G. DIACONESCU

Department of Economic Cybernetics
Academy of Economic Studies

An Interactive Method for Coding of Entities

Starting from the necessities of setting up a statistical data base, in the first part of the study, the author presents a description of the functions of an "ideal" coding system for large files.

Further, algorithms with an original inter-active method for the creation and maintenance of codes are described.

A. INTRODUCTION

As is known, the problem of creating and maintaining code lists represents one of the essential aspects of the implementation of a data base. As a rule, designers choose the simplest though not the most efficient solution that of pure (undescriptive) codes solely defining the instances of the coded entity and which cannot be used in complex retrievals, according to some characteristics.

In case of descriptive codes, due to the inexistence of a retrieval facility at a character level, the retrieval according to fragments in the code is excluded which makes difficult the utilization of the descriptive codes in a conversational context (they use sorting, sequential scanning etc.).

The difficulties appearing in the implementation of a solution or anothe are generated by the FUNCTIONAL COMPLEXITY of the code lists, manifested in the existence of a big number of functional demands, very often contradictory in a context or another.

A review of these functions can outline the qualities of an "ideal" list system.

1. THE IDENTIFICATION FUNCTION — of exact disociating among the instances of the coded entity;

2. THE COMPACTIFICATION FUNCTION — of reducing redundancy is linked with the necessities for an efficient use of the information physical devices;

3. THE FUNCTION OF STABILIZING THE REPRESENTATION —That function refers the code's permanency in the initially established form. A "good" code will have to appear with the instance of the entity to be coded and "change" only if the respective achievement has disappeared;

4. THE DESCRIPTIVE FUNCTION — of describing analytically some of the essential characteristics of the entity to be coded;

5. THE RETRIEVAL FUNCTION -- of association — strictly necessary within the large files in which codes appear and also within the lists themselves. Under simple aspects it is linked only with identification; the complex retrieval appears in connection with the descriptive function, for an efficient use of the descriptive characteristics in the code.

6. THE SEMANTIC FUNCTION — of facilitating the representation in the USER'S CONSCIENCE of the instance of the coded entity. Generally, the semantic function for the code lists appears simply as a NAME or COMPLEX DENOMINATION attached to the code itself.

The functions described are contradictory; thus, a coding system in which the descriptive function is well represented may have problems concerning compactification, etc.

The functions are contradictory, in the first place because they are the result of the name-code contradiction, that is if a system of signs mainly contains functions 1, 2, 3 and less 4, 5, 6 it is a system of codes and inversely, f a preponderence of functions 6, 5, 4 exists to the detriment of functions 1, 2, 3 we have a system of names.

In reality, pure codes (1 = worker; 2 = engineer; 3 = technician) and the pure names (in which the semantic function prevails, as in the musical terminology, "maiestuoso", con molto grazia") exist only in theory, because any pure code through a repeated utilization acquires function 6 (facilitates the representation in the user's conscience of the instances of the coded entity). In this case, the user is said to know the codes "by heart", that is they have become names.

At the same time, a name — as long as it does not allow the facilitation of representation in the user's conscience of the instance of the coded entity — is, in fact, a code for example, the words of a foreign language, in the stage of learning. In this respect, a list of codes is an INTERFACE BETWEEN NAME AND CODE, and a dictionary for a foreign language is a list.

The order in which the functions were presented is, in fact, the order of the difficulty of implementation. In the simple solutions, functions 4, 5 and 6 appear in the most elementary form, that is undescriptive codes with an identification retrieval and with a name attached to the code.

In the large lists of codes (effectively, the files of lists) these solutions are no longer satisfactory.

Especially, the introduction of the descriptive function, without great difficulties of retrieval is not entirely efficient; at the same time, it is extremely important that the use of the denomination part in the list (linked with the semantic interface) becomes more and more flexible, approximating as much as possible the current use of the denominations in the interhuman communication. In connection with this last demand, it appears natural to introduce facilities which should enable the computer to detect the identical semantic content of some syntactically distinct denominations.

An attempt in this direction was made by the author in the CODEN code list system described in [1] based on the reduction of the semantically equivalent denominations, as in the examples:

— copper scaled sheet
— copper sheet scaled
— sheet of scaled copper

to a standard internal denomination (copper scaled sheet) in which:

1. the order of words is lexicographically established;
2. the synonyms are replaced by a preferred synonym;
3. the linking words (semantically insignificant) are abandoned and the delimiting space is reduced to a character.

This system is currently exploited in a practical application at the technological data base for the Electromagnetica plant (however, the system has neither descriptive nor retrieval complex facilities).

Naturally, the real take over of the semantic function by the computer is extremely difficult and is far from being solvedby expedients of the sort of the previously described one.

A much more productive solution is that of an interactive man-computer system in which functions 5 and 6 are taken over within a dialogue. In this case the dialogue may become a real semantic interface.

In the following, we shall present such a solution, in which the basic principles are the following;

— the major retrieval tasks fall on the computer;
— the human operator makes at the display, simple retrieval operations at several screens;
— the establishment of the semantic content in the list is entirely incumbent on man.

The implementation means are based on the techniques algorithmized by the author; the stable descriptive coding and the character inverting of files. [2], [3], [4].

B. STABLE DESCRIPTIVE CODING

The properties of a stable descriptive code are:

1. The entity code is not an elementary but a group of elementary fields;
2. The elementary fields of the entity code are divided into two categories:

— coding characteristics — THE DESCRIPTIVE PART.
— The SERIAL (the series number of the descriptive part);

3. The coding characteristics must be:

— descriptive, defining certain properties of the coded entity;

— stable, that is not to be changed from the creation of the entity's instance to its deletion — IDENTITY CHARACTERISTICS (for example, SEX — the characteristic of a person), or in very rare cases, they are once or twice changed from the creation of the entity's instance to its deletion (for example, the number of birth certificate for a person) these are PSEUDO IDENTITY CHARACTERISTICS.

— coded alfanumerically;

a) in a MNEMONIC code — the code suggests the coded value by abbreviations, for example:

W — For Woman, M — for Man, OL for steel.

b) in an order code, the code indicates in a lexicographical order a natural order in the coded characteristic, for example 1 — worker, 2 — technician, 3 — engineer (the qualification order).

c) in the ORDOMNEMONIC code — the code suggests both the order and the coded value. For example in Romania

A — for pre-school

G — for general education

L — for high school education (liceum)

S — for higher education

U — for post-university education

d) in pure code — the code describes only one entity (1 technician, 2 worker, 3 engineer).

e) uncoded — for example the birth-day.

Of course, the most advantageous code is the ordomnemonic one but it cannot be always introduced.

4. The maximum limit of the serial (LMS) should be as small as possible (ideal 0 — the inexistent serial) so that the retrieval part left to the human operator in the inter-active system be as small as possible.

5. The entity code must not be superdimensioned. The construction of the stable descriptive code is made by successively adding stable descriptive characteristics until LMS (the maximum limit of the serial) attains the desired value (0, tens or hundreds). A subsequent addition of characteristics is said to superdimension the code.

The design and construction of CDS (stable descriptive code) can be algorithmized as follows:

a) The coding characteristics are chosen.

The selection criteria in the PRIORITY ORDER are the following:

1. *stability*; can be maximum for an identity characteristic; very great for a pseudoidentity characteristic and great;

2. *the uniformity of the distribution* over the instances of the entity to be coded (those with uniformly distributed values contribute to a greater extent to the diminishing of the serial);

3. *the shortening of the characteristics code* (a short code of characteristics contributes to the compactification of the entity code);

4. *the type of the characteristic's coding* in the following priority order; in clear (uncoded), ordomnemonic, mnemonic, ordo, pure.

5. *the statistical independence* between the characteristics.

B) The maximum limit of the serial (LMS) is calculated. The most probable (frequent) value within the entity given for the time period in which

the entity is considered, is established for each characteristic. Correspondingly, the probability (relative frequency) associated with the most probable (frequent) value is established and ΠP^i_{max} — the product of the probabilities previously established for each coding characteristic is calculated.
Thus,

$$\text{LMS} = \text{NRE} \prod_{i=1}^{n} P^i_{max}$$

where NRE represents the number of distinct instances of the entity for the considered time period and n — the maximum number of characteristics.

The registration number of the employees in an enterprise with 20,000 employees with a maximum fluctuation of 10% annually may be considered as an example of calculus of LMS for CDS. For this, a 10 year period is considered in which NRE can be estimated at 30,000. The coding characteristics SEX, BLOOD GROUP, BIRTH-DATE = MM DD are taken into account.

The enterprise has a feminine preponderence (60% women) and the most frequent age group (30 years) appears with a "probability" (relative frequency) of 0.05. We have LMS = 30,000 $\times$ 0.6, 0.4, 0.05. $\frac{1}{12} \cdot \frac{L}{34}$.

After making the calculations, the results are rounded to ten, to hundred, etc.

Practically, the estimation of LMS, due to the final rounding does not require a very exact calculation of probabilities but a sufficiently good estimation. The demonstration of the above algorithms is based on the fact that

$$\text{LMS} = \frac{\text{NRE}}{\text{minimum number of distinct values of the descriptive part.}}$$

As the minimum number of distinct values of the descriptive part is obtained by considering the product between the minimum number of distinct values for each characteristic (the hypothesis of statistical independence) and as the minimum number of distinct values in a characteristic appears if we take into consideration $\frac{1}{P_{max}}$ resulting is

$$\text{LMS} = \frac{\text{NRE}}{\frac{1}{P^1_{max}} \quad \frac{1}{P^2_{max}}}$$

where $\text{LMS} = \text{NRE} \prod_{i=1}^{n} P^i_{max}$.

Another example of a stable descriptive code might be the code of localities with the coding characteristics: county and the name of locality. The coding of characteristics can be made as follows: for the county ,we consider the usual mnemonic coding on two letters (a letter and a number for the Bucharest sectors).

For the name of the localities, the following mnemonic coding can be established — the first 2 letters of the name of the locality. The two characteristics of modification are of pseudoidentity (an extremely rare modification).

The LMS calculation is made by estimating the maximum number of distinct communes (villages) for a 10 year period (NRE). Afterwards, the maximum relative frequency of the characteristic of the county (it appears at the county with the biggest number of localities) and the maximum relative frequency of the first two letters in the name of the locality are established.

Practically, we come to LMS = 99.

C. USING THE LIST IN THE DIALOGUE

The list is based on the existence of three files.

1. *OBJECT* LIST

A registration includes:

— the stable descriptive CODE (CDS)

— denomination

2. *INVERTED* LIST character inverted on the descriptive part of CDS

3. THE CDS SYNTAX and the CDS SEMANTICS/ a text describing the model of the CDS code and then the list of values and the code for each characteristic.

In this case, the creation and maintenance of the list is done by successive up-dating, according to the method schematized as follows:

The user knows one of the semantically equivalent denominations of the achievement of entity and the *value* of the coding characteristics *(they often result from the very denomination).*

— a DISPLAY of file 3 is carried out, the user being instructed about the coding system;

— the user forms the descriptive part of CDS;

— with the aid of file 2 the retrieval is carried out in file 1 of all "related" positions (which have a common descriptive part) .The list appears on the display's screen together with the maximum serial which was until that moment granted to the "relatives";

— the user compares the denomination of the realization of the entity to be coded to the denominations existing on the screen (fig. 1).

In case the list is exploited for retrieval, a simple situation is that appearing in establishing the denomination when CDS is known.

More frequent, however, is the use of the *coding characteristics* (of their value) for finding out the code. In this case, the scheme is similar to that above, but, the file updating is not practised. The detail solution also comprises the 4th file, the file of PROBABLY NECESSARY CODINGS which results from retrievals and which are carefully examined before being entered in the list.

Practically, there are frequent cases when the coding characteristics appear explicitely or implicitely in the denomination. In these cases, the retrieval is used to ESTABLISH THE CODE BY STARTING FROM ONE OF THE EQUIVALENT DENOMINATIONS.

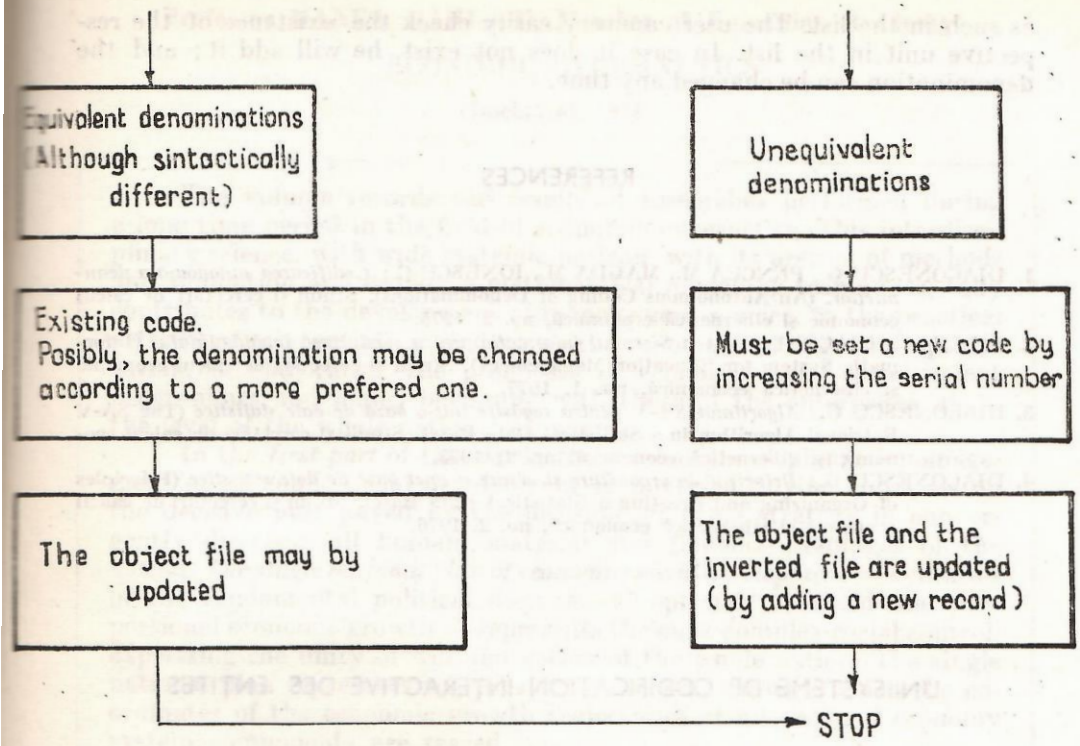

Fig. 1

D. AN EXAMPLE OF UTILIZATION

The code of the educational institutions can be considered as an example of stable descriptive code.

The coding characteristics are:

— the level of institution with the code values: A, G, L, S (ordomnemonic);
— county — code on 2 letters (mnemonic code);
— locality — the first 2 letters of the name of the locality (mnemonic code).

LMS — the maximum limit of the serial is 99.

A separate serial for the locality and another for schools can also be used.

The above text would represent the information from file 3 (syntax and semantics).

Supposing that the institutions :

— the part-time day nursurey attached to the "Avîntul" factory in Timișoara;

— the nr. 23 Kindergarten Timișoara;

— the "Hawcks of the homeland" kindergarten-Timișoara represent one and the same thing after their names are changed, they will be treated

as such in the list. The user can very easily check the existence of the respective unit in the list. In case it does not exist, he will add it; and the denomination can be changed any time.

REFERENCES

1. DIACONESCU G., PENCEA M., MAGDA M., IONESCU C.: *Codificarea autonomă a denumirilor.* (An Autonomous Coding of Denominations). Studii şi cercetări de calcul economic şi cibernetică economică, no. 3, 1975.
2. BILCIU C., DIACONESCU G.: *Sistemul informatic pentru conducerea învăţământului* (Informatic System for Education Management). Studii şi cercetări de calcul economic şi cibernetică economică, no. 1, 1977.
3. DIACONESCU G.: *Algoritmul SA-V pentru regăsire intr-o bază de date statistice* (The SA-V Retrieval Algorithm in a Statistical Data Base). Studii şi cercetări de calcul economiç şi cibernetică economică, nr. 1, 1975.
4. DIACONESCU G.: *Principii de organizare şi creare a unei baze de date statistice* (Principles of Organizing and Creating a Statistical Data Base). Studii şi cercetări de calcul economic şi cibernetică economică, no. 2, 1976.

UN SYSTEME DE CODIFICATION INTERACTIVE DES ENTITES

Partant de la nécessité de créer une base de données statistiques, l'auteur présente, dans la première partie de l'article, une description des fonctions d'un système "idéal" de codes pour les grands fichiers.

Dans la suite, il décrit plusieurs algorithmes à l'aide d'un procédé itératif original de création et de maintenance des codes.

СИСТЕМА ИНТЕРАКТИВНОГО КОДИРОВАНИЯ СУЩНОСТЕЙ

Исходя из необходимости создания банка статистических данных, автором даёсят в первой часто рабиты описание функций «идеальной» системы кодирования для крупных файлов.

Далее описываются алгоритмы с оригинальным интерактивным способом создания и сохранения дерэв.

24. IX. 1977.