Some Thoughts on Scientific Information Dissemination

Most of us have been aware for some years of a crisis in scientific information processing, caused by the large mass of available information products, and reflected by the present difficulties in providing the right information to the right persons at the right time. The existing facilities in our libraries and information centers are overtaxed by ever-increasing user populations, and the processing capabilities are choked by mountains of materials, including the conventional books and documents, as well as the newer types of information stored in data banks, or recorded on tapes, films, or cards. The result is a chronic inability on the part of most information processing centers in meeting current budget figures and in operating within the existing physical constraints, accompanied by a breakdown of the established operations for acquiring, indexing, classifying, storing, and retrieving the stored information items.

There exist, unfortunately, no attractive solutions to the library problem within the presently prevailing technical and administrative framework. In the end, even the largest libraries will be forced to abandon their attempts to remain comprehensive and complete in every subject area. Compacts and cooperative ventures will very likely be introduced based on selective and shared acquisitions procedures—each center acquiring only certain types of materials; common cataloging and indexing methodology may have to be implemented, so that catalogs and search tools may service users at many different locations; common storage facilities will eventually house pooled sub-collections from many centers; and cooperative search and information dissemination facilities will insure that users can obtain access to many different information resources no matter where located.

Obviously, such a restructuring of the existing information facilities cannot be contemplated unless standardized operations are agreed upon and some of the existing autonomy within each center is given up. In return for abandoning the dozens of different cataloging standards and the many incompatible information acquisition and circulation systems, it will be possible to introduce meaningful and financially viable mechanized operations, and the resulting increases in efficiency and processing capability will be reflected in an elimination of much duplicated processing, a streamlining of library operations, and a drastic reduction of the demands presently made on the professional personnel in charge of information handling.

It would be nice to assert with confidence that the changes outlined above are just around the corner. However, human nature being what it is, one may expect that ten, or perhaps twenty, more years may pass before any meaningful reorganization of information handling operations actually takes place.

One may then ask whether in the interim all of us—as professional people, readers, users, authors, editors, referees, or managers of information processing activities—might not help ourselves in coping with the present information crisis, even while awaiting the important changes in processing methodology previously outlined. I suggest that we might, and I would propose a four-point program which could be implemented at once, and would go a long way in helping us withstand the on-

slaught:

- (a) let us concentrate on that material which is really worth studying and preserving;
- (b) let us de-emphasize the alleged requirement for speed in the publication and dissemination of information;
- (c) let us abandon the many information dissemination gimmicks which supposedly help us cope with the information explosion;
- (d) let us make a discriminating use of computers in the information processing field.

First, we might concentrate on those information items which are essential to our professional endeavors, and abandon the publication—or at least ignore the existence—of the thousands of marginal items which overwhelm us. It is an astonishing fact that the existing information dissemination facilities have become less, instead of more, selective as the number of circulating items has increased. In some cases, the coverage has been extended to thousands of journals and pamphlets in a given technical area, and it becomes impossible for the average user to isolate the few worthy items from mountains of trash. It may be unrealistic to suggest that we abandon the publication of marginal items—including, for example, most conference proceedings—and rely for our information on the 2 or 3 or 4 major primary journals in each area, together with appropriate review publications, but it is a fact that everyone would be far better off. If the energies now devoted to the introduction of questionable publication schemes were directed instead to the strengthening of editorial and refereeing procedures of the main primary literature, a principal part of the information crisis would automatically disappear. (One may add that a drastic reduction in the number of new publications may become mandatory whenever the libraries decide to curtail standing orders for all newly published materials.)

A reliance on the primary literature unfortunately raises the spectre of extensive publication delays, and resulting potential losses in research and development activities. I would suggest that the “speed of publication” argument is overworked and, indeed, counterproductive under current conditions. There is no point in rushing to buy information when almost everything that one might acquire is not worth having. It is regrettable that the reviewing and editorial processes require an extra year—after the many years already spent in preparing and finalizing a given research or development topic—but it is essential for authors and readers alike that a minimal stamp of approval be obtained before the resulting reports are circulated to the public. And let us not forget that other impediments—for example, secrecy—impose far greater restrictions on the publication process than lack of speed alone.

In the hope of alleviating the information problem, some organizations have implemented new forms of information dissemination methods, including extra fast alerting services and preprint exchange programs. These services provide weekly, or monthly, printouts for the user population on a continuing basis, including, for example, papers recently issued by a given author, published in a specified journal, citing a given bibliographic reference, or containing a given word in their titles. Such services may be extended also to preprints and to informal notes before publication has actually taken place.

While such alerting services often deliver items which properly match a user's interest “profile,” they help in disseminating uncontrolled material—particularly

when used in conjunction with "early publication," or preprint, distribution systems—and the material so distributed often reaches the customer at the wrong time in the wrong context. In view of the many demands already made on the user's time and energies, it appears preferable to emphasize retrieval endeavors undertaken in response to specific user queries—so called "on-demand" searches—instead of attempting to distribute materials on a continuing basis in answer to standing orders.

Concerning, finally, the intelligent use of computers in information handling, it appears that many of the most interesting and rewarding applications have not been considered in practice. It is possible at the present time to eliminate manual indexing and cataloging of library material, with all their inherent inconsistencies, and replace them with automatic procedures based on text analysis. Documents can be classified automatically into affinity groups, and information searches can be conducted in such a way that only the most likely document groups are actually searched.

Whereas in standard library practice, user queries and document classifications are considered sacred—almost never is a subject heading altered after it has been assigned—it is possible in an automatic environment to change query formulations as well as document identifications during the search process. Thus, items which are jointly identified as relevant to certain user queries may be brought more closely together, so as to be more easily retrievable in answer to subsequent queries in the same area; contrariwise, items that are never wanted may be slowly moved away from the active search areas, and eventually removed entirely. We should neither attempt to store or circulate every existing item, nor should we keep every item once acquired.

To summarize, we do not need bigger and better computers in the information field, capable of circulating more and more unwanted material. Rather, we need a change of attitude on the part of the people in charge of information activities and of the budget officers responsible for financial support. Procedures must be standardized and resources shared to a much greater extent than at present, and the questionable emphasis on speed and volume must be abandoned. When the budget officers become convinced that many of the expensive endeavors and grandiose national schemes are in fact useless and unhelpful to the intended user population, we may be able to begin again with smaller budgets, and the prevailing brute force approach may be replaced by a more discriminating and sophisticated use of the tools at our disposal.

—G. Salton