"The differentiation that is made between the two types of indexing is that word indexing is inextricably tied to the words in a text: If a word appears it gets indexed as such; if it does not appear it does not get indexed. Concept indexing, on the other hand, has an element of abstraction in it: Words may either be indexed as such or may be converted, either by themselves or in combination with other words, into concepts which may not bear a direct resemblance to the words or combinations of words that evoked them in the indexer's mind."

Machine techniques such as those of Luhn's KWIC, like the early Uniterm systems, look no farther than the words used by the one author himself. Techniques such as those of Maron, Swanson, Borko, Meadow and Williams, among others, look specifically to relationships between words as used by one author to patterns of word usages in a given subject area or given document collection. They may also look to these patterns as in turn related to prior human analytic judgments of the "aboutness" referrents of items in the collection. In this sense, they at least attempt replication by machine of assignment indexing.

There is no real question but that machines can in fact derive words from text provided that it is in machine-readable form. This machine procedure may involve direct extraction of all words as index entries, as in a complete concordance. It may involve the extraction of only those words which survive a "purging" operation in which articles, conjunctions, adjectives, and other "common" words are first deleted. Various machine-controlled modifications to such "derivative" indexing are also available. The case for machine achievement of assignment indexing for any but limited special cases is not so clear.

2. INDEXES COMPILED BY MACHINE

A first and obvious use of machines in indexing processes is in the manipulation of index entries, previously selected on the basis of human analysis, to produce various orderings, duplications and listings of these entries. The power of machine techniques to speed and economize the sorting, ordering and listing operations in the preparation or compilation of indexes was recognized quite early, both in the field of library science and in the consideration of potential areas of application by specialists in machine potentialities.

In particular, two specialized types of index, at least in the broad sense, are such that their compilation would be almost prohibitive in terms of time and cost were it not for the use of machines. These are, respectively, the case of the complete index, the index to all words of a text in their various contexts, which is a concordance, 1/ and the case of the "citation index", which has been used in the field of law for many years but has only quite recently been suggested for literature search purposes related to scientific and technical information.

1/ See, for example, Doyle, 1963 [162], p. 11: "Without data-processing machinery, concordances are prohibitively expensive to generate for most uses except in those cases where it is well known that a given volume of text is going to be used again and again, by large numbers of people over a long period of time. As we know, clergymen have made use of manually prepared concordances of the Bible since the 12th century".

In machine-compiled indexes, no item or entries are eliminated by the machine, whereas in even the most rudimentary of machine-generated indexes, such as KWIC, various reductive or extractive operations are automatically applied as a part of the machine procedure. We shall be concerned in this section with brief discussions of machine-compiled indexes and related devices, specifically, concordances, card or book catalogs mechanically prepared, citation indexes, and special indexes such as Tabledex. The use of machines to compile, sort, duplicate and list index entries can only be considered to be mechanized indexing in a relatively trivial sense. We shall consider, therefore, only a few representative examples, emphasizing early work and some of the pioneering instances.

2.1 Concordances and Complete Text Processing

When as early as 1856, Crestadoro proposed the use of permutations of the words in titles as a subject-content index the only "machines" available for the processing operations were people acting in a strictly clerical way. Precisely such clerical operations have been used for centuries in a process that is, in the special sense of full representation of document contents, an index-producing operation--the making of concordances. 1/ The task of listing each separate word in a book in all the contexts in which it appears is incredibly time-consuming and tedious when carried out by manual means. There are those who have spent the major part of their lifetimes at this task. For example: "It took James Strong thirty years to compile his exhaustive Concordance of the Bible..." 2/ The use of machines capable of processing signals which represent and preserve information offered a potentially revolutionary change, and with the advent of the electronic computer even more radical possibilities of very high speed processing were opened up.

As early as 1949, J. W. Mauchly (the co-inventor of ENIAC and UNIVAC) envisioned the use of computers for documentation and library science activities. He suggested that the full information contents of the Library of Congress collections could be recorded in machine language, stored in this form on magnetic tape, and searched by machine in a procedure which would match words or other selection indicia occurring in the recorded information to the specified words or selection criteria of a query or search prescription. Specifically, he estimated that the entire collection, then amounting to 10,000,000 books, could when transcribed to binary-code representation 3/ be serially searched in 20 hours. 4/

1/ See, for example, Black, 1962 [65], p. 314: "The oldest book in the world has had such an index for many years--the concordance to the Bible;" Markus, 1962 [394], p. 19: "The ultimate in permutation for indexing is a published concordance;" Linder, 1960 [363], p. 99: "We know of a concordance prepared in the 13th Century;" Simmons and McConlogue, 1962 [555], p. 3: "Complete indexing has been used of course for centuries in the preparation of concordances."

2/ Carlson, 1963 [101], p. 211.

3/ That is, markings which have one of two values (thus, binary digits or "bits"), can be used to distinguish between 2^n different other symbols such as alphabetic characters by using log 2^n of such markings. A binary code for the 26 letters of the English alphabet requires a five-bit representation for each letter. If numeric digit characters are also recorded, (26+10), a six-bit code representation is required.

4/ Mauchly, 1949 [406], p. 295. See also "Report to the Secretary of Commerce on the application of machines..." 1954 [620], p. 67.

Mauchly's suggestion was, in effect, the idea of a complete index that could be searched by machine. We should note, however, that although subsequent technological advances could significantly decrease his original time estimate, the crucial questions that remain are those of what, assuming one-to-one representation of document text, one would search for. 1/ Natural language searching by machine, in the sense of full text inspection, is a "pay-as-you-go" concordance technique. It is, however, a technique which must be aided and abetted by various forms of synonym reduction, syntactic normalization, homograph resolution and other special processing operations if it is to be in any sense an effective tool for selection of clues to be retrieved.

Gardin, in a series of recent lectures on automatic documentation, (Gardin, 1963 [207, 208]) 2/ refers to the opinions of some investigators that it should be possible to "jump" the stage of indexing and to search the natural language texts directly. The problem, he points out, then shifts to the determination of all the various ways in which the possible answers to a question may have been expressed in these natural language "complete indexes". Instead of carrying out reductions or condensations of the documents, as in normal indexing procedures, amplifications of questions are required. "Reductive" indexing of the source documents can only be eliminated at the expense of "expansive" indexing of questions. Gardin concludes that the gain from this is very doubtful.

There is also the presently staggering burden of time and cost to convert full texts to machine-usable form. As of February, 1961, it was estimated that the natural language text material available for machine processing amounted to little more than the words contained in the Harvard Classics five-foot shelf (Stevens, 1962 [567]). Perhaps up to ten times that amount is now available, notably in the 6,000,000 words of the statutes of Pennsylvania 3/ and in several million additional words that have since been keypunched at the Center for Automation of Literature Analysis, Gallarate, Italy. 4/ A very recently

---

1/ See, for example, Yngve, 1959 [657], pp.978-979: "We will have to find formal connections between widely divergent ways of saying essentially the same thing. In addition there is much that we will have to learn about searching. If we had today a complete grammar of English which was capable of rendering explicit all the relations and distinctions implicit in the document, I doubt that we would know how to use it effectively in a machine search situation. We would be embarrassed by the very wealth of the information available. Much more must be learned about search situations."

2/ See also Bar-Hillel, 1962 [35], p. 415: "Could not the stage of clue assignment be completely skipped and the request topic be directly compared with the original documents? It is very natural that such a thought should have arisen, but it must be stressed that there is nothing in our knowledge of the workings of communication which would indicate that such a proposal is, or ever will be, practical."

3/ See various references by J. F. Horty, W. B. Eldridge and S. F. Dennis, E. M. Fels, R. Wilson.

4/ R. Busa, data reported at the NATO Advanced Study Institute on Automatic Document Analysis, Venice, July 1963.

completed study made by the TRW Computer Division, Thompson Ramo Wooldridge, involves the investigation of the possibilities for a center to provide text in machine-usable form. The report gives a total figure of approximately 50,000,000 words of text so available as of February 28, 1964, but this includes non-scientific text, such as newspaper and popular magazine materials (Mersel and Smith, 1964 [415]).

Mersel and Smith also report on the estimated requirements for machine-usable text for various research groups, averaging over a million words per year per group. Yet, at present keypunching costs of one cent or more per word, is it reasonable to assume that any of these research groups can provide a budget of over $100,000 per year for this purpose alone? Moreover, this budget would provide for the conversion of no more than a thousand 1,000-word items or a hundred 10,000-word items at costs, respectively, of $100 or $1,000 per item. For the present, therefore, the conclusion is inescapable: either indexing or search based upon full text processing is not yet practical. Even the most enthusiastic proponents of "searching full natural language text" (Swanson, 1960 [589]) and "maximum-depth indexing" (Simmons and McConlogue, 1962 [555]) generally agree as to the present impracticality of full-text mechanized indexing except for special limited cases.

The two problems of determining what to search for, given full text, and of feasibility of conversion of text into machine-usable form thus combine to limit "complete indexing" largely to the special cases of providing corpora for studies in the field of computational linguistics and of compiling the traditional scholarly tool--the concordance to all the words in a given literary work or works. Apparent exceptions, including experimental work with abstracts only and the law statutes studies, are usually cases in which the selective principle of disregarding common words (and hence the bulk of the actual text) is applied automatically either on input or in subsequent processing (Cleverdon and Mills, 1963 [131]). These cases, therefore, may be considered machine-generated indexes rather than machine-compiled. Moreover, it should be noted that:

> ". . . . . The law, itself, is an appropriate field for data retrieval. The statutes, especially, are written in relatively clear, concise language. At least, this is their intent. Practically, this means that input and output can both be relatively short and that retrieval of legal information will be involved with fewer semantic difficulties." 1/

In the area of concordance-making, however, the potentialities of machine compilation have been put to good use. The pioneer efforts in this area are unquestionably those of Father Roberto Busa, S.J., of the Gallarate Center. As early as 1946, Busa proposed to his superiors that a card file recording all the words used in all of the works of St. Thomas Aquinas should be set up, and he began his actual experiments using IBM punched card equipment in 1949 (Busa, 1953 [87], 1960 [91], and 1958 [92]; Secrest, 1958 [540]). 2/ Appearing in 1951, his Sancti Thomas Aquinatis Hymnorum Ritualium Varia Specimina Concordantiarum is the first known example of a complete word index that was compiled by machine techniques. The early Gallarate work was carried out on standard punched card equipment, but from the time of the concordance to the Dead Sea Scrolls, computers have also been used (Tasman, 1959 [595], [596], and [597]). The major continuing task is still to other works of St. Thomas. Other machine-compiled concordances produced by Busa's Center include one to Goethe's Farbenlehre, Bd. 3.

1/ Asher and Kurfeerst, 1963 [24], pp.1-2.

2/ See also Scheele (ed.), 1961 [522], pp.206-209.

Other relatively well-known examples of machine-compiled concordances include those to the Revised Standard Version of the Bible (Ellison, 1957 [186]; Cook, 1957 [139]) and to Matthew Arnold's poetry (Painter, 1960 [461]; Parrish [467, 468]). The Cornell Concordance Series, under the general editorial supervision of Parrish, includes investigations of Old English, such as The Anglo-Saxon Poetic Records (Bessinger, 1961 [59]).

The November 1962 issue of Current Research and Development in Scientific Documentation, No. 11, [430], lists several concordances compiled by machine including the work of Sebeok [533, 534] and associates at Indiana University on Cheremis folksongs, the work on the National Vocabulary of the French language under Quemada at the University of Besancon, 1/ the preparation of glossaries and concordances to the works of Kant at the University of Bonn 2/, and concordances to medieval German texts being compiled by Wisbey at the University of Cambridge (Wisbey, 1962 [646], [647]). At the University of Gothenburg in Sweden, work has begun on mechanical linguistic analysis of English language texts, using the machine-readable teletypesetter tapes used for the printing of paperback books (Ellegård, 1960 [184] and 1962 [185]). 3/ Another recent example is that of the work at the Summer School of Linguistics, University of Mexico (Grimes and Alvarez, 1961 [243]). By 1963, Marthaler writes that "Compiling concordances with the aid of a computer is already standard routine to such an extent that it needs hardly be described in detail." 4/ As of January 1964, a general-purpose computer program for the IBM 7090 which can compile various types of concordances has been announced as available from the Mechanolinguistics Project at the University of California. (1964 [95]). 5/

The major advantage of using machines to compile concordances is, of course, the enormous difference in the time required to complete the work. Thus, only 120 hours were required on the UNIVAC computer to prepare the 800,000 words of the Concordance to the Revised Standard Version of the Bible (Cook, 1957 [139]; Ellison, 1957 [186]). 6/

1/ See "Actes du colloque sur le mecanisation...", 1961 [1]; Quemada, 1961 [485] and 1959 [486]; Centre d'Etude du Vocabulaire Francaise, "Specimens de Travaux lexicographiques...", 1960 [106].

2/ National Science Foundations CR&D Report No. 11 [430] p. 316

3/ Ibid, p. 321.

4/ Marthaler, 1963 [399], p. 14

5/ "California Concordance Program Available", 1964 [95]

6/ Carlson, 1963 [101], p. 211.

In the use of the IBM 705 for the concordance to the Summa Theologiae, Fr. Busa reports that only 60 hours were required to arrange in alphabetical order 1,600,000 words. 1/ This advantage of speed, with the concomitant benefits of both economy and timeliness, is illustrated by Tasman as follows:

> "... It has been estimated that it would take 50 scholars 40 years... to manually index the 13 million or so words of St. Thomas Aquinas' complete works. IBM punched card machines would produce the indexes and concordances much more accurately and would take ten scholars about four years. Large-scale data processing techniques would reduce the time to about 25 percent...(or)...ten scholars to do the job in less than a year." 2/

Other advantages stem from the facility with which further machine processing can be introduced. Once the text is in machine-readable form, a number of valuable byproducts can be derived. Examples are statistics on the number of words that have 2, 3, ... n letters, frequencies of letter usage; printouts of occurrences of specified words or groups of words; and lists alphabetized on terminal rather than initial letters. Added advantages of computer processing are further exemplified in the options available with the California concordance computer program (1964 [95]), some of which are as follows:

(1) The user may obtain a restricted rather than a full concordance by supplying a list of words for which no entries are to be made.

(2) The user may obtain a selective concordance by supplying a list of words for which, and only for which, entries are to be made.

(3) Each entry word may be centered with its preceding and succeeding context, up to the limits of one full line of 131 characters, or each entry word may be listed together with the full sentence or verse in which it occurs.

(4) Text with interlinear information such as grammatical symbols can be used and selective concordances can be compiled on the basis of such interlinear information.

(5) The citations of an entry can be listed in order of textual occurrence, in an order determined by preceding or following words in its context or in an order determined by accompanying interlinear symbols.

2.2 Card Catalogs, Book Catalogs, Bibliographies and Subject Index Listings Prepared by Machine

The use of machines such as punched card equipment for the preparation and processing of library card catalogs and of index listings was advocated by a few far-sighted documentalists at least as early as the 1930's (Parker, 1938 [463]; Dewey, 1959 [153]).

1/ See his statement in Scheele, 1961 [522], p. 209.

2/ Tasman, 1958, [596], p. 11.

McCormick's bibliography on mechanized library processes (1963 [407]) lists a number of early suggestions, notably those of Fair in 1936 [187], Shera in 1938 [547], and Gates [225] and Callander [96, 97, 98] in 1946. Cox, Bailey and Casey proposed the use of punched card equipment for the preparation of bibliographies in the field of chemistry in 1945 [142].

By 1946, Gull claimed that:

> "...Punched cards and present equipment offer new possibilities right now for solving the problems of the indexes to Chemical Abstracts. These indexes are large undertakings in themselves, and the work of arranging, cumulating, and printing them can be simplified by placing the index information on punched cards at the time the abstracts are made. With current indexes on punched cards, two or three cumulations of the author index during the year will greatly reduce the work required in using current issues from that approach. Cumulations of the subject, patent, and formula indexes immediately become possible for intervals more frequent than once a year." [245]

The following year (1947) saw a summary by Gull of potential applications of punched cards in special libraries [247], and Becker surveyed some of the then discernible prospects for library mechanization, as a student in the Library School of Catholic University. He stressed such advantages as flexibility in the processing of new material for abstracting, indexing, filing, and interfiling purposes and the printing out of various listings in any format. 1/

The potential use of machines for library science and documentation had not actually been recognized, however, for many years after the invention of punched card equipment. Both the punched card developments (beginning with Hollerith and Powers in the 1880's) and the electronic computers developed from 1946 onward were first applied to the automatic manipulation of information in the sense of statistical, mathematical, or engineering data, rather than to information about data or information about other information. Dr. John Shaw Billings, himself a librarian of note, was apparently the first to suggest to Herman Hollerith the idea of recording information as holes punched in cards which could then be sorted mechanically. 2/ Larkey comments: "It is not known if Billings ever thought of applying the principle to bibliographic work, but it would seem eminently fitting that it might be so utilized." 3/

Larkey himself as head of the Army Medical Library Research Project at the Welch Medical Library, Johns Hopkins University, was certainly one of the pioneers in such utilization, but this was almost 70 years from the date of the Billings-Hollerith conversations. The Army Project, begun in late 1948 or early 1949, had as its contract

---

1/ Becker, 1947, [43], pp. 11-12: "From the flexible arrangement of the cards, bibliographies become readily available by subject, author, and title. In special libraries, where material on one subject is concentrated, the research possibilities of gathering, sorting, filing, and printing information are almost limitless. Continuous machine interfiling permits keeping current with new entry additions."

2/ "With the masters...", 1963 [648], p. 18.

3/ Larkey, 1953 [351], p. 34.

objective "to explore existing and projected methods, emphasizing machine methods, applicable to such pilot projects as may be necessary" (Larkey, 1949 [348], 1956 [349], and 1953 [351]). Also as of 1949, the Library of the Department of Agriculture is reported to have "conducted an experiment in the use of electronic data-processing machines to produce the author and subject indexes to the 'Bibliography of Agriculture'." 1/

It is not until the early 1950's, however, that punched card machine techniques were actively put to use for the preparation of card catalogs, book catalogs, bibliographies and various index listings. Then, a number of independent but largely concurrent applications were tried out on at least an experimental basis, including in addition to the work of the Welch Medical Library Project pioneering efforts in mechanized book catalog production (Griffin, 1960 [242]; Martin, 1953 [400]; Berry, 1958 [58]) and what is claimed to be the "first successful non-experimental punched-card catalog of periodicals", the Serial Titles Newly Received (now New Serial Titles), as published by the Library of Congress from 1951 onwards. 2/

The work at the Welch Medical Library continued for several years, the final report being issued in 1955 [234]. Beginning in 1951, the project maintained in punched card form the subject heading authority list used for the Current List of Medical Literature (Larkey, 1953 [351]; Garfield, 1953 [217] and 1954 [220]." Garfield has stated that this work "clearly demonstrated the ease of converting alphabetic subject heading lists to categorized or classified lists of terms by the use of punched card equipment." 3/ That is, each heading or subheading had assigned to it a numeric code reflecting its appropriate position in the classified system, which could then be used by machine for sorting, ordering and listing. Ingenious use was made of the IBM 101 Statistical Machine in the preparation of printed subject indexes (Garfield, 1953 [218] and 1954 [216]). Other subject heading lists maintained by punched card techniques by 1953 or earlier included those of the U. S. Patent Office and the Technical Information Division of the Library of Congress. 4/

The first loose-leaf printed book catalog to be produced by machine methods was apparently that of the King County Public Library in the State of Washington in 1951, and the following year the Los Angeles County Library inaugurated a similar system for the distribution of a master book catalog prepared by mechanized techniques (Berry, 1958 [58]; Griffin, 1960 [242]; Martin, 1953 [400]; Alvord, 1952 [4]).

The work on mechanized preparation of lists of periodicals at the Library of Congress has been reported as follows:

> "In 1951, the Library began publishing, at monthly intervals, Serial Titles Newly Received. In 1953, its title was changed to New Serial Titles... Ever since its inception, the fundamental ingredient of the publication has been the IBM punched card...

1/ U.S. Congress, Senate Committee on Government Operations, 1960 [619], p. 147.

2/ Dewey, 1959 [153], p. 36.

3/ Garfield, 1959 [221], p. 471.

4/ Garfield, 1954 [220], p. 1.

"Two important advantages of the punched-card method were foreseen when the publication began. First, it would be possible to print lists from the cards at will, without any further editing or proofreading, once the information was in punched-card form. Second, there was the possibility of mechanically preparing special lists of titles, selected on the basis of subject, country, or language." 1/

Thus, by 1953, "a number of instances of printed indexes prepared by machine" could be claimed. 2/ The use of punched cards to sort, to prepare tabular listings for various drafts and revisions, and to interfile corrected or revised entries greatly facilitated the preparation at Battelle Memorial Institute of the subject index to the Proceedings of the International Conference on the Peaceful Uses of Atomic Energy, 1955 (Lipetz, 1960 [367]).

Developments in the use of punched card machine techniques in bibliographic operations of these types, beginning in the 1950's, have by no means been limited to the United States. For example, Remington Rand punched cards have been used in the preparation of a national union catalog of Italian libraries, 3/ and Mikhailov reports for the All-Union Institute of Scientific and Technical Information (VINITI) as follows:

"The development program for machine production of indexes has been underway at the Institute for a number of years... In fact, operational use of Soviet-made punch-card machines to compile the author indexes for some of the series of our Abstract Journal has been practiced at the Institute since 1957." 4/

In France, at the Centre d'Etudes Nucleaires, Saclay, a program has been developed for mechanization of the production of biweekly and cumulative indexes and for demand searches (Chonez, 1960 [116, 117, 118]).

With the advent of automatic data processing systems, the speed, the flexibility and the capability for multiple-purpose processing buttress the claim that the card catalog can be "replaced or supplemented by book catalogs made with the aid of mechanized equipment". 5/ It is further claimed that "The printed catalog produced by means of automatic equipment combines the best features of the conventional card catalog and the traditional printed catalog, and adds to both new dimensions that would have been unbelievable a generation ago." 6/ A joint project is under way by the Medical Libraries of Columbia,

---

1/ U. S. Congress Senate Committee on Government Operations, 1960 [619], p. 85.

2/ Larkey, 1953, [351], p. 38.

3/ Berry, 1958 [58], p. 287.

4/ Mikhailov, 1962 [410], p. 50.

5/ McCormick, 1963 [408], p. 195.

6/ Vertanes, 1961 [625], p. 242. This is with reference to the LILCO Library Printed Catalog, which is prepared by sorting and processing information on titles, authors and titles-by-subject-groupings serving as indexes to the holdings at the Long Island Lighting Company.

Harvard, and Yale Universities for computer preparation of book catalogs for books published from 1960 onward (Kilgour, et al 1963 [324]). Another recent illustrative example of the production of printed book catalogs by means of computer compilation is that of the Boeing "SLIP" System (Weinstein and Spry, 1963 [633]).

Along with recognition of computer-processing potentialities there has emerged increased awareness of the desirability of taking advantage of one-time recording of information to serve multiple purposes: the principle of by-product data generation. The advantages for the library and document collection are that a single recording of bibliographic information in machine-usable form can lead to a variety of products, specifically including printed book catalogs, 1/ recurrent and demand bibliographies, the requisite number of copies for conventional card catalogs, card catalog sets or catalog listings for the personal use of the individual worker, input to mechanized selection and retrieval systems, and machine-manipulatable data for such other purposes as circulation control.

Turner and Kennedy report, for example, the initial use of a Flexowriter to prepare library catalog cards and the by-product generation, via a 1401 computer, of bi-weekly listings of unclassified report titles at the Lawrence Radiation Laboratory, the "SAPIR" System (Turner and Kennedy, 1961 [615]). Chasen discusses a change from a previous punched card system for circulation and recall at General Electric's Missile and Space Division Laboratory to a combined Flexowriter and G. E. 225 computer procedure to provide mechanized retrieval, compilation of desk catalogs, computer updating of catalogs and files, and the maintenance of subscription lists (Chasen, 1963 [108]).

Fasana describes a system at the Air Force Cambridge Research Laboratory Library where typing indications in the tape are used as boundary codes. He reports:

> "Input tapes are currently being processed on a computer to automatically produce catalog card sets, circulation control records, and book form indexes. Original input tapes now being accumulated will form the basis of a machine-searchable file to be used in the future for more sophisticated printouts and searches." 2/

For such applications, Durkin and White make the following typical claims:

> "The system described has permitted the IBM Command Control Center Engineering Library to produce its catalog cards and library bulletin both faster and cheaper. Since a by-product of this process is the preparation of all catalog information in

---

1/ See for example, Olney, 1963 [458], p. 42: "During the past few years a number of libraries have initiated a program of mechanization... by punching on IBM cards or paper tape some of the bibliographic information normally given on catalog cards. Recording this information in machine-readable form makes it very easy to prepare printed book catalogs..."

2/ Fasana, 1963 [195], p. 326. This system involves the "Machine-Interpretable Natural Format" and procedures developed for AFCRL by Itek Corporation; see also Lipetz et al, 1962 [368].

punched card form, it has also permitted the establishment of a circulation control system, the publication of overdue notices and reading lists, and the eventual institution of a computer retrieval program" (Durkin and White, 1961 [173]; White, 1963 [638]).

Heiliger reports for the library of the new Chicago Campus of the University of Illinois as follows:

"The type of bibliography the computer can produce does make greater use of LC card information than do present card catalogs. With the computer programmed with a set of library filing rules and a set of symbols that describes for the computer the various parts of the bibliographic unit, it can print-out, for instance, a list of books published in a given country, between certain years, on a certain subject (or combination of subjects), that are illustrated and have bibliographies. It will also be possible to permute on individual items in LC subject headings in the same fashion that Chemical Titles does on titles. This index has been dubbed POSH (permuted on subject headings)." 1/

Some recent experimental work at Inforonics, Inc. puts major emphasis on by-product data generation, beginning with the actual preparation of manuscripts for publication. Tape typewriter processing of manuscript for journal articles is being studied from the point of view of producing machine-usable text. This text, together with coded identification of the separate items in the text, is so prepared that computer programs can produce from the single-input automatic typesetting tapes for the article itself, author and subject index entries, and the like. Computer text transformations can also produce entries for citation indexes, abstract journals and search files (Buckland, 1963 [83, 84]).

Other computer-produced indexes or special indexes involving compilation rather than selection by machine include indexes to Nuclear Science Abstracts (Day and Lebow, 1960 [151]), the Current List of Medical Literature (Chonez, 1960 [116, 117, 118]), the Retrieval Guide to Thermophysical Properties Research Literature, 2/ and the Research and Development Abstracts of the USAEC (Sherrod, 1963 [541]). At the Atomic Energy Commission also, a modification of this RDA computer program is used for author, corporate author, number and subject indexes for the Engineering Materials List, which includes announcements of blueprints and drawings. 3/ In several instances, machine processing capabilities are used for permuted listings under various assigned indexing terms. 4/ Special cases of machine permutation operations involve compilation and organization of chain indexes, used to reflect the various key entries in faceted classification systems (Dowell and Marshall, 1962 [159]; Foskett, 1962 [199]; Olney 1963 [458]).

1/ Heiliger, 1962 [259], p. 475.

2/ Markus, 1962 [394], p. 19; Touloukian, 1962, 1963 [607].

3/ Davis, 1963 [150] p. 237.

4/ See, for example, reports on the SWIFT program for NASA's STAR (Newbaker and Savage, 1963 [438]); the AIMS System (Heller, 1963 [260], and the SPINSTRE System (Wheater, 1963 [639]).

A final special case of a computer-compiled index should be noted. This is the work of Schultz and Shepherd with reference to the annual meetings of the Federation of American Societies for Experimental Biology (FASEB) (Schultz and Shepherd, 1960 [532]; Schultz, 1963 [527]; Shepherd 1963 [545]". 1/ The indexing terms are generated first by the authors of the papers but are then run against a computer program, which by thesaurus-type look-up eliminates synonyms and supplies syndetic devices in addition to formatting the subject index for printout.

The machine-readable thesaurus developed for this project presently performs the following four basic functions (Schultz, 1963 [527]):

1. It accepts words from titles and indicia supplied by the authors without modification if they match acceptable indexing terms.

2. It recognizes certain other words as acceptable if modified and modifies them accordingly, for example, by "use" directions for synonyms and near-synonyms.

3. It adds additional indexing terms when certain words occur, an example being " 'penicillin', use also 'antibiotics'."

4. It deletes certain words if they do not occur in the context of an acceptable indexing phrase.

2.3 Tabledex and Other Special Purpose Indexes

The uses of machine techniques in index compilation so far discussed represent instances in which conventional tools of bibliographic control can be prepared at lower cost or more rapidly, or both. In addition, however, certain new and unconventional types of index have been or are being produced with the aid of computers.

The Tabledex method, as proposed by Ledley in 1958 (Ledley, 1958 [352], Zusman, et al, 1962 [661]; O'Connor, 1960 [442]), involves coordinate indexing in bound book form, with special features to facilitate search, conserve space and display index terms co-occurring with a given term for a given item. 2/ A major advantage claimed for this method is that by the use of computers bibliographies and book-form indexes can be organized, compiled, and printed in page format within a matter of hours.

A Tabledex index typically consists of a bibliography proper, in which each citation has been assigned an identifying number; an alphabetical list of the indexing terms used,

---

1/ These investigators claim the first production of a conventional subject index by computer.

2/ See, for example, O'Connor, 1960 [446], p. 241: "Ledley approximately halves the average size of the document descriptions required by imposing an order on the vocabulary of indexing terms. When a document description belongs in a term subset, only those terms of the description need to be recorded which come later in term order than the term of the term of the subset. This illustrates another type of storage organization."

which may also have numeric codes; and a set of indexing tables. These tables contain item numbers in the leftmost column, and either the names or the codes for indexing terms assigned to an item along the row. There is one such table for each distinct term used in indexing the items.

To facilitate searching, only those terms which are of higher numeric or alphabetic order than that for the term for which the particular table is compiled are recorded in the rows. Thus to make a search on several terms, the user turns to the table for the one of these terms that has the lowest term value, which table records all items to which the term has been assigned, and checks the rows of the table for the second lowest ranking term, the third, and so on. Variations in the Tabledex method allow for the automatic assignment of numeric codes to the indexing terms based on relative frequency of use within the collection. Ledley also discusses methods for finding articles associated with all except one, all except two, or all except n of the given words in a search prescription. 1/

A first example of a computer-compiled Tabledex index was that to a bibliography prepared by the Library of Congress for the International Geophysical Year (Zusman et al, 1962 [661]). 2/ The computer program for the IBM 7090 carried out the operations of assigning accession numbers, extracting index terms and compiling the term lists, determining frequencies so as to assign frequency numbers to the terms, organizing and preparing the tables, and developing an author index. Two formats were used, one giving terms by numeric code and the other spelling out the terms as normal words. The latter feature provides a measure of browsability in the system. 3/ A Tabledex compilation program is also in use at the Applied Physics Laboratory of Johns Hopkins University (Olmer and Rich, 1963 [454]).

Another coordinate index search tool, making use of what is in effect a document-descriptor matrix with special codes and column arrangements to save space and facilitate rapid scanning, is the Scan-Column Index suggested in 1960 by O'Connor [449]. He further suggested the use of computers for compilation, as follows:

"A computer can organize information about documents into a scan-column index. The input needed consists of the document identifications and their accompanying

---

1/ Ledley, 1959 [352], pp. 1235-1239.

2/ See also National Science Foundation CR&D No. 11 [430], pp. 130-131.

3/ Zusman, et al 1962, [661], p. ii: "... The word tables have the advantage that browsing can be accomplished and possible associations made during the search... Such 'browsing' can be enhanced by including at the end of each row in a table all the other words also associated with the article of that row".

index terms... and an indication of either the number of columns desired or the column density desired. The computer will determine the frequency of each term, the positive and negative correlations of terms, and the quantity of these correlations by counting or sampling key figures, such as the average number of terms per document. It then can assign column-character codes accordingly."[1/]

In 1961, Costello described the use of computer techniques for compilation and computer printout of a dual dictionary for a coordinate indexing system using links and roles at DuPont's Polychemicals Department. After manual analysis, term-role assignments are keypunched, the cards are listed for editing including the elimination of synonyms and the indication of appropriate postings to more generic terms, and re-keypunched for conversion to magnetic tape. Tapes for posting of items and links to term-roles are merged by computer with tapes giving alphabetical equivalents of term codes and with appropriate syndetic indications for final output on an IBM 407 high-speed printer [141].

Still another instance of a coordinate index, modified to show pre-coordination of terms as compiled by computer, is that ot the Electronic Properties Information Center (Johnson, 1963 [301]). The system consists of abstract cards maintained in accession number order, together with machine printouts that pre-coordinate descriptors within nine major categories. The listings of pre-coordinated descriptors are arranged in three different indexes; alphabetically arranged within each category, alphabetized without respect to category but with code indication of the category reference, and a non-categorized listing arranged alphabetically in reverse order. Advantages of machine processing include the ease with which various statistical counts can be made, such as the average number of items in the system for a given material and a specified property. Summary indications of the state-of-the-art in the field of interest can be obtained, "for the system will indicate not only areas where research has been done, but also areas where gaps in the literature occur, and a measure of the growth of research activities in the field can be developed." [2/]

2.4 Citation Indexes

"A citation index is a directory of cited references in which each reference is accompanied by a list of source documents which cite it." [3/] This is a relatively new

---

1/ O'Connor, 1962 [449], pp 18-49.

2/ Johnson, 1963 [301], p. 296.

3/ Sher and Garfield, 1963 [546], p. 63.

type of bibliographic search tool that would be almost impossible to compile without the use of machines. 1/ In at least one case, moreover, the availability of mechanical devices was itself the inspiration for the idea of a citation index to the scientific literature. Garfield states in a 1954 paper that he was led to the idea of "Shepardizing" from an earlier concern with the development of citation codes or "coden" 2/ that would facilitate machine processing of bibliographic and index entries. 3/

The value of Shepard's Citations in tracking down precedents and decisions has been recognized in the legal field for many years. 4/ The desirability of a similar tool for literature searchers in the fields of scientific and technical information was suggested about a decade and a half ago, when Seidell and others proposed its use for patent searching (Seidell, 1949 [541]; Hart, 1949 [255]). In 1954, the Bush Committee in its considerations of the potential applicability of machines to Patent Office problems received a proposal from the Atlantic Research Corporation of Alexandria, Virginia, which was to cover "the development of a Patent Citation Index, comparable to Shepard's Citations". 5/In the period 1954-1956, both Garfield 6/and Fano 7/independently advocated the development of a citation indexing tool for scientific and technical literature. As

---

1/ See, for example, Atherton, 1962 [25], p. 4: "The volume of data to be processed is so massive that processing machines are a necessity"; Garfield 1954 [210], p. 4: "Where such large volume of data is to be handled it must be expected that mechanical devices of high speed and versatility... would probably be a determining factor in the system's success."

2/ That is, brief codes, often mnemonic, for journal title abbreviations and other clues to publisher and date of publication.

3/ Garfield, 1954 [210], p. 2.

4/ How to Use Shepard's Citations [281] has been published periodically by Shepard's Citations, Inc., Colorado Springs, since 1873.

5/ U. S. Dept. of Commerce "Report to the Secretary of Commerce...," 1954 [620], p. 27.

6/ Garfield [210, 211, 212]. Adair, writing in January, 1955, specifically acknowledges a suggestion of Garfield's (for 1955 [2], p. 32) but Garfield in turn credits Adair, (1963 [214], p. 290).

7/ Fano, 1956 [191], p. 3: "Let us accept, at least for the sake of this argument, the conclusion that linguistic associations between documents cannot lead to a satisfactory definition of a bibliography. Then the only other type of association for which evidence is available is that provided by simultaneous references in the literature, by the concomitant use of documents by experts as evidenced by library records, and by other similar joint events."

of today, there are at least five or six instances of citation indexes that have been produced, several different experimental investigations are under way, and new interest has been generated by the considerations of the Weinberg Panel. Thus:

> "Of the newer approaches to the indexing of scientific documents, the Weinberg Panel was particularly impressed with the citation index as a promising bibliography tool. In order to learn more about this approach, the National Science Foundation is currently sponsoring the compilation and publication of extensive citation indexes for the fields of genetics and also for statistics and probability; and is supporting two kinds of experiments to evaluate different techniques for using citation data in indexes and searching systems in the field of physics." 1/

In general, the principle of citation indexing is based upon the hypothesis that the bibliographic references cited by an author provide significant clues to the subject content of the author's own paper and/or that there is a certain commonality in subject between papers that cite the same references or that are co-cited. 2/ The principle can be applied to the compilation of bibliographical or indexing tools in several different ways. First, there is the method of citedness, which groups for a given item the identifications of subsequent items that have cited it. The converse of this is, of course, the bibliography or reference list of a given item. 3/ In the first case, we are concerned with "descendants," and in the list of references with "ancestors". 4/

1/ Committee on Scientific Information, 1963, [135], p. 16.

2/ Compare Adair, 1955, [2], p. 32, with respect to Shepard's Citations itself: "Since all of the cases listed under a given case have cited it, it follows that they must all be, more or less, pertinent to the case cited." See also Kessler, 1963, [320], p. 1: "This method ... originated in the hypothesis that the bibliography of technical papers is one way by which the author can indicate the intellectual environment within which he operates, and if two papers show similar bibliographies there is an implied relation between them."

3/ See Salton, 1962, [520], p. III-3: "A citation index consists of a set of bibliographic references (the set of 'cited' documents), each being followed by a list of all those documents (the 'citing' documents) which include the given cited document as a reference. A citation index is to be distinguished from a reference index which lists all cited documents under each citing document."

4/ See, for example, Tukey, 1962, [611], p. 5: "Any user's greatest need is likely to be for access to the latest information rather than to the oldest, but the latest items are children, not ancestors. Genealogy is important, but progress requires tracing descendants Iung and Vandeputte, 1960, [291], p. 11, make a similar distinction between "histoire" (antecedents) and "filiation" (successors).

A second method, implied in Fano's suggestions for the use of relative frequencies of association between items found in the literature, is one of citingness, which groups together items that cite one or more identical references. This method has been developed by Kessler and his associates as the technique of "bibliographic coupling" (Kessler, [317] through [323]. The purpose here is to identify groupings of related items where relatedness is defined in terms of the number of references shared by each of the members of the group with some given test paper or with each other. It is noted that where the citedness index and the reference list typically give the bibliographic references themselves as the searching or retrieval tool, the bibliographic coupling technique seeks rather to define groups of similar papers. 1/ A third method, and one which may be combined with either of the other two, is to derive indexing terms for a given paper from the overlay of indexing terms previously assigned to any papers which it cites. Salton 2/ further suggests that:

> "... Citation indexes could be used to extend a given set of index terms by starting with the terms attached to a given document or document set, and adding to them the 'related' terms obtained from new documents which cite the original ones."

The suggested advantages of citation indexing include the claims that this tool does not require trained indexers, 3/ that it is highly susceptible to mechanization (Garfield, 1955 [213], 1956 [212], 1957 [211]; Atherton, 1962 [25]; Becker and Hayes, 1963 [45]), and that it may cost significantly less than subject indexing. 4/ A major advantage claimed is responsiveness to user, rather than indexer, interests and view points. 5/ Some of the representative claims with respect to this factor are as follows:

1/ See Atherton and Yovich, 1962 [26], p. 3: "Kessler's method, however, does not retrieve the references cited by a paper. Instead these references are examined to determine the 'bonds' between papers; e.g., if two papers share six references, in common, they are said to have a 'coupling strength' of six. By applying either of two criteria of coupling, one can 'filter out smaller groups of papers' related to a given paper."

2/ Salton, 1962 [520], p. III-8; see also Lesk, 1963 [356].

3/ Atherton, 1962, [25], p. 3.

4/ See Atherton and Yovich, 1962 [26], pp. 3-4: "Garfield estimates cost of abstracting and indexing 200,000 articles in one year to be $3 million. He estimates the cost of a citation index for these same articles (approximately 3 million citations) to be $300,000." See also Doyle, 1963, [162], p. 8: "The editing labor, the input preparation cost, and the automatic processing time are all so small that it's very likely citation indexing is destined for a great surge of popularity in the immediate future."

5/ Committee on Scientific Information, 1963 [135], pp. 55-56: "Because the indexing is based on the author's rather than on an indexer's estimate of what articles are related to what other articles, citation indexes are particularly responsive to the user's, rather than to the indexer's viewpoint."

"The most feasible scheme for alerting individuals to what is of interest in their own field requires an on-going up-to-date citation index. For each narrow field of interest of an individual there are, it is believed with good reason, three to five to ten key items such that:

(c1) If he knew that a new item referred to one of his key items, the individual would be glad to skim the new item,

(c2) An individual who skimmed all new items referring to one of his key items would be adequately alerted to the newest results in his own specialties." 1/

"A research worker who finds one article several years old can relate later developments by locating all subsequent articles that have referred to it. Corrections and errata can be brought together by a citation index." 2/

"Citation indexing will overcome artificial dividing lines that are drawn in various abstracting services." 3/

"It is believed that citation indexes will be useful...in bringing together related materials in different fields where the interrelationships are not readily identifiable from other types of indexes." 4/

"Since the end product of a citation indexing is a listing which collects in one place the bibliographical descendants of a given cited author, bringing these titles together helps to illuminate for the searcher the extent and nature of information association patterns employed by other authors who had a similar or related interest to his own. Its development, therefore, serves as an approach to the user's frame of reference, not the indexer's." 5/

The importance of being able to pick up more than the principal subject matter clues is indeed an advantage of citation indexing. Garfield, commenting on the potential cross-breeding of interests, gives an example of a personal search for more information on the RCA electronic scanning pencil in which he was led to one of Busa's reports on machine use in philological analysis and to an article of interest in the field of information theory. 6/ Garfield further points out that the cross-breeding can extend across

---

1/ Tukey, 1962 [611], p. 9.

2/ Atherton, 1962 [25], p. 2. See also Garfield, 1955 [213], p. 1.

3/ Atherton and Yovich, 1962 [26], p. 3.

4/ Brownson, 1963 [82], p. 3. See also Garfield, 1957 [211], p. 4.

5/ Becker and Hayes, 1963 [45], p. 137.

6/ Garfield, 1954 [210], pp. 4-5.

changes of terminology with time, [1] and Lipetz suggests that it can break down barriers with respect to use of foreign literature. [2]

Other claimed advantages relate to the usefulness of the citation index for purposes other than those of direct literature search. Such other purposes include identification of significant research by "equating frequency of citation with relative significance of subject matter", (Salton, 1962 [520]), determinations of the number of references cited in a given field or by journal or publication date (Atherton, 1962 [25]), evaluation of the relative importance of various scientific journals (Westbrook, 1960 [636]; Kessler, 1961 [322]), tracing of trends in the history of ideas or in a particular field of literature (Brownson, 1963 [82]; Salton, 1962 [520]) [3] and empirical studies of the frequencies of self-citation, multiple authorship, and the like (Atherton, 1962 [25]).

A number of disadvantages of the citation index are to be noted, however. First is the obvious lack of consistency between authors in terms of whether or not they cite the prior literature at all and in terms of the completeness and correctness of the citations they do make. [4] Atherton quotes Westbrook as saying:

> "Science is subject to changing fashions of interest that lead to a distorted number of published papers in a given subject and an inordinately high level of citations to any one who reports first on the fashionable subject. The method will not appraise work performed but not published." [5]

1/ Ibid, p. 6: "Changes in terminology are to a certain extent overcome through the citation approach, since the author who makes a reference to a paper that is forty or fifty years old is making the jump in terminology for us." See also Garfield, 1956 [212], p. 11.

2/ Lipetz, 1963, [366], p. 265: "It is reasoned that availability of a citation index derived from Soviet physics journals and approachable through familar American references should stimulate utilization of the Soviet physics journals in the United States."

3/ See also Reisner, 1963 [497], p. 71: "Citation indexes are receiving increasing attention as bibliographic aids and as sociometric tools. As sociometric tools, they are being used to explore the flow of information across national boundaries and from pure to applied fields, to determine the structure of a field, and to determine the 'value' of documents or authors."

4/ See, for example, Doyle, 1963 [162], p. 8: "The disadvantages of this kind of indexing is, of course, that it depends on authors providing ample and suitable references"; Salton, 1962 [520], p. III-7: "In many cases personal preferences are evident both as to number and types of papers cited; authors have varying backgrounds, and there may also exist a tendency toward self-citation regardless of relevancy"; Thompson, 1963 [600], p. II-1: "The difficulties...are largely due to the extreme variability of format and to the lack of standardization which prevails in the publication of citations."

5/ Atherton, 1962 [25], p. 4, citing J. H. Westbrook.

An author not cited frequently enough or not cited within a given time period will not appear in the citation index. Doyle points out that there are "many kinds of documents we would like to retrieve where it is not customary to provide citations at all". 1/ In the bibliographic coupling method, both those papers which make no references to any other paper and those papers which do not share at least one reference with some other paper in the system are automatically excluded. 2/

Other disadvantages of the citation indexing technique relate to difficulties of the lack of standard practices in the citing of references and to problems of recognizing whether one citation is or is not equivalent to another. These are, of course, related to the normal difficulties arising from non-standardized formats and practices in descriptive cataloging, in use of journal abbreviations, in transliterations of foreign language titles and names, and the like, but they are now aggravated by the present prospects for direct machine processing. As Lipetz points out:

> "Author's names may be cited in somewhat different ways, and there is no simple mechanical procedure for bringing together the different versions. For example, an author's name may be cited both with and without initials; it would take a comparison of the additional information on the cited reference to establish that these authors are the same. Even more difficult are the problems of mechanically determining that a misspelling has occurred." 3/

Both the disadvantages of incomplete and disproportionate coverage and of failures to equate equivalent citations are quite readily obvious to the user of a citation index if he is reasonably familiar with the subject field or document set that is covered. Thus, the use of the citation index as the exclusive tool for literature search is subject to defects of both oversight and 'over-cite' which are cumulative and which are often easily recognizable. Atherton and Yovich emphasize that: "Knowledge of these weaknesses tends to prevent anyone from trusting the system's ability to retrieve the pertinent literature." 4/

In general, however, the citation index has not been proposed as an exclusive means for literature search and retrieval, but rather as one of a set of tools or as a supplement to other indexes. 5/ In this connection, it is of interest to note that a manual technique of literature search tested at The Thermophysical Properties Research Center

---

1/ Doyle, 1963 [162], p. 8.

2/ See Atherton and Yovich, 1962 [26], p. 39; Marthaler, 1963 [399], p. 23.

3/ Lipetz, 1962 [364], p. 262.

4/ Atherton and Yovich, 1962 [26], p. 39.

5/ See, for example, Tukey 1962 [611], p. 10: "The citation index, in its retrieval and pursuit uses, is not something to be used alone. Rather, it is the tool whose presence makes all the other tools more effective."

while not using a citation index as such, makes use of a supplementary citation tracing technique both to shorten manual search time through abstract journals and to follow up additional search leads (Lykoudis, et al, 1959 [387]; Cezairliyan, 1962 [107]). The technique is briefly described as follows:

> "One starts searching the abstracting journal beginning with the most recent issue and going back through a number of years, a. Next, the bibliographies of the papers located in these a years are searched for new references. The references found in this second step of the search will, in general, cover a period of years (b - a). Then one reverts back to searching through the abstracting journal again for another period of a years starting with the year b. This cyclic procedure of alternate searches through the abstracting journal, followed by searching the bibliographies of uncovered papers, is repeated until the total number of desired years of search is covered." 1/

In a sample search on the thermophysical properties of metals, the results showed that the cost of the cyclic procedure was only 65% of the cost of conventional manual search using the abstract journals only.

Recent efforts in the development and use of citation indexes proper include experiments in evaluation at the American Institute of Physics, 2/ an extensive compilation and processing program at the Institute for Scientific Information, 3/ and a cooperative program between the Statistical Techniques Research Group of Princeton University and the Bell Telephone Laboratories (Tukey, 1962 [611] and [612]). Reisner has reported work on the compilation of a citation index to 30,000 patent disclosures and its experimental evaluation in progress at IBM's Thomas J. Watson Research Center (1963 [497]). Goodman is concerned with a citation index to the literature of new educational media, especially that on programmed learning and teaching machines (1963 [235]).

At the Centre d'Etudes Nucleaires de Saclay, a citation index to papers in the field of thermonuclear fusion and plasma physics is being prepared. 4/ Lipetz is carrying on work in the preparation and evaluation of citation indexes, begun at the Itek Corporation, as an independent worker and consultant to the A.I.P. project. 5/ Carroll and Summit report that citation indexing is under consideration at Lockheed's Missile and Space Division, (1962 [102]). Kessler and associates at M.I.T. 6/ and Salton's group at

---

1/ Lykoudis et al, 1959 [387], abstract, p. 351.

2/ Atherton and Yovich, 1962 [26]; National Science Foundation's CR&D Report No. 11, p. 12.

3/ Ibid, pp. 27-28.

4/ Ibid, p. 76.

5/ Ibid, p. 181.

6/ Ibid, p. 128.

the Harvard Computation Laboratory (Salton, 1961 [512], 1962 [513], 1963 [514] and [515]), are concerned with citations as a basis for grouping and categorizing sets of related documents.

Early examples of citation indexes that have been produced include the precedents in the fields of statistics and information theory listed by Tukey. 1/ Tukey also refers to early experimentation involving manually manipulated card files by J. L. Hodges, Jr., Charles H. Kraft, and William H. Kruskal. 2/ Goodman (1963 [235]) describes the use of Termatrex cards showing for each item other items cited by it.

Examples of machine-compiled citation indexes, however, are those of Garfield and Sher in the field of genetics (1963 [546]), Lipetz's experimental index to the citations in the proceedings of the two United Nations conferences on the peaceful uses of atomic energy, (1961 [364], 1960 [365]), and the citation index to references listed in the "Short Papers" submitted for the 1963 Annual Meeting of the American Documentation Institute (Luhn, 1963 [377]). As of January, 1964, the first five volumes of Science Citation Index are available from the Institute for Scientific Information. These volumes are reported to have 2,250,000 lines of copy representing the computer-compiled citation trails for 102,000 articles published in 1961. 3/

Preliminary evaluations of the citation indexing principle have, as noted previously, been carried out in an American Institute of Physics project supported by the National Science Foundation. One experiment involved the selection of a single paper from the December 1, 1961 issue of The Physical Review and the tracing of references and citations through that journal for the period 1956 to 1960. A bibliography of 64 papers was produced as a result. This was then evaluated by a nuclear physicist, who found that the titles alone were an insufficient basis for judging whether or not these papers should all have been included, and who commented critically that there was no way of knowing if all the papers really relevant to the subject of the test paper had indeed been found. A further check by search of the subject index did in fact reveal six pertinent papers which had been missed by the citation indexing technique.

A second experiment at the American Institute of Physics involved application of Kessler's "coupling strength" criteria to 41 of the 64 papers selected in the first experiment, the remainder being excluded because they shared no references with any other paper. The resultant groupings of presumably highly related papers were also evaluated by a subject matter specialist, who found them relevant to each other but the selection incomplete. Atherton and Yovich, reporting these A.I.P. experiments, concluded that: "More work will have to be done before the usefulness of citation indexing can be accurately determined." 4/

1/ Tukey, 1962 [611], pp. 23-24.

2/ Ibid. p. 24.

3/ See news note, Special Libraries, Jan. 1964, p. 58.

4/ Atherton and Yovich, 1962 [26], p. 22.

Kessler himself and his associates have also conducted some experiments in comparative evaluation of indexing aids derived from citation data on the one hand and from conventional subject indexing on the other. The basis for evaluation was a total of 334 papers published in The Physical Review in 1958. The study involved detailed comparison of the ways in which these papers fell into related groups according to the "analytic subject index" used by the journal's editors and according to the method of "bibliographic coupling". The essentials of the latter method are described as follows:

> "a. A single item of reference used by two papers is called one unit of coupling between them.
>
> "b. A number of papers constitute a related group G_A, if each member of the group has at least one coupling unit to a given test paper P_O.
>
> "c. The coupling strength between P_O and any member of G_A is measured by the number of coupling units (n) between them." 1/

For the 334 papers, 73 categories of the Analytic Subject Index (ASI) had been used. For the bibliographic coupling method, each of the papers was in turn considered as the test paper and groups were formed for any of the 333 other papers that shared one or more citations with it. In general, it was concluded that there was good correlation between the groupings of papers achieved by the two methods. It should be noted, however, that 44 papers fell into no groups at all on the basis of the bibliographic coupling criterion. 2/

Salton and associates at the Harvard Computation Laboratory are also concerned with the citation indexing principle as a possible basis for grouping similar documents. They are also concerned with evaluation of results so obtained by comparison with document groups obtained by subject indexing means. In the comparative experiments, data were first compiled for a closed document set of 62 items as to similarities with respect to both "citedness" and "citingness". The same items were manually indexed and similarity coefficients between these items were derived from overlappings of assigned index terms. When the two measures of similarity were compared with each other and with document associations obtained by random assignments of "citations" and "terms", the conclusions reached were as follows:

> "The similarity coefficients obtained by comparing overlapping citations for a sample document collection with overlapping, manually generated index terms are much larger than those obtained by assuming a random assignment of citations and terms to the documents; relatively large similarity coefficients are generated for nearly all documents which exhibit at least a minimum number of citations; little seems to be gained by using citation links of length greater than two; for early documents, citedness furnishes a better indication than the amount of citing, and vice versa for recent documents; for documents which can both cite and be cited, equally good indications seem to be obtained by comparing citing and cited documents." 3/

1/ Kessler, 1963 [320], p. 1, footnote.

2/ Ibid, p. 5.

3/ Salton, 1962 [520], p. III-42.

In the Salton project, tests of the value of citation links for the assignment of index terms have been made by comparing the citation pattern of an "unknown" document with those of other documents in the collection to derive a set of five "related" documents, where relatedness is decided on the basis of the magnitude of the similarity coefficients for the citation links. Any index term that appears at least twice in the set of terms previously assigned to the five related documents is then assigned to the new item. In general, approximately 50% of the terms so assigned were also assigned to the same "new" items by human indexing procedures. 1/

As we have previously noted, however, the advantages of citation indexing are likely to be most effectively applied when used as part of an array of other tools. Tukey suggests, in particular, that permutation indexes of titles, as in KWIC systems, would be of great value as "starter" and "re-check" mechanisms for the use of citation indexes. 2/ Brownson reports:

> "Consideration is now being given to the possibility of experimenting with a 'hybrid' type of index that would combine permuted titles, authors, and citation data. Such an index might be more useful than any of the individual types of indexes issued singly; and, since no human indexing judgment would be involved, it could be prepared largely by machine and issued rapidly." 3/

Williams, while at ITEK, proposed a hybrid integrated index combining listings by authors, corporate authors or author affiliations, keywords-in-context from title, and references to works cited by and to works citing an item, and she also developed a sample format for selected items from several journals in the field of philosophy. 4/

Precisely such a hybrid tool was provided with the Short Papers for the A. D. I. Annual Meeting 1963, and it was indeed issued rapidly. A brief period of only two or three weeks elapsed between receipt of many of the manuscripts and the distribution of two automatically typeset volumes. The second of these volumes contains a KWIC and an author index to these papers themselves, a bibliography and citation index to all papers referenced by them, and KWIC and author indexes to the cited papers, all computer-compiled within this time period. 5/

1/ Ibid, See also Lesk 1963, [357], p. V-8.

2/ Tukey, 1962, [611], p. 12.

3/ Brownson, 1963 [82], p. 4.

4/ T. M. Williams, private communication, dated January 4, 1962.

5/ Luhn, 1963 [376], and [377], pp. 353-382.

2.5 Machine Conversion From One Index Set to Another

A final possibility in the general area of machine compilation of indexes and machine use to improve the availability of indexes is as yet in a highly speculative stage. This is the possibility of converting from one index set to another by machine look-up procedures. In the Welch Medical Library project, mentioned earlier, use was made of punched card techniques to convert from one index arrangement to another, 1/ but machine-recognizable identifiers for both arrangements were explicitly encoded in the material. In recent studies at Datatrol, however, preliminary investigations have been conducted looking toward machine lookup of index-term equivalence tables in order to convert, for example, DDC descriptors to corresponding subject headings used in the AEC vocabulary.

Hammond and Rosenborg (1962 [250] and [252]) report on the compilation of a unilateral table of "indexing equivalents" between approximately 7,000 DDC descriptors and those AEC subject headings judged by them to be identical, synonymous, or "usefully" equivalent, such as one or the other being subsumed by a broader or more generic term. Findings showed 23.8% of the terms of the DDC vocabulary presumably identical to those of AEC, 38.1% of lower generic level, 7.4% of higher generic level, and 10.9% for which no useful equivalents could be found. A sample table of indexing equivalents was prepared for DDC-to-AEC conversion, but not in the opposite direction.

Since, in general, convertibility of indexing vocabularies would be desirable wherever duplication of cataloging and indexing effort is likely to occur (that is, where two or more different documentation organizations receive at least some of the same material as inputs to their systems), the results of these preliminary studies are provocative and appear to merit the further study that is being sponsored by an Interagency Task Group on Vocabulary Study of the Committee on Scientific Information, under the Federal Council for Science and Technology.

There are many substantial difficulties, however. When applied to actual indexing of the same items by the two agencies, it was found that for 277 items indexed by both AEC and DDC (then ASTIA):

> "ASTIA used a total of 2,571 descriptors, and AEC 840 subject headings... of these, 392, or roughly half of the AEC terms, were either completely or, for all practical purpose, identical." 2/

Painter (1963 [460]) made further studies of equivalency in her investigations of duplication and consistency of subject indexing at several Government agencies. For 200 items indexed by both AEC and DDC, she found 20% DDC equivalency, 67% AEC equivalency, and 30% similarity of actual indexing. She concludes, in part:

> "In considering these solutions and the statistics revealed by the studies it should be concluded that with a maximum of only 69 percent equivalency, or convertibility, and a minimum of 28 percent, there is still a large proportion of terms which will

---

1/ Garfield, 1959 [221], p. 471.

2/ Hammond 1962 [250], p. 4.

necessitate some other form of retrieval. This is the proportion which is involved with the problem of generics, where a term in one system subsumes two of another ---and vice-versa. An additional problem evolves in attempting to reconcile two different subject concepts, one, the subject heading which usually has a single access point and one, the uniterm or descriptor which has multiple access through coordination. Thus the practicality of a system made up of many units supplying information indexed differently, using as a basis for retrieval a table of equivalents, is questionable." 1/

Moreover, the results of tests of inter-indexer consistency rates within the same agency were not encouraging. Thus Painter further concludes:

"The study, in combining the results of the equivalency analysis and the consistency of indexing within each system and an equivalency of only 30 percent within the broadest system, a table of equivalents is at present of little value in either a manual or a machine system. In order to apply a table of equivalents efficiently, both a high degree of consistency and a high degree of equivalency is essential." 2/

She therefore stresses that the possibilities for conversion by machine techniques from one indexing set to an equivalent set for another vocabulary are adversely affected by the generally poor rates of inter-indexer consistency. With reference both to the Datatrol Studies 3/ and to corroborative findings of her own, she states:

"The value of equivalency studies and most particularly the table of equivalents presuppose the consistency of indexing. Convertibility between systems is thus dependent on the consistency of indexing. Without consistency, the vocabularies as units are not sound; equivalencies cannot be drawn or effectively used for convertibility." 4/

1/ Painter, 1963 [460], p. 104.

2/ Ibid, p. ix.

3/ Hammond, 1962 [250]; Hammond and Rosenborg, 1962 [252].

4/ Painter, 1963, [460]. p. 109. Note that these estimates of inter-indexer consistency may be quite optimistic, as discussed on pp. 157-160 of this report.