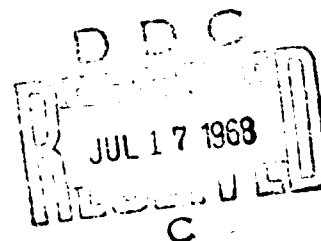


APR 1969

SOME THOUGHTS ON MACHINE INDEXING

Andrew E. Wessel

June 1968



P-3869

This document has been approved
for public release and sale; its
distribution is unlimited.

SOME THOUGHTS ON MACHINE INDEXING

Andrew E. Wessel*

The RAND Corporation, Santa Monica, California

I want to share with you some of my thoughts concerning the problems associated with machine indexing, to offer you some opinions, and to discuss some specific cases of attempts to accomplish at least partial solutions to some of these problems. If I were asked to express the substance of my views in a few words, it would be something like this:

No matter how feasible one views the long-term possibilities of fully automatic indexing, significant effort must be allocated to what I have termed "machine-aided indexing;" we should not expend all of our efforts in the attempt to achieve fully automatic machine indexing.

Now let us see what exactly is meant by this remark and why I feel that other individuals working in this field should feel similarly obligated.

For one thing, machine-aided indexing is not only feasible and practical in many real applications, it can also be regarded as the development of an important series of initial steps toward the achievement of more fully automatic processes. I say "more fully

* Any views expressed in this paper are those of the author. They should not be interpreted as reflecting the views of The RAND Corporation or the official opinion or policy of any of its governmental or private research sponsors. Papers are reproduced by The RAND Corporation as a courtesy to members of its staff.

This is background material that was used in a talk at the STAC I ICIREPAT meeting in Munich on April 22, 1968.

The author is a consultant to the Studiengruppe für Systemforschung, 69 Heidelberg, Bergstr. 143, West Germany.

automatic" rather than "fully automatic" because I am increasingly led to believe that fully automatic indexing will prove to be a will-o'-the-wisp, or to be less colloquial, to be similar to attempting to square the circle. And I don't really think that approximate solutions convincing enough to apply to real world data (excluding certain oversimplified cases) are available.

On the other hand, the development of machine-aided indexing techniques which keep the human indexer integral to the indexing process seems quite feasible, as I said in my talk to the STAC 1/STAC 2/ICIREPAT meetings this April in Munich. Machine-aided indexing asks of the machine only that machines do well what they can do in providing help to the human indexer in doing well what the human can do.

The machine can provide on-line and direct error-checking and error-corrective guidance; it can provide consistency checks; in fact, if we so desire, the machine can help to enforce consistency among different human indexers. The machine, if appropriately programmed, can also provide a learning or instructional period for training indexers as they begin to use a new index list. Based upon certain procedural rules which I believe can be made explicit, the machine can help to bring about more thorough indexing by human indexers in less time than with purely manual methods. But the machine cannot, I happen to believe, index real world material all by itself. No matter how clever we may think we are, the machine is simply not so clever, and while with much ingenuity we may teach this idiot of a machine to count expressions and to use statistical algorithms, to locate and associate so-called key words, etc., we are still left with results that at best

are the output of a well-trained idiot. It is not so much a question of whether indexing is an art or science, it is rather that even if indexing could become a science, it will not become a science for idiots.

Yet it is quite clear, and will become even more so as more and more automated data banks come into existence, that machine-aided indexing will become an absolute necessity to achieve anything like the expected utility of such data banks. And by "expected" utility I mean the utility claimed for such automated data banks when sold to the users. I am much concerned that if we devote all of our research efforts to achieving fully automated solutions of indexing problems, we will be unable to offer any help to such users with their burdens of data preparation for entry into data banks.

An abstract argument as to why fully automated solutions will fail can be derived from the experience of early 20th century attempts within the logical positivist philosophical camp to develop an acceptable linguistic structure for mathematics and science. Early in this game Russell and Whitehead were confronted by the implications of Gödel's incompleteness proof in their attempt to develop a logically sound and rigorous basis for mathematics. Later Carnap et al attempted a similar approach for the language of physics and other sciences. While much clarification was achieved by such attempts, their overall goals continued to recede faster than the development of syntactic and semantic technology within modern logic. For me, at least, the end result was the rather interesting but somewhat neglected proof of Professor Nelson Goodman in the 1950's that no two uses of the same word could be regarded as synonymous in meaning even with a formal

linguistic structure. This should cause some deep soul searching on the part of those who hope to instruct computers to index raw text without significant human participation. If humans, using the most powerful logical tools available, are unable to develop acceptably rigorous explications of the language of the sciences, I think it rather optimistic to believe that with the relatively weaker logical tools available for computer syntactic analysis, useful machine indexing can be achieved for material far less precise and rigorously structured than the language of physics. Our data banks will be dealing with what to the logician, if not to the layman, is ordinary language. By this I mean that the material--the texts--required for our data banks will be in languages for which a rigorous and formal syntax cannot be developed and for which a formal and rigorous semantic interpretation does not exist. The fact that computers can "count" very rapidly does not seem to help us very much in this case. The problem, unhappily, is not one of mere semantic interpretation. If it were, perhaps we could legitimately entertain the slim hope that by syntactic techniques (counting, associating frequencies of appearance of words and expressions, statistical techniques, etc.) we could sufficiently determine semantic meaning to permit valid indexing. Yet even this slim hope rests upon the dubious assumption that syntactic and semantic meaning exhaust the total meaning of the terms and expressions in the material we shall have to use for our automated data banks. When we introduce the potential system user, when we remember we have him to deal with, and that he is what the system is all about, we are continually confronted by the blunt fact

that no necessary (or even discoverable) relationships exist between syntax and semantics on the one hand and pragmatic meaning (the use of the data and material) on the other hand.

Now I admit to having indulged in some rather abstract philosophical issues, the relevance of which may be thought to be questionable, but I have done so because the statistician, engineer, computer programmer, etc., are sometimes too arrogant to pay much attention to the implications of the theoretical limitations of their tools. They sometimes behave as if they believed that because their techniques have been shown to be useful in some contexts, and because they have some kind of picture image in their heads as to what the term "automatic indexing" may mean, that all we need do is to somehow bring these techniques together with their picture image of the problem and Lo and Behold, a solution to the problem of automatic indexing will, given time, energy and, of course, funding, appear. Furthermore, it has been all too easy to construct at least one example of an automatic indexing case to prove the point that success must be just around the corner, conveniently either forgetting any and all counter-examples or explaining such counter-examples away by one rationalization or another.

It might not be so bad an idea to continue this game if the development of the computer state of the art had remained where it was only a few years ago. But we now have the hardware capability, the memory speed and capacities (or will have in only a few years) to make large automated data banks more generally available than we had supposed if (and this is a large if) we can make the data preparation and entry tasks economically feasible and within the range of

organizational capabilities. If we simply produce and deliver computerized systems for automated data banks which continue to impose upon potential users laborious and uneconomic burdens in order to prepare and maintain the data for such systems, we will continue to see the system's utility severely compromised by the real world restrictions and limitations upon data entry. And sooner or later, somebody might get just a bit disillusioned with the whole process. In particular, the results of our attempts to achieve automated indexing have been so dismal to date that many potential users have rejected all attempts at machine indexing. I believe such a total rejection is an error of pessimism equal to the earlier optimistic error. The real danger then seems to me to be that such past failures and present difficulties could lead to the rejection of attempts to develop machine-aided indexing. If this takes place we are going to have a large number of white elephants on our hands and appropriately disillusioned users.

I want now to attempt to illustrate these general remarks by discussing some practical experiences with three different kinds of machine-aided indexing tasks. As soon as one gets down to real tasks or applications one discovers that there is no such thing as the general problem of machine indexing. These three cases present quite distinct problems and require equally distinct and unique solutions. They pose another blunt warning to those searching for a general solution to the problem and point out a recurrent feature of scientific development and application. The questions we ask are often more important to success than the solutions we seek.

Case No. I - A Dissemination Problem

Several years ago I was involved with a problem in the dissemination of various kinds of news-type reports to analysts of such material. A bottleneck existed with what I will call the daily (or even weekly) take, i.e., material gathered from various sources, such as news reports, magazine articles, government releases, etc. Such material "ages" rather soon and in this case useful analyses depended upon rapid and timely dissemination to the appropriate analyst. The context was one with which we are becoming increasingly familiar. The bottleneck consisted of a small, highly skilled, central group of human indexer/disseminators who received all such incoming material, and after indexing for dissemination via content analysis sent the now indexed material to the appropriate analysts often six months after initial receipt!

Now several research projects attacked this bottleneck by posing the question: "Why can't we machine-disseminate this material and do almost as good a job as the humans but do it faster?" There was no need to do it as well, for obviously moderately good dissemination of the material while still "warm" is better than perfect dissemination after it is "cold".

This sounded rather plausible. The only trouble was the interesting assumption that the best solution of the bottleneck was to do almost as good a job as the human disseminators much faster by machine indexing. Now, sadly, it is years later and the machine has yet to be taught how to do such an "almost as good" job. At the time it seemed to me that this assumption was more interesting than some people thought. For example, why even attempt to do by machine what the human indexers did

by human content analysis? Would not the bottleneck be resolved by another "almost as good" approach? In short, a better question, I thought, was to ask: "Why can't we machine-disseminate this incoming material to one analyst hopefully the most appropriate one, who could use it, and let the manual/human dissemination process continue to function for dissemination in depth?" There were twenty or so different analysts who might possibly use the incoming material for perhaps twenty different reasons. The existing dissemination system was set up to see that all the analysts who might have some need for the material would get it. For this, indexing in depth by content descriptors was required.

But suppose all we required the machine to do was to select just one of these analysts to receive immediate dissemination of selected material. Could we teach a machine to do that? This question paved the way to a successful approach to the solution because it made us notice facts some of which were essentially irrelevant to indexing by content analysis. These facts were:

- a. The material received was a large number of reports, each report containing a variety of not necessarily connected items.
- b. While no acceptable indexing code based upon content description of these items seemed to be in sight, the individual items were already identified by a code or codes indicating that such an item had been obtained to satisfy some established user requirement.
- c. Although such pragmatic or user-requirement codes could not be used to determine the content of the item, the codes were descriptive of particular analysts' needs as they had been developed over time. It then turned out that the requirement codes could be put in matrix form with the names of the desks (tasks) of the analysts and that by doing so the number of analysts who could use the coded items for immediate analysis was reduced from the twenty or so possible to two or three.

- d. Finally, if we matched the set of coded items in a given report to the set associated with the individual analysts, an optimal match would, with about 90 percent of the reports, identify exactly one analyst. For the other 10 percent, no choice could be made between two analysts. This did not mean that other analysts would not be interested in the report or in certain items, but it did mean that the most appropriate analyst could be selected about 90 percent of the time and be given the report with the items of immediate interest to him indicated by the requirement codes. A coin flip could do for the remaining cases where the choice was narrowed down to two.

Furthermore, the same task--to pick only the analyst you think most requires the material--was accomplished less successfully by the human disseminators than by this mechanical process. Content description failed to permit the human disseminators to narrow the choice down to one analyst. Although the humans did an excellent job of dissemination in depth, it was slow when they were overwhelmed with new input, which was usually the case. In short, here was a case where pragmatic indexing (based upon use) was successful for the specific dissemination task chosen and where there seemed to be no determinable relationship between semantic indexing (based upon content) and the pragmatic.

Case No. II - Mental Health Records

Now let us consider a different case. The Reiss-Davis Clinic in Los Angeles has a grant to study the decision processes involved in the diagnosis of mentally disturbed children. It was decided to build an automated data bank of case records used to decide if treatment was required, and if so, what kind of treatment. For our purposes here, we can assume a working search and retrieval system permitting various research-type questions to be asked concerning the diagnostic process. But to prepare such case records for the computer was a back-breaking,

highly expensive task involving high-level analysis of texts and laborious coding and indexing of data. A data description classification procedure was successfully created which required the use of skilled editors. The resultant coded material was then entered by punch card, which required a second cycle of error-correction procedures and re-punching. Although the search and retrieval system was highly successful and adaptable for use by similar clinics, the cost of data preparation ruled out more general application of the system.

Here, then, is another case where waiting for a solution involving fully automated indexing would have meant no further application of the system for the foreseeable future. The question became one of whether anything at all could be done. A positive answer was achieved once it was accepted that the human editors must remain integral to the process but that they could be significantly aided in their analytic, coding, and data description tasks by placing them on-line with a computer. The computer could be programmed to present the raw text to the editors, offer the editing rules and categories, provide immediate error-checking procedures, and finally to accept the edited and coded text for immediate entry into the files, avoiding the time-consuming and error-prone punch card operation. Paper simulation indicated a saving of edit time by a minimum factor of 10 to 1 with significant improvement of editing consistency. Furthermore, the technique proved useful in instructing new editors more easily and with better results than the normal instruction period. Funds are now being requested by Reiss-Davis to demonstrate this machine-aided process for editing such text.

Case No. III - Patent Examination

A third case differed from the Reiss-Davis case in that explicit procedural rules for editing of text had not yet been developed for manual operation.

Currently a RAND/U.S. Patent Office project is developing rules for indexing U.S. patents in the fluidics field. Based upon these indexing rules and procedures, JOSS* programs are now being produced to offer machine-aided, on-line guidance to the patent examiners who are indexing this field. We will machine-demonstrate this process by August or September of this year both in Washington using remote consoles and in Santa Monica. This system will then be used in conjunction with the RAND/U.S. Patent Office search and retrieval system for patent examination. Latest reports from my colleagues at RAND indicate that all is going well.

My conclusion was given to you earlier when I tried to formulate my views on this subject in as few inches as possible. Perhaps a repeat of that comment will now take on more meaning:

No matter how feasible one views the long-term possibilities of fully automatic indexing, significant effort must be allocated to what I have termed "machine-aided indexing;" we should not expend all of our efforts in the attempt to achieve fully automatic machine indexing.

* JOSS is the trademark and service mark of The RAND Corporation for its computer program and services using that program.