5 March 1958

SUBJECT: Comments on a User's View on OCR's Document Handling Capacity.

1. Country Files

The paper recommends the establishment of country files in support of analysts with political equilibria problems.

This approach dispenses with all of the information system elements - a code structure, coders, and search instruction writers.

Country files forgo all the advantages of multi-area document indexing. Near East problems, for instance, are generally multi-national. The user will probably need to go through a number of lengthy files and will be presented with new memory problems in making his correlations.

CIA documents cannot be filed by country since they are not so identified as to source. Filing of multi-copy to cover all areas or Blocs mentioned as subjects would appear impractical from many points of view.

Search of a country file involves the handling of high percentages of extraneous material.

This paper supports the maintenance of analyst's "gem" files. The criteria for such files appear purely subjective - could not be policed - would be no different in content or quality than they are today. "Gem" files are apt to be consulted by "seekers" in preference to library country files.

The writer completely underestimates the physical problem of searching country files of documents on microfilm in aperture cards.

2. Inconvenience of Intellofax

25X1

[redacted] expands the Library Consultants discussion of this subject with his judgment that the system does not help the analyst consolidate data bearing on complex estimating problems.

Consolidation must take place in time. Today's document adds unevaluated information which confirms, expands, contradicts or ignores yesterday's report on the given developing situation. No coder can retrieve and adjust his coding of yesterdays' document in the light of today's development without becoming an analyst engaged in substantive activities. Thus contradictions, confusions and error must appear inevitably in the search product and only the analyst - not the coder - will have the competence to recognize and dispose of these in his "weighting of opposing forces and inferring of strategies".

~~C-O-N-F-I-D-E-N-T-I-A-L~~

The paper does not examine the evidence of the "ten horrible cases" and 25X1 [redacted] response thereto. It does not examine the ten new cases for further evidence of important deficiences. What is the evidence that the system is not already reasonably efficient in the handling of "easily recognized logical categories" (par. 16)?

In the Annex - par. A6, the author states that document analysis is and should be primarily a matching problem. But, he says, a separate part of the code is required to deal with derived concepts, paraphrases, nuances and complicated relationships. What is his solution?

A solution is to be seen in the indexing policies of the New York Times. There highly trained indexers do deal with indexing through time, by writing what are in effect index summaries of complex stories which developed day by day through the period covered by the annual index consolidation.

The paper does not comment on the major revision of the ISC code now underway under community auspices. One of its prime objectives is to introduce new focus on general categories of observable data and to lead the coder (or searcher) from general to particular sub-categories with full utilization of qualifiers or "aspect" codes - cf par. A4. There appears to be general agreement among all interested parties concerning the objectives of the ISC revision.

25X1 [redacted] raises the problem of guaranteeing homogeneity of coding. There is no argument on the point that a logical, precisely defined code will promote logical coding, however, the best open-library classification system in existence offers no guarantee of homogeneity of interpretation (by two catalogers processing the same book; or one cataloger processing similar books at different points in time). The standard library answer is to install a senior cataloger-reviser to inspect the daily performance of catalogers and to measure it against a shelf-list in which is recorded all previous cataloging. [redacted] 25X1 does not acknowledge that the Document Division has introduced the revision concept in its establishment of senior coder positions. The establishment and operation of the shelf-list for fragmentary raw information reports in very large volume is judged to be an economic impossibility as well as a dubious guarantee of the desired homogeneity in handling this data.

25X1 Can [redacted] be more precise in identifying the homogeneity problem? Attache information reports writers do not possess the journalistic competence of N.Y. Times correspondents. Their product shows it without getting into the complexities of the appreciation problems they face in observing foreign situations - many in subject fields in which they have no competence. (One might argue that the reports should be collected and written in the field by a committee of specialists as well as that a committee of specialists should code them (par. A9) at headquarters.) Does not the answer lie first of all with the imagination, ingenuity and persistence of the researcher, secondly with the highly trained reference librarian - both human resources, not machine or systems resources? If a "general" indexing system such as Intellofax turns up large populations of documents as its search result does the analyst not have the moral obligation to examine each and every item and the library the obligation to tell him as much as possible about each document - by expanded title, or abstract - to help him in deciding where to start. (A 300 item answer out of a total population of 1,000,000 documents is already an enormous economy in search time! If the first search is fruitless, surely there is an obligation to broaden the search to marginal codes! If time does not permit, analysts files are obviously the main line of defense. The time limitation is present in <u>every</u> search and "where to stop" is always partly an economic decision since the answer might be found by application of <u>more</u> searchers to outside libraries, files of unindexed open literature, or collection effort in the field.)

3. <u>Role of the search instruction writer</u>

25X1A9a In his recommendation that Intellofax be retained on probation and that Minicard be pushed, [redacted] increases the importance of search instruction writing. The point of the Library consultants was not that the Agency should go to a manual system because of the lack of a simple unifying logic, or of an adequate cross reference policy, or of high level subject analysis. They argued that we must go to an open system <u>to obtain adequate results</u> on these three counts. The point is that a printed bibliographic service for documents and an open card-catalog virtually eliminate the role of the search instruction writer as in public library catalogs. The analyst obtains thereby the maximum of freedom in making his own correlations of information recorded in the system. The dilemma not fully appreciated by the Library consultants is the scale of the problem. No one solution is convincing. A combination of approaches seems most promising.

4. <u>The economics of use of indexing resources</u>

The paper suggests, par. 21, that <u>demand</u> for the given CR service is not sufficient justification for the provision or denial of the given activity. Yet there <u>is</u> a problem of economics in CR. The spectrum of existing service represents allocation of scarce means to alternative ends. There are innumerable intelligence research projects and services which the Office cannot afford to undertake, particularly in a period when it is undergoing reduction in staff.

Does not the inauguration of CRAG provide a proper and sufficient device for the disposition of proposals for marginal services?

5. Minicard

25X1A9a

[redacted] recommendations on Minicard are best characterized as an act of faith. Before the most extensive testing has taken place there is no basis for identifying its true capabilities in solving a variety of information retrieval problems. The present CR system of punched cards and film in aperture cards is undoubtedly superior to Minicard in some respects.

Minicard as a machine system must be "debugged". This has already consumed a very substantial amount of time.

Minicard at the present stage dispenses with the bibliographic step in information searching. There will be no list of citations for initial screening by the analyst. Initial retrieval of entire documents for screening purposes poses many problems - time, cost and analyst orientation.

Some of the potential of Minicard appears to lie in its capability of storing much larger index sets for documents. The scale of indexing operations necessary to generate these index entries is unknown.

The capability of Minicard in meeting the Library Survey charge of blindness (directed repeatedly towards Intellofax) is undetermined. In principle, Minicard goes far beyond Intellofax in separating the analyst from the traditional open card catalog or printed bibliography. There is a distinct possibility that the complexities of Minicard operation will disqualify it for very current operations.

6. Conclusions

The preceding paragraphs discuss differences of emphasis or, in a few cases, fact. Overall there is general agreement on objectives and on aspects of the system where efforts to improve should first be concentrated.

There appears to be general agreement on the objectives to be pursued in revising the subject classification scheme.

There is complete agreement on the desirability of improving the competence of the personnel operating the system.

As to priorities, this comment places greater emphasis on the benefits that might derive from (1) enlarging the role of reference librarians in the research community and (2) in so explaining the system that analysts will accept greater personal responsibility in coping with the problems of ordering masses of raw information. As next best steps, programs that might be undertaken to improve "system" or machines will cost more and yield less return than equivalent effort to improve human performance.

A statement by Bar-Hillel on this subject seems particularly pertinent:

> "...it should still be clear that in general no information retrieval system, and be it as detailed and encyclopedic as can be, will be able to replace reading the original literature, a procedure the value of which consists mostly in the stimulation it provides through its general line of argumentation and method rather than its wealth of factual material".
> A Logician's Reaction to Recent Theorizing on Information Search Systems, by Yehoshua Bar-Hillel, p. 105, American Documentation, 1957.

25X1A9a

[redacted]