APPENDIX B. PROGRESS AND PROSPECTS IN MECHANIZED INDEXING

A working paper prepared for the Symposium on
Mechanized Abstracting and Indexing, Moscow,
28 September - 1 October 1966

Mary Elizabeth Stevens
National Bureau of Standards
Washington, D. C. 20234

The term mechanized indexing can be interpreted in two different ways: as involving the use of machines to produce indexes once the index entries have been pre-determined manually, or as involving the use of machines to select the index entries as well as to prepare the indexes.

The first interpretation, that of machine compilation of indexes is perhaps best represented by the progressively more sophisticated mechanization used for the production of Index Medicus from manual "shingling", through sequential card camera operations, to the computer-based system using a high-speed phototypesetter, the Photon GRACE 1, 2/. As noted elsewhere in this report, machine capabilities have made practical the preparation of citation indexes. In general, however, machine-compiled indexes work with the results of human intellectual efforts as applied in the subject content analysis of documents. We also find machines used to provide aids to the indexer. Two different tools may be employed to improve the quality of indexing. There are prescriptive aids in the sense of limiting and rigorously defining the scope of index terms to be used, and there are suggestive aids in the sense of provoking ideas about additional terms that might be used.

The first type may involve a mechanized authority list or thesaurus used to normalize proposed index term entries, as has been demonstrated by Schultz 3/ and Schultz and Shepherd 4/ from 1960 onward. The potential value of this technique is indicated by further investigations of Schultz et al 5/ in which it was found that index terms proposed by authors agreed more with terms employed by more than one member of a typical user group than did terms available in the document titles. Another example of developments in the use of a mechanized thesaurus is the system at Lockheed Missiles and Space Division, Palo Alto 6/.

This type of tool is used to check proposed indexing terms against the terms of the system vocabulary, to prescribe choices between synonyms and different levels of specificity, and to supply syndectic devices such as "see also" references. Computer manipulations of thesauri can also be used to diversify search questions and to provide useful groupings of terms previously used in the system. The mechanized thesaurus can thus serve as the second type of aid by suggesting to the human indexer additional terms he might use. In effect, such a thesaurus provides a display of prior term-term, document-term and document-document associations observed in a particular collection, such as was demonstrated in the form of special purpose equipment in Taube's "EDIAC" 7/ and the "ACORN" devices at A. D. Little 8/.

The associational thesaurus can also be used to aid in the resolution of ambiguities of natural language and to provide for updating in the light of changing terminologies or changes in the subject scope of a collection. What are the prospects for automatic updating and revision of a mechanized thesaurus? Luhn 9/ has suggested that a record of the number of times words and groups are looked up would be "an indispensable part of the system for making periodic adjustments based on the usage of words or notions as mechanically established."

Another suggestion for the development of mechanized aids in human indexing procedures has been made by Markus 10/. This is to "explore the possibility of applying programmed teaching to indexing, with or without machines."

Machine-compiled indexes rest upon the efficacy of human indexing and there is increasing reason to doubt that this will be "good enough" for the future. It appears that there is a growing consensus with respect to inadequacies of present scope and coverage of indexing services. Cheydleur 11/ emphasizes that: "The cost of manual classification and abstracting of all the articles in the world's hundred-thousand technical periodicals would be fantastic. The practicality of carrying it out in a coordinated and timely way by manual methods is unrealizable. There is also a pressing need to extend the coverage of a myriad of unpublished working papers. Hence, there is an utter necessity for automatic indexing, abstracting, and summarization by electronic data processors."

Secondly, little confidence can be attached to routine, manual operations to produce subject-content selection indicia for subsequent selection and retrieval of stored documentary items for the following reasons:

1. Wide variations of intra- and inter-analyst consistencies occur in the assignment of content-indicia, even with respect to well-established client-interests and index term vocabularies.

2. Potential clients may or may not be inclined to use the system, regardless of whether or not it provides efficient content-indicator-clue and selection criteria mechanisms.

3. Future queries cannot, in general, be effectively predicted in advance, except for the cases of specific author or title retrieval requests.

The problem of intra-indexer and inter-indexer inconsistency is of special interest because the degree of inconsistency will seriously affect search and retrieval effectiveness and because serious questions are raised with respect to the evaluation of any indexing system in terms of prior or independent human indexing.

With respect to the effect of indexer inconsistency upon subsequent search effectiveness, O'Connor 12/ considers the possibilities of overassignment (i.e., the assignment of indexing terms to an item that a subsequent searcher would not consider pertinent to that item) in the case where a search is specified by index terms A, B and C, each term is over-assigned with ratio 1.0, and assignments and overassignments by the recognition rules are statistically independent: "Then only one eighth of the papers selected by the conjunction of A, B and C would correctly have all three terms."

The complementary disadvantage of missing relevant references on search, because of indexer failure to supply all the appropriate indexing terms that a searcher would have considered relevant to a particular document would imply that, for a three-term query, assuming independence of term-assignments and a consistency level of 50 percent, only 12.5 percent of the documents that the searcher would consider relevant would be retrieved if someone else had indexed these items.

We have previously reported 13/ on the results of 700 simulated 3-term searches based upon both manual and machine indexing of approximately 20 items with respect to a fixed vocabulary of less than 100 allowed descriptors. These results show, that if indexer A assigns to a given document the term "A" as indicative of subject content, then his subsequent chances of retrieving that document with a query for term "A" are 58.4 percent if the item had been indexed by someone other than himself, and 55.8 percent if indexed by an automatic indexing procedure developed at NBS, called SADSACT" (Self-Assigned Descriptors from Self And Cited Titles) 14/. For three-term searches, any one searcher would be able to retrieve 26.4 percent of the items he would consider relevant to his query if they had been indexed by any of the other user-indexers, and 24.7 percent if the items had been indexed by the machine technique.

Tinker 15/ provides evidence on the relationships between inter-indexer inconsistency and retrieval efficiency, assuming that a given indexer is a potential querist, with average chances of retrieval ranging from 6.5 to 36 percent. Additional evidence on the generally unsatisfactory state of manual indexing consistency has been reported as follows:

1. Korotkin and Oliver 16/ report that five psychologists and five non-psychologists indexed 30 items with three descriptors per item. The task was repeated two weeks later with the aid of an alphabetized list of "suggested" descriptors derived from the data acquired in the first session. Mean percent consistency results were as follows:

	Session I	Session II
Group A (Psychologists)	39.0%	53.0%
Group B (Non-psychologists)	36.4%	54.0%

2. Evaluations of relevancy of selected items to a given search request have been explored by Badger and Goffman 17/ as follows: "Each of three evaluators was asked to dissect the output into relevant and non-relevant subsets... A chi-square test was applied to the observed evaluation as compared to those expected if the three evaluators were in complete agreement. The chi-square test of 81.57 was very significant, indicating that there was an absence of agreement."

3. Greer 18/ reports on investigations of the interpersonal agreements between subjects asked to list the search words they would use in posing queries in the field of information storage and retrieval systems. He found "a mean percentage consistency agreement of 26.1 among subjects in stating search words."

4. Hammond 19/ provides a sampling of the use by NASA (National Aeronautics and Space Administration) and DDC (Defence Documentation Center) of a common set of indexing terms to index an identical set of 996 technical reports. In considering 3-term searches against the variant indexing shown in Hammond's tables, sample calculations show a 25-30 percent failure to retrieve potentially relevant items.

5. In terms of intra-indexer consistency, Rodgers 20/ reports that: "A consistency of .59 in selecting words to be indexed on two different occasions is not sufficiently high to give us great confidence in expecting a stable store when human indexers are used."

For these reasons, increasing consideration should be given to the second interpretation of the term "mechanized indexing", that is, to machine generation of index entries, or automatic indexing. This typically involves machine processing of some natural language text, with severe problems of input. The first of several solutions involves use of automatic character recognition techniques to convert printed text to machine-usable form. This approach holds considerable future promise, but there are many current limitations and difficulties.

A second possible solution, manual keyboard operations to produce a machine-useful transcription of a text, is plagued by high costs (i.e., at least $0.01 per word for unverified keypunching), and also by limitations of available time or manpower.

A third alternative is suggested by current developments in computerized typesetting or tape-controlled casting or photocomposition machines. However, while such techniques promise major improvements for the automatic indexing of textual information to be published in the future, little can be done for already available literature, even with respect to the bibliographic citation information alone. Today's difficulties are emphasized by estimates of a cost of 30 million dollars to convert the present Library of Congress catalog to machine-readable form 21/.

Assuming, however, that the input processing problems have been solved, we may ask what machines can do with respect to words in texts, or in portions of texts, that are available in machine-useful form? The machines can "read" the words for purposes of shifting and sorting and can copy or reproduce the words in some desired order, as in a machine-prepared concordance. Machines can match input words with words already in store and thus exclude input words from further machine consideration (as by stoplists in KWIC (Keyword-in-Context) and other forms of derivative indexing) or stress certain input words with reference to a selective "inclusion" dictionary.

Next, machines can tabulate and count, so that both absolute and relative word frequency data may be applied to either indexing or search-selection algorithms. Measurements of sequential distances between selected words in the input text may also be applied. Machine look-ups against a master vocabulary can provide automatic supplying of syndectic information, synonym reduction, lexical normalization, generic-specific subsumption, data with respect to previously observed word-word or word-subject co-occurrences. In addition, information can be provided as to the possible syntactic roles of input words.

In the light of such machine capabilities, what can be said of the present state of the art in automatic indexing? Automatic indexing in the sense of machine-prepared indexes that are generated by the automatic extraction and manipulation of keywords, especially from titles, is of course widely used in KWIC indexes such as Chemical Titles and many others both in the United States and elsewhere.

Fischer 22/ provides a retrospective view of KWIC indexing concepts, including variants like KWOC (Keyword out of Context) and WADEX (Words and Authors Index to Applied Mechanics Review). She stresses the potentialities of linking such extraction indexing to selective dissemination systems and concludes: "Plans for using the 'Echo' satellites to link information centers around the world, in a world wide drive toward immediacy in information dispersion, will surely provide a place for KWIC indexes and for the KWIC concept." Warheit 23/ also reports that consideration is being given to combining selective dissemination systems and KWIC. Fundamental questions remain: How useful and how much used are KWIC and other machine-generated indexes based upon the extraction of words from a limited portion of the author's own text?

These questions relate to an important distinction between two quite different types of indexing. The distinction is that whereas "derived" indexing takes as index entries the author's own words in the title, the abstract or the full text, in "assignment" indexing an index term, descriptor, subject heading, or classification code is assigned to a document as an indicator of content and the term assigned does not need to be identical with any of the author's own words.

We can report continuing progress in use of derivative indexing techniques such as KWIC, and also in experiments with automatic assignment indexing and automatic subject classification. Timeliness of index production is certainly one of the major virtues of KWIC. A similar timeliness is promised for automatic assignment indexing techniques provided that requirements can be kept sufficiently low with respect both to keystroking and computer processing.

Intermediate results may be achieved by pre-editing, normalization, and post-editing techniques. Manual pre-editing to modify and supplement keywords in title, abstract, or portions of text has been used in permuted title and KWIC-type indexing from the punched card system that began operation in 1952 24/ to the "notation-of-content" system developed for NASA 25/. Kreithen 26/ suggests a combination of derivative and assignment indexing, as follows:

> "The combination of these two automatic indexing methods, whereby a number of indexing terms would be assigned to a document on the basis of its category dependency, and the rest extracted from text, might be a desirable solution."

Automatic assignment indexing, with clue-words in the input textual material used to determine the proper assignments of indexing terms to incoming items, is generally equivalent to automatic classification techniques that assign a single classification category to items, again on the basis of clue-words in the input text, because a minimum cut-off level in the automatic assignment procedure, combined with a sufficiently generic vocabulary, can achieve classificatory as well as indexing results. The present state of the art in automatic assignment indexing and classification is marked by intriguing demonstrations of technical feasibility for the relatively small samples so far investigated. Present difficulties associated with automatic assignment indexing or classification techniques, however, relate to problems of input processing requirements, computational limitations, the special purpose nature of results demonstrated to date, and problems of evaluation.

A listing of automatic classification and assignment indexing experiments as of 1964 is provided in Table 2, pp. 101-103, of the text of this report. To this we should add more recent results of our own as well as additional results reported by O'Connor 27/ and Williams 28, 29/, Dale and Dale 30, 31/, and others.

In the SADSACT method, we start with a "teaching sample" of items representative of our collection, to which indexing terms have previously been assigned. We then derive the statistics of co-occurrences of substantive words in the titles and abstracts of these items with descriptors assigned to them, ending with a vocabulary of clue words weighted with respect to prior co-occurrences with various descriptors with which they have been associated.

Then, for new items, we look up each word of input (typically consisting of 100 words or less: title and up to 10 cited titles, or title and brief abstract, or title and first or last paragraphs) and derive "descriptor-selection-scores" based upon the prior ad hoc word-descriptor associations. The highest ranking descriptors, in terms of the accumulated selection scores, are then assigned, at some appropriate cut-off level, to the new item.

To date, machine first-choice assignments (corresponding to performance figures reported for other automatic classification and indexing experiments) have been checked for 213 test items either against prior DDC indexing or against user evaluations, or both, with 72.3 percent mean overall agreement.

Our most recent results involved 150 test items. Machine assignments of descriptors to items were checked by having up to five actual users of our collection rate the relevance to a given one of 14 descriptors of items whose titles were listed under that descriptor by the machine assignment procedure. A total of 451 pairings of user-relevance-ratings with the machine has now been analyzed, with a mean relevance rating of 74.9 percent. With respect to machine first-choices, there were 206 pairings with 85.4 percent of the machine assignments rated as at least somewhat relevant.

Checks have also been made of SADSACT results as compared to which of these same documents would be directly retrievable if a KWIC or some other title-only index were to be used. For the first 50 machine assignments rated as "highly relevant" in user-evaluations, a check was made to determine whether or not the same item would be retrievable by lookup under the name of the descriptor in a KWIC index. There were 9 such cases, or 18 percent. In 48 percent of the cases, a part of the descriptor name occurred in the document title. For 17 cases, or 34 percent, there were no title words identical with any part of the descriptor name.

One evaluator was also asked to review the titles of 150 test items and to indicate which, if any, he would wish to retrieve under each of 14 descriptors. He requested in all 353 items and 209 of these were retrieved on the basis of the SADSACT assignments, for a recall ratio of 59.2 percent. Of these, 167 had been previously evaluated by the same user for an overall relevance ratio of 81.4 percent.

Summary accounts of automatic classification and assignment indexing experiments have been provided by Schultz 32/ in the form of an "imaginary panel discussion" (in which, hypothetically, Borko, Schultz, and Stevens discuss their respective systems), and by Black 33/ who concludes: "Provided that overall effectiveness is nearly equal, the system that depends less on the human element would clearly seem to be more desirable from a standpoint of reliability and efficiency, and perhaps even from a standpoint of economics as well."

Additional work has been reported by Dale and Dale 30, 31/, Damerau 34/, Dolby et al 35/, Kreithen 26/, O'Connor 27/, and Williams 28, 29/, among others. Borko's 36, 37/ more recent papers on this subject consider problems of reliability and evaluation. He reports comparisons of automatic and manual classifications of 997 psychological abstracts into 11 categories, factor-analytically derived from 65 percent of these abstracts used as source items. He concluded that it was possible to determine that the percentage of agreement between automatic classification and perfectly reliable human classification could reach 67 percent.

O'Connor's 1965 report 12/ provides further promising results of his "machine-like indexing by people" studies and also discussions of other techniques and of difficulties and limitations in automatic indexing experiments to date. Using Merck, Sharp and Dohme indexing data, O'Connor tested additional recognition-of-clue-word rules based on syntactic emphasis, a first sentence and first paragraph measure, a syntactic-distance measure, negations forbidden near clue words, and words naming substances or types of operations being required in close proximity to clue words.

He reports considerable success with these new rules as follows: "The computer rules selected 92% of 180 toxicity papers. Allowing for sampling error, these rules would select between 88 and 95 percent of the toxicity papers. Thus the computer rules would be roughly comparable to, or perhaps superior to, MSD indexers in identifying toxicity papers."

With respect to the difficulties to be observed in automatic indexing experimentation, O'Connor questions the adequacy of samplings of subject specifications, documents, and collections, the size of clue word vocabularies, and the human judgments used as standards in many of the studies that have been made.

The question of sampling adequacy in terms of the representativeness of clue word vocabularies as related to index terms or classification categories may be particularly critical for methods using small teaching samples. Spiegel and Bennett 38/ report that: "There seems to be no simple relation between the size of the corpus and the size of the vocabulary but after a certain point vocabulary size increases very slowly."

Findings by Williams 29/ are encouraging. Working with teaching samples of 35, 70, and 140 items respectively, he reports that in the first 10,000 word tokens processed from the text of 2,700 abstracts 1,800 different word types were encountered but that in the 80,000 to 90,000 range only 255 new types appeared. He found further that "an increase in sample size beyond 140 would not appear to offer any significant increase in classification performance."

Williams found an average correct classification of 62 percent for 474 test items automatically assigned to one of four solid state categories 28/. In other tests, 2,754 solid state abstracts were classified into three primary and three secondary categories, using a computer program capable of handling up to 50 clue words, 10 subject categories, and any number of documents. Performance effectiveness ranged from 62 to 88 percent correct by comparison with the original classifications at the more generic level and from 67 to 92 percent correct at the more specific level.

Further progress in the application of statistical association, clumping and syntactic analysis techniques have also been reported. Statistical association techniques are concerned with correlations and coefficients of similarity assumed to exist between items or objects sharing common properties. In documentary item applications, document-document similarities are calculated for sharings of the same index terms or for common patterns of citing the same references, of being cited by the same other documents, and the like. Word-association techniques include the development of absolute or relative frequencies of co-occurrence in a given set of documents, such as those representative of a specific subject matter field. Various normalizing procedures can be used to remove effects of tendencies for certain words to occur frequently in general. Spiegel and associates 38/ at Mitre Corporation have explored means of normalization to eliminate effects of length of text strings, relative positions of words in a string, and vocabulary size.

Ernst 39/ reports that at Arthur D. Little: "We are ... seeking to provide a working retrieval system which will incorporate associative features. The objective will be to make use of automatically computed index term associations as a basis for detecting and presenting an appropriate list of near-synonyms for the concepts desired by a user --- essentially the automatic generation of a limited thesaurus in response to individual user requests." In Switzer's model 40/, co-occurrence statistics of index terms consisting of words from title or text, author's names, and words and author names from cited titles, are used. Significant probabilities for such co-occurrences are then derived.

Methods that group objects or items in terms of co-occurrence data for their properties or characteristics are involved in the "clumping" techniques as proposed at the Cambridge Language Research Unit. Further investigations into the development of the basic CLRU approach have been conducted at the Linguistic Research Center at the University of Texas, by Dale and others 30, 41/. In this work, simulation of associative document retrieval by computer gave results for 260 computer abstracts, using the same 90 clue words as previously used by Borko: "The recall ratios in the test requests were high (i.e., very few relevant documents were not retrieved); relevance ratios were characteristically smaller (of the order of 10 percent). However, since the output lists are ordered, it is interesting to note that the relevance ratios are significantly much higher in the upper portions of the output lists (roughly between 25 percent and 50 percent in the upper fourth of the output lists), and that recall ratios are still of the order of 50-70 percent."

In 1964 a report of the Astropower Laboratory 42/ outlined a "semantic space screening model" based on the assumptions that keywords or phrases have quantifiable 'values', that by itemizing the keywords in a document sufficient information is obtained for its classification, and that by adding the values for the keywords in a document the pertinence of that document to a particular subject field can be determined. A training sample consisted of 120 abstracts drawn from six subfields of electrical engineering. Results showed successful classification of source items, using four different classification formulas, as ranging from 49 to 96.3 percent. Results with test items ranged from 32.9 to 69.0 percent accuracy.

The automatic indexing, selective dissemination and retrieval system design developed by Ossorio 43/ is based on a system vocabulary subsequently used for the automatic assignment of new items to appropriate locations in a pre-established "classification space". An "attribute space" may also be developed to identify the kind of information found in a document, e.g., that it deals with concepts such as weight or physical size rather than with mathematical or space and time concepts.

Both types of "space" in this system are constructed through the use of factor analysis applied to previously established relationships between the terms in the system vocabulary (approximately 1,450 terms) and 49 subject fields and to relevance ratings of attributes with respect to items. Then, "documents are indexed by being assigned a set of coordinates in the classification space by means of the classification Formula and the system vocabulary."

With respect to the use of linguistic techniques in automatic indexing and classification, methods of computational linguistics may be used to derive measures of the probable significance of words in document texts. Damerau 34/ reports experimentation with word subset selection for indexing purposes based upon word occurrence frequencies significantly larger than expected frequencies (following Edmundson and Wyllys, in part), with encouraging results. Findings by Black 33/, Simmons et al 44/, Spiegel and Bennett 38/, and Wallace 45/, among others, suggest the need for continuing investigations in the area of proper discrimination between significant clue words and non-informing words for a particular corpus or collection. Extensive computer processing and analyses such as Dennis 46/ has applied to the legal literature are needed for other subject matter fields. The latter investigator warns that neither raw word frequencies nor the numbers of documents in which a word occurs provide good criteria for distinguishing between trivial or non-informing and significant or informing words. She suggests, instead, that "discrimination increases with the skewness of the word distribution in the file".

Baxendale has suggested that certain types of phrase structures and nominal constructions, as determined by relatively unsophisticated machine syntactic analyses, are useful in revealing appropriate subject-content clues. A recent example is provided by Clarke and Wall 47/: "The hypothesis is that the importance of nominal constructions in selection of index unit candidates places emphasis on the bracketing of all noun phrases." Baxendale's continuing work 48/ further suggests that "through the methods of statistical decision theory it is hoped to formulate quantitative measures that will separate informative index terms from noninformative." Continuing use of syntactic analysis principles is provided as an option in the SMART system (Salton 49/) and possibilities for choosing index terms automatically by syntactic criteria have been explored by Dolby et al 35/.

Closely related to automatic classification or indexing experiments involving linguistic factors are document and word grouping investigations for homograph resolution and subject field identification purposes, such as those of Doyle 50/ and Wallace 45/. Doyle used a Fortran computer program developed by Ward and Hook for iterative automatic groupings of 50 physics and 50 non-physics documents. He was able to show clear-cut separation of two meanings of words such as "force" and "satellite".

A case involving overlaps of word memberships in more than one subject class has been investigated by Wallace 45/. Using word frequency data, he found 48 words in common on the first 100 word-frequency rankings for psychological and computer literature abstracts, with function words predominating. However, using a word rank sum criterion, he was able to separate 50 psychological abstracts from 50 computer abstracts with 78 percent success.

We may thus conclude that the progress and prospects of automatic indexing, as of September 1966, are both provocative and challenging. They are "provocative" because so much in terms of both practical and theoretical accomplishment has already been demonstrated, and "challenging" because so much remains to be done. Further, what remains to be done will in all probability require serious, intensive, and imaginative investigations of a wide variety of questions from the relative usage and acceptability of a KWIC index through possible changes in author and editor practices to the fundamental questions of semantics and human judgment.

Nevertheless, when the results of automatic classification or automatic indexing procedures reach levels of 70 percent or better mean agreement either with human indexers or with potential users evaluating the relevance of items retrieved by such indexing, then the machine methods should be preferred to routine, run-of-the-mill, manual indexing wherever the costs are at least commensurate.

The technical feasibility of achieving such performance levels for a relatively small number of classification categories or a relatively small vocabulary of index terms has already been demonstrated experimentally. There remain unresolved questions of the extent to which it will be possible to apply such techniques to the larger vocabulary requirements and the practical operating considerations in actual collections.

Assuming that we can solve these problems, however, many advantages will accrue. First is the speed with which many items can be indexed --- in a few minutes or hours at most for, say, 10,000 items. Secondly, there are advantages of timeliness and the ease with which an entire collection can be re-indexed or re-classified. A third advantage is the consistency of the machine procedures, especially as compared with the <u>in</u>consistency to be noted in available data on tests of comparative performance among indexers.

The advantage of ability to re-index quickly, easily, and inexpensively (because most input costs will have been incurred previously) is of major importance in terms of overcoming present barriers to the introduction of improvements in operating systems (since, as Kyle 51/ points out, "The most common reason for not trying new and/or improved techniques of classification and indexing is the difficulty of reclassifying and re-indexing large collections") and in terms of dynamic revision and up-dating (as Borko 37/ emphasizes).

Another advantage, particularly of methods using teaching samples is (as suggested by Mooers as early as 1959 52/), the capability for making assignments of indexing terms in, say, an English language system to items whose texts are written in other languages: French, German, or Russian. This type of advantage can point the way to greater international collaboration in indexing and document control procedures.

A further possibility is suggested by the convergence of automatic indexing techniques based upon teaching samples with adaptive selective dissemination systems and client feedback possibilities, especially those involving "more-like-this!" requests. If we assume a large-scale, multiple-access system with adequate personalized files for the typical client, the <u>common</u> data bank of document identificatory and selection criteria, condensed representations, and full text (if available) can be selectively accessed by him on the basis of automatic indexing generated by his own choice of selection criteria and his own choice of exemplar items for each such criterion.

He may provide a standing-order interest profile with respect to patterns of his own selection criteria, with weighting indications as to relative degrees of interest. Dynamic re-adjustments to standing requests and weightings can be made in accordance both with his responses to notifications and with any "more-like-this" requests received from him. System accounting and usage statistics can provide a feedback warning system as to the adequacy of his selection-criteria set and enable him to initiate re-processing of those documents in the collection likely to be of current interest to him.

We must close, however, with a caveat: if machines have not yet mastered us, neither have we yet the requirements of the machine to the degree of advanced planning that will be required, especially for those information processing operations involving the analysis of content and not merely the manipulation of records: for here we are faced with the great challenges of human communication, human decision-making, and human-problem-solving.