G. DIACONESCU
M. PENCEA
C. IONESCU
M. MAGDA
G. MITRAN
Department of Economic Cybernetics
Academy of Economic Studies

Automatic Coding of Denominations of Raw Materials and Materials in View of Implementing a Data Base for the Production Management

1. INTRODUCTION

In the process of creating, processing and transmitting of the information in the administrative, economic, technical and scientific practice, etc., there appear information consisting in the denominations of some extremely varied entities. From one instance to another these denominations can be names of materials, books, people, etc. The informational processes tend to utilize, owing to efficiency reasons, the complete forms of the denominations only in the initial stages of data processing and/or in the final ones (of display). In all the intermediate stages one works with their codified forms. The objective backing of this trend resides in the fact that, on one hand, the semantic level of a denomination is useful and necessary only in the stages which presuppose the direct participation of the human operator, and on the other, that the denomination can be processed, transmitted or stored, as a non-redundant information, set up exclusively at the syntactic level. An arbitrary denomination uses a short text, containing maximum 10 to 20 words (the instance of denominations of materials) whose grammatical structure renders evident, together with the relevant words, a series of link words. In some instances, the multitude of denominations attached to the various samples of a given entity (the multitude of the names of all the people in a plant, the multitude of all the works in a big town or the multitude of the materials in a plant, etc.) makes use of the combinations of words which can be grouped into a vocabulary of a limited scope. One of the most frequent instances is that, in which each word of the vocabulary occurs with exactly the same sense in each denomination, irrespective of

its inflexional forms. Accordingly we will consider, in the following pages, that the flexional forms of a word are semantically irrelevant. This assumption is verified in the case of the denominations of materials, where the forms "of the sheet", "of the sheets", "the sheet" type can be, for instance, identified with "the sheet" without the denomination changing its identity. From the empirical coding systems (abbreviations) up to the most advanced systems of codification the target pursued is to set up a one-to-one correspondence (within space) and a permanent one within time, between each sample of an entity and a certain representation of this under the form of symbols. The retrieval within time and at long distances of the correspondence is a question of information transmission which makes it necessary to deal the codification problem with specific means (error detecting codes, self-checking codes, cyclical codes, etc.), while the correspondence retrieval in the storage space (conceptual space, consisting of information units, perfectly distinct within the physical space) is a pure informatic problem. This latter modality of considering the coding question is tackled in the following lines. Moreover, for a clearer delimitation of the problem we will point out that there are involved neither the codification of numerical values (the standpoint of the arithmetic of the computer), the symbol coding (EBCDIC code for instance), nor the coding in view of data security. We will present a system which gives the computer the possibility of making the distinction between one denomination and the other (the distinction being made at the semantic level, not at the syntactical one) and to grant identity to denominations. The retrieval of this identity, materialized in a number, is possible for any of the possible equivalent forms of the denomination. For instance, in the case of materials, if there has been admitted that the identity of the material

PICKLED␣␣␣SHEET␣␣␣OF␣COPPER is 493

(with␣the space has been marked as typographical sign), the system has also to identify the following denominations:

PICKLED␣COPPER␣␣␣SHEET
SHEET␣OF␣␣␣COPPER␣␣␣PICKLED
PICKLED␣␣␣SHEET␣␣␣OF␣COPPER
PICKLED␣␣␣BRASS␣␣␣SHEET

etc.

as well as material of whose code is 493.

In any coding system besides the strictly necessary function of identification, the condfication of denominations should also include the compactification function (code display with minimum number of symbols).

To clarify the aspects put into discussion it is necessary for us to make a distinction between the physical sample of an arbitrary entity and the name, the code respectively, of this. For instance, in the case of the "material" entity all its samples are physically distinct (sheet, bar, wire, etc.). To each of these samples there is associated a name which consists of a sequence of symbols (letters and/or figures and special signs) arranged into words. The correspondence established between a physical sample of the entity and the associated number is however in general nonsingle-valued, the name having only a nominating role, of "label". The sense, the semantic content of a name being in fact the single-valued correspondent. One can say that

in the instance (1) although the names are syntactically distinct, the sense which is associated to each name is identical. Therefore, the essential function of the name is the semantic function and less the function of identification at syntactical level.

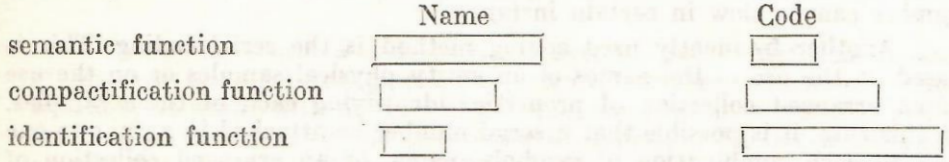
The strengthenning of the function of identification of the name is sometimes effected by adding new words to the name, such as

JOHN DOE 1
JOHN DOE 2
BRASS SHEET STAS 1054
BRASS SHEET STAS 2455

When the identification function of the name is strengthened so, one can say, by abuse of language, that the name is a code or, more correctly, that a name-code is involved.

On the other hand, from among the code functions, the most important is the identification function, the compactification function coming next; in some cases there is also a semantic function (descriptive codes).

In relation with the share of the three functions (semantic, compactification and of identification) one can speak of "pure" codes, of "pure" names or of "code-names" and "name-codes", etc.



An example of "pure" code is the serial code of a product (serial number, expressed eventually in binary system) where the identification and compactification functions are strongly emphasized, while the semantic function is lacking. The name "pure" can be exemplified with the help of certain musical terms: ("Maiestuoso...", "Con molto grazia", etc.) and in general with the metaphor-names where the semantic function prevails, the identification function being practically absent. The name-code may be illustrated by the usual names of people where the identification function is stressed.

A coding system, related to the purpose for which is designed, involves in reality a wide variety of functions (error detecting function, self-checking function, code safety function, etc.). The aspects presented above aimed however to more accurately define the context in which the suggested solution lies.

2. PRESENT SOLUTIONS

The existing methods of codification are using either the already existent names of the physical samples of entities or a characteristic feature or an arranged group of characteristics which describe in a one-to-one way an execution aiming at the improvement of the identification and compactification function.

Recently, for people coding, one uses the firm descriptive codification based on durable characteristics in time (sex, blood group, date of birth), the identification function being effected with the help of the introduction (attaching) of a synonymy number, to each arranged group of characteristics (sex, blood group, date of birth). Another example of firm descriptive code is the EBCDIC code for the display in bits of the graphical characters where as it is known, the first two bits show the class of the character, the next the subclass and the last 4 bits are analogous to a number of synonymy. As long as the firm characteristics are involved (as number for all entity samples and also within time, for each entity sample the descriptive coding offers a major advantage in the sense of the semantic improvement of the code. The descriptive code is however disadvantageous in the case when at the setting up of the coding system there cannot be provided for all the characteristics which describe the physical samples of the entity, the system remaining open. Thus, is the case of a descriptive coding of the materials it is perfectly possible for new materials to appear, which are synonymous in comparison with the properties taken into consideration at the initial establishment of the system, but which differ in reality through new properties (for instance, colour). Moreover the setting up itself of the synonymy number can be slow in certain instances.

Another frequently used coding method is the serial coding. This is based on the use of the names of an entity physical samples or on the use of an arranged collection of properties identifying each of these samples. In this case, it is possible that a serial number be attached in an one-to-one way to each combination of symbols (name, or an arranged collection of symbols), different from those existent. This method offers the advantage of displaying a certain autonomy in relation to the coding object (codified entity), the method being applied irrespective of the type of this. The major disadvantage of this method lies in the fact that is connected with a rigid syntactical structure of the denomination or the arranged collection of properties.

Other coding methods combine the descriptive coding method with the serial one. So, they bring in control signs, allow a less rigid syntactical structure, etc.

3. GENERAL DESCRIPTION OF THE AUTONOMOUS CODING SYSTEM

In the suggested system the coding source is the collection of all the denominations of entity samples. One can deliberately admit that for each physical sample there is a whole class of equivalent denominations at semantic level and that each physical sample is in an one-to-one correspondence with a class of equivalent denominations.

The result of the codification, the code respectively, is a number ranging between 1 and the maximum number of distinct physical samples, this number not being serially fit (in the order of entrance in the system) but randomly and corresponding one-to-one with the distinct physical samples of the entity (Fig. 1).

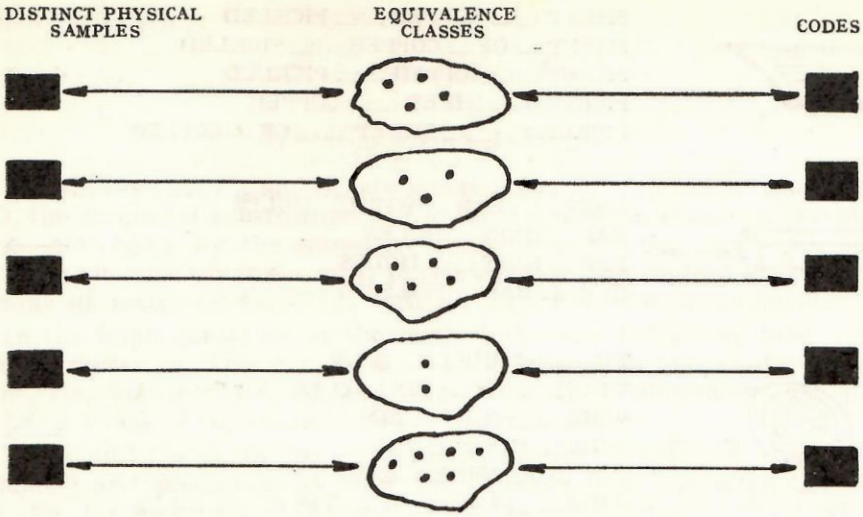


Fig. 1.

The method uses a code retrieval system, one of the denominations of physical sample being given. The retrieval system is strictly necessary because one works with millions of possible denominations and tens of thousands of physical samples as in the case of materials. The retrieval is based on separate storage, on magnetic disk of the:

(1) vocabulary (list of words)
(2) standard denominations (a denomination from each class of equivalence)
(3) synonymous words
(4) link words

the access to these files using the technique of direct addressing (HASH-ADDRESSING).

To illustrate the above we will consider the example of entity material/raw material with three distinct physical samples and 15 syntactically distinct denominations (Fig. 2).

It has to be mentioned that in the suggested system one considers that two denominations make part of the same class of equivalence if:

(a) differ by the order of the words in denomination
(b) differ by the number of blanks between words
(c) differ by the used link words
(d) differ by the pairs of synonyms (copper-brass-CU., or stanum-tin-SN., etc.)
(e) differ by the pairs of words that are flexional variants of the same word (SHEETS SHEET, etc.);
(g) do not differ by other characteristics but the above ones.

The equivalence so defined allows the coding system to be autonomous not only in relation to the sense of the words, respectively of the properties which they express but also in relation to the above properties. The available code can remain internal to the system, the human operators using further the denominations as a namecode, namely there is neither manda-



SHEET' ␣␣ COPPER ␣␣ PICKLED
SHEET' ␣ OF ␣ COPPER ␣␣ PICKLED
SHEET' ␣␣ COPPER ␣␣ PICKLED ⟷ 3
PICKLED ␣ SHEET' ␣␣ COPPER
PICKLED ␣␣␣ SHEET' ␣␣ OF ␣ COPPER



IRON ␣␣ BAR ␣ WITH ␣ HOLES
BAR ␣ IRON ␣ HOLES ⟷ 1
BAR ␣ IRON ␣␣ HOLES
IRON ␣ BAR ␣ HOLES



TIN ␣␣ WIRE ␣␣ Ø 1.5
WIRE ␣␣ OF ␣ TIN ␣ Ø 1.5
WIRE ␣␣ Ø 1.5 ␣ TIN
WIRE ␣ TIN ␣ Ø 1.5 ⟷ 5
TIN ␣␣␣ WIRE ␣ Ø 1.5
WIRE ␣ Ø 1.5 ␣␣ OF ␣ TIN

Fig. 2.

tory nor reccomendable to substitute in the human environment the denominations of a sample by the standard denomination, or by the code supplied by computer. In the case of materials, after the system is set up, the denominations can be further given freely enough, in keeping with the rules a-f, the system assuming all the tasks connected with the identification of the material, its classification, etc.

4. NOTATIONS AND DEFINITIONS

To write the denominations of raw materials and materials one uses the multitude of all symbols (for printing). Therefore, in practice, the alphabet consists of the Latin capital letters, and figures, to which some special signs are added, inclusively Greek letters for units of measure (Ω — ohm, etc.). For the implementation of the system it is necessary to reduce the real alphabet to the alphabet (set of types) available in the key board of punch devices (it will be written therefore MIU instead of μ, OHMES instead of Ω, DIAM instead of Ø, etc.). In the contemplated system a privileged role, among the elements of the alphabet, is, held by the separators of words respectively the blanks (typographical spaces). Thus, an interrupted sequence of various types will form a word. In the implementation of the system it is necessary to clarify problems such as: the units of measure have to be separated from the size which they measure or not (2000 MM or 2000 M M, etc.?). It has to be given since a clear definition to the words where ambiguities can arise.

Let consider the denomination D_i consisting of the words (C_{i1}, C_{i2}, ... C_{ik} ... C_{in}). C_{ik} will be a link word if its absence in any of the D_i is semantically irrelevant. In this sense, the word WITH can be, in the context of the denominations of materials, considered a link word, while the word WITHOUT cannot be considered a link word.

And now considering:

$$D_i = (C_{i1}, C_{i2}, \ldots C_{ie} \ldots C_{in})$$

and

$$D_j = (C_{j1}, C_{j2} \ldots C_{jk} \ldots C_{jm}),$$

We will say that C_{jk} and C_{ie} are synonymous if no matter which will D_i and D_j the reciprocal substitution of C_{jk} and C_{ie} is of no effect on the physical sample shown by the respective denominations (the reciprocal substitution is semantically irrelevant). In this sense, in the context of the denominations of materials COPPER BRASS CU. BR. are synonymous.

In the implementation of the method the elimination of homonyms is strictly necessary. This can be done by marking the abbreviations with a point, etc. For instance, AL link word and AL. — aluminium, the problem being a task of the analysis before implementation.

The multitude of all the syntactically distinct words (as a sequence of symbols) and present in at least one denomination, excluding the link words, for the multitude of the denominations which cover all the physical executions of an entity will be named by us vocabulary of the system.

There is named order of synonymy of a word, noted with OS, the synonymy number of the word in the vocabulary. For the word without synonyms OS = 0, for the word with 1 synonym, OS = 1, with 5 synonyms OS = 5, etc.

The collection of all the words in the vocabulary which have OS = 0, the collection consisting in equivalence classes generated by the synonymy property is named matrix of synonyms. In each equivalence class there is a word and only one which we name it PREFERRED SYNONYM. In the display of synonym matrix, owing to reasons of retrieval facilitation, the preferred synonym will always figure the first in its equivalence class. For example:

WITHOUT, W.
COPPER, BRASS, CU., BR.
TIN, STANUM, SN., SN.

There is named word-variant any word with OS $\neq$ 0.

A denomination with the following properties:

(1) the link words are eliminated;

(2) all the words-variant are substituted by the preferred synonym,

(3) all the words from denomination are arranged in the EBCDIC sense (alphabetical order, with numbers after the alphabet)

(4) all the longer separators are reduced to a single blank. It is named standard denomination (SD). With the aid of the notion of standard denomination all denominations will be separated into equivalence classes, each equivalence class being characterized by a distinct denomination.

In order to learn which equivalence class a certain denomination belongs to, there are effected all the above operations on the source denomination. Two denominations will be equivalent if their standard denominations are *identical character by character respectively they will differ if* the standard forms are different.

There is named code of a standard denomination its number of order in the file of standard denominations. The manner of obtaining this code is based on the HASH-ADDRESSING technique which will be briefly presented, as an example:

The maximum number of words in a denomination (for materials can be considered according to each case 15, 20 or 30) will be named the lenght of denomination. The *average* length can be appreciated at 4 to 5 words in case of denominations of materials. We will name the code range its maximum value. Taking into account the conditions of new physical samples which must be codified, the range has to substantially exceed the number of physical samples at a given moment. The requirements imposed by the HASH-ADDRESSING technique presuppose, moreover, that this range should be an equal number with 2^r. In practice r can be 16, 15, 12 (respectively tens of thousands of materials for a big enterprise).

To obtain the code from the standard denomination the latter is "packed" (the representation in 8 bits of each type is dropped to the last 4 bits) and from the packed standard denomination (DSI) there is successively picked a bit from each word, the order number of the bit in the word being equal to the order number of the word in DS. The picking continues till r bits are obtained. When the average length of the denomination is smaller than r (a frequent instance) the method is cyclically resumed from the first word numbered in surplus with one, as againts the last word. The collection of r bits obtained in this way is named HASH-SOURCE and provides the raw material from which the code is built. In reality, the method of building the hash-source may differ from application to application. The aim of a technique or another is always to obtain a pseudo random method of generating a sequence of bits within a number which corresponds to the range.

For access to vocabulary (obtaining the order number of the word in the vocabulary), the hash-source from each word is built in the following way: the word is "packed" and a series of bits is built, the k number of bits corresponding to the range of vocabulary (maximum number of posts in vocabulary $= 2^k$). The transition from the hash-source to the code is made by considering the value (in binary) of the hash-source as an address. The conflicts between different SD which have identical hash-sources is solved by special techniques. Finally, after the conflicts are solved, each SD will have in the ONE-TO-ONE correspondence an address (NO OF ORDER) which will be the internal CODE.

5. DIAGRAM FOR THE IMPLEMENTATION OF THE METHOD

To implement this method there are required four data files. Thus, two files can directly be addressed to by HASH-ADDRESSING, one through direct record number (alike in FORTRAN) and one consisting of one record. According to the succession of the quantity of the necessary memorizing space, the four files are:

(1) *STANDARD DENOMINATIONS FILE* has direct access, records of fixed length equal to the maximum length of a denomination. This is

organized like a dispersed table. The post (address) number in which each standard denomination is stored constitutes the INTERNAL CODE of the physical sample identified by the respective standard denomination.

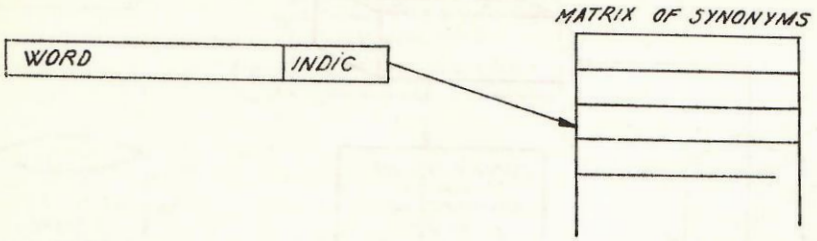


Fig. 3

(2) *VOCABULARY* has direct access based oh the address obtained by the selection of several bits from the EBCDIC coding, of the word. Each post includes, besides the recording of a word also a pointer to the record number in the matrix of synonyms, which incorporates all the synonyms. The records are of fixed length equal to the length of the longest word plus the length of the area for pointer (Fig. 3).

For the words which are not words-variant the pointer SHOWS O.

(3) *MATRIX OF SYNONYMS* is usually a small file with records of fixed length equal to the product between L_{max} word and MAX NO words from denomination. The access is direct and based on the record number. Each record includes the list of all synonyms associated to a preferred synonym.

(4) *LIST OF LINK WORDS* is a simple record of equal length to L_{max} word $\times$ NO of link words. The link words are recorded as a vector in this record.

The use of the four tables is done in a different way in the moment when the files are loaded and in the moment of their current operation. In the implementation of the system there are to be distinguished three stages:

1. Manual checking (see Fig. 4);
2. Loading of files (see Fig. 5);
3. Current operation

— Updating (code additions)

— retrievals — obtaining of internal code according to denomination.

In the figure 5 the block 2 is sophisticated enough, consisting of a whole series of programmes. It includes the sequential reading of the list of denominations and the control for each word from the each denomination against the link words. The link words are ignored. For synonyms the word is inserted into vocabulary with the respective pointer. After each word is detected and introduced in the vocabulary, the whole denomination is inserted in keeping with the hash-addressing algorithm in the code file under the standardized from described before.

The third stage assumes that the system is currently used either for the retrieval of certain internal codes, the denomination being given in one of the three forms: denomination → coding system → internal code, or for the addition of new standard denominations in the code file. The first sub-stage

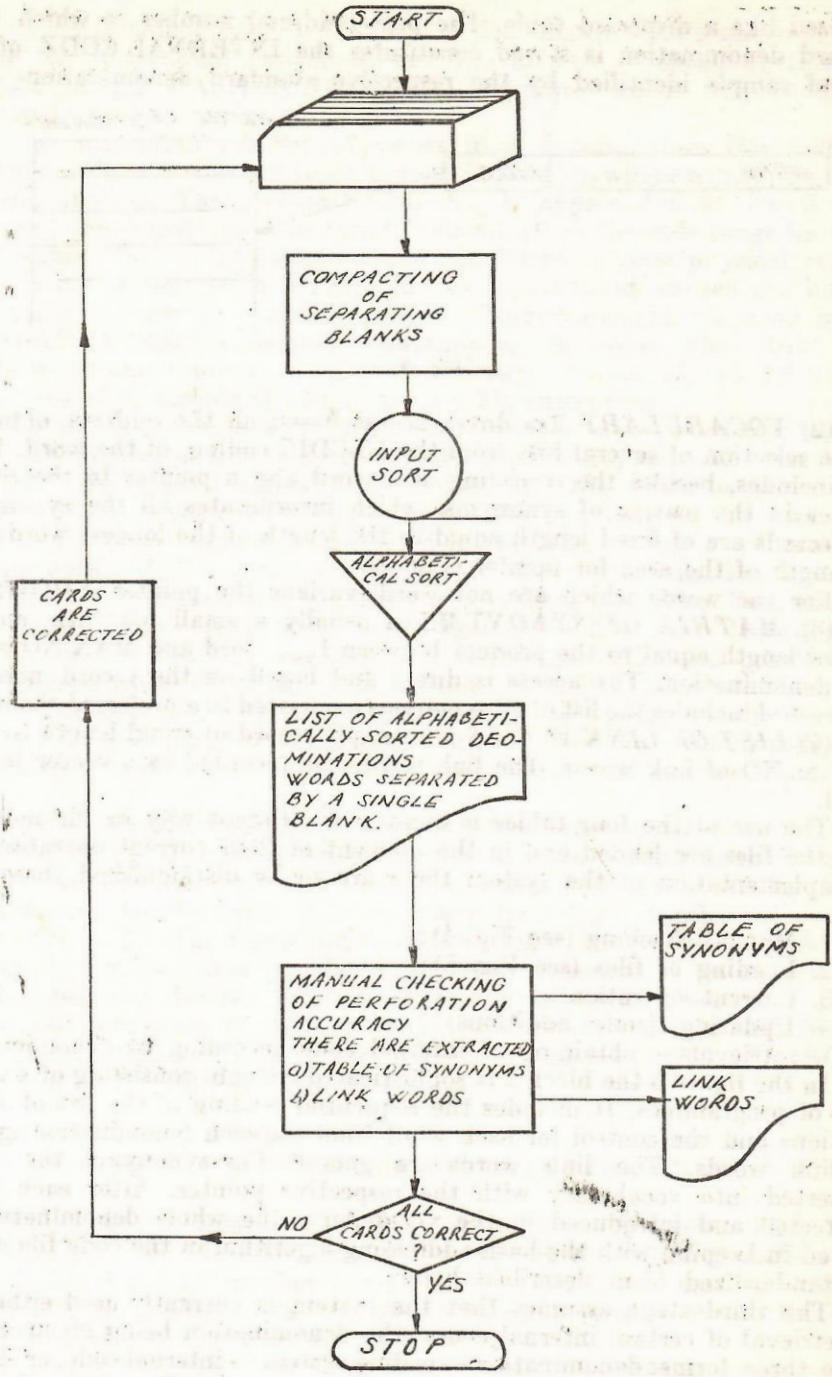

START
COMPACTING OF SEPARATING BLANKS
INPUT SORT
ALPHABETI-CAL SORT
CARDS ARE CORRECTED
LIST OF ALPHABETI-CALLY SORTED DEO-MINATIONS WORDS SEPARATED BY A SINGLE BLANK.
TABLE OF SYNONYMS
MANUAL CHECKING OF PERFORATION ACCURACY THERE ARE EXTRACTED a) TABLE OF SYNONYMS b) LINK WORDS
LINK WORDS
ALL CARDS CORRECT ?
NO
YES
STOP

Fig. 4

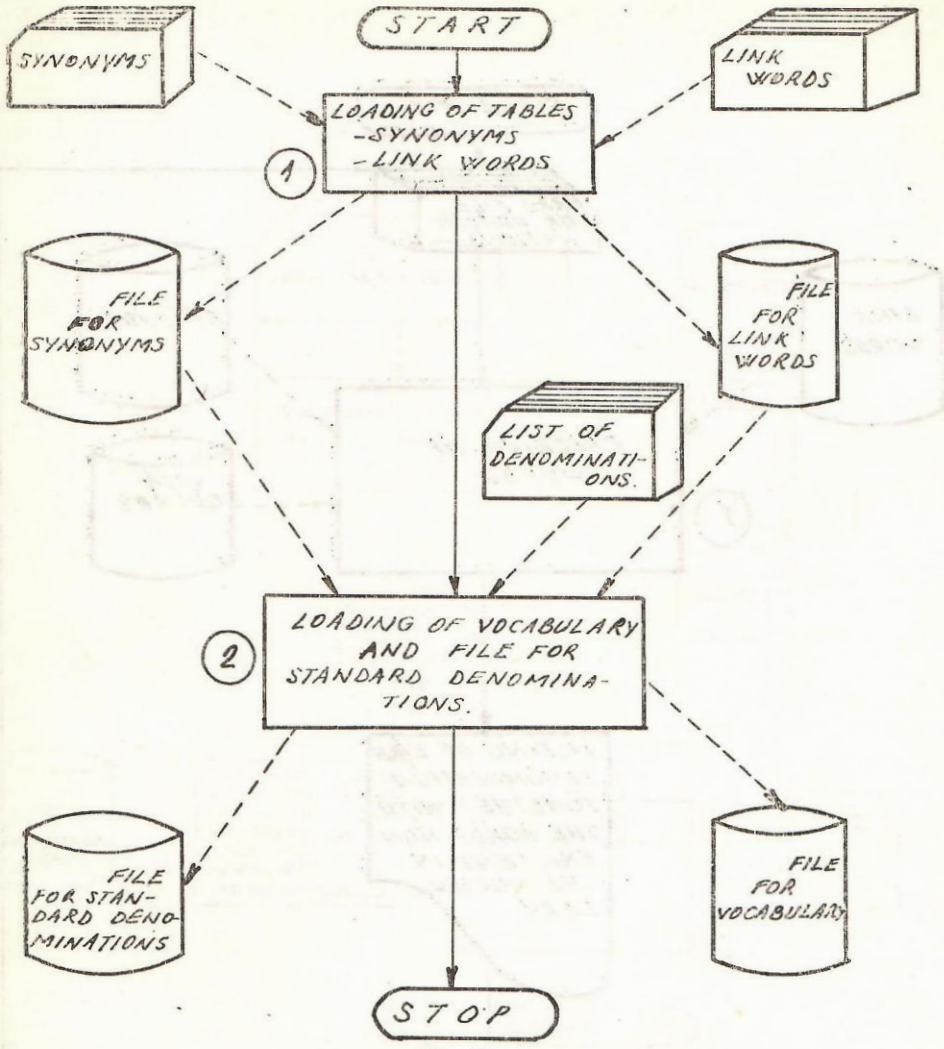


Fig. 5

from the IIIrd stage is carried out in compliance with the flow-chart from Figure 6.

In this first sub-stage the user decides whether the new words are new from the vocabulary standpoint or mere punch errors.

The second sub-stage from the IIIrd stage is carried out according to the flow-chart from Figure 7.

The method implementation flow-chart can in principal be improved by admitting the updating also in the third stage of the tables of synonyms and link words, but this can be considered in each separate case depending on the stability of the vocabulary.

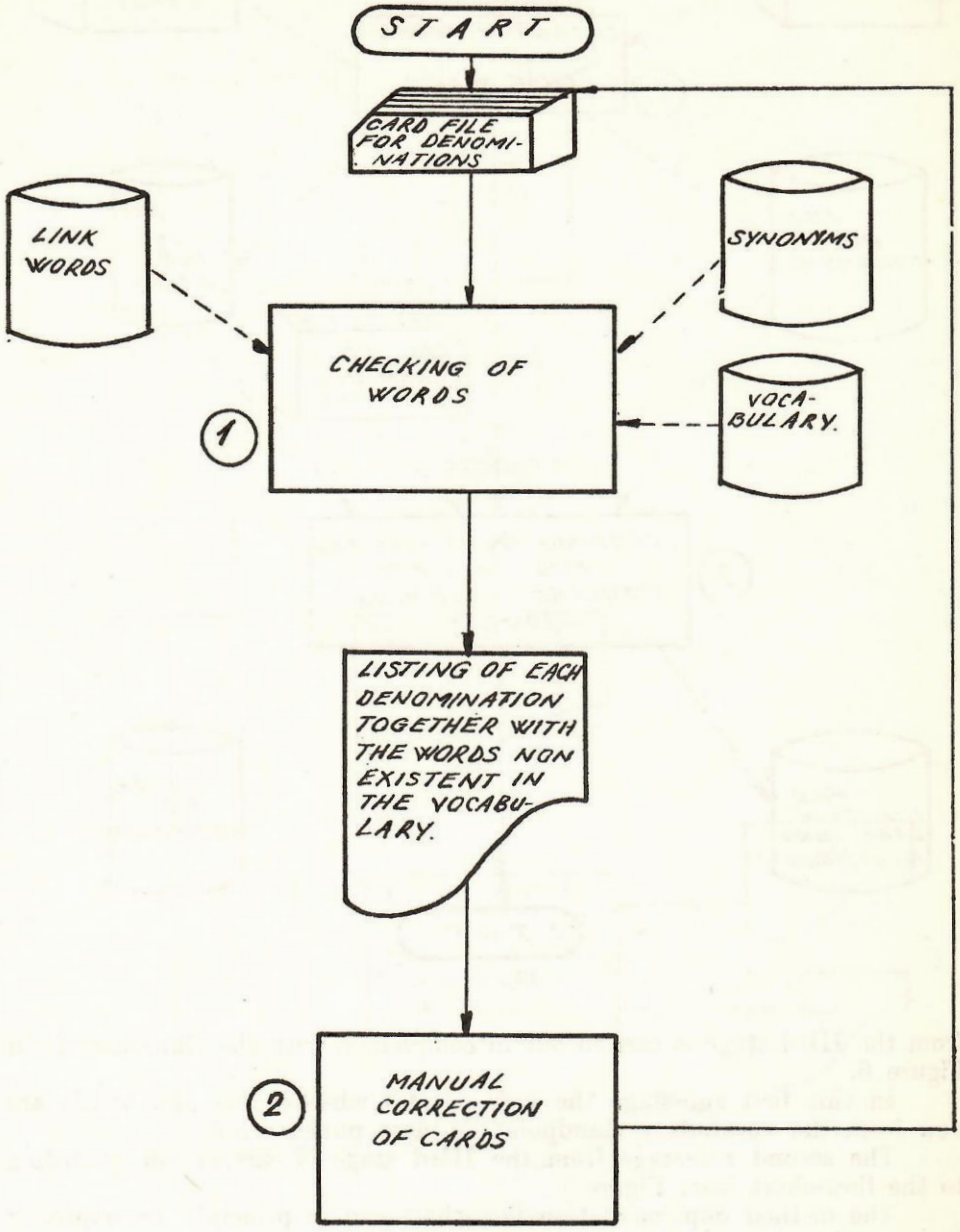

START
CARD FILE FOR DENOMI-NATIONS
LINK WORDS
SYNONYMS
CHECKING OF WORDS
1
VOCA-BULARY.
LISTING OF EACH DENOMINATION TOGETHER WITH THE WORDS NON EXISTENT IN THE VOCABU-LARY.
2
MANUAL CORRECTION OF CARDS

Fig. 6.

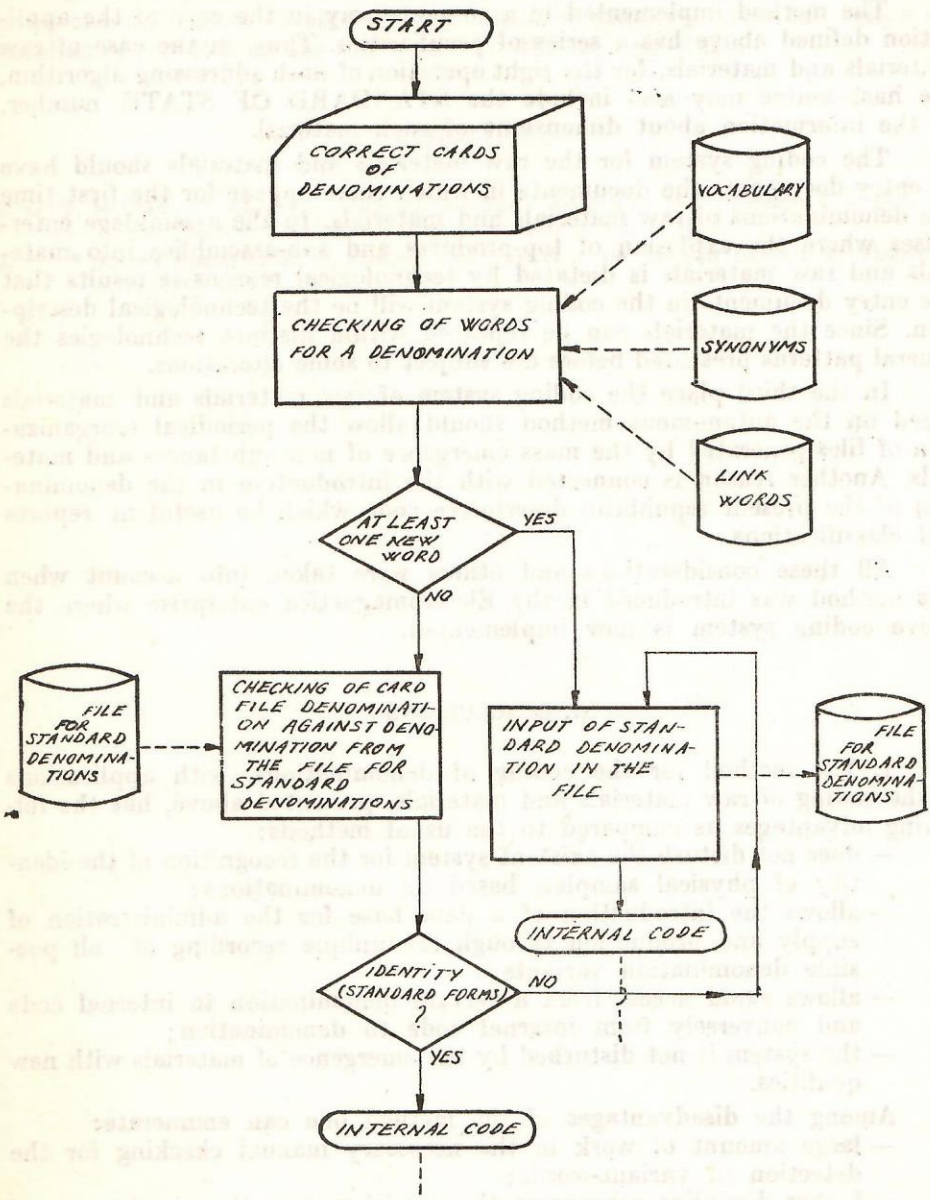

START
CORRECT CARDS OF DENOMINATIONS
VOCABULARY
CHECKING OF WORDS FOR A DENOMINATION
SYNONYMS
LINK WORDS
AT LEAST ONE NEW WORD
YES
NO
FILE FOR STANDARD DENOMINATIONS
CHECKING OF CARD FILE DENOMINATION AGAINST DENOMINATION FROM THE FILE FOR STANDARD DENOMINATIONS
INPUT OF STANDARD DENOMINATION IN THE FILE
FILE FOR STANDARD DENOMINATIONS
INTERNAL CODE
IDENTITY (STANDARD FORMS) ?
NO
YES
INTERNAL CODE

Fig. 7.

6. PECULIARITIES OF IMPLEMENTATION IN THE CASE OF RAW MATERLIAS AND MATERIALS

The method implemented in a concrete way in the case of the application defined above has a series of peculiarities. Thus, in the case of raw materials and materials, for the right operation of hash addressing algorithm, the hash-source may also include the STANDARD OF STATE number, or the information about dimensions of each material.

The coding system for the raw materials and materials should have as entry documents the documents in which there appear for the first time the denominations of raw materials and materials. In the assemblage enterprises where the explosion of top-products and sub-assemblies into materials and raw materials is dictated by technological reasons it results that the entry documents in the coding system will be the technological description. Since the materials can be repeated within distinct technologies the general patterns presented before are subject to some alterations.

In the third place the coding system of raw materials and materials based on the autonomous method should allow the periodical reorganization of files generated by the mass emergence of new substances and materials. Another reason is connected with the introduction in the denomination of the present republican descriptive code which be useful in reports and classifications.

All these considerations and others were taken into account when this method was introduced in the Electromagnetica enterprise where the above coding system is now implemented.

7. CONCLUSIONS

1. The method for the coding of denominations, with applications to the coding of raw materials and materials, presented above, has the following advantages as compared to the usual methods:

- — does not disturb the existent system for the recognition of the identity of physical samples, based on denominations;
- — allows the introduction of a data base for the administration of supply and production through the unique recording of all possible denomination variants;
- — allows rapid access from a certain denomination to internal code and conversely from internal code to denomination;
- — the system is not disturbed by the emergence of materials with new qualities.

Among the disadvantages of the method one can enumerate:

- — large amount of work in the necessary manual checking for the detection of variant-words;
- — higher discipline concerning the words' syntax, than in the manual system;
- — slower sorting and selection of standard denominations in comparison with sorting after a descriptive code.

Taking into account the advantages and disadvantages, the method is adequate to big enterprises with tens of thousands of materials.

Its nationwide generalization has to be tackled with caution being given the difficulties of setting up a table of synonyms and the range of vocabulary, as well as the difficulties of the loading stage.

REFERENCES

1. G. T. MURPHY: *Coding of Materials for Their Adaptation to the Processing of* Data in *the Electronic Computer.* Direction et gestion, no. 5, 1970.
2. * * * *General Classification of Products and Services in the Socialist Republic of Romania,* volume C.S.P., 1973.