. There is solid honest work being done to cope with the paper panic; there are solid honest things yet to be done; experimental systems should and can be tested in the laboratory and the pilot plant (and the accounting office) before being committed to full-scale production. Scientists do, somehow, manage to get most of the information they need in time for it do them some good even today. Computers are getting cheaper and people dearer year by year.

There is at least one fundamental paradox in human activity. The little things you and I do day in day out always take longer than we plan; the big things catch us by surprise. Are there any bets that we won't see some sort of National Library of Science System in the next 8 years? Our problem, as responsible professionals, is to make sure that when we do get one it will work, efficiently and economically.

Meanwhile, I seem to be fresh out of broken beer bottles.

By Gerard Salton

These notes are written in an attempt to provide a technical summary of the issues raised in the original report by Arthur D. Little (ADL) on "Centralization and Documentation," and of the amplifications and rejoinders contained in the correspondence which precedes.

The reader who has followed the discussion up to this point may perhaps be astonished to hear the present writer's opinion that the original ADL report was basically an interesting piece of work, which on the whole was not badly executed. In fact, a centralized, coordinate search system is considered in that report; a mathematical model is then constructed which purports to represent such a coordinate indexing system; using the model, evaluation measures are calculated which tend to show that the effectiveness of the system lessens as the collection size increases; the conclusion is finally reached that large-scale, centralized search facilities cannot be expected to render a useful service.

In an area in which too few attempts are made to furnish technical answers to technical problems, the use of an original mathematical model is in itself an interesting development, regardless of the appropriateness and correctness of the model; furthermore, some of the work on the evaluation of retrieval performance contained in the original report appears to this writer to be completely beyond reproach. Had the writers of the report therefore chosen to let their study stand or fall solely on its technical merits, it is not at all clear that the debate surrounding it would have been quite so thick and prolonged, and that the critics would have been quite so severe.

Unfortunately, the writers chose not to leave well-enough alone, first by including in the original report some wording and certain recommendations which could not reasonably be justified on technical grounds, but appeared to everyone to be politically motivated, and then by publishing the letter which is now in front

of us and which, they assert, restores "a technical perspective." Even a cursory reading of some of the telling comments which precede, by Messrs. Kent, Pucinski, Taube, and Wooster, reveals that the preparation of the letter by the ADL writers was an error in judgment. There surely was no need to renew the battle at this point, and to infuriate the critics to the point where they feel obliged to answer as acidly as does Harold Wooster in his "Forum on Centralization," or as cleverly as does Allen Kent.

The technical questions surrounding the issue may be examined under six main headings: the coordinate indexing system; the exhaustive search assumption; the mathematical model; the recall-precision calculations; the associative indexing recommendations; and the "burden of proof" argument. These are now taken up in order.

Consider first the basic restriction of the model to *coordinate indexing systems*. This appears to the present writer to be an eminently sensible decision, since in fact the majority of the mechanized search systems are at present—and may be expected to continue to remain for a while—of that type. Dr. Wooster quarrels with the notion that if a centralized system were to be implemented, it would have to be based on some form of *mechanized* search operation, calling this "a flagrant example of myopic, xenophobic, Cantabrigian, computer-intoxicated, commercial and intellectual parochialism." He cites a few examples of manually operating systems using, for example, the UDC classification, which Dr. Wooster says operate "very nicely." This may be so; nevertheless, the present writer agrees with the notion that mechanized systems should be the main object of study.

Unfortunately, Messrs. Ernst, Giuliano and Jones then go on to restrict the particular coordinate indexing systems to be considered to those primarily used for searches of an *exhaustive* nature, in which the user attempts to retrieve *all* relevant items. This restriction is not properly motivated even by making appeal to the problem of prevention of duplicate research, and it is at the root of the unfortunate model with which the authors work. This model is lacking in three main respects.

(a) The assumption that a *single* search operation could ever be used to retrieve *all* documents relevant to a given search request is obviously unrealistic; a sequence of iterated search steps, possibly involving user feedback, should have been considered instead;

(b) The assumption that any exhaustive, coordinate search system could operate without adequate vocabulary normalization procedures, possibly in the form of dictionaries, scope notes, and the like, is obviously untenable;

(c) The assumptions which lead to the mathematical equations relating vocabulary size with size of the document collection are faulty, and the formulas which suggest that vocabulary size increases indefinitely with document collection size have been disproved by subsequent theoretical studies (see E. Wall, "Further Implications of the Distribution of Index Term Usage," Proceedings of the American Documentation Institute Annual Meeting, Spartan Books, October 1964), and by actual data collected from existing systems (see D. L. Drew, R. K. Summit, R. I. Tanaka, and R. B. Whiteley, "An On-Line Technical Library Reference Retrieval System," Proceedings IFIP Congress-65, Vol. 2, Spartan Books, to appear).

The situation may then be summarized by stating that the model proposed by the writers from ADL could not possibly be expected to perform adequately in practice. When the writers proceed to show by a series of interesting, and perfectly legitimate *recall and precision measurements* that their system in fact loses effectiveness with increased size, they are proving nothing that could not have been predicted in advance. An inadequate model invariably leads to useless results.

Dr. Taube's criticism of the "relevance-recall mathematics" introduced by the Aslib-Cranfield studies and taken over by the ADL writers is misplaced, because it is based on a confusion on his part between two different uses of the term "relevance." The introduction of the evaluation calculations provides in fact one of the main reasons for the legitimate interest in the original ADL report. On the other hand, Dr. Taube is hard to refute when he speaks about the inappropriateness of the model as follows: "since no serious authors have proposed the creation of (such a centralized system) . . . the recommendation that such centralization not be supported is supererogatory."

A word must be said about the recommendation concerning the implementation of a statistical *word association system*. This recommendation is clearly out-of-order, as Dr. Wooster properly points out. ADL's statistical association work has nothing to do with the present issue, and should not have been permitted to intrude on the discussion. Vocabulary normalization procedures of many kinds—including, possibly, associative schemes—should have been considered.

A final word may be reserved for the assertion, contained in the original report, and repeated in the letter, that "it seems reasonable to accept tentatively the results of our evaluation, placing the *burden of proof* of feasibility on those who would seek radical expansion in documentation center scope and capacity." Each one of the commentators, including also Congressman Pucinski, complains about this transparent attempt to befuddle the issue. It is obviously improper to draw conclusions derived from an inadequate model, and then to withdraw, claiming that the issue was now settled.

It would have been smarter, originally, not to draw any conclusions, and to let the report stand as one interesting contribution in the field of systems evaluation. Subsequently, the authors might have remained silent, instead of reviving the issue by producing the "manifesto" which is more objectionable than the somewhat inadequate original.

Response by Ernst, Giuliano and Jones

On reading some of the remarks and comments printed above, we were surprised to find such great emphasis devoted to criticism of our two-year old study and report, and so relatively little attention paid to the substantive issues raised in our letter relating to what steps are needed for future progress. As to the criticisms, it is impossible even to begin to rebut them within the 500 words allowed us here by the editor. Moreover, the strong emotional undercurrents which pervade some of the remarks would tend to make further debate in the present vein not only sterile but embarrassing to all parties concerned. A few of the technical criticisms are well taken and deserve to be acknowledged, but a surprisingly large number of them are either not valid or not relevant—as can be observed through study of the report and its supporting appendices. We feel that the same is even more true of the criticisms of purpose, scope and findings of the study.

The aspect of the present exchange which is most interesting to us is that it has served in no uncertain terms to bring to the surface and into print some of the strong emotionalism connected with issues which affect large-scale centralization of information resources. Based on the present exchange, it appears that Harold Wooster's view of our report being considered by some as a "broken beer bottle" is correct. Perhaps, now that there has been something resembling a minor barroom brawl, we can proceed with the very real tasks of planning, research, design and systematic evaluation necessary for improvement of our nation's resources for communication of scientific information.