The bases for the aggressive U.S. approach to documentation [redacted]

ON PROCESSING INTELLIGENCE INFORMATION

Paul A. Borel

The cycle of organizational activity for intelligence purposes extends from the collection of selected information to its direct use in reports prepared for policy makers. Between these beginning and end activities there lie a number of functions which can be grouped under the term *information processing*. These functions include the identification, recording, organization, storage, recall, conversion into more useful forms, synthesis and dissemination of the intellectual content of the information collected. The ever-mounting volume of information produced and promptly wanted and the high cost of performing these manifold operations are forcing a critical review of current practices in the processing field.

Storing and Retrieving Information

Efficient and economical storage and retrieval of information is by all odds the toughest of the processing problems. Millions are being spent on it by the research libraries of universities, of industry, and of government. Even as we meet here today, an international conference is under way in Washington at which new means of storing and searching for scientific information are being discussed.

For intelligence, storing and retrieving information is a particularly vexing problem. Our Document Division alone processes daily an average of some 1,500 different intelligence documents, received in an average of 15 copies per document. This is exclusive of special source materials, cables, newspapers, press summaries, periodicals, books, and maps. Since these reports come from scores of different major sources, the daily volume fluctuates and shows lack of uniformity in format, in reproduction media, in length and quality of presentation, and in security classification. As they come in they must be read

with an eye to identifying material of interest to some 150 different customer offices or individuals.

We have a general library of books and periodicals, whose operations approximate those of the conventional library. We have several registers (in effect special libraries) through which we handle special source materials, biographic data on scientists and technicians, films and ground photographs, and data on industrial installations. Most of these materials are subject to control through indexes of IBM punched cards.

We have a collection of two million intelligence reports miniaturized by microphotography. Short strips of film are mounted in apertures on IBM punched cards filed in numerical sequence. Access to these cards, from which photo reproductions can be made, is obtained through an organized index of IBM cards now numbering eight million. Thus access to the document itself is indirect, through codes punched into the index cards to indicate subject, area, source, classification, date and number of each document. The data on index cards retrieved in response to a particular request is reproduced on facsimile tape and constitutes the bibliography given the customer. This system—which seeks to fit a given request with the relevant "intelligence facts" on hand—we call the Intellofax system.

These then are our assets. I'll say no more at this time about problems in connection with the general library, or those of operating our registers, since they are in many respects variations on the theme of our concern with the effective operation of the Intellofax system.

Demands made on our document collection stem from three types of requests:

Requests for a specific document to which the analyst has a reference or citation;

Requests for a specific bit of information in answer to a specific question;

Requests for *all* information relevant to a subject which may or may not be well defined.

Our major difficulties are almost all connected with the last of these three, the one which requires a literature search. In searching unclassified literature we rely on commercially produced reference aids, but in searching classified materials we

use the Intellofax punched card index. This index we would use to retrieve, for example, information responsive to a request for "anything you have on the movement of iron ore from Hainan to Japan between 1955 and 1958, classified through *Secret*, and exclusive of CIA source material."

Intellofax is a high-cost operation. Only 10 to 15 per cent of the questions put to the information section of our Library are answered by literature search; yet some 30 people are used in the necessary coding, and another 50 to 60 in IBM and auxiliary operations exclusively in support of Intellofax. On the other hand, some portion of this cost would be incurred in operating any alternative system even at minimum level; and Intellofax makes possible the organization of bibliographic material in various forms and at speeds which would not be practical under a manual system.

Search results, however, are not uniformly accurate. We recently tested the accuracy of the Intellofax system by having a task team of three analysts from a research office conduct a controlled experiment. Five subjects, corresponding to common types of reports produced by that office, were selected. The test indicated quite conclusively that the system does an efficient job of retrieving documents referring to specific objects or categories (trucks, factories, serial numbers), but that it is less satisfactory in handling a more general subject, such as industrial investments in China. A comparison with the analysts' own files showed very satisfactory Intellofax performance in retrieving documents placed in the system, but some documents in the analysts' files were not retrieved. Reruns with the same code patterns yielded consistent results.

The inaccuracies of the Intellofax system reflected in the above and other tests can be reduced by revising procedures and improving supervision, but they cannot be eliminated altogether. In literature search a set of symbols assigned to incoming documents is used to provide the searcher with a clue to the pertinence of any document to the request he is servicing. This set of symbols is in the nature of an index, but different people viewing these symbols may give them different interpretations. This makes the problem complex, for the determination that there exists a meaningful relation between even two pieces of information depends on many differ-

ent, often subtle criteria which elude unequivocal symbolic representation.

The solution of the accuracy problem would appear to turn on the ability to develop a master set of symbols, a Code, large enough to cover an extremely wide variety of subjects and areas and small enough to be contained on an index card, one applicable to diverse documents containing fragmentary, fugitive and often seemingly unrelated information, and at the same time conducive to uniform application initially by those coding incoming documents and later by those seeking to retrieve them. To prepare such a Code is a tough assignment today. The job is not likely to be easier for some time.

It is relevant at this point to invite your attention to the views on this subject of the Working Party organized last year [redacted] to examine the possibility of establishing a common reference service: [redacted] *: books of reference and finalized intelligence reports. It would be impracticable to try and include the welter of documents from which such finished reports are built up; even if it were practicable, it would be an immense task beyond our resources.*[1]

I disagree. Not as to the difficulty of the task or its relatively high cost, but as to its impracticability. I believe the solution lies in a) selectivity in identifying those documents to be held by the Center, and b) the organization of those documents into discrete collections, each controlled by an index suitable to its particular requirements. This is the approach we have taken, more by accident than by design. Such an approach makes it possible to cope with small problems, even though the big problem may still be unmanageable.

Reference Service and the Research Function

Where central reference services have been organized independent of research offices, it soon becomes evident that the functional line of demarcation between them and the research units is not clear. This becomes important when it results in

[1] [redacted] *Modern Methods of Handling Information*, 15 Oct. '57 (Confidential), para. 6.

duplication of effort or, worse, in non-use of reference materials by the researcher laboring under the misimpression that he has all relevant documents in his possession. Today's researcher, like his predecessor, feels insecure without files which he can call his own. In such a situation we must have a proper regard for tradition, but sometimes it is difficult to distinguish tradition from inertia. Recently our Biographic Register, receiving a report published by a research office, found that failure on the part of the author to check the Register files had resulted in some one hundred errors or omissions.

It must be decided whether a reference service is to be active or passive, dynamic or static. To take a simple case, a passive approach to reference service would mean that reference personnel would merely keep the stacks of the library in order, leaving it to research analysts to exploit the collection. Under the active approach, on the other hand, reference analysts would discuss the researcher's problem with him and then proceed, as appropriate, to prepare a bibliography, gather apparently pertinent documents, screen them, check with colleagues in other departments for supplementary materials, make abstracts, have retention copies made of popular items in short supply, initiate a requirement for supplementary field service, or prepare reference aids. In CIA we aim at active rather than passive reference service. How active we are in a particular case is a function of the customer's knowledge of our services, his confidence in us, and how pressed he is to get the job done.

Once a separate facility has been set up to provide reference services it is not long before it publishes. This comes about for several reasons, the least controversial of which is that a customer has made a specific request. Thus our science analysts may call for a compilation of biographic data on the individuals most likely to represent the Soviet Union at a forthcoming international conference on the peaceful uses of atomic energy. We call this type of publication a research or reference aid. Some are quite specific; others are more general, being prepared in response to a need generally expressed. A number of different customers may, for example, make known that it would be very helpful to have a periodic compilation of all finished intelligence reports and estimates for ready refer-

ence. Or the need may be implied rather than expressed: the reference analyst may note that over a period of time the demand on him for biographic data about Soviet scientists is heavy, many requests calling for much the same information furnished earlier to others. The result: the production of a major reference aid along the lines of our "Soviet Men of Science." And naturally it isn't long until a revised edition is called for.

Criteria for determining when and when not to summarize information holdings in a general reference aid are elusive. It is similarly difficult to define the proper scope of the general reference aid. How far can it go before the researcher considers it an infringement on the research activity for which he is responsible? This question has implications beyond those readily apparent. Quite basic is the feeling among research personnel that they and their mission are a cut above the reference officer and his role. A manifestation of this attitude is the steady flow of competent people out of reference into research, with only a trickle coming the other way. I doubt whether the inconsistency of this position is appreciated in view of the joint effort required by research and reference activities to provide the soundest base possible for the research effort.

In my view the legitimate limits of the reference aid can best be arrived at in terms of the highest level of service expected of the reference officer. Stated simply it is this: to make known the availability of services and information the existence of which may be unknown to the researcher; and, given a task, to make the preliminary selection of materials to meet the particular need of a particular user. This may involve bulk-reduction operations (such as abstracting) to leave a smaller quantity of material containing everything pertinent to the user's problem, or conversion operations (such as translation) to get information in usable form. I would even say that the reference function includes evaluation, evaluation of the reliability of information. To the researcher must be left the determination of its significance for the present; to the estimator its significance for the future; and to the policy-maker the indicated course of action.

Machine Application to Documentation Problems

In processing intelligence information, increases in efficiency may depend upon the adoption of techniques involving automata. This is especially the case when savings of time are sought. But as soon as you consider automation, that is, the inclusion in your processing system of a machine as an integral part of it, you are faced with the need to make decisions different in nature from those made with respect to the desirability of expanding staff or restricting functions. It is a difficult problem to achieve an optimum balance between man and machine. Among the many considerations involved there are two important ones which ought to be, but seldom are, fully explored before you commit yourself to a particular machine—you should accurately determine the *net* gain or loss in terms of time, space, manpower, and money; and you should be fully aware of the limitations of the machine and of its use by man. It is often more important to know what cannot be done with the machine than to look wholly to what can.

Nevertheless, I would again incline to disagree [redacted]

In view of the great initial investment needed to launch [a mechanized reference system], the very large and persistent requirement for coding, maintenance and other supervisory skill and the inevitable limitations of machinery when applied to intelligence processes, we do not think the introduction of such a system merits further examination.

No one would argue that large investments should be made in schemes unless they hold promise of relieving major problems. And the demands of a mechanized reference system for special skills are admittedly both high and persistent. However, these factors should be weighed in terms of the relative costs, not only the cost of alternative ways to solve the particular documentation problem, but also the cost of not solving it at all. We take exception to the conclusion that the limitations of machinery when applied to intelligence processes are "inevitable." We also believe it unwise to categorically dismiss the introduction of machinery as not meriting

further examination. Limitations there are today and will continue to be. But those which are inevitable are fewer than is generally supposed. Only by daring and risking will we come to know how few are the real limitations of a mechanized approach to documentation. This philosophy is yielding promising developments in the fields of microphotographic storage, automatic dissemination, abstracting, and translation, all fields of particular concern today.

Microphotography. Both Air Intelligence and CIA are testing a system developed by Eastman Kodak known as Minicard. This system in essence substitutes a 16 x 32 mm film strip for the present CIA system of IBM punched index cards corresponding to hard copy or film in the document storage file. Self-indexing Minicard document images are read electronically, not mechanically as IBM cards are. The characteristics of Minicard make possible a reduction of space requirements by a factor of 4, and an increase in speed of handling by a factor of 2. The new system is capable of a level of information manipulation and a degree of coding sophistication which gives promise of radically augmenting the contribution of the information fragment to the solution of reference problems requiring a search of the literature. And, contrary to present practice, the integrity of the file is maintained at all times.

Automatic Dissemination. Air Intelligence is testing a Document Data Processing Set designed by Magnavox. This is a general-purpose computer especially designed for problems requiring close correlation. Requests for information form the reference file against which incoming documents must be compared. Up to 20,000 words specifying the subjects and areas of interest, other qualifying data (such as evaluation or type of copy desired), and user identifications are stored to define the requirements of 160 users. When a document is to be disseminated, its subject and area coverage, previously coded and punched into paper tape, is fed into the machine. The machine searches its file of requirements and prints out a list of those who have requested such a document, the total number of copies needed, and the form in which it is wanted. Speed and uniformity of performance rather than financial economy is what the Air Force is after in this case.

Automatic Abstracting. Army intelligence and IBM are working on means for producing, entirely by automatic means, excerpts of Army field reports that will serve the purposes of conventional abstracts. At a recent demonstration the complete text of a report, in machine-readable form, was scanned by an IBM 704 data-processing machine and analyzed in accordance with a standard program. Statistical information derived from word frequency and distribution was used by the machine to compute a relative measure of significance, first for individual words and then for sentences. Sentences scoring highest in significance were extracted and printed out to become the "auto-abstract." Adoption of this method of producing abstracts of overseas reporting would require the use of a flexowriter in the field. When the original report is typed on stencil, a flexowriter tape would be produced simultaneously as a byproduct and would accompany the report to headquarters. There tapes in sequence would be fed into a computer and auto-abstracts printed out.

Mechanical Translation. The only successful Free World demonstration of machine translation to date took place on 20 August 1958, when a continuous passage of 300 sentences taken from Russian chemical literature was translated by the Georgetown University research group, under CIA and National Science Foundation sponsorship. An IBM 704 computer was programmed with the appropriate grammatical, syntagmatic and syntactic rules, and a Russian-English vocabulary was introduced into its memory system. The machine alphabetized the text, determined the lexical equivalents of the words, reconstructed the text, performed the necessary logical operations, and printed out the English translation. Only minor stylistic editing was required to make the product compare favorably with a translation made by a linguist. The rate of translation was about 24,000 words per hour. With improved input equipment (reading machines), rates up to 100,000 per hour are foreseen as possible. Research has already started on mechanical translation from Polish, Czech, Serbo-Croatian, French, Arabic, and Chinese. Soviet research in this field is considerably ahead of ours.

Outlook

In closing this general review of aspects of the intelligence documentation problem, we should look briefly at certain trends which affect us all. First, channels for procuring publications and techniques for storing and retrieving the physical document are extensive and well developed. The immediate outlook is for no basic change in ways and means in this field, but rather an expansion and intensification of present methods.

Second, the type of reference or information service coming to be required will demand action primarily in preparing reference personnel to give assistance of higher quality than is given today. Reference tools will need to be improved also, but this is likely to follow if there is a more sophisticated reference officer to create a demonstrable need for them. The increase in amount and kinds of material available will call for more intense exploitation of it by the research analyst; he in turn will by necessity rely increasingly on the reference officer for first-cut selection and evaluation. Reference officers will therefore need greater subject competence, more language ability, and a wider training and experience in all aspects of intelligence documentation. Already a number of American corporations are using information specialists as members of research teams. This approach deserves testing in intelligence.

Third, in the field of literature searching, specialized schemes will be developed to fit the needs of specialized users. While general theory will continue to be developed, pragmatic approaches to problems based on an analysis of the way users employ services and exploit materials will play an increasingly important role. Proved systems employed by reference centers will be simplified and adapted for use by the individual analyst to enable him to control the literature he requires in his immediate possession. The analyst in turn will provide the central system with the means of subject retrieval in his specialized field as a by-product of the way he controls his files. In this field, machines will long continue to play a secondary role.

Fourth, the present and future demands for reference service will lead to increased use of machines where these can be

introduced without jeopardizing the performance of essential intellectual operations. This fact and the increasing volume of information which must be processed will bring about more centralization. The problem then becomes one of insuring that central reference is at least as responsive to research needs as the reference facility which is an integral part of the research area. The solution is to be found in an approach which integrates the information-processing activities, wherever performed, into a single system within which collection, processing, and user components operate along well-defined lines.