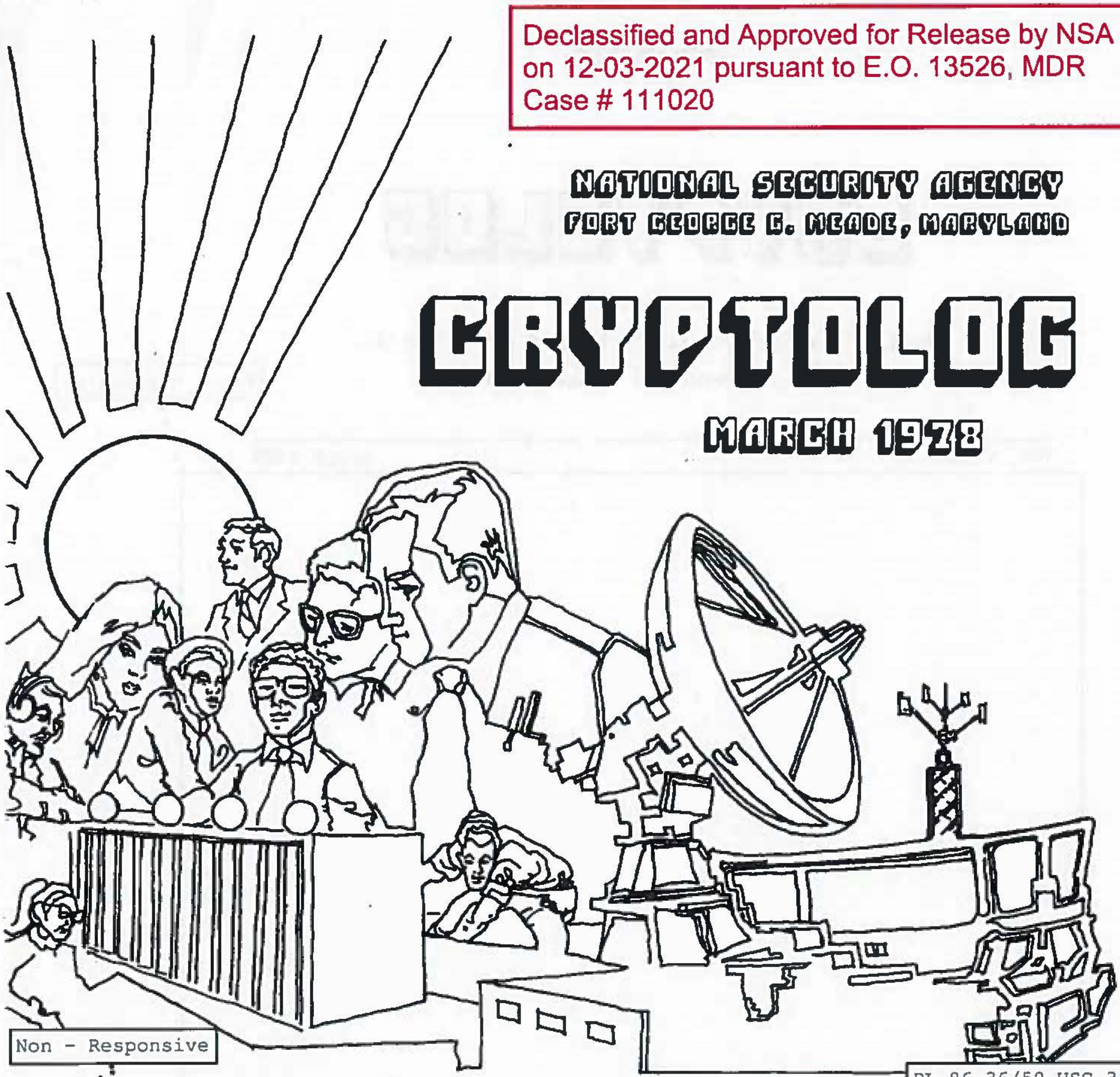
NATIONAL SECURITY AGENCY
FORT GEORGE G. MEADE, MARYLAND

CRYPTOLOG

MARCH 1978



Non - Responsive

PL 86-36/50 USC 3605

~~THIS DOCUMENT CONTAINS CODEWORD MATERIAL~~

~~TOP SECRET~~

~~Classified by DIRNSA/CHCSS (NSA/CSSM 123-2)~~
~~Exempt from GDS, EO 11652, Category 2~~
~~Declassify Upon Notification by the Originator~~

~~TOP SECRET~~

CRYPTOLOG

Published Monthly by P1, Techniques and Standards,
for the Personnel of Operations

Non - Responsive

VOL. V, NO. 3 MARCH 1978

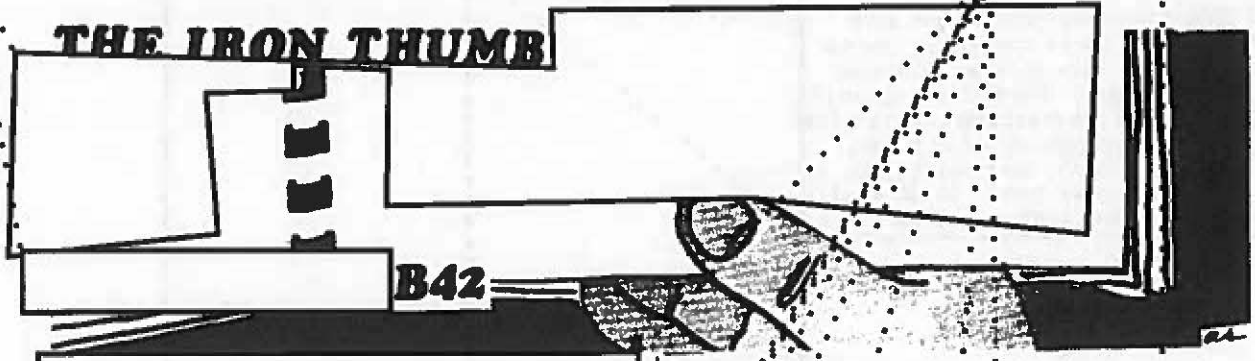
[redacted]

~~TOP SECRET~~

PL 86-36/50 USC 3605

EO 3.3b(3)
PL 86-36/50 USC 3605

THE IRON THUMB



[redacted]

[redacted]

The specific impetus for this paper, however, is discussion-centered around the definition process. There are managers who feel that dictionary lookup for each word [redacted] is unnecessary because linguists should know it all anyway. And there are linguists who think that they do know it all. On the other hand, there are managers who think that, since the machine provides a "translation," they can save money by using nonlinguists, or at least lesser-quality linguists. And there are linguists who are under the impression that the machine is supposed to offer a "translation" and reject the machine because the "translation" is inadequate. Since machine-assisted translation and linguist utilization are topics with some currency, this paper is an attempt not only to correct misapprehensions about [redacted] in this regard, but also to introduce a method of operation to interested readers.

And now the entire process, beginning with the beginning.

Chinese Language

A long time ago and far away, Chinese characters were invented to represent what may have been a monosyllabic language. Today, most people have the impression that Chinese is still monosyllabic -- that is, each character represents a "word." This is only partially true; many words do use one character (*gas, vehicle, middle, heart*), others do not (*car, center*). Words like *car* are called compounds and are made up from two or three individual characters. The word *car* combines the characters for *gas* and *vehicle*. Thus, *gas* and *vehicle* are words only if *car* is not meant.

But the convention for writing Chinese is that all characters are given equal spacing. One of the fundamental problems of reading Chinese, therefore, is the question of which characters combine into compounds in the context at hand and which are to be read singly. [redacted] [redacted] this problem even occurs between phrases, clauses, and sentences, since punctuation is usually omitted. For example,

[redacted]

Ⓐ ① ② ③

[redacted]

Ⓑ ① ② ③

There are a number of options available in this fragment. If one reads [redacted] as part of the compound [redacted] one gets interpretation A. But if one reads [redacted] as part of the compound [redacted] one gets interpretation B (which is less concrete). Moreover, one can read 1 as the end of one clause/sentence (reading 2 + 3 either as a complete sentence, or as the beginning of the next sentence); or one can read 1 + 2 as a sentence (with 3 beginning the next sentence). In combination, these options yield four possible translations:

1.
2.
3.
4.

[redacted]

This problem of identifying words and phrases bears directly on how much time is spent in a dictionary. With text *A, B, C, D,* a linguist may know all four characters singly but have to look up *AB*, *BC*, and *CD* as compounds. Or he may know *AB* and *CD* but have to ensure that *BC* is not relevant in context. The manhours spent in dictionary lookup can be cut drastically by relying [redacted] To use the phrase coined by Norman Wild, it is the "iron thumb" that never gets tired of turning dictionary pages: it gives all possible combinations and their meanings.

[redacted]

[redacted]

Analytic Factors

The users [redacted] respond to a wide range of [redacted] requirements. The following list of short-term product titles was selected from reports issued during the few weeks preceding the writing of this paper:

[redacted]

EO 3.3b(3)
PL 86-36/50 USC 3605

[redacted]

Reporting in B4 and its predecessors (B25, B37, B5, B24) has been affected by several factors. [redacted] Most other reporting has been concerned, on the one hand, with current development and, on the other hand, with long-term trends in fairly broad terms. Third, in response to this, the work style that has developed is one in which linguist/analysts read, understand, and absorb large amounts of data in the original language. They then produce narrative reports which are a synthesis of significant developments or trends in their area of responsibility.

Fourth, in such a situation, for a synthesis to be broader or deeper or better, the analyst must first absorb more data (the quality of the analyst not being under discussion). This means [redacted] Thus there is an opportunity for the system to contribute greatly to the speed and sophistication with which an analyst digests his material. At the same time, the system serves as a channel through which each analyst contributes directly to the understanding of every other analyst [redacted]

[redacted] Sixth, as a corollary of the above points, there is a pressing need to retain and disseminate widely the results of linguistic research.

Finally, since such a large proportion of [redacted] a flexible and responsive retrieval data base is required.

[redacted] is the product of B Group's continuing effort to place the most complete information possible at the fingertips of the linguist/analyst in order to make maximum use of scarce human resources. To expand on the simple definition of the functions given in the introduction, [redacted]

EO 3.3b(3)
PL 86-36/50 USC 3605

[redacted]

EO 3.3b(3)
PL 86-36/50 USC 3605

[redacted]

EO 3.3b(3)
PL 86-36/50 USC 3605

[redacted]

history. As a means of easing the transition from the academic community, which naturally concentrates on characters, to an organization [redacted] and as a means of maintaining proficiency in characters for the linguists in such an organization, such a feature is highly desirable. The cost of adding Chinese characters to a high-volume job has, however, been greater than was worthwhile, in terms of both money spent and time lost in processing. Fortunately, in this instance, computer evolution is moving at a rapid pace. Another review of this question has recently taken place and cost-effective, fast character-printing technology may soon be available.

[redacted] output is available to analysts by category on paper or microfiche at their option. In addition, print of any individual category may be turned off or on at any time. [redacted] may also be suppressed by category.

[redacted]

[redacted]

[redacted]

[redacted]

"Reading" the Text in English

Managers are constantly looking for ways to get jobs done at less cost, whether in salary level or in training investment. An obvious question, therefore, is why can't a nonlinguist "read" the English dictionary lookup? There are several reasons why this is impractical [redacted] [redacted] These reasons, noted individually in passing above, are here recapitulated with an example.

[redacted]

[redacted]

- Fourth, there is a tendency among inexperienced linguists or nonlinguists to take the English dictionary definition as truth. The experienced linguist knows that the dictionary lookup is only a tool and treats it accordingly.

[redacted]

EO 3.3b(3)
PL 86-36/50 USC 3605

[redacted]

[redacted]

[redacted]

Retrieval

[redacted] daily is forwarded to two retrieval data bases. First, a 60-day file is maintained on-line in OAK on a one-day-on, one-day-off basis. [redacted] This file also fills a gap between the daily program and the long-term retrieval file. Data is added to the long-term file month by month. This file resides on STRONGBOX and contains 10 years of data on a one-year-on, one-year-off basis. [redacted] taken off STRONGBOX remains available in the tape library for special requests and extends back to 1956. Both retrieval systems use the [redacted]

In the 12-month period ending in August 1977, more than 23,900 individual requests were made in 60 batches run against an average of 7 years of data; [redacted] The results of the retrieval scan is presented in statistical form to the analyst, who then makes "print/no-print" decisions based on his anticipated workload. [redacted] output can be either paper or microfiche, with or without full dictionary lookup. [redacted]

[redacted]

EO 3.3b(3)
PL 86-36/50 USC 3605

[redacted]

[redacted]

Conclusion

[redacted] has existed unchanged in concept since 1962, which is to say either that its current sponsors are badly behind the times, or that its creators were very farsighted. Having seen the system respond dynamically to changing environments and requirements over the years, I think that the latter is the case. The system is not perfect -- there are useful improvements that can be made to it right now -- but I believe that the concept will remain valid for the foreseeable future.

~~(SC)~~

[redacted]

[redacted]

Non - Responsive