Information Analysis Center Viewpoint on Wordage Problems: Amount, Languages, and Access*

G. S. SIMPSON, JR.
Battelle Columbus Laboratories, Columbus, Ohio 43201

Received August 13, 1971

Scientists and engineers have solved their wordage and most other information problems through the creation and use of information analysis centers (IAC). IACs are growing in numbers not only in the U. S. but throughout the world, as it becomes increasingly clear that to overcome the amount of, languages of, and access to information requires subject specialists who perform an intellectual service for a peer group. IAC's synthesize, analyze, and compress pertinent information so that knowledge transfer remains effective.

There are over 100 IACs supported by the U. S. Government, and an unknown but probably greater number supported by the private sector. Each IAC, by its very nature, should be considered unique, not only because of its distinctive area of technical responsibility but also because of the way it looks at information problems depends on the type of input it uses to satisfy its user-audience needs.

Grist for the processing mill of an IAC is information, data, or combinations of information and data. At first, such a differentiation sounds like hair splitting. But it really isn't.

What is information then?

Information is a written record. It is subjective (obtained and communicated by a human being). Information usually, but not always, contains data. Examples of information media are letters, trip reports, journal articles, informative abstracts, extracts, and technical reports.

Data are the record of an output of a machine or observation and, in order to become information, must be interpreted by a human being. Examples of data are photographs, tensile or hardness test numbers, chemical reactions, and mensurations.

I will concentrate on IACs whose major emphasis is on information. Conceivably, and hopefully, there should be some agreements between the views of IACs that concentrate on data—e.g., NSRDS—and mine. Most assuredly there will be differences.

WORDAGE PROBLEMS-AMOUNT

Clearly by virtue of the definition I used earlier, information is made up of words, sentences, and paragraphs, accompanied by charts, graphs, pictures, and data. With respect to individual items selected for input, problems that relate to the amount of words are: those potential inputs that contain extraneous words and those that are too brief. This sounds like no one can please us no matter what.

When the information input is only a title or an indicative or telegraphic abstract, then there are not enough words to give us the precision of selection we need in an IAC. We must then search out the original, and if we can get it, we can rescreen and exercise a judgement as to whether we wish to process the information. Of course, this requires quite a bit of time, and, since timeliness is so important to an IAC, we dislike items with too few words.

On the other hand, some authors write in excess detail. Their papers, when of value, are simply extracted so that the pertinent details are preserved.

In essence then, amount of words is not a critical problem to an IAC, but quality of writing is, especially when the paper deals with new concepts or with data, its evaluation, and the circumstances surrounding its collection.

LANGUAGE PROBLEM

The major problem that IACs must cope with is time: the time it takes to identify a pertinent article (usually through a secondary publication), the time it takes to track down the original document, the time it takes to ascertain whether a translation exists, and if not, the time it takes to get the document translated.

Fortunately, since secondary publishers are covering much of the high value world's literature through their abstracting programs, coverage is not a severe problem. Time is the major problem.

An obvious way around the time problem is to utilize scientists and engineers with a capability for reading foreign languages. Unfortunately, such scientists and engineers are rare, and even the rare ones are usually competent in but one or two languages. A number of our IACs follow not only the European literature but also Slavic as well as Oriental. Where are the practicing U. S. scientists and engineers who can read Japanese, Italian, German, Russian, and Hungarian?

Fortunately, many of the foreign high value journals are available as English translations. These are very valuable and are excellent back-ups to the secondary publisher products.

*Presented at the ACS 3rd Northeast Regional Meeting, October 11, 1971, Buffalo, N.Y.

ACCESS PROBLEM

Access is an interesting problem in and of itself. Access refers to the ability of an IAC scientist or engineer who has need of some information to get it—when he needs it, as he needs, with a minimum of muss, fuss, bother, and bureaucratic nonsense. The relevance recall ratio should be excellent, otherwise he is likely to complain about irrelevant material. And, if he is forced to go to many places to obtain the coverage he demands, he will voice a strong objection.

The IACs I'm familiar with suffer little from access problems. This is due to the intimate and continued participation of the user in the collection, storage and retrieval, and using phases. However, when the IAC information operation utilizes more general collections—i.e., relatively broad information stores—access can be a severe problem. As seen through our IAC eyes, the access problem of the general collection is caused by the idiosyncracies of indexing; specifically, an indexing operation designed for other than our specific areas of technology.

The major access problems faced by the types of IACs we operate stem from two conditions—that access problem caused by proprietary interests, and that access problem caused by national defense interests. Both of these interests are understandable and justified. No ready solution to these problems has been offered.

End of Symposium

Amortization of Indexing (Input) Costs*

JULIUS FROME
U. S. Patent Office, Dept. of Commerce,
Washington, D. C. 20231

Received September 3, 1971

A formula is given that enables one to determine the parameters for amortizing an information system within a year. The formula is illustrated with three mechanized information systems for patents: (1) mold and mold coating composition class 106, subclass 38.2–38.9; (2) synthetic fibers class 264–210, 288, 289, 290; and (3) abrasives class 51, subclass 298–309. The most critical parameter is shown to be indexing time.

The cost of information systems resolves itself into several main factors:

1. System design
2. Machine costs
3. Indexing (input)
4. Retrieval (output)

System design as a one time proposition is a minimal cost in the over-all picture. Machine cost is also not of too great importance since in many instances computers and punch card equipment are already present for other purposes or if not the acquisition of these will lead to so many other uses that the cost of the machines would be a small part of the over-all cost of the system. Retrieval costs are minimal. The major cost of an information system is the indexing of the file, particularly in the time of the scientist or professional. Clearly, the quality of search by a well designed and indexed machine retrieval system is certainly good or better than a manual search.[1-3] This paper describes a technique of lowering the cost of indexing and possibly amortizing the main cost of indexing in one year.

THE PROBLEM

Most organizations operate on an annual budget. It would be desirable, therefore, if the cost of the information system could be amortized in the period of one year. That is, the benefit could accrue in the same period of time as the cost of the system (one year). As indicated supra, the main cost of an information system would be in the time used in indexing. Thus, if the time saved in using a system for search would be equal to or greater than the time spent in indexing, the cost could be amortized in one year. This would be very important for management to be able to plan effectively within the annual budget.

CRITERIA OF SECTION OF ART

To achieve our purpose, the art must have the following criteria:

1. The art must be sufficiently active
2. The manual search must take a fair amount of time
3. The important or commonly used terms must be able to be developed
4. The amount of time saved per search times the number of searches needed to be made in one year should be equal to or greater than time needed to index each patent times the number of patents—i.e.,

$$T \times N_s = I \times N_D$$

T = time saved per search (machine over manual)
N_s = number of searches per year
I = indexing time per document
N_D = number of documents in system

This formula can be used for example as follows:

Let us say there are 2000 documents in an art. The time saved per use of the mechanized system is about six hours. Let us say that there are about 50 searches in a year. Therefore in accordance with the formula:

*Presented at 3rd Northeast Regional Meeting, ACS, Buffalo, N. Y., Oct. 11, 1971.