Outline of a proposed community system linking diverse mechanized files of biographic information, with particular attention to the problem of name variants.

A NATIONAL NAME INDEX NETWORK

Walter Jessel

The development of automatic processing systems has now opened up the possibility of constructing a central facility that would provide quick access to information on foreign personalities stored anywhere in the intelligence community. The establishment of such a center would not require pooling the information itself: compartmentation and need-to-know security should be maintained by linking electrically the indexes of individual agencies housed separately in computers programed and operated by their own personnel. The new permanent facility would be only a switching and message center with a medium-sized computer storing names in alphabetical and phonetic look-up tables. Intelligence officers and systems engineers representing their parent agencies would form a committee to keep programs and equipment compatible and the look-up tables up to date. If work were to commence in all agencies by the middle of 1962, the center could be in partial operation by 1965.

A proposal that we take advantage of this opportunity has been presented to the USIB Committee on Documentation. If we are to do so, we need a long lead time and much preparatory effort to reap the evident benefits. In particular, we need to begin immediately to capture in machine language, as on punched cards or paper tape, the typing of biographic index cards or forms in agencies that consider themselves potential participants in the system.

The scope of beneficial participation cannot be outlined in this paper with any precision. The writer's perspective derives from the counterintelligence field, in which files are most frequently checked for security information, but he sees no reason why the principles of index organization and information handling outlined below might not equally be applied to positive biographic intelligence. In practice, the dividing line be-

tween the two is elusive; we have long since given up the quest for a durable definition of the term "derogatory information."

The great bulk, however, of name searching activity in the intelligence community—responding to thousands of requests daily—is initiated to determine whether, in the records of any one of a number of government agencies, there is information of a nature to preclude employment, the receipt of a grant, the issuance of an immigration or a visitor's visa, entry into the United States, an invitation to a conference, etc. Only in a fraction of name searches is such information discovered; most people are quite all right.

Using modern communications and data processing tools, the community can rid itself of at least that mountain of paper work which now piles up merely to determine that there is no pertinent information of this kind. It should be feasible to get such negative answers routinely in an hour or less. Much government business would then proceed promptly which is now held up for weeks. Many people the world over who want to serve or visit the United States would be relieved, impressed, and grateful.

The proposed community coordination, furthermore, should raise the quality as well as the speed of name tracing. Present methods, as we shall see, are not as effective as may be generally believed: because of inevitable shortcomings in manpower and qualifications and through the inroads of human error, there is a fair amount of information in the files which cannot be found. But techniques which can produce more comprehensive and reliable replies to inquiries about people are becoming available; and these techniques, once developed and applied to a single large biographic index, can be simply and inexpensively adapted to others without violating the rules of security compartmentation.

The CIA Counterintelligence Index

CIA's counterintelligence responsibility includes the maintenance of biographic information on persons of security interest to the United States abroad. An extensive system of dossiers and reports with card indexes to them, begun by CIA's predecessor organizations, has been built up over the years. As the collection grew it tended to become unmanageable both

with respect to maintenance and as to the speed and quality of responses to inquiries. These difficulties led to changes in procedures and to intensive studies and experimentation on methods of mechanization. Work on automatic systems both for indexing and for document storage and retrieval has been under way since 1957.

Here we are concerned only with indexing. Under a systems contract with IBM, engineers aware of prospective developments in automatic equipment have cooperated closely with intelligence and systems personnel in CIA, so that the study has not been limited to equipment already on the market. The resulting planning calls for mechanization of the biographic index by 1964, using equipment which will be available by that time.

The mechanized index replacing the present many million cards will store between two and three billion alpha-numeric characters. Random access will permit a computer search to proceed directly to any position in this storage.[1] There has lately been an average of 1,000 inquiries a day into the card index; the mechanized system is designed to handle 2,000. The cards are arranged alphabetically; the computer makes practical other techniques of index organization that will render findings more reliable.

Organization by Phonetic Group

From the *Washington Post* of Dec. 16, 1961:

> **NEWSPAPER STUFF**
> Reuters said that Nazem Kudsi had been elected President of Syria. The Associated Press wrote it Nazam el Koudsi. UPI said it was Nazim Kudsi. Our world desk checked with Syrian diplomatic representatives here. They said the correct transliteration is N a z e m Coudsi. We took their word for it.

[1] For a general description of the use of computers in automatic data processing, see Joseph Becker's "The Computer—Capabilities, Prospects, and Implications" in *Studies* IV 4, p. 63 ff.

As this little story suggests, the key problem in organizing an index of references to people anywhere on the globe, more often taken from a foreign language than from English, is the variant ways in which identical names may be reported, spelled, transliterated, or translated. It takes a strong linguistic background to make an effective search of an alphabetically organized file. Phonetic rules may be reasonably useful in straightforward transliteration from a foreign language, but they fail when, shall we say, the name of a Polish member of the Laotian truce commission is reported by a Lao through a French intermediary to our embassy in Vientiane.

If name variant searching is an intricate task for trained experts, how can a computer be taught phonetic rules? We attempted to develop a phonetic system suitable for programming. It bogged down quickly as we learned that logical rules could not be formulated to deal with the unpredictable name spellings in many languages in the index. A grouping scheme which is controlled by the judgment of expert linguists was then adopted. This is how it is set up:

When we began to think about the possibilities of mechanization some years ago, we took an essential first step to make the use of computers possible. The section which produces our 3x5 index cards was equipped with Flexowriters, electric typewriters which punch a communications-type paper tape in the course of normal typing. This paper tape can be converted to other media carrying machine language, including the magnetic tape commonly used for the input to computers.

From this machine language version produced simultaneously with our 3x5 cards, a computer extracted surnames and given names separately, pulled them together by area of origin, and alphabetized them. Then it punched out one IBM card for each spelling of every surname and of every given name originally typed. We thus obtained one card each, for example, for Mueller, Smith, and Kim, and one each for Fritz, William, and Alexander.

The linguists working on our grouping scheme go through these IBM cards for like-sounding surnames and for equiva-

lent given names. They pronounce the names, often aloud, and those that to their ear and mind belong together they assemble behind a card bearing a serial number. This assigned group number provides the linkage among name variants that is needed in filing and searching. The grouping process is a vast game of solitaire in which little writing takes place, merely an arranging of punched cards. Machines do the processing and printing, leaving the linguists to use their expertise free from boring clerical routines. The approach is completely pragmatic: only names which have actually been reported are included in the groups. Theoretical possibilities do not interest us; we have no need to search for them.

Tables 1 and 2 below contain groups of surname and given name variants, respectively, each designated by the collective serial number assigned it.

003626	CHEVCHENKO
	JEVCHENKO
	SCHEVCHENKO
	SHCHEVCHENKO
	SHAVCHENKO
	SHCHEVCHENKO
	SHEUCHENTCO
	SHEVCHENKO
	SHEVSCHENKO
	SHEWCHENKO
003630	CHIGVINTSEV
	TSIGVINTSEV
	TSIGVINTSEVA
002170	KRCMAR
	KRETSCHMAR
	KRETSCHMER
	KRETZSCHMER

TABLE 1

Z63228	ALEX	
	ALEKS	
RELA	Z00029	ALEXANDER
	Z00116	ALEKSEI
Z00029	ALEJANDRO	
	ALEKSANDR	
	ALESANDRO	
	ALEXANDER	
	ALYA	
	OLEKSANDER	
	SANDOR	
	SASCHA	
	SHURIK	
RELA	Z63228	ALEX
Z00116	ALEGSEI	
	ALEKSEI	
	ALEKSAY	

TABLE 2

In the computer these names, each with its group number, are reordered alphabetically, as in Table 3, so that the machine in beginning a search can get the group number from an alphabetic look-up. (For explanation of the search procedure see p. 10 below.)

003625	CHETVERIKOV
003625	CHETVERIKOVA
505716	CHETVERNIA
520151	CHEVALLIER
505717	CHEVARIN
003626	CHEVCHENKO
520152	CHEYRON
000414	CHEYSHVILI
000414	CHEYSVILI
003610	CHIBANOFF
526539	CHIBAS
505718	CHIBIKOV

TABLE 3

The name variant problem is of course most complex when transliteration from other alphabets is involved. Linguists, in their minds, retransliterate what they see to the original spelling, and this spelling governs their selection of variants to form a group. No effort is made to designate a correct transliteration; it has no bearing on the performance of the system, which accommodates any transliteration and even outright misspellings like the JEVCHENKO in Table 1. The groups of variants are flexible. They may be split or combined as experience and professional understanding of the system's needs may dictate.

Variants of the names of public figures who write them in the Latin alphabet, on the other hand, are unimportant. In looking up the German chancellor, there is little point in subjecting the name ADENAUER to a phonetic variant treatment. For such searches the system will permit bypassing the name grouping feature so that the machine yields only records of names which exactly match the spelling in a request. The extent to which such an option will be used by requesters will depend on the amount of irrelevant information otherwise produced by the system.

The grouping of names in an empirical phonetic order is hardly original; the British MI-5 index, for one, was organized in this manner decades ago. The method is eminently suited

to the counterintelligence problem. As people move around they adjust the spelling of their names to new surroundings; phonetic reporting is normally unreliable; and in transliterating from one language to another, several different standard and non-standard schemes are frequently used.

The Nonsense Bloc

We collect information about people in all parts of the globe. The chain of individuals who do the reporting and filing—agents, intelligence officers, foreign officials, typists, communications personnel, headquarters analysts, more typists, index clerks—can readily have a weak link, someone who lacks adequate knowledge of foreign languages and customs and who may thus confuse a surname with a title, an occupation, an honorific, or a given name. When this happens, the index reference is misplaced, and the information lost. Table 4 contains some of the words which have crept into our index as surnames.

RECHECK TRIGGERS	LANG	MEANING
O ALIAS	ENG	ALIAS
N BINBASHA	TURK	MAJOR
O EHEFRAU	GER	WIFE
N HAJ	ARAB	PILG. TO MECCA
O NADPORUCIK	SLAV	FIRST LT
O RECHTSANWALT	GER	LAWYER
O OBERSTLEUTNANT	GER	LT. COL.
N TERCERO	SPAN	THE THIRD
O VOPO	GER	E. GER. POLICE

TABLE 4

At first glance, this seems to show a sorry state of affairs. But when you look at it, how is an index clerk, a well-educated one, to know that NADPORUCIK is Serbo-Croatian for First Lieutenant and cannot occur as a surname? Perhaps he should know that FRAU means Mrs. and could hardly be someone's surname, but then a little knowledge could lead him into quite a trap: there are seven people listed in the Washington phone book by the name of MISTER, and 20-odd by the name of HERR. Anyway, we cannot expect to have a reporting

chain of universal linguists, one of whom will not occasionally let a SONDERFUEHRER creep into the index.

Our linguists have now tagged all such words, summarily labeled "Titles," under two headings: "N" for those which may also occur as surnames (HAJ, PRINCE, GRAF), and "O" for those which may not (ALIAS, OBERSTLEUTNANT). The system can therefore recognize them and take precautions against their false entry as follows.

The name tables act as gates for new entries. The surname in a new entry to be filed must match a name in the alphabetic table, pick up its assigned group number, and then enter itself as a new member of the entire grouped set of references. If it finds no match, the machine prints out a notice to this effect to the editor, who assigns a group number after consulting his tables or the expert concerned and feeds it back into the machine. But if it finds a match in one of the words our linguists have identified as a title, the machine prints out a notice:

> EDITOR. THIS WORD IS A TITLE IN THE XXXXX LANGUAGE, MEANING XXXXX IN ENGLISH. HOWEVER, IT CAN ALSO OCCUR AS A SURNAME. RECHECK REQUIRED.

Or:

> EDITOR. THIS WORD IS A TITLE IN THE XXXXX LANGUAGE, MEANING XXXXX IN ENGLISH. ITS OCCURRENCE AS A SURNAME IS EXTREMELY UNLIKELY. REVIEW AND REWRITE ENTRY.

Because name order is often confused in reporting names in foreign languages, we are in a similar danger of losing information by filing index cards under given names instead of surnames. The machine therefore contains tables of common given names in all languages against which the surnames of all new entries are checked. If there is a match, the editor will receive a print-out as follows:

> EDITOR. THE NAME IS A COMMON GIVEN NAME. GROUP NO. XXXXX CONTAINS ITS EQUIVALENTS. ADVISE RECHECK ON NAME ORDER.

Particles which occur in surnames are tagged as such by the linguists, and thereafter they are ignored in the machine's internal processing. Thus, in one group, we will place

```
002874  /DE LA/ ROSA
        /DE/ ROSA
        /DE/ ROZA
        /LA/ ROSA
```

By this device reporting inconsistencies are reduced to the common denominator—the group number. But while

```
506784  WAGNER
        WAGONER
        /VAN/ WAGONER
```

will be found in one group,

```
507865  VAN NGUYEN THAN
```

is listed without tags enclosing VAN, which in this case is not a particle. A computer rule to treat all occurrences of VAN as a particle would cause trouble.

Neither the phonetics of last names nor the translation or transliteration of first names can be sorted into such clear-cut groups that all ambiguities are avoided. Furthermore, some groups tend to become so large that the reference output from them would be unwieldy. To cope with these problems, we have introduced through the concept of "related groups" a mechanized cross-reference scheme. It looks like this:

```
007878  'ASIM
        GHASHIM
        GHASIM
        JASIM
        KASEM
        KASSEM
        KASSIM
        QASEMI
        QASIM
        QASSIM
  RELA  008495  KAZIM
```

TABLE 5

Retranslated into Arabic, KASSEM (and variants) is not the same as KAZIM. In transliteration, however, these are easily confused. The "see also" technique illustrated in "RELA 008495 KAZIM" draws attention to this possibility. This approach is useful in compiling tables of given names as well. In Table 2, ALEX abbreviates ALEXANDER as well as ALEKSEI, but the use of a single group for all of these would have been confusing.

Machine Search Procedure

When the system is established—beginning we hope in 1964—the name trace procedure will run as follows. The computer will first check the surname to be traced against the alphabetic table. If it finds a match, it picks up its group number and switches over to the element of the magnetic storage carrying that number. In the magnetic storage, which corresponds to the present 3x5 index cards, all records pertaining to all surnames in the same group are filed and searched in the same element.

The records under the surname group are then automatically sifted to match up given names or initials, age range, a country or countries of residence, sex, citizenship, document date and source, etc. All index entries that meet these criteria are now printed out in full text. The system is designed to complete one average search in about five seconds.

In addition to references, the print-out gives all name variants in the group searched, so that analysts using the index can suggest changes in group composition. The variants in related groups, if any, are also reproduced, but the records under these are not initially searched.

Having reviewed the output, the analyst decides whether to request a further search based on the content of related groups. At the service speeds we envision, he can afford to do this without loss of efficiency. Instead of taking the browsing approach to which manual searchers are often addicted, the machine searches according to strict rules, supplying only such alternatives as experts have previously decided are possibly relevant. The aimless and inhumanly dull blind groping through a card tray from end to end will, we hope, come to a stop.

The advantages of such a machine search over clerical manipulation of an alphabetically organized collection of index cards are readily apparent. The index clerk, using all the knowledge and imagination at his command, can still hardly be expected, when asked for a check on RAHMAN, consistently to cover such variants—which have actually occurred—as RACHMAN, AB'ALRAHMAN, ABD AL RAHMAN, ABDAR-RAHMAN, and so on. The requesting Near Eastern analysts, knowing this, therefore tend to maintain elaborate crutch indexes to assure them of proper results independently of the central index. Professional personnel spend much time in this manner, and even they are hardly in a position to retrieve a reference in which a name has been misspelled or mistyped somewhere along the line.

The machine, without invading the province of human judgment, thus becomes a tool for improving the quality of our collection and our work in general. Expert judgment, once rendered, is repeatedly applied, no matter how much or how little an individual analyst or clerk may know, until a better expert comes along to make a change. The usefulness and the effect of expert knowledge are vastly broadened in this way.

The careful reader will have discerned a fair number of problems for which a solution is not apparent in this brief description. Much work remains to be done. One of the most complex matters is the treatment of Far Eastern names. In these, original ideographs are often represented by four-digit numbers, the Telecodes. We are likely to find that Telecodes can be used as the equivalents of group numbers, to supplement phonetic spellings. The complication is that many Far Eastern names are reported without Telecodes.[2]

The grouping process, which will go on in parallel with the conversion of the several million index cards to machine language, takes a great deal of effort on the part of linguistic experts. We are fortunate in having associated with us linguists whose knowledge spans the globe. We take satisfaction from the thought that, once done, this effort will not only lighten our own chore of running 1,000 name traces daily, but may be

[2] For an outline of the basic complications of this problem see Guy P. Webb's "Machines and the Chinese Name," *Studies* V 1, p. A29 ff.

turned to advantage by others in the intelligence community as well.

Functions of the Network Center

The term "network" is used to make it clear that the system proposed in this paper does nothing more than relay inquiries about names and yes-no answers to them among the mechanized files developed individually by participating agencies. The network center does not deal with substantive information; it does no professional work except the maintenance of name look-up tables; it does not even see substantive information developed by a participating agency in response to a request.

A Pool of Name Tables

The gist of the proposal is to capture as a by-product of typing and machine processing a machine language version of all surnames occurring in the indexes of the participating agencies and to pool these in central computer tables. They are thus divorced from all other information held by the agencies concerned. The Center can then perform two functions: take care of the name variant problem centrally in the manner outlined above; and address inquiries to those agencies which, according to the central tables, have filed information on someone with either the exact surname queried or one of its grouped variants. Participants could, of course, direct their inquiries specifically to one or several individual agencies if they wished.

Look-up tables of given names and other elements like nationality would also be stored in the Center's computer. Their purpose is to supply the number or code that stands for the whole group of given name equivalents in each of the participating agencies' computer indexes and the codes designating other elements in each.

The grouping of names in the CIA counterintelligence index is apt to reach a plateau sometime in 1963, after which relatively few name spellings that have not occurred before will have to be dealt with. From then on, the Agency would be in a position to supply surname and given name magnetic tapes to the central computer in both alphabetic and group number order. Surnames filed in the computers of other network agencies could then be matched against those from CIA, and

those that match tagged with a symbol for the filing agency. If there were no match on a name from another agency, a punched card would be produced which linguists would use to assign the appropriate group number, thus rounding out the Center's tables. In return for its list of surnames the Center would furnish the contributing agency computer materials—tapes and programs—for the organization and storage of its own index entries in groups corresponding to the surname tables at the Center.

A Search through the Center

The chart on the next page illustrates the Center's modus operandi. We may anticipate these steps:

1. Using a predetermined common format, a member agency teletypes its search request to the Center. It contains the usual elements—name, date and place of birth, sex, citizenship, residence, occupation, etc. It includes information for the member agencies' analysts to use when the computers have produced possibly relevant references—the purpose of the search, its intended depth, and any additional information about occupation, geography, events, etc. which will help in fulfilling the request.

2. When the message is in, the Center's computer first assigns the request a serial number which accompanies the processing of all its elements until an answer is returned.

3. Then it separates out the request elements that require matching against the Center's tables. The match on a surname yields a group number and a roster of agencies in whose indexes information on persons with such a name may be found.

4. Next the machine loads the complete request, together with the group numbers it looked up in the tables, into program converters for the member agency indexes that are to be checked. These converters are sections of the computer in which the common request format is transposed into that used by a participating agency in its own machine index. There is thus no need for total systems uniformity among the members.

5. Message assembly, transmission to the members' computing installations, and bookkeeping are the remaining steps in processing the search to the members' mechanized indexes.

SECRET
REQUESTERS
NETWORK CENTER
PARTICIPATING COMPUTER INDEXES
DICTIONARY TABLES
PROGRAM CONVERTERS
CIA
STATE
ARMY
NAVY
AIR
FBI
OTHERS
REQUEST FORMAT EDIT
REPLY CHECK & TRANSMISSION
SURNAME ALPHA
SURNAME GROUP
OTHER NAMES ALPHA
OTHER NAMES GROUP
OTHER ALPHA (OCC'N. TITLES)
OTHER GROUP (OCC'N. TITLES)
MESSAGE ASSEMBLY & TRANSMISSION
"NO INFORMATION"
"INFORMATION TO FOLLOW"
SUBSTANTIVE INFORMATION
SECRET

6. An answer will be expected from each of the indexes addressed; a log entry in the "Reply Check" section of the Center's computer will call a Center employee's attention to delays beyond a reasonable time, say more than an hour. When the returns are in, the computer composes a message to the requester and transmits it. These answer messages are of two types only:

(a) No pertinent information available, or

(b) Possibly pertinent files are being reviewed; further information will come through normal, direct channels.

There are two possible sources for the completely negative (a) answer: One is the member agency's index computer itself, drawing a blank on the search criteria. The response in this case is automatic. If, however, the search yields possibly pertinent references, an analyst should rapidly scan these merely to determine whether it might be useful to consult documents or files. If not, the search ends at this point. A simple instruction from the analyst to the computer will give the requesting agency the answer (a) or (b).

Is It Worth Doing?

Mountainous computers have now labored and brought forth one of two mice—either a negative or a promise of something, which may in turn be negative, to come. Otherwise put, all we are suggesting is an immediate show of ignorance or an expedited review of pertinent information leading to the same sort of positive reply at present produced by analysts in all agencies. Will this cost too much?

The economy of computer systems is often a touch-and-go matter. With no information from other agencies available, we have, of course, no proper basis on which to make a calculation. However, let's play with some numbers anyway.

We do over a thousand name traces daily in CIA headquarters alone, and most of these are run as well in several other agencies. The community surely makes 15,000 trips a day to its Washington name indexes and files, not counting look-ups in overseas repositories (which our plans for the future include tying into the headquarters computer facility through electrical communications).

It appears to take us about an hour under our present manual system to do the average name trace. The community's aggregate manpower commitment to name tracing is thus apt to be around 15,000 man/hours daily. However, the 10 or 15% of traces that require file review or memo writing probably take up half of this time.

The labor for negative fruit that machines can take over is thus on the order of 7,500 man/hours a day, or the full time of around 1,000 people who cost the Government some $10,000,000 annually in salaries and overhead. The aggregate computer budget should remain below this figure. It should then be worth doing in the name of economy, as well as for the sake of long-term quality and service gains.

A Stitch in Time

We have listed neither all the benefits nor all the costs of such a system. Among the latter the conversion of existing indexes to machine language is prominent, a job so arduous that it will have to be spread over a long period of time.

We in CIA, as we have said, began typing our index cards with a machine language by-product in the fall of 1957. By early 1964, when our index computers should begin to operate, we will probably have half—the recent, most useful half—of our index entries in machine language. Then begins the drive on the rest.

What if we had not begun in 1957? We would be overawed by the conversion problem, instead of merely impressed and annoyed. And the prospect would appear less pleasant every day. Optimists might point to the development of mechanical print-reading techniques; we have little faith in the possibility of applying these to a collection of heterogeneous 3x5 index cards. Every index or reference card now produced on an ordinary typewriter by a USIB agency which later decides to use a computer—whether in the proposed network or by itself—will probably have to be retyped. To make a beginning, therefore, index and reference typists need to be re-equipped with keypunches or tape-producing typewriters.

Those members of the intelligence community that have taken this step will be in a position to link their biographic indexes in a network. Consider the trends: Information vol-

umes are going up, the number of trace requests is increasing. Clerical manpower needs and labor costs are going up correspondingly. On the other hand, computer flexibility is improving, and computer costs—per unit of work done and information stored—are coming down. And as we read current history, we see no visible trend toward a lessening of pressure on the security of the United States here and abroad which might reduce the Government's need to use its intelligence and counterintelligence tools as effectively as possible.