NATIONAL SECURITY AGENCY

MECHANIZATION IN SUPPORT OF COMINT

PHASE 1

WARNING

THIS DOCUMENT CONTAINS CODEWORD MATERIAL

TOP SECRET CONTROL NUMBER 301514
COPY 11 OF 38 COPIES
PAGE ____ OF ____ PAGES

NATIONAL SECURITY AGENCY

Washington 25, D. C.

MECHANIZATION IN SUPPORT OF COMINT

Phase I

Compiled by

R/D Personnel

October 1954

TABLE OF CONTENTS

TABLE OF CONTENTS

INTRODUCTION

This study is an evaluation of the analytical machine and intercept equipment phases of the present Research and Development (NSA-30) program in light of the present and projected problems of Communications Intelligence. The conclusions drawn from this evaluation (should) result in a Research and Development program consistent with Comint mechanization and traffic requirements.

The study has four phases. The following paragraphs serve as an introduction to the first two phases only. The present objectives of phases III and IV are included, but may change considerably as phases I and II mature.

Phase I attempts to uncover the areas where mechanization is, or could be, of use to the Comint Effort. In phase I no attempt is made to evaluate previous or present efforts at mechanization in these areas.

Phase II considers only the areas uncovered in phase I, and attempts to give a quantitative estimate of the success of the present effort in each of these areas and thus reveal the shortcomings in the present R and D program.

Phase III will consider each of the shortcomings discussed in phase II, and will list possible procedures,

techniques, etc. which can be brought to bear on these areas.

Phase IV will analyze the relative merits of the procedures, techniques, etc. set forth in phase III and synthesize the results to formulate an R/D program which will satisfy COMINT requirements, be within technological limitations, and be consistent with agency and government policy.

The collection of the information contained in the first two phases of this study was an effort directed by R/D personnel. Key personnel from Production have willingly and ably assisted in the actual writing of this report and in offering pertinent information thru discussions and conferences where and when needed. While PROD personnel assisted in the formation of this report, the interpretations contained herein are those formulated by R/D personnel. It is R/D's hope that these interpretations conform with those of the other Offices.

That there are many errors of omission and commission in this report should be "a priori" knowledge. The R/D Office will be most grateful to persons bringing these errors to its attention so that a more accurate and more complete picture may be obtained.

I. COMINT Collection Activities

The COMINT Collection Problem divides into two phases: the interception of the various signals discussed below, and their subsequent transmission to NSA Headquarters and distribution to the interested analysts.

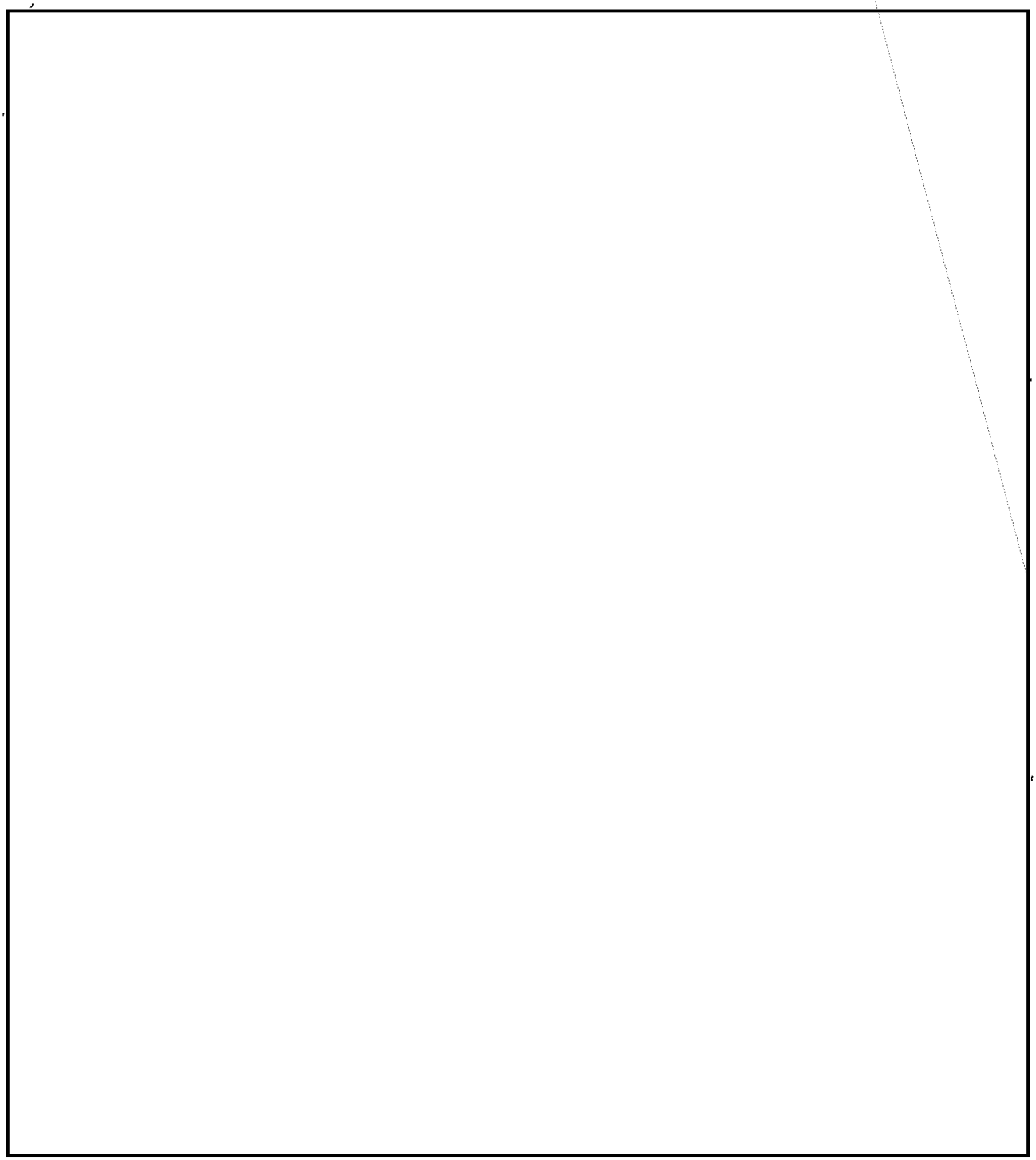
Sometimes analysts require information about a signal over and above that provided by a complete transcription thereof. For example, it may be required that an indication of the quality of interception of enciphered traffic be supplied with the traffic on a character-by-character or a signal element-by-signal element basis. Another requirement is that of providing the analyst a record of time intervals between successively transmitted characters. Still another requirement is the inclusion of information which will serve as a means of providing transmitter identification and in some cases the identification of individual operators.

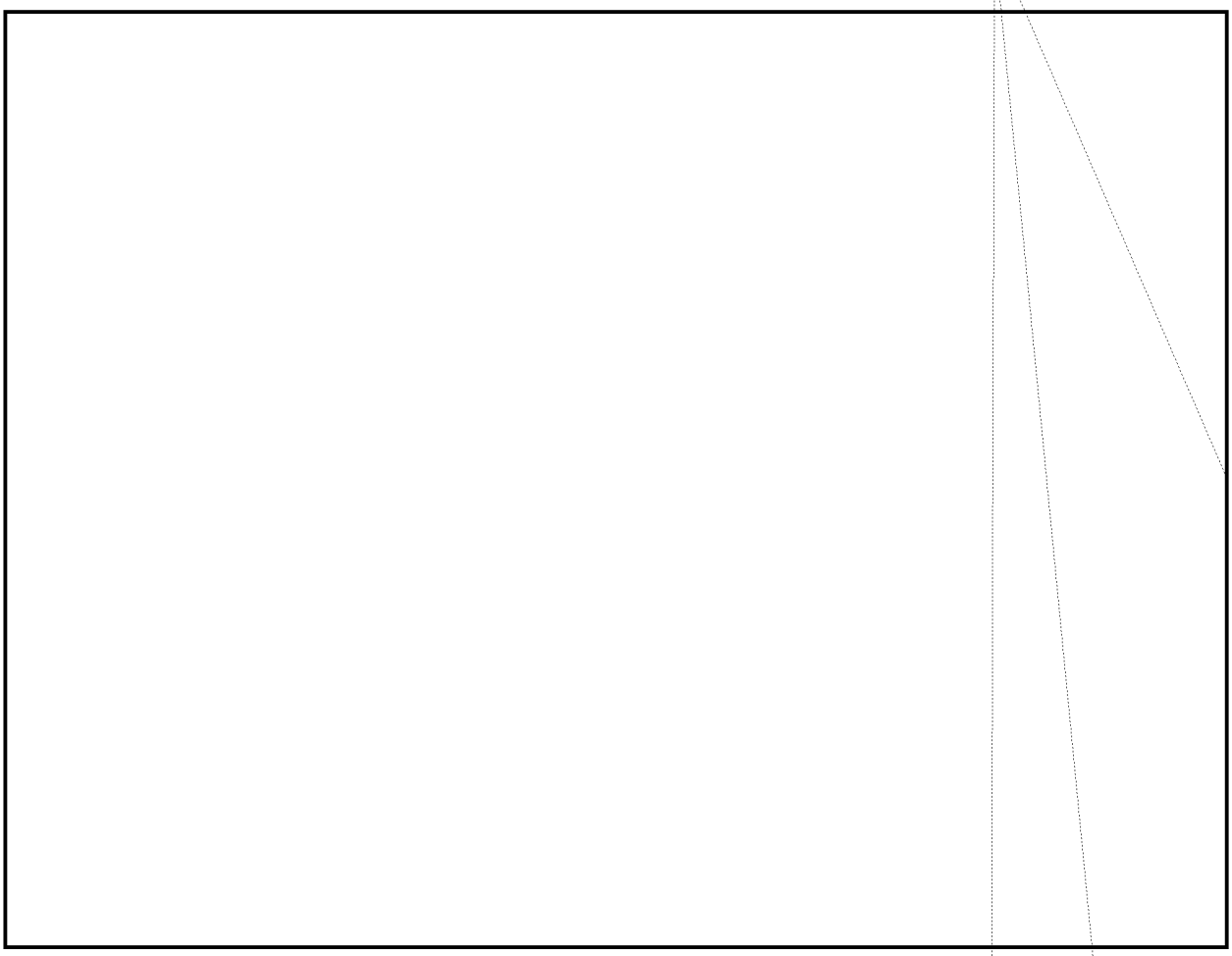
These requirements apply generally to several of the intercept situations.

(A) Functional Categories of Intercept Problems

The various transmissions of interest to NSA may be broken down into functional categories. These functional categories provide certain degrees of information about the characteristics of the transmissions and thus some

information regarding the intercept problem. Internal networks often create substantial problems in obtaining good intercept sites.





3. Government Communications Services

This classification is used here for internal networks of various government organizations such as border guards, secret police, police, agent transmissions, and the like. These nets are generally widespread, have low traffic densities, and [redacted] irregular schedules. [redacted]

4. Military Tactical

Military Tactical traffic (sometimes referred to as low-level traffic) is usually intercepted and often processed, to some extent, by the supporting service COMINT organizations. This material is also forwarded to NSA. This traffic is usually low powered, and transmitted by simple communication systems with irregular schedules.

5. Military Strategic

This category consists of high-level military communications up to the very highest echelon. High traffic densities and regular schedules characterize these transmissions.

6. Support Communications

This category comprises those communication services used to support various types of operations. These may be either broadcast or point-to-point transmissions. Weather nets and broadcasts, and navigational aids and services are examples of support communications.

These transmissions appear on regular schedules.

[redacted]

(B) Major Technical Intercept Problems

1. [redacted]

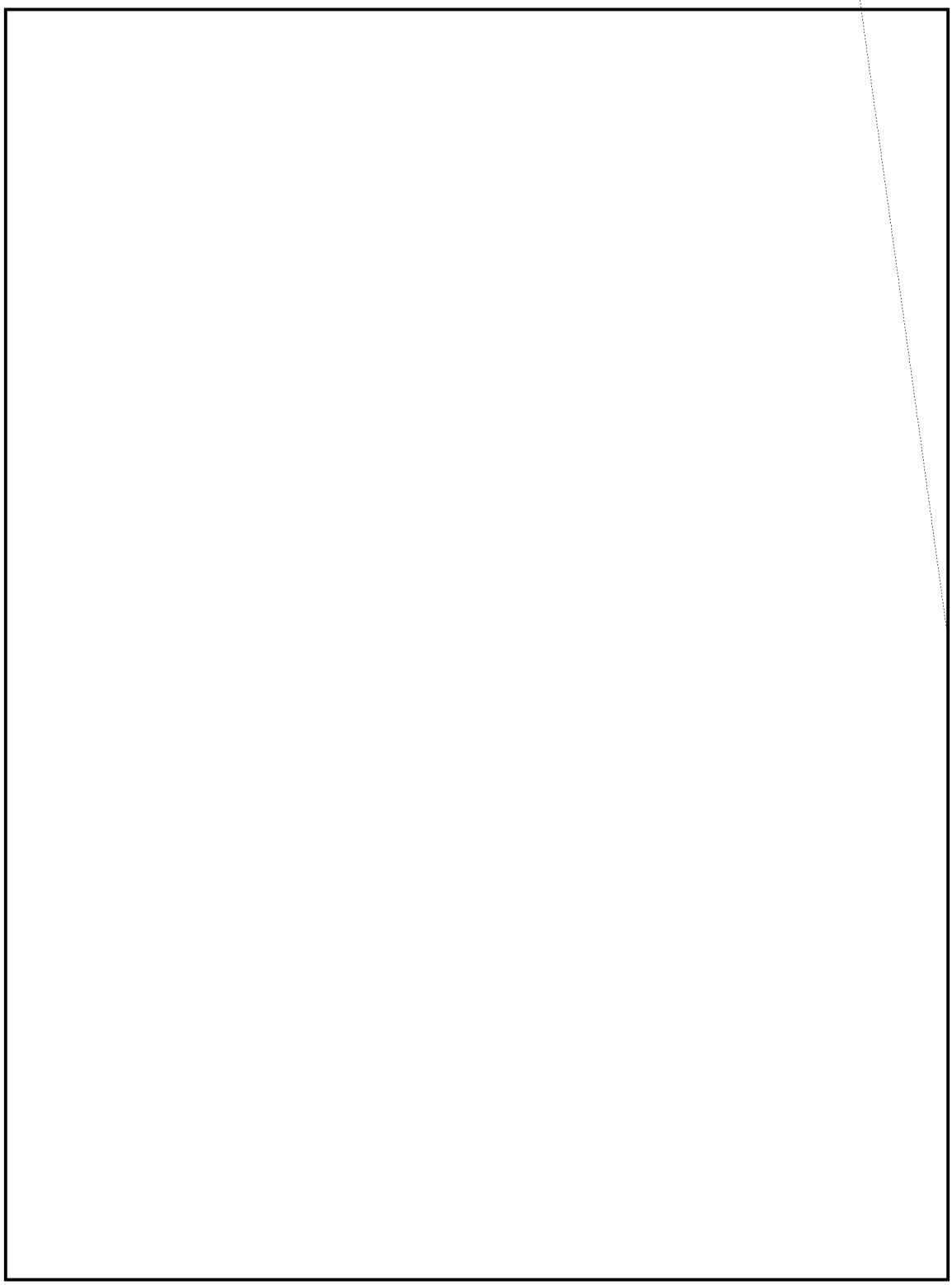
Early in 1952, a [redacted]

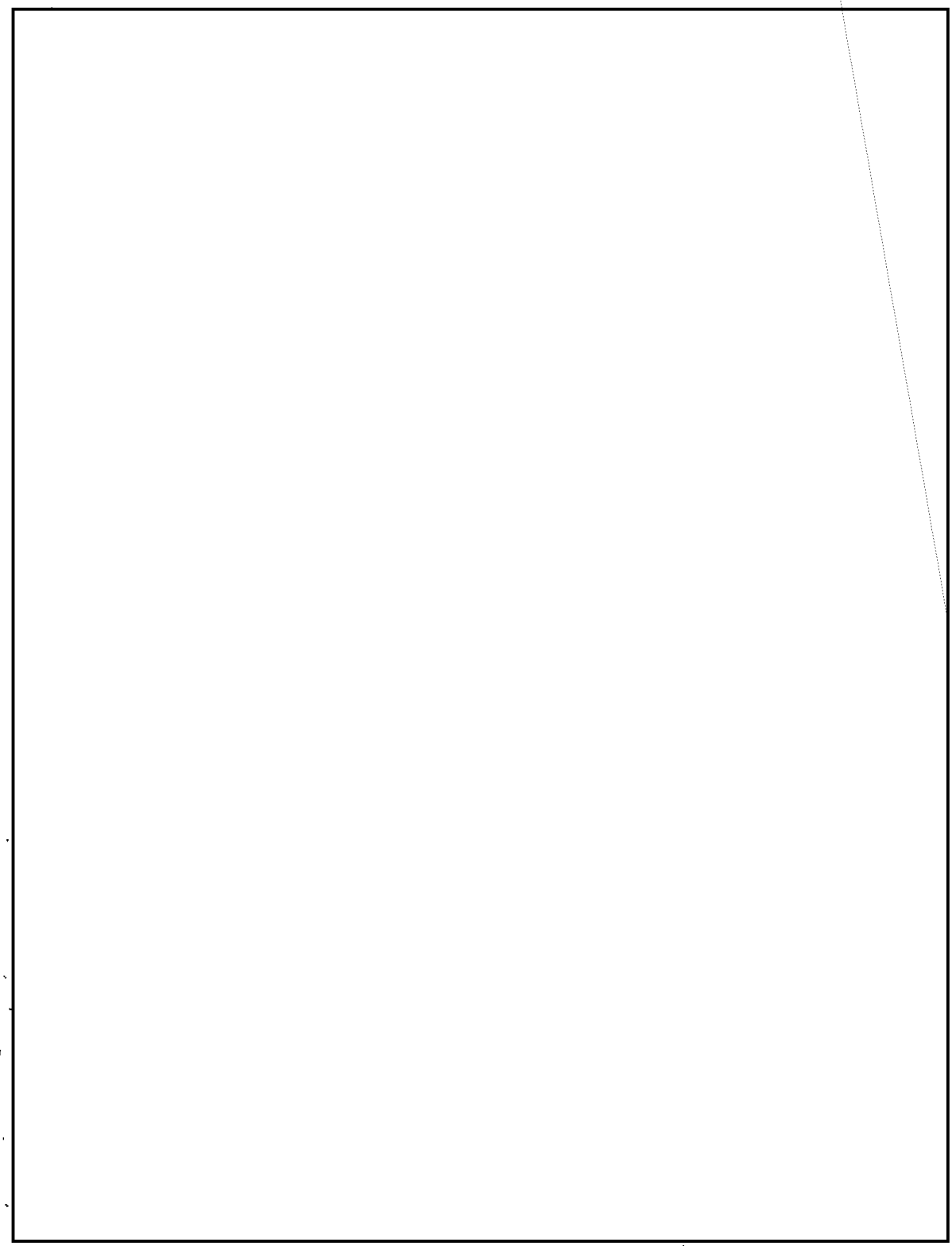
a. Naval Single Channel

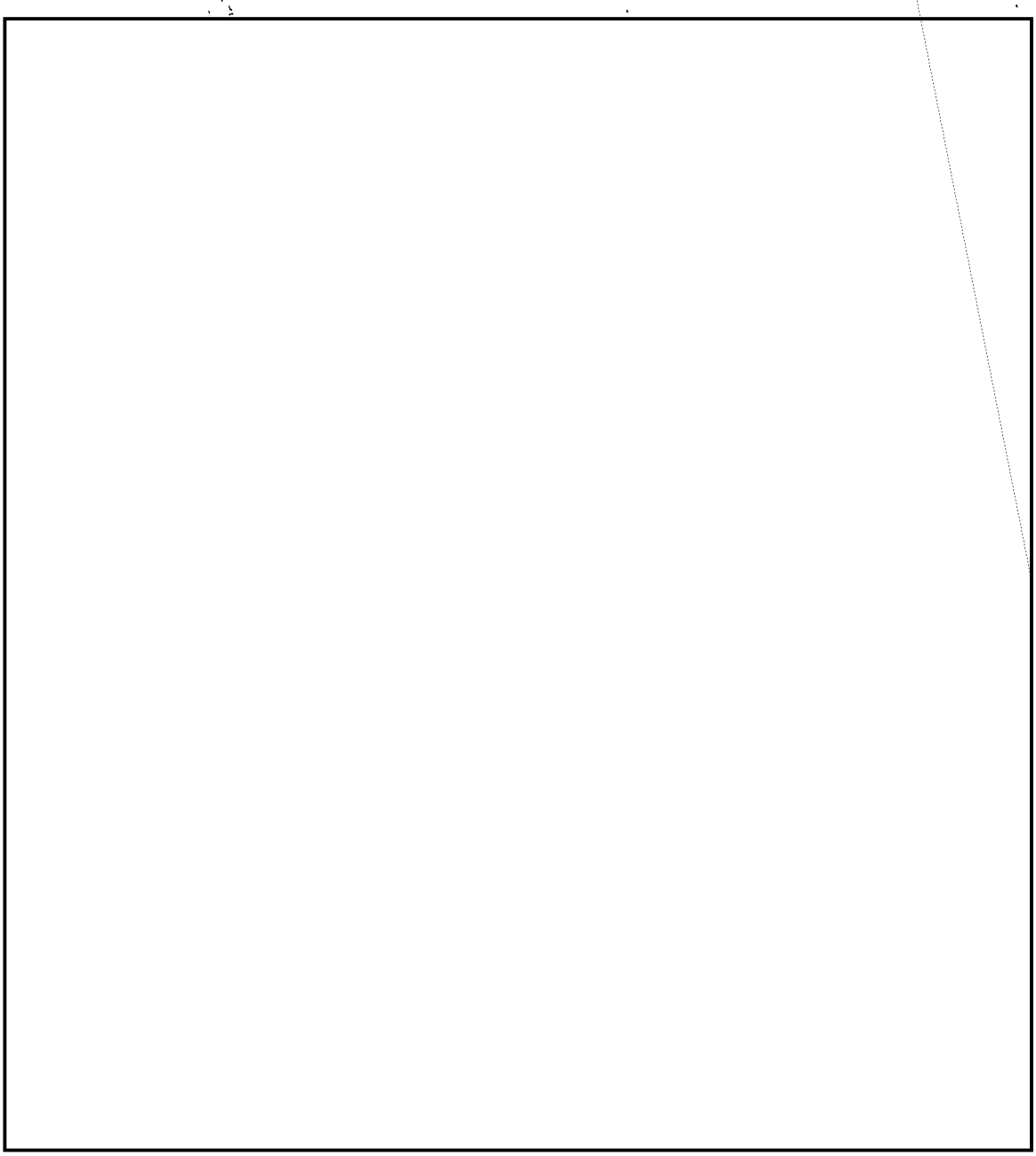
The first problem is to [redacted] This knowledge is required by the analyst in his analysis of the traffic.

b. [redacted]

Here the problem is to achieve good quality intercept. [redacted]







II. Operations on Traffic Independent of System

(A) Preliminary Processing

Many operations must be performed on traffic

arriving at NSA prior to its distribution to the users. These operations fall under the general headings of (1) Logging and Editing, (2) Data Conversion and (3) Data Storage and Recovery. The operations performed during these preliminary processing stages are independent of the crypto-system (if any) employed on the traffic by the sender prior to transmission. These operations performed on the intercepted traffic and the final forms and numbers of copies required are governed by the needs and desires of the analysts, machines, etc., which must further process and analyze the traffic. The problems in this area which appear to be amenable to mechanization are as follows:

1. Logging and Editing Problems

Generally these problems encompass the optimization and mechanization of the logging and editing functions which involve the logging, processing, division of page copy into identifiable systems, and the deletion of uninformative material.

Specifically the problems involve the keeping of a record of all traffic and the sections of analysts which are responsible, the sorting of messages in an order pre-determined by a format, and recording the messages in a log book by hand or by some machine

process. Duplicate messages are noted at this point and messages are given worksheet numbers.

The editing problems include the deletion of uninformative material, reordering of information, discrimination between textual and non-textual groups, correction of group length and run-together groups, indentification of duplicate messages, etc.

2. Data Conversion Problems

Traffic arrives at NSA in as many as four different forms: hard copy, perforated paper tape, magnetic tape, and occasionally, punched cards. Multiple copies or copies from which unwanted material is deleted may be required in more than one of these forms. Because of this the general problems of this area center about devising procedures and effecting mechanization for converting the original traffic into the various desired forms.

The specific problems involve the preparation of hard copy where it does not exist, and the preparation of edited versions of these data in punched cards, perforated paper tape, and magnetic tape as required as inputs to the various analytic machines.

3. Data Storage and Recovery Problem

The general problem here involves the

selection of permanent media on which to reproduce all had copy received, the storing of information of use to the intelligence analyst, and the maintenance of files containing all copy handled.

Particular problems are: the reproduction on a permanent medium, such as microfilm, and the numbering in the order of arrival of all had copy traffic; the storage and filing of collateral information of use to the intelligence analysts; keeping the files current by continual addition of both new subjects and new information under old subjects; and the immediate recovery of material from storage or files as required.

(B) Traffic Analysis

Traffic Analysis (T/A) is that area of NSA primarily concerned with the problems of obtaining communications intelligence from non-textual information. A problem of this area is the reconstruction of communication networks which carry communication intelligence of interest. A second problem involves the attempt to determine the "order of battle" of the communicator, i.e., the location and strength of the communicator's armed forces and strategic material and the identification and location of key personnel.

The specific problems are those of performing the sorting and indexing processes and other clerical operations of traffic analysis. Large volumes of

traffic must be scanned to uncover frequencies, call signs, message serial numbers, transmittal dates and times, page and pad numbers, addresses, and operator mistakes. After studies are made of these results, communication nets are reconstructed or modified.

Techniques which are applicable in certain situations to network reconstruction, are the recovery by cryptanalytic processes, of call signs (transmitting station identification) which have been enciphered. These cryptanalytic processes are similar to some extent to those which are applicable to textual analysis.

(C) Textual Analysis

Textual analysis deals with the problems involved in the selection of messages of potential intelligence value from the large volume of plain text and plain code messages handled by NSA. In plain text messages the communicators make no attempt to conceal the information content. Thus the problem is one of scanning messages and selecting those messages which contain certain elements as messages of intelligence value.

In plain code messages an unenciphered code is employed for privacy and brevity. Here the problem is essentially that encountered in the scanning of plain text, with some complicating and some simplifying factors.

The particular problems in this area are: to scan rapidly an extremely large volume of plain text messages to see whether they contain certain words, addresses, signatures, etc.; to print out the plain text messages containing the elements sought for; and to categorize the selected messages by subject matter using words, addresses, etc. as the criteria.

The problems listed above are quite similar to some discussed previously.

For plain code messages the same problems exist except that the problems are simpler in that code groups are usually uniform in length, but complicated by the fact that many codes can appear over a communications link.

A problem unique with plain code messages is the automatic decoding and printing, when the code book is known.

A side problem is one of determining whether a secret code or a known commercial code has been employed.

III. Diagnostic Operations

(A) Search for and Statistical Evaluation of Phenomena

Where messages appear which are neither plain text nor in some readily readable enciphering system, we must, in the absence of outside information, analyze the

text of the messages in order to determine the cryptosystem involved. The analysis may be of the texts themselves (Identity) or upon derivative forms of the texts (Latent). The problems of this area encompass the procedures in and the mechanization of: (1) the operations by means of which original text is converted to the desired latent form, (2) the diagnostic operations involved in the search for and statistical evaluation of phenomena which arise in the original or derived text.

We list the problems to be considered under the two headings Identity and Latent according to which form of text is being considered.

(1) Identity Problems

Here a problem is one of searching for the exploitable intrinsic characteristics of plain language and its encipherments. Examples are the widely varying frequencies at which the individual letters occur in literal text and the cohesion of plain language as exhibited by the rough frequency distribution of pairs of letters (digraphs), triples of letters (trigraphs), etc. Under encipherment the intrinsic characteristics of plain language are partially or totally destroyed. However, new characteristics may become prominent and these may be employed

in the characterization of the enciphering system. An example is the Enigma machine in which no letter may be enciphered into itself.

Once a type of message having particular characteristics is identified, a problem is to collect all messages of that type so that the benefits of a large statistical population may be employed.

(2) Latent Problems

Sometimes a characteristic of cipher text is found, not in the text as sent, but in something derived from the text. For example, we can assign the numbers from 1 to 26 to the letters from A to Z. A will be considered to follow Z, as in a cyclic arrangement. By the use of this we introduce a measure of distance between letters of the alphabet. We sometimes "difference" the letters in a message by replacing each letter by its assigned number and then subtracting each assigned number from its predecessor in the message. In some cases the distance from a given letter to its next occurrence is used to replace the letter.

The problems met in the processing and analysis of Latent text are essentially those of Identity text. However, the problem of transforming the original text into the desired Latent form and the preparation of the Latent data for machine processing exists and is not fully included in the problems of Identity.

(B) Test of Specific Hypotheses

On hypothesizing the enciphering process employed in the production of the messages under analysis, more specific tests may be employed to determine the validity of the original hypothesis. The hypotheses fall into two general classes: (1) encipherment was performed by a machine system, (2) encipherment was performed by a non-machine (hand) system (unenciphered and enciphered codes are included here).

Each encipherment system produces certain distinct and observable characteristics which are discernable message-wise or over a selected group of messages. The problems here concern the mechanization of the tests, operations, etc. employed in searching for and establishing the existence of these observable characteristics.

1. Machine Systems

Certain machine systems produce cipher text having marked characteristics. For example, the Enigma type machines cannot encipher a letter into itself. This phenomenon tends to make Enigma cipher text non-random, in the sense that high frequency plain letters tend to occur less frequently in cipher text.

The specific problems here involve the processing of varying amounts of text to ascertain whether

characteristics are present. The process usually includes statistical computations.

2. Hand Systems

Whenever it is believed that a new unenciphered code is in use, messages employing this code must be separated from others sent over the same communications link. This can be done by comparing all groups within the message against themselves for coincidences to obtain the internal "cipher"-group repetitions. A problem here is to perform the coincidence tests and obtain the internal "cipher"-group repetitions.

When code structures are partially or fully known, all code groups in a message may be compared with the known code groups and scored statistically. The problem here is to perform the comparisons and do the statistical scoring.

When codes are enciphered by means of an additive pad, reuses may occur on a different communications link at a later time. If an additive pad has been recovered from the first use, other messages may be deciphered with the recovered key and the resulting texts inspected as in the first paragraph to see whether it looks like either known or unknown unenciphered code.

Some codes have a "built in" checking feature

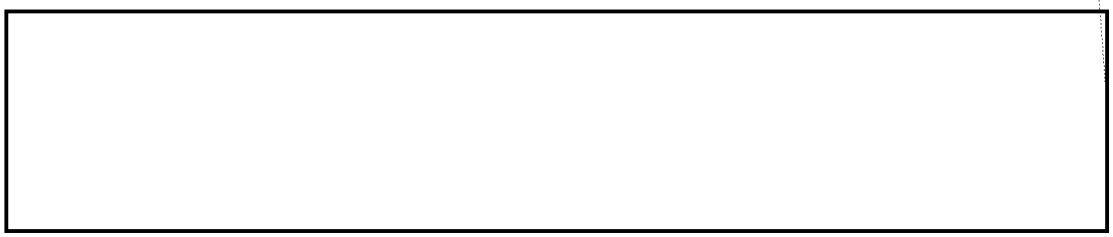
which allows the message receiver to check his text as to garbles, lost portions of code groups, etc. One such checking feature is that the last digit of the code group is the modular sum of the other digits of the code group. Such characteristics can be utilized to identify codes. The problem here is one of finding and exploiting these checking features.

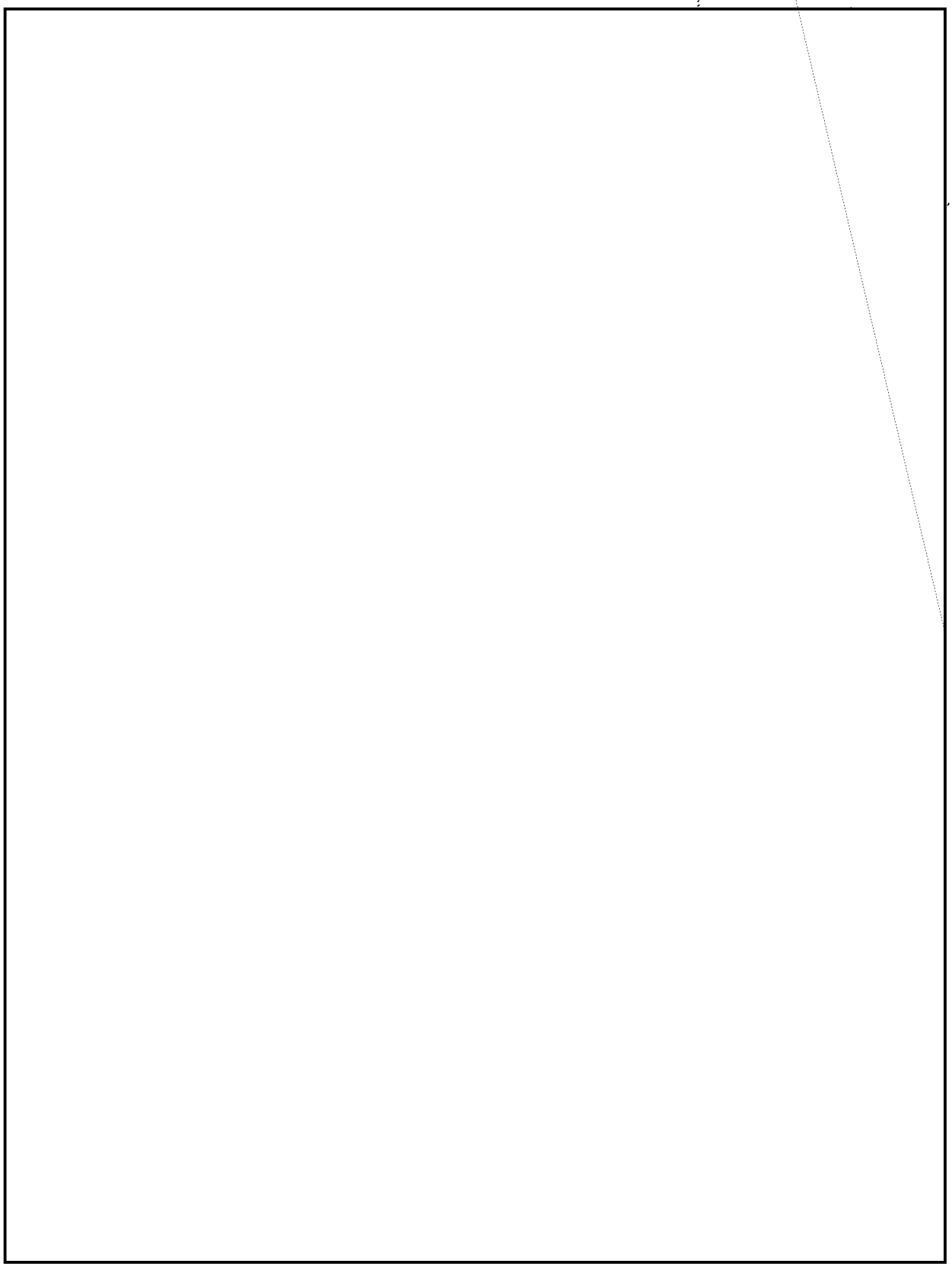
IV. Operations Based on Knowledge of the General System

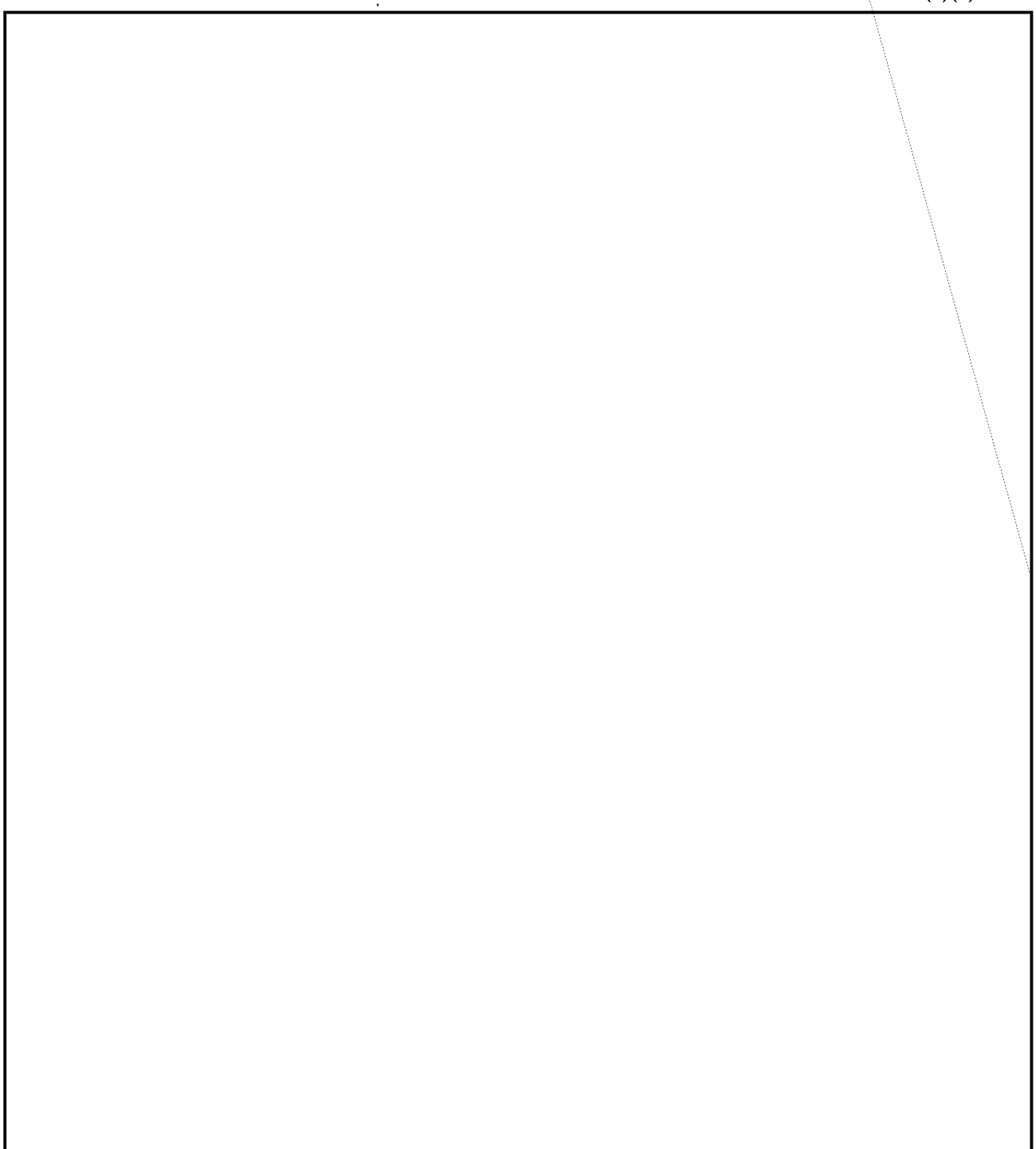
Encipherment of text is performed by means of a machine or hand system. In some cases we are able to assume with a high degree of certainty that a particular message or set of messages was enciphered (encoded) by means of a specific machine or hand system. If we possess statistical, partial or complete knowledge of the variable elements of the system involved, then certain statistical, exhaustive, or analog attacks may be made to recover the messages and some of the remaining unknowns of the cryptosystem.

For convenience we list the problems under two headings: (1) machine systems and (2) hand systems. However, a great deal of overlap exists as will be seen below.

(A) Machine Systems







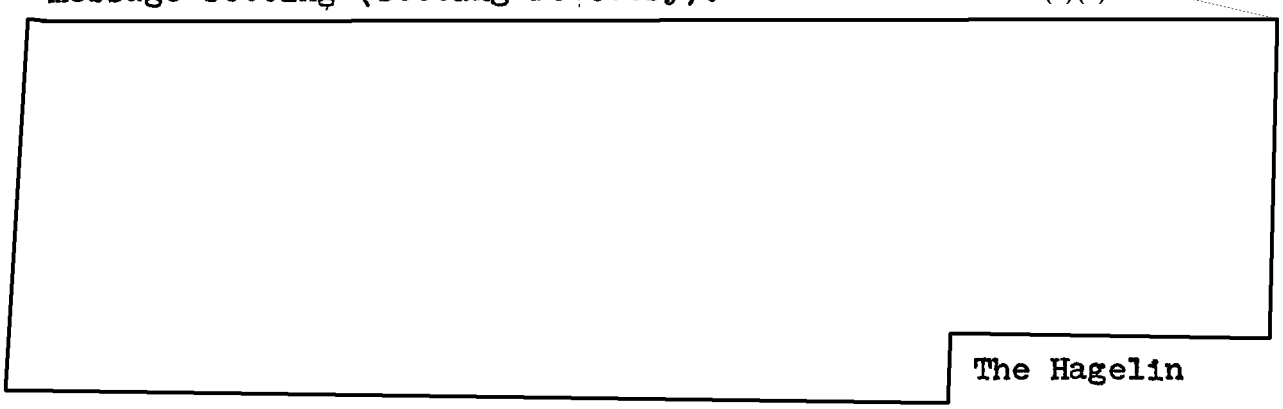
2. Machine Recovery and Setting

Cipher machines are usually set up on two bases. The first is an arrangement of the internal elements

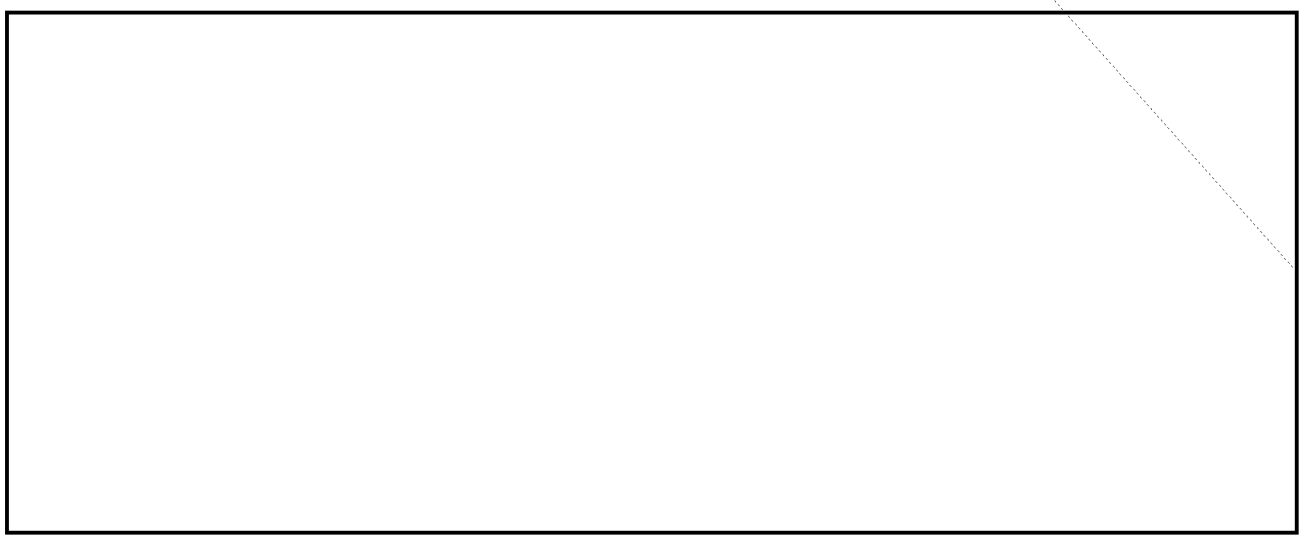
of the machine which may be kept unchanged for a relatively long period (days, months or even years). This arrangement is called the internal setting. For example, the wiring of the rotors, the order of rotors, the plugged substitution between the keyboard and the input side of a rotor maze of the Enigma machine, and the pins on the Hagelin wheels fall into this class.

In addition, there are usually means for starting the machine at some point in its cycle by setting certain marks (usually letters or numbers) on the rotors opposite their respective bench marks. This is called the message setting.

The internal and message settings are considered as the variables of the machine. The problem here is to recover the internal setting (machine recovery) and the message setting (setting recovery).



The Hagelin and Enigma offers an example of the former and latter respectively.



A problem which involves a combination of logical and exhaustive trial techniques follows. Let us assume that an Enigma machine enciphered the message. Let us assume also that we are given a crib (a length of plain text corresponding to a certain stretch of cipher text) free of garbles. In addition, certain of the internal variables are assumed. This narrows the field covered by exhaustive trials to the remaining unknown variables. We also require an analogue to the cipher machine which supplies us with the Enigma motion and its internal elements. Then by assuming one or more features of each of the remaining unknown variables, we employ the crib and machine analogue to derive a set of implications. If the implications are contradictory, then our original assumptions were incorrect. We run through the set of possible assumptions and keep only those which lead to non-contradictory statements and which pass a certain predetermined threshold.

The number of problems involved in the attacks on machine systems is too large to list here. These problems are listed in Appendix A to Phase I.

3. Decryption

At times, a cryptographic period is completely broken, i.e., all the daily settings are recovered, and the message settings are recoverable as part of the message. We are then in the same situation as the intended addressee. Hence the problem is to decrypt the traffic as it becomes available and print the resulting plain text. However, the high deterioration rate of intelligence makes it necessary to perform this decryption and printing as rapidly as possible.

(B) Hand Systems

PL 86-36/50 USC 36
EO 3.3(h)(2)

1. Additive Encipherment

Successful recovery of additively enciphered messages depends on the existence of [redacted] Sometimes the key is predictable because the text of some known book (like a table of logarithms or a novel) is used as though it were a sequence of symbols. At other times, the symbol sequence is generated by a typist striking keys in what he believes to be purely random fashion. The problem is to provide the analyst with pairs of groups, one of which has a probability of occurring as a fragment of plain text and the other of which

has a probability of being produced as typewriter random, and paired in such a way that if both had occurred, they would have produced a group of the actual cipher text. It is desirable that with the pairs of groups, the product of their probabilities of occurrence be also presented. Again, the sequence may be generated by a machine built for the purpose. Once a large enough sample is obtained, the problem is to construct an analogue of the machine so that we have a means of generating key similar to the key which will be used. This last problem is exactly like the problem of machine recovery.

The problems of additive analysis fall into four classes.

(1a) The first class of problems deals with system discrimination which tries to identify the elements making up the enciphering process. Both the steps in the encipher-process and the enciphering elements are sought.

(1b) The second class of problems are those of indicator recovery, or identifying and exploiting the decipherment information sent in the message such as pad number and starting point.

(1c) The third class of problems encompasses the search for depths, or the discovery of messages enciphered by the same pad. The search for isologs (messages having the same plain text but enciphered by different pads made up of exploitable additive) is included here.

(1d) The fourth class of problems are those of exploitation, or the actual recovery of the plain text or plain code using methods previously discussed.

2. Exploitation of Statistical Phenomena

The case of carelessly generated key which was discussed above is subject to attack. In the case of the typist discussed above, we can generate key similar to that which he produces. We then try it out at various positions in a message and examine the resulting decryption to see if it looks like plain text. A unique problem here is the generation of likely key. The problems of placement of key in the message, the decryption of the message employing this key, and the recognition of "good" plain text have been encountered above.

In another case we have a different phenomenon. Studying previously recovered key of the system, we notice that not all key values occur with equal frequency. In this case we say that the distribution of key values is statistically rough. Furthermore, we may notice other statistical properties which more exactly describe the distribution than the mere statement of its roughness. Let us say that some central station using this system sends the same message to a number of its outlying stations, but enciphers this message with a different pad for each

transmission. The practical reason for this is that recipient station A does not have the same one-time pads as station B or C, nor do B and C hold like pads, because a one-time pad is usually made in only two copies, one for each of a pair of correspondents.

Now we have a sort of reverse "depth", a set of cipher messages in which the texts are identical but the key for each is different. This set of messages can be broken in the sense that the single plain text and all keys can be recovered. To affect the solution we assume some plain text and derive key. For each message the correct assumption should produce key which has the previously noted statistical properties. In fact, good assumptions in various portions of the text would be consistent and yield better information about the statistical nature of key population which in turn is used to improve or direct assumptions in other portions of the text.

The problems here include the placement of plain text in the alignment of messages, the deciphering process to obtain the underlying key, the statistical tests made on the key, and the identification of good results.

3. Additional Procedures

One case of this is the solution of a columnar transposition. The cryptography consists of writing a message, in the usual left to right manner, into a

rectangular box. Then the columns of letters are written out, but the columns are chosen in a mixed order, not merely from left to right. In cryptanalyzing, we take advantage of our knowledge of language letter frequencies. We assume that a certain stretch of message was a single column. Then we try other stretches of the message as the column which was originally to the right of the selected column at the time the message was written into the rectangle. In the correct case, the pairs of adjoining letters produced will have plain text digraphic properties. Furthermore, since we know the statistics of the language, we weight each pair by the probability of occurrence and choose the correct match by the best total score.

Then a third column is added, and our knowledge of letter triples forms a basis for scoring, etc. Thus in a system of this type the problem is one of storing probability weights and data handling to make and score assumed matches of stretches of text.

A second example is the case where it is known that additive key was not furnished to the encipherer as a one-time tape or reusable tape or pad, but was generated by a known complex manual process. Some specific variable or variables of the process are not known. Here it is necessary to duplicate the steps followed by the enciphering clerk, making assumptions about the unknown variables. Each assumption results in "trial" additive which can be

applied to the message text. The result, called "pseudo-plain", can be examined to see if it is plain text, or has the statistical properties of plain text.

V. Support Functions

These functions embrace a number of different activities. These activities are somewhat general to all cryptanalytic work and are not based on any particular traffic or system.

A. Linguistic and Statistical Aids

A number of special dictionaries and statistical studies based upon various languages of interest are required as aids to cryptanalysts. The dictionaries and statistics may be based upon particular traffic decryptions or upon general samples of the language. Dictionaries arranged in special ordering (such as ordered alphabetically on last letter of the words) may be desired. Frequent revision of dictionaries and statistical studies are required. Besides linguistic studies large numbers of special mathematical and statistical tables are prepared. These have included special Poisson, Binomial and Multinomial tables, and others.

B. Generation of Crypto-System Data

It is sometimes desired to provide the analyst with listings of data pertaining to a particular cipher machine system. For example, tables which enable computation of cycle distance between specific machine settings, and lists of successive settings of a cipher machine. In certain hand

systems in which the combining of texts is done in several steps, tables showing the end results of the several steps for various variables may be desired.

C. Desk Aids

There are a number of small devices which may be provided for the individual cryptanalyst to use directly along with his work. These include commercial adding machines and desk calculators, individual cipher machines, analogues of portions of various crypto-processes, tallying counters and the like. Such devices may be useful, providing that they are actually available to the person while working.

D. Cryptanalytic Research

There are times, in the course of current cryptanalysis, during which elements appear which are not subject to regular periodic change, but are such that a change in them would obviate current methods of attack. Again there are times when information as to new cipher machines or systems not in use, but offered for sale, becomes available. In this sort of situation a substantial effort to prepare to solve a problem which may never materialize is justified.

E. Collateral

Cryptanalysis is not performed in a vacuum. Often what helps an attack on messages is some notion of what they are likely to be about. Sometimes a name recovered from

one message gives clues as to the subject matter or personalities probably referred to elsewhere in the message, or in some related message. Sometimes direct references to an unread message by date or serial number in the messages that are being read helps provide a wedge into the referenced message.

Similarly, a "Who's Who" file may furnish footnotes to a solved message which will increase the amount of information furnish to an intelligence analyst.

Extensively cross-indexed files are kept in operating sections and at higher echelons to furnish aids of this sort. This type of information is called "collateral".

Here we have a large scale data handling problem which includes filing, indexing, cross-indexing, looking up, abstracting, etc.