Information Systems for Research, Practice and Education

William Goffman, Professor

The recent nuclear accident at Three Mile Island and the subsequent reaction concerning the efficacy of peaceful uses of nuclear power suggested to me an interesting parallel between the use of the atom for peaceful purposes and the use of computing machines for automated information storage and retrieval. Although this analogy on the surface would seem to be far-fetched, there are several intriguing similarities. For example, we all know that power is the basis of our modern industrial society, yet it has been said, and with some justification, that the greatest power source of all is information. In fact, the age in which we live has been characterized at various times as both the nuclear age and the information age. Both nuclear power and mechanized information systems came into being at about the same time as byproducts of World War II. In the former case, it was, of course, the atomic bomb which paved the way for the program of peaceful uses of atomic energy whereas in the latter case, it was the critical importance of military intelligence in the conduct of the war. Interestingly, Dr. Vannever Bush was associated with each activity as a member of the Manhattan project team and as author of the now famous article in the Atlantic Monthly in 1945 in which he warned of a coming crisis in scientific communication. Finally, both movements were thought to hold enormous promise for the future, it being predicted that the atom would in the next twenty-five years be the source of most of our peacetime energy needs and that mechanized information systems would not only serve the traditional mundane informational needs of society but would contribute to the creation of new knowledge in its own right by identifying and aiding in the solution of scientific problems.

Some 30 years later, however, we see that in both cases this promise has not been realized. Granted, there has been progress. We do have nuclear power plants and we do have large mechanized retrieval systems. However, nuclear energy provides only a very small percentage of our needs and in the wake of the Three Mile Island accident, has an unknown but precarious future. On the other hand some of the mundane informational needs of society such as airline ticketing, personnel filing and inventory seem to be well served by mechanized systems. Yet, the predicted nuclear and information revolutions have not come to pass. Before we condemn the atom and the computing machine and cast them aside, let us remember that both are still by far the most effective mechanisms for producing those products for which they were originally developed, namely bombs and computations. The peaceful use of atomic energy and large scale mechanized information systems, we must also remember, are secondary applications, and these have not been too successful. Why is this so? The obvious answer is that we do not yet know enough about the problems involved. In the case of nuclear energy this seems to be the effect of radiation. In the case of information systems, the reasons are a little more complex, and to address this issue, I believe it would be useful to take a brief retrospective look at the history of this movement.

As I previously mentioned, the origins of large scale mechanized information processing goes back to the war during which the efficient and knowledgeable handling of masses of information was so vital. Because these tasks were generally carried out by undermanned staffs of human beings, it was natural to believe that the solution to problems relating to information processing lay in the supply of necessary man power to carry ou a sequence of clerical tasks.

With the immediate post war proliferation of scientific publications which to no small degree resulted from the demonstration of the value of science in the war, it was no wonder that some scientists began to feel that an information explosion was taking place and that critical communication problems were arising in the scientific community as a result. Thus, the Bush article and the ensuing program.

Simultaneous with the attitude was the fact that digital computing machines were becoming accessible to the scientific community. Consequently, based on war time experiences, it was believed that solutions to problems created by the information explosion were obtainable by replacing large staffs of human processors by computing machines which could carry out the required clerical tasks more accurately and more efficiently. It was also natural that the major effort was directed at the scientific literature, a situation which has not substantially changed in the past thirty years. This is, of course, not inappropriate if we accept the notion expressed by Prof. J. Ziman, the eminent British Physicist, that results of research become completely scientific only when they are published, hence the only legitimate base of scientific information is the primary scientific literature. Discontent with the scientific literature by members of the scientific community is not a recent phenomenon and has been voiced by every generation. What is a modern phenomenon is the introduction of mechanization in achieving a systematic approach to the problem. The roots for such an approach precedes the war itself and was best expressed by the British Crystallographer J.D. Bernal in the thirties. "It is clearly no longer sufficient to see that every new observation and discovery is published. The problem has to be looked at from the other end; we need to be sure that every scientific worker receives just that information that can be of the greatest use to him in his work and no more". Bernal went on to say that "The problem is essentially a technical one of selecting units and arranging for their proper distribution and storage, a problem which is every day solved in large business houses and mail order stores. The kind of organization we wish to aim at is one in which all relevant information should be available to each research worker and in amplitude proportional to its degree of relevance". So based on Bernal's hypotheses and spawned by Bush's warning, a major effort was launched. For over a quarter of a century there has been a great deal of activity, supported mainly by the National Science Foundation, National Institutes of Health and the Department of Defense aimed at producing large scale, mechanized information systems, which it was believed would solve the problems in scientific communication posed by the vast amount of scientific literature.

This effort has produced many remarkable computerized data bases, most notably the MEDLINE system of the National Library of Medicine. Thus, modern technology has been applied successfully to gathering vast amounts of bibliographic material into sophisticated computerized data bases. However, the question is: Has this effort been effective? Although such data are very difficult to accumulate and assess, a recent report by Donald King entitled Statistical Indicators of Scientific and Technical Communication prepared for the National Science Foundation, indicates that only 0.9 percent of useful information retrieved by scientists was obtained via computerized systems. These data were based on a survey in which authors identified those channels they actually employed in obtaining articles which they cited and the frequency with which those channels were used. Furthermore, a National Library of Medicine study of its MEDLINE system showed that the percentage of physician and student use among all of its users declined from 1973 to 1975.

What are the reasons for the apparent lack of success of large scale computerized information systems? It certainly isn't due to lack of effort or lack of financial support. In my opinion, it is due to four basic flaws in the underlying assumptions upon which these systems are based.

The first flaw is the assumption that the problem is merely a technical one. This is clearly not the case, otherwise our superb modern technology would already have come to grips with the problem. The fact is it is an intellectual problem, and not a trivial one. That is, "to be sure that every scientific worker receives just that information which is of greatest use to him in his work and no more" cannot be accomplished by technical means alone. In no way can selecting and distributing relevant information among scientists, and selecting and distributing inventories of business houses be considered comparable, as Bernal suggests. In the former case, what constitutes relevance is a complex determination; in the latter, it is relatively simple. Clearly, those large scale mechanized information systems which are successful, e.g. airline ticketing, etc., are those for which the relevance problem is trivially solved, and can be handled by technical means alone. To program a computer in such a way that it can separate from the vast and continuously expanding universe of scientific information, the important and valuable contributions and match them to the interests of the individual scientific worker requires a much deeper understanding of the process of how information is stored and processed in the brain than we now have. Only with such knowledge can we reduce the process to a series of simple clerical tasks which a computer is capable of carrying out.

The second flaw relates to the fact that, as far as I know, all large scale mechanized information systems are quantity rather than quality based. That is, they do not have the capability of filtering out from among the vast quantity of information that which is of high quality. It has been estimated, for example, that any where from fifty to seventy-five percent of the scientific literature is of questionable value. The philosopher W.V. Quine has recently characterized this situation in a brief article aptly entitled "The Paradox of Plenty" which appeared in the journal Daedelus as follows: "The mass of professional journals is so indigestible and so little worth digesting that the good papers, though more numerous than ever, are increasingly in danger of being overlooked".

The physical problems are so vast that attention has been focused on the quantitative aspects of information to the detriment of an understanding of its qualitative aspects. So, although the present systems seem to have been reasonably successful in coping with the mass of bibliographic material, they have not been able - nor have they attempted - to cope with the qualitative aspects. But if establishing relevance is a difficult problem, filtering the relevant material for quality would seem to be even more elusive. For a quality based information system is one that will deliver the information needed, when it is needed, in only the quantity required, so that some judgement relative to reliability, accuracy, and so forth can be made. Is it possible that such a task can be delegated to a mechanical information system? If a system is expected to fulfill this function, and I feel that it should, then that system must encompass more than the technology. It must include the prime resource - the people who work with that technology, namely the users of the system. A purely quantity based mechanized information system will probably deliver more product to the user, but that product will usually contain much non-relevant material of low quality. Consequently, much of the user's time will be spent on unproductive activity.

The third flaw relates to the use of the literature as the primary input and output resource. Although there can be little argument that the literature is the most reliable information source despite its qualitative inconsistency, it is questionable whether the user of an information system is best served if the output of such systems is in the form of documents or, as is generally the case, only titles to documents. Even though these documents might contain the needed information, the user is required to first obtain them, then read them and assimilate the information which they contain. When we add to this the fact that most of the retrieved output will consist of documents of questionable relevance and quality, the figure of 0.9% usage of computerized information systems in the King report is not surprising.

The notion of document outputs, of course, reflects the orientation of such systems towards the researcher who in the natural course of his activity is involved with documents both as producer and consumer. It also reflects the fact that these types of outputs are relatively easy to generate. This leads us directly to the fourth flaw.

The community of users of scientific information can be roughly divided into three groups, namely researchers, practitioners, and students. Heretofore, mechanized information systems have almost exclusively been directed towards the researcher. Yet one may argue that of the three groups, the researcher is the least likely to have an information problem requiring the aid of a large scale computerized system. This would seem to follow from the fact that researchers are involved in the creation of new knowledge whereas the practitioner and the student are involved in the use of known knowledge. Thus, in most cases the researcher whose interests are narrow would have no need to consult a formal information system since in general he could only hope to learn what he already knows. This in general would not be the case for the practitioner or the student. Thus, it would seem to be those two segments of the user population that would derive the greatest benefit from formal information systems, yet very few have been designed with them in mind. Furthermore a document retrieval system would seem to be of secondary importance to practitioners and students although clearly of greater value to students. What these users need, in particular the practitioner, are fact retrieval systems with capabilities of providing intelligence as well as the raw facts. There is an amusing TV commercial for ITT showing a French physician examining a patient, looking puzzled, punching some buttons on a keyboard which we are told will put the physician in direct contact with the NLM which we are shown, from which the physician immediately receives the information needed to treat the patient, all thanks to ITT. We are not told, of course, that the only information the physician can receive is a list of bibliographic citations which under the circumstances would not have been immediately helpful.

Of course, the information fed into any fact retrieval system must be derived from the literature, but the documents themselves besides providing the raw data would only provide supplementary reading material which the user can avail himself of at leisure. Moreover, the relevance and quality problems would be much less severe for systems of this type since the system would not have to anticipate key work as would be the case for a researcher oriented system but need only assess that knowledge which has met the test of time; such as assessments being carried out by subject experts and university faculties.

In the past few years, I believe, there has been a slow but steady change in attitude. For example, there are more and more fact retrieval systems emerging of which the nutrient data base is one of the best examples. The NLM is experimenting with a prototype on-line system for hepatitis which is filtered for quality with capabilities of producing facts, data, and other sorts of intelligence which would seem to be useful to all three segments of user populations. Moreover, this system utilizes a panel of representative user experts whose responsibility it is to carry out the overall assessment of the information before it enters the system. At present, this prototype is expensive and cumbersome but shows that the NLM is beginning to move in the right direction.

The Rockefeller Foundation is greatly interested in promoting research activity relating to the quality issue and has already sponsored one conference and will hold two others later this year to focus on this issue. The WHO is also very interested in these issues particularly as they relate to Less Developed Countries.

Hence, there seems to be cause for cautious optimism and perhaps much of the unfulfilled promise for mechanized information systems will yet come to pass. As for the peaceful use of nuclear energy, who knows?