FOR PARTICIPANTS ONLY

REFERENCE DOCUMENT
DDR/27
30 April 2001

ENGLISH
ORIGINAL: SPANISH

ECLAC
Economic Commission for Latin America and the Caribbean

First meeting of the Statistical Conference of the Americas
of the Economic Commission for Latin America and the Caribbean

Santiago, Chile, 9-11 May 2001

COORDINATION OF THE NATIONAL SYSTEM OF NOMENCLATURES

Computerized Codification System (SiCI)
for economic and sociodemographic operations

This document was prepared by Mariano Lanne and Mara Riestra, coordinators of the SiCI project at the National Institute of Statistics and Censuses (INDEC), Argentina.

01-2-188

Introduction

The creation of a computerized codification system first arose as a concern in connection with the coordination of the National System of Nomenclatures at the end of 1998. Prior to that time, the concern had been to obtain classifications more appropriate to national needs. Explanatory notes and dictionaries were also developed to facilitate classification procedures while documenting the decisions taken after various consultations. However, the concern remained that many of the manual codification procedures were tedious, that they did not provide sufficient experience for the codifiers, that long working days were spent codifying the same thing over and over again, and that there were discrepancies between the criteria applied by different codifiers. There was also little time left to discuss the more difficult cases. In the particular case of massive operations such as the censuses, these difficulties resulted in a high demand for human and monetary resources and extremely long periods of time for codification, so that the information was slow in reaching the users.

The coordination of the Census 2001 was just the framework needed for the above-mentioned concern to find a response. In April 1999, the Group for Nomenclature Application 1 was established in connection with the working methodology suggested by SiNN, made up of people working in the Activities and Products department, the Programme for Measurement and Analysis of Occupational Structure and in other areas such as the Directorate of Statistical Methodology, the Department of Cartography, the Directorate of Informatics and the Census Team. We believe that this was the main factor which made it possible to move forward and achieve the results that we have today. It was this multidisciplinary union that succeeded in creating a system which is far from complex. The main aim is to obtain a high level of quality in the network of dictionaries that feed the system.

The project basically consisted of conducting a very detailed study of the codification methodology applied in the manual processing of each of the variables to be codified. For this, a working system was designed to help the SiNN "codifier" to work in a methodical and standard way, and which, as well as clarifying all the steps involved in assigning a code, would produce the tools needed to design the SiCI (Computerized Codification System), namely, the dictionaries.

The present document summarizes the experience gained to date. It is divided into three parts. The first is an introduction to the SiCI, followed by a review of its objectives, and a third section presents the work programmes to which it is applied. The second part refers to the conceptual model, defining the basic concepts and explaining the stages involved. The third and final part presents the results obtained when testing the SiCI with the Pergamino Experimental Census.

1. - About Computerized Codification

As already mentioned in the introduction, the project for design, development and implementation of the SiCI requires a significant initial effort as the first step for any system of this kind is to "model" the process which is to be systematized. The SiCI is thus a system which uses various methods to recreate the whole set of "intellectual" procedures that the codifier goes through when he reads, interprets, analyses and assigns a code to the phrase before him.

If we consider carefully the way in which we refer to the codification process, we note that automatic and computerized codification is not the same, as the latter has a broader meaning. We use the term automatic when a code can be assigned without the intervention of a person;[1] whereas computerized codification includes the latter, as the code may be assigned in an automatic, assisted or semi-manual way. The cases we deal with can not always be resolved in a standard way, and some do not even appear frequently as we only find them in universal censuses. For this reason, those cases which cannot be modelled require semi-manual codification. Once the codification has been resolved it can be done automatically in future surveys, depending on the context of the response.

The automatic codification procedure is based on the application of a set of previously codified phrases, in such a way that cases that recur are resolved in the same way. A basic tool required for this procedure is a DICTIONARY, which is a set of previously codified cases. We shall see that there is actually not just one dictionary, but a set of dictionaries, which interact in the codification procedure.

2. - Objectives

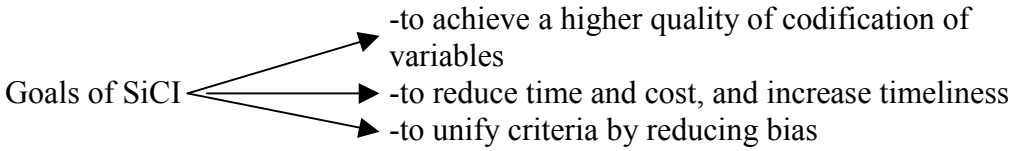
The main objective of the SiCI is to codify various variables for statistical use. The variables to be codified are the so-called "open" responses, that is, those for which there is no pre-codification in the form and where the informant responds with his own words to an item in the questionnaire. This means that different people who have the same occupation, carry out the same tasks and work for the same company may describe their activity and occupation in different ways. The codification of "closed" variables does not generally require too much effort, as they already contain a kind of codification. In many cases, however, the closed variables will be used as a complement for codifying the open variables. As an example, some of the variables to be codified are:

- activities
- name of occupation
- description of task
- geographical variables
- university courses

[1] In fact the codification was carried out by the team from the National System of Nomenclatures during the stage of dictionary generation.

We can see that there is a significant difference between the variables for activities and occupations on the one hand and the geographical variables and university courses on the other. This difference lies in the infinity of possible responses that could be received in the first case; whereas in the second the number of different possible responses is much more limited. For this reason, the computerized codification of the variables of the first group may be more complicated than for the variables of the second group. In this document we shall refer mainly to the computerized codification of activities and occupations, but the conclusions reached are generally applicable to the codification of other variables.

The objective of the SiCI is not only to codify these variables, but also to reduce codification times and to unify interpretation criteria. Differences in the criteria adopted by each of the codifiers is one of the problems that adversely affects the quality of the results of a survey or census. However, by applying uniform criteria in a computerized system, all similar cases can be treated in the same way; while it also allows rapid recodification if the criterion needs to be changed. This same task, if carried out manually, would require an excessive degree of effort and expense. In a computerized system, the cases which have different possible interpretations can either be codified by applying a specific criterion, or automatically grouped for subsequent codification. In summary:



3. - Applications

3.1. - The 2001 National Population and Housing Census

It is calculated that this census will cover approximately 37 million persons, 12 million of whom are working. The figures are too high to consider manual codification, as a very large number of codifiers would be needed, or else a sample would have to be taken, as in the 1991 census, or else timeliness of the results would have to be disregarded, or, worst of all some of the questions would have to be removed. Under these circumstances, computerized codification becomes a necessity. However, the 2001 census is not the only purpose of the Computerized Codification System, as it is already planned to use it in various work programmes, the main ones being the codification of the Permanent Household Survey, the National Economic Census and the National Directory of Economic Units.

The task of coordinating the codification of the Census does not refer only to producing appropriate codification software, but also to a set of tasks related to codification which

influence to a significant extent the quality of the data. This requires constant interaction with specialists in various areas related to the census, including in:

The company responsible for optical reading of the forms. The SiCI supplies the dictionaries of words, and updates them on a daily basis during the reading period.

Statistical methodology. Another task related to codification is to help decide on the method for measuring data quality, a task which is carried out jointly with those working in statistical methodology. They in turn are responsible for developing the codification method known as "scores", which will be explained below.

The codifiers. This stems from the need to train personnel in the use of the software as well as in managing the classificatory tools and in the actual criteria for codification. This has to be done as team work, in order to detect errors of interpretation and to distribute the work in the most efficient manner.

Analysis and consistency. In some cases codification is not possible unless certain data are already consistent, and in other cases the same level of consistency requires prior codification. One example is that geographical variables require prior codification to be consistent.

Computer science. Finally, it would not be possible to develop the software without constant interaction with computer specialists, both at the stage of system development and during codification of the census in order to ensure continuous updating of some of the dictionaries, as well as feedback and calibration of the system.

3.2. - The National Economic Census (CNE)

Another large-scale operation that is a challenge for application of the SiCI is the codification of the variables "activities" and "products". On the one hand, this is because although the number of responses can be reduced to about 1,500,000 cases, the level of disaggregation required for codification increases substantially. This necessarily implies increasing the level of detail of the dictionaries and also the speed of processing and codification as the results have to be available within four months following the census operation.

For this application, the primary sources for compiling the codification dictionaries, in addition to the 1994 National Economic Census, are the original records collected by the Directory and the surveys carried out by other sectors of the national statistical system, such as the industrial survey, the national industrial register and the surveys of the Secretariat of Agriculture, Fishing and Food, to name a few.

Continuing with the classification of activities, in this operation, the name of the company is an important piece of information at the time of defining the codes, as it might be possible to pre-codify the companies prior to going into the field, so that the number of cases for codification is significantly reduced.

Finally, the questions on products have created a new task for the SiCI, which is to incorporate product classifiers, and requires the development of new dictionaries.

3.3. - The Permanent Household Survey

The Permanent Household Survey is the main source of descriptors for almost all the variables which it is planned to codify. However, it is now one of our main "users". A continuous survey means that the codification system is in constant use. In addition to the benefits it brings in terms of the rapid availability of the Survey's results, it makes it possible to calibrate the system, and establishes the background materials needed for the census operation. This continuity will also compensate for the efforts made in the last few years. Currently, a pilot test of the SiCI is being carried out on the last available Permanent Household Survey. It is hoped to have the results in the second half of February.

3.4. - The National Directory of Companies

The National Directory of Companies is, together with the 1994 National Economic Census, the main source of original records for codification procedures for activities taken from economic surveys. The SiCI will help the Directory to collect information on activities and products, and to the extent that this information is in the dictionaries, the codification will be automatic. Otherwise, the procedures for assisted and semi-manual codification will be activated, which will result in an improvement in the SiCI codification dictionaries while enhancing the quality of the codification of those variables in the Directory, thus reducing the supervisory work involved.

4. - Definitions

The SiCI was an "original" creation in the sense that, owing to the lack of bibliographic materials, a system had to be developed from scratch, including the terminology used. For this reason, despite the limited scope of this document, the present section has had to be included on definitions of the terms to be found in the text.

SiCI: a network of dictionaries of different kinds, interlinked by linguistic and codification procedures. Using this system, the set of records containing the original texts with the variables to be codified are transformed into descriptors to which different codification methods are applied in order to assign to each one a single appropriate code.

According to this definition, the three basic elements of SiCI are:

- Dictionaries
- Linguistic procedures
- Codification procedures

4.1. Dictionaries: are ordered lists of words or phrases which make up the basic tools of SiCI and are based on the empirical responses collected in each of the operations serving as a source. Two types of dictionaries coexist in the system: those which are used to process the words, and the codification dictionaries.

Dictionary of spurious words (E): a set of words which do have a literal meaning, but are irrelevant for the purposes of codification. Examples of words in this dictionary are: numbers, proper names (except for the names of companies, which may be used to assign a code); geographical names; adjectives which are not relevant for the purposes of codification such as colours, sizes, adjectives relating to places or shapes, etc.; isolated letters and roman numerals; and other words which have a meaning but are not necessary for codification. These usually occur with low frequency.

Dictionary of deleted words (A): a set of words which have no meaning. They are the result of typing, reading or writing errors and it is not possible to correct them by substituting another word. They are usually formed by the division of a word. For example: let us suppose that the word "computer" is divided into two parts: "compu" and "ter". The first part can be useful as it seems to relate to computing, and so it will not be in the dictionary of deleted words. The second part "ter" cannot be related to anything specific or with very much at all, and so it will be part of the dictionary of deleted words. Unlike the spurious words, these "quasi-words" do not have a meaning and are therefore not part of the reading dictionary.

Dictionary of connectors (C): a set of articles, prepositions, and other words which are used to form a phrase, but are not relevant for codification purposes. Examples of connectors are: **and, the, with, by**, etc. On the other hand, the connectors **not, for** and **except** are relevant for codification and are therefore not included in this dictionary.

Dictionary of exceptions (X): a set of connectors whose presence in a phrase may affect the codification of the phrase and therefore are not part of the dictionary of connectors. At present it consists of three words: **not, for** and **except**.

Correcting dictionary (R): a set of relationships between incorrect and correct words. The incorrect words may be due to typing or spelling errors, abbreviations or other causes, but **are always related to one and only one correct word**. For example: the incorrect word "foodstoffs" will have the equivalent correct word "foodstuffs"; the incorrect word "gral" will be replaced by the correct word "general". One case in which it is not possible to maintain an incorrect-correct word relationship is "art", as in one context it may be understood as the abbreviation of "article" whereas in other contexts it may be the abbreviation ART, for Aseguradoras de Riesgo de Trabajo (Work Risk Insurers).

Dictionary of correct words (D): a set of words that are written correctly, are relevant for codification and therefore are not included in any of the previous dictionaries.

Reading dictionary (L): is formed by combining the following dictionaries: spurious, connectors, exceptions and correct words.

$$\mathbf{L = E + C + X + D}$$

Codification dictionary: is the list of phrases and words associated with each code, and is used to calculate the heuristic weight used in the scores method, which is described below. The component elements are: the dictionary of correct words (D) and the dictionary of exceptions (X).

4.2. - Linguistic procedures: are procedures which modify the text of the phrases to be codified, by simplifying the vocabulary and reducing the number of words involved. The term **text** or **descriptor** refers to the original response provided by the informant, whether this was a person or an economic unit (company, local unit, etc.). The linguistic procedures applied to the texts include:

Normalization procedure: consists of removing non-valid characters from the phrases in the database for the three variables (activity, occupation, and task) and converting the phrases to upper case.

Semantic fields or family grouping: consists of assigning to a word taken as a reference word (referred to as a father), a list of words which are accepted as synonyms (referred to as sons).

Standardization procedure: consists of processing all the words in the dictionary and eliminating the indications of number and gender, as appropriate, in order to produce a dictionary of unique terms (without repetitions).

Without standardization: when the previous procedure is not carried out.

It is important to emphasize that the linguistic procedures are not codification procedures.

4.3. - Codification procedures: these operate in different ways depending on the case to be dealt with. They are procedures which are based on the "modelling" of the analytical procedures carried out by the codifiers when assigning a code.

Macroprocedure: is a set of instructions which are modelled by computerized statements and which make it possible to divide the universe to be codified into large groups. This division then makes it possible to choose the range of possible codes. Examples are the macroprocedures for "boss" in occupations and for "sales" in activity, as will be explained in the relevant section.

Microprocedure or automatic procedure: is a set of instructions which are modelled by means of computerized statements and which make it possible to codify a particular text without the intervention of codifiers. Strictly speaking, unlike macroprocedures, these are methods of codification.

Third Generation: a form of codification, an element which indicates which variables were used in the codification of a specific variable.

Autophrase: is a method of automatic or direct codification which allows a single code to be assigned without the intervention of codifiers. For this, a codification dictionary is used that is made up exclusively of phrases which have a single coding option and are independent of the other variables in the questionnaire.

Scores: is a method which combines two elements. On the one hand there is the "specificity" of each word with respect to the different codes. For example the word "milk" is more specific than "fabrication", as a small number of codes are associated with the first word whereas the second is used more generally in all branches of industry. The specificity of each word in the dictionary is measured by the so-called "heuristic weight" which is also part of the dictionaries together with the texts and the codes. On the other hand, the score also analyses the relationship between the phrases in the dictionary and the phrases to be codified. When a phrase is to be codified, the "score" makes it possible to choose "candidate phrases" within the "range" offered by the dictionary. These "candidates" are chosen taking into account the greatest number of common words between the phrase to codify and the phrases in the dictionary. The greater the coincidence between both types of phrases the greater the "score".

Autoword: is an automatic or direct codification method which allows a single code to be assigned without the intervention of the codifiers. For this, a codification dictionary is used which is formed exclusively by words and the associated codes according to the phrase from which they come.

Assisted: is an indirect codification method which allows a single code to be assigned with the intervention of the codifiers. In this case, the SiCI offers a choice of a small number of options that are automatically generated.

Semi-manual: is a codification method which allows a single code to be assigned with the intervention of the codifiers. In this case, in view of the large number of alternatives that are possible, the SiCI offers elements to assist the codifier without making automatic suggestions.

5. - STAGES OF THE SiCI

In order to simplify the explanation of the different stages of the SiCI, the 2001 National Population and Housing Census has been taken as an example. The general diagram shown on the next page may thus be slightly different when other operations are being processed, such as the Permanent Household Survey or the National Economic Census. There follows a very brief overview of the SiCI, and more details are given from section 5.1 onwards.

There are two main periods of work in the implementation of SiCI that are quite different.

First period: dictionary development. This is the stage prior to carrying out the operation which is to be codified. One of the characteristics of this system is the importance given to spelling correction. For this, different dictionaries are generated with the sole aim of correcting the descriptions that appear in the field operation to be codified. Naturally, spelling mistakes, abbreviations, acronyms and different styles of writing are such that phrases with the same meaning are not quite the same and therefore cannot be codified automatically. The system incorporates all of those corrections which re-occur in order to enhance the interpretation of the descriptions. This stage is one of the most tedious as it basically consists of codifying, correcting and establishing connections between most of the words and phrases from different sources in order to obtain the DICTIONARIES at the end of the whole process.

Second period: preparation of the databases to be codified and codification. When the file is received from the operation to be codified, a series of stages takes place to prepare the databases for codification. As the dictionaries obtained in the first period are used for codification, it is important to complete a set of tasks that will give the file to be codified the same characteristics as the dictionaries. The figure on the following page shows the sequence of steps that occurs:

1. The words not contained in the reading dictionary are identified. The aim of this step is to send the up-dated **reading dictionary** to the company that reads the cells.

2. It is verified that the structure of the databases is as required by the SiCI and that the phrases are normalized. This consists of eliminating extraneous characters and characters not valid for codification (commas, double spaces, full stops, etc.).

3. The spelling is automatically corrected using the correction and deletion dictionaries.

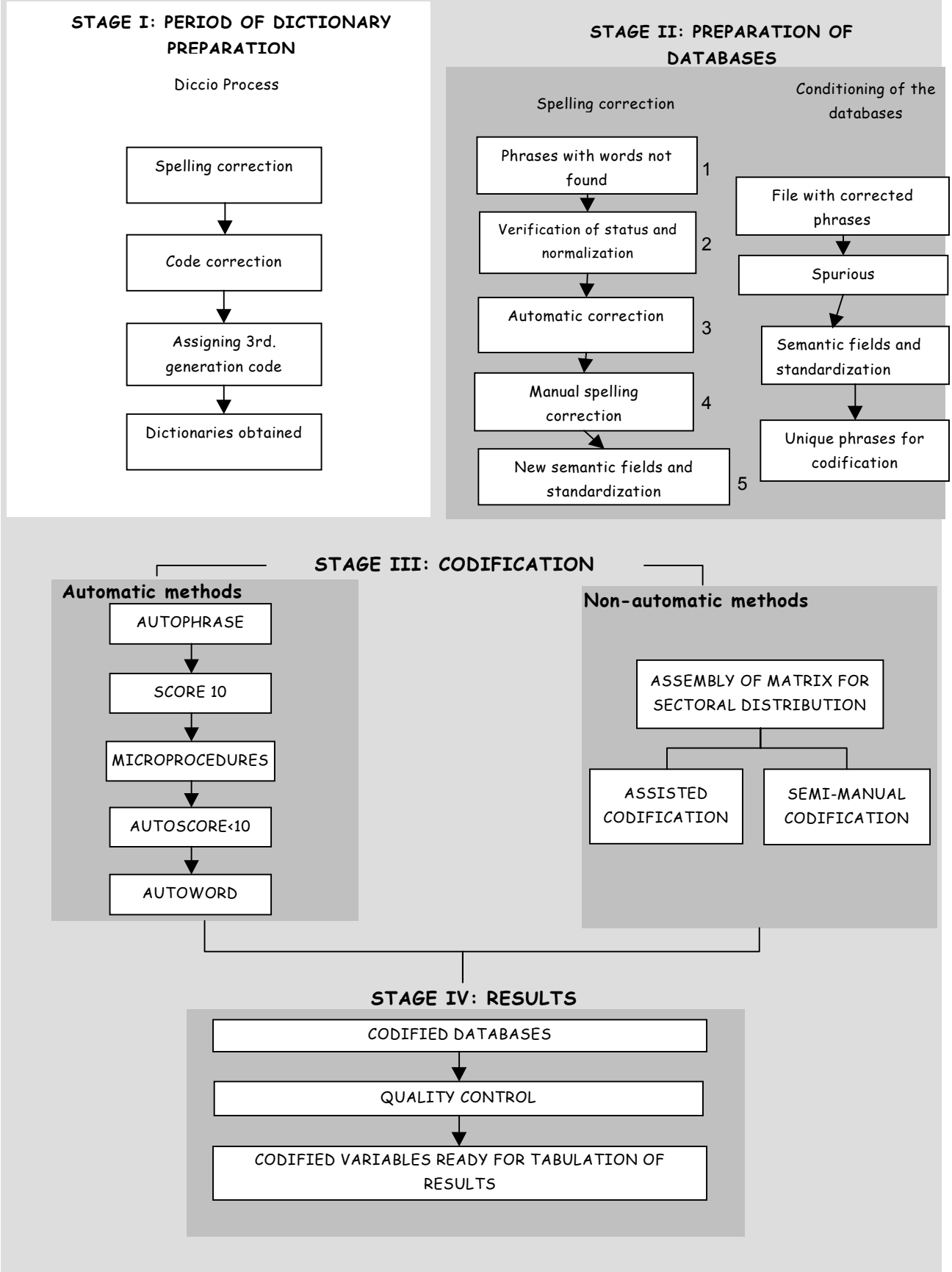
4. Spelling which could not be corrected automatically is corrected manually, beginning with the most frequent errors.

5. If necessary, new semantic fields and standardization procedures are created for new words.

6. The file is assembled with the phrases corrected in the descriptive stages and the stage begins of “preparing the databases for codification”.

7. Words which are not useful for codification purposes are eliminated using the **dictionary of spurious words**.

8. The phrases to be codified are modified by simplifying the words. The semantic fields are applied, that is, different words which have the same meaning for codification purposes (son words) are linked to a general word (father word). Standardization also takes place, that is, the genders and numbers of the words are removed, leaving only the root. Finally, if some words within the phrase are repeated as a result of the previous modification, there is a simplification procedure (see section II.2 for more details).


STAGE I: PERIOD OF DICTIONARY PREPARATION
Diccio Process
Spelling correction
Code correction
Assigning 3rd. generation code
Dictionaries obtained
STAGE II: PREPARATION OF DATABASES
Spelling correction
Conditioning of the databases
Phrases with words not found
1
Verification of status and normalization
2
Automatic correction
3
Manual spelling correction
4
New semantic fields and standardization
5
File with corrected phrases
6
Spurious
7
Semantic fields and standardization
8
Unique phrases for codification
9
STAGE III: CODIFICATION
Automatic methods
AUTOPHRASE
SCORE 10
MICROPROCEDURES
AUTOSCORE<10
AUTOWORD
Non-automatic methods
ASSEMBLY OF MATRIX FOR SECTORAL DISTRIBUTION
ASSISTED CODIFICATION
SEMI-MANUAL CODIFICATION
STAGE IV: RESULTS
CODIFIED DATABASES
QUALITY CONTROL
CODIFIED VARIABLES READY FOR TABULATION OF RESULTS

9. Many of the phrases to be codified will be the same as a result of the above procedures. A file is then obtained of unique phrases (without repetitions) to be codified. In practice, the database of activities has been reduced by about 67 per cent. This means that 35 000 phrases have been reduced to 11 000.

Once the database of unique phrases has been obtained the stage of actual codification begins. This consists of applying sequentially the different methods of codification, which are explained in detail below, beginning in paragraph 5.3. The fact that the methods are applied sequentially does not only imply that there is an order of application, but also that if one method has successfully assigned a single code, the phrase is already codified and will not be codified again. Thus, each method codifies the residual database it receives from the previous method.

Finally, at the stage of autoscore < 10, the system offers more than one code option. If there are three or less options, the "assisted" method of codification is used; if there are more than three the base is sent on for "semi-manual" codification.

The codified database obtained as a result of all these stages is submitted for quality control and finally the data are transmitted to the census processing office.

CODIFICATION ONLY TAKES PLACE AT STAGE III

5.1. - Stage I: Dictionary development

The dictionaries are the foundation of the computerized codification system as they are used in all stages of its operation. This chapter focuses on the method for creating these dictionaries in a sistematized form.

An error in a dictionary will be reflected in the codification of an activity or occupation as many times as it appears for codification. It is thus important to have an error-free dictionary. Unique codification criteria can be applied through the dictionaries in order to avoid the use of different interpretations. An error in the dictionary is multiplied by automatic codification. In any case, whether the error is in the dictionary or in the decision to adopt a criterion other than the one in the dictionary, the latter can be corrected and the codification system can be applied again to those records requiring modification. At the beginning it was mentioned that there is not just one dictionary, but a set of dictionaries; we shall see how they emerge.

5.1.1 Sources for the dictionaries

When a dictionary is used for codification, the base to be codified and that of the dictionary should be as similar as possible. The dictionaries should therefore be assembled from field records which have similar responses to those in the base to be codified. For example, if a sociodemographic survey is to be codified it is appropriate to use mainly records that come from programmes with a similar range. This does not mean

that records from economic surveys cannot be used, and certainly does not mean that they would not add to the dictionary, but it is more likely that the responses in two surveys in the same area will be more similar. Also, parallel dictionaries can be created, that is, one dictionary which uses records from one source to codify a particular type of survey and another which is made up of records from other sources to codify another type of survey. The following sources are used to create the dictionaries:

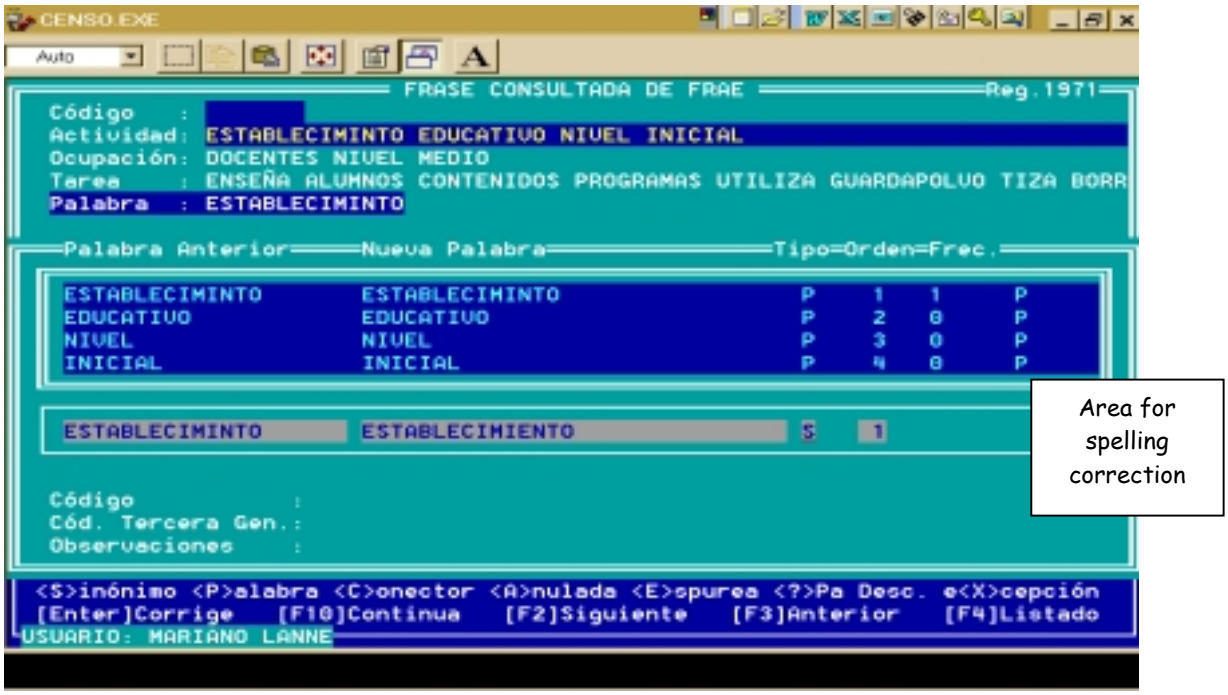
- the Permanent Household Survey;
- the third and fourth pilot tests for the 2000 Census;
- the sample from the 1991 Population and Housing Census;
- the National Directory of Economic Units.

5.1.2 "Diccio" procedure

In order to assemble the dictionaries from the sources mentioned above, a screen has been developed to systematize the codification, spelling correction and application of the third generation code.[2] This screen is not part of the computer codification as such, but is part of a stage that occurs prior to codification. This stage was called the "Diccio" procedure and it is during this procedure that time and resources are invested in the correction and coding of the databases which will then be part of the dictionaries. This is one of the longest and most tedious stages, as it requires the revision or coding of the databases chosen to form the dictionaries. The decisions taken at this stage serve as a guide for the SiCI on how to act in specific cases.

The work carried out by means of this screen (diccio procedure) could be carried out in any table or file. However, it is advisable to use a screen where the tasks are systematized and there is automatic management of the databases that make up the dictionaries. It also offers greater security when using the databases. The next page shows a model of the screen for the correction module.

[2] The third generation code indicates which elements were taken into account when assigning the code to each codified phrase. See paragraph 5.1.2.3.



PHRASE FROM THE FRAE (DATABASE OF PHRASES)

Code:
Activity: ESTABLSHMENT EDUCATIONAL LEVEL INITIAL
Occupation: TEACHERS LEVEL MEDIUM
Task: TEACHES PUPILS CONTENTS PROGRAMMES USES WHITE COAT CHALK
Word: ESTABLSHMENT

Previous word	New word	Type	Order	Freq	
ESTABLSHMENT	ESTABLSHMENT	P	1	1	P
EDUCATIONAL	EDUCATIONAL	P	2	0	P
LEVEL	LEVEL	P	3	0	P
INITIAL	INITIAL	P	4	0	P
ESTABLSHMENT	ESTABLISHMENT	S	1		

Code :
Third Gen. Code :
Notes :

<S>Synonym <P> Word <C>Connector <A>Deleted <E>Spurious <?>Unknown W e [X]ception
[Enter] Correct [F10] Continue [F2] Next [F3] Previous [F4] List
USER: MARIANO LANNE

5.1.2.1 Spelling correction

The aim of the spelling correction is to obtain "correct" dictionaries, although when correcting the records which make up the dictionaries, these will then be different from the records to be codified, which contain spelling errors. This is the reason for the first dictionary, which we shall call the "Correcting Dictionary". This is made up of a set of ordered pairs of words (an incorrect word and the corresponding correct word), which are acquired during the process of spelling correction. The spelling correction is important for reducing the size of the codification dictionaries.

It is important to distinguish the corrections which may be made using the context of the phrase to help define the correct word associated with the incorrect word, from those cases where the incorrect word corresponds uniquely to one correct word regardless of the context of the phrase. For example, in the case of the word "establshment", there is no doubt that it corresponds to "establishment", and so it can be generalized and automatically corrected in all cases where it appears. But if the word "art" appears, the context of the phrase may indicate that it can be related to the word "article". However, in other cases it is a correct word "ART", the abbreviation for Administradores de Riesgo de Trabajo. The correcting dictionary should only consist of those cases where the word can be corrected automatically.

The deleting dictionary is composed of deleted words, which are those that do not have any meaning because they do not actually exist. It does not include those words that contain spelling errors and cannot be included in the correcting dictionary because they correspond to more than one correct word.

5.1.2.2 Codification

The source databases used to create the dictionaries have in many cases already been codified by the entities that provided them. The quality of that codification then has to be reviewed in order to guarantee the accuracy of each code contained in the dictionaries, and to resolve any problems arising from differences of criteria. This stage of codification correction takes place in the same screen as shown above and is subsequent to the spelling correction.

As we are referring to the Population Census, the classifiers used are the 1997 National Classification of Economic Activities for activities and the National Classification of Occupations for the occupations. Then, in order to keep the commitments made in the context of the Mercosur agreement, dictionaries were obtained using the respective tables of correspondence for the CAES (Classification of Economic Activities for Sociodemographic Surveys of Mercosur) and in the groupings of the International Standard Classification of Occupations (ISCO).

5.1.2.3. - Third generation

The third generation is a code in itself, which shows how a particular variable was codified. It indicates whether it was sufficient to read only the variable to be codified or whether another variable contained additional information that was used to determine the code. To illustrate this, the possible third-generation codes for activities are given below:

- A - the code was assigned on the basis of the information for the variable "activity";
- O - the code was assigned on the basis of the information from the variable "activity" and "occupation";
- T - the code was placed with the information from the variable "activity" and the description of the task;
- Ch - this is a specific case to denote persons who do casual work, without reference to where the information was read (in the variable "activity", "occupation" or "task");
- Am - this is a specific case to denote itinerant vendors, without reference to where the information was read (in the variable "activity", "occupation" or "task");
- ? - insufficient information (a code could not be assigned).

The third-generation codes make it possible to create another set of dictionaries. An activity phrase for codification that is identical to a phrase from the dictionary which has the third-generation code "A" can be codified without problems in a totally automatic way; but an activity phrase for codification that is identical to a phrase from the dictionary which has the third-generation code "O" indicates that in order to assign the code the occupation has to be considered. Therefore, for the same activity phrase there are different possible codes. To summarize, the correction module enables us to carry out the following tasks:

- CORRECT WORDS
- DELETE WORDS
- CORRECT CODES

} CORRECT PHRASES

- ASSIGN CODES
- ASSIGN THIRD-GENERATION CODES

} ASSEMBLY OF DICTIONARIES
- correct
- correcting
- deleting
- dictionary of third-generation phrases

5.2. - Stage II: preparation of the bases

Reading the forms: although reading the forms is not one of the tasks of the SiCI, this is relevant to a certain degree because of the creation of the reading dictionary. The initial plan was to use an ordinary general dictionary, which was sure to contain all the words used in the Spanish language. This idea was discarded, as a dictionary of that size would delay the reading of the forms. By creating codification dictionaries, the words which are

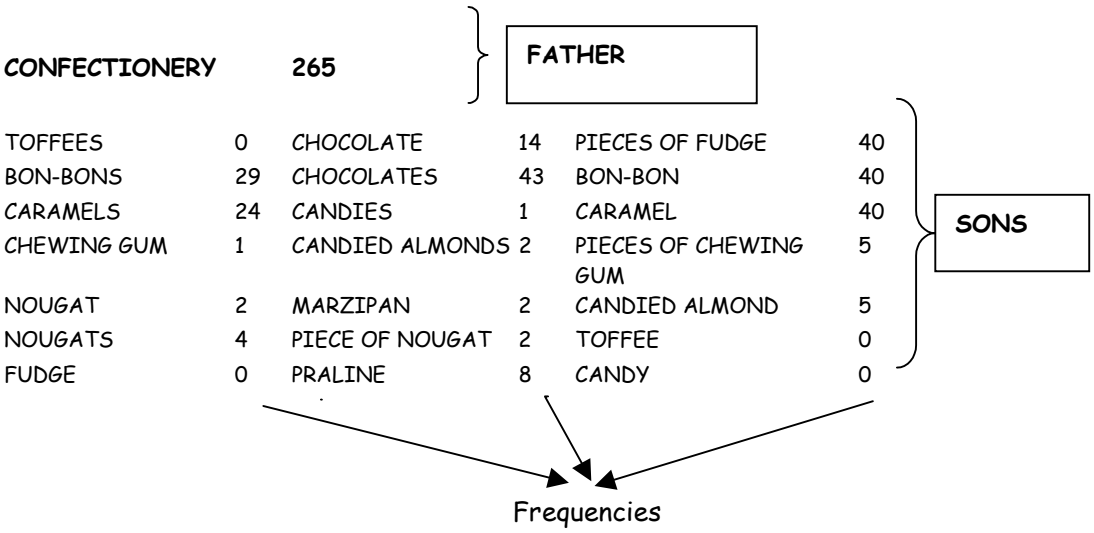
most commonly used to respond to the variables under consideration have been collected, and to date are no more than 10 000 words. Nevertheless, new words are incorporated as they appear during reading, if they are considered to be correct.

Normalization: once the file is obtained of phrases which could not be interpreted during the optical reading, the first thing to do is to verify the status of the databases. This consists of determining whether the structure of the databases is compatible with what has been established by the System. A second step is normalization, which consists of removing all the characters which are not useful for codification. For example:

"="	is replaced by " " (empty space)
")"	is replaced by " " (empty space)
"1º"	is replaced by " " (empty space)

Conditioning of the bases

The main idea at this stage is the linguistic procedure which we refer to equally as "semantic field" or "family grouping". For clarification, we can briefly note that a semantic field is a set of words (referred to as sons) that are semantically different but for codification purposes can be reduced to a single word (referred to as a father). For example:



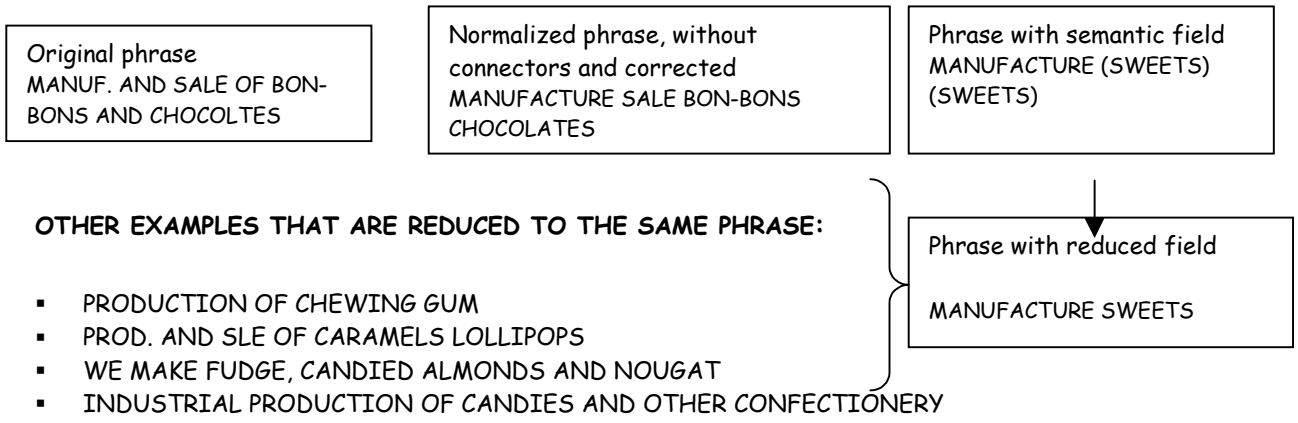
By applying the semantic fields, the father word acquires a higher frequency, as it can replace any of its sons. In the same way, the frequency of the phrases which contain the father word is increased. This is very important when calculating the heuristic weights and the scores.

The family includes son words with the same root as the father word. For example:

Father: MANUFACTURE
Sons: MANUFACTURES, MANUFACTURING

Thus the family is no more than a special case of the semantic field. Having clarified this, the semantic field procedure acts on the base to be codified in the following way:

A complete example of the stage of preparation of the databases



A table of results follows below in order to give an idea of the effects of this linguistic procedure.

RESULTS OF PRE-CODIFICATION

Operation	Original number of phrases	Number of unique phrases after pre-codification
Experimental census	35.567 (100%)	11.745 (33%)
Permanent Household Survey (October 98)	34.047 (100%)	9,984 phrases to codify (29%)

5.3. - Stage III: Codification

The codification stage is the most important, as it is here that a solution must be found for rapid and accurate codification. Different codifying strategies have thus been devised to unite the following three disciplines in one:

- Classifying standards (standardization framework for sectoral skill classification and codifying practice);
- Computer science (logic and development of the system);
- Statistical methodology (scores and quality control).

The resulting codification system thus includes the following methods:

Autophrase: very simply, if the dictionaries of codified phrases contain a phrase such as "milk production", which has a third-generation code "A" as it does not require any other variables to be considered, then any operation which includes the description "milk production" (or its equivalent in terms of semantic fields) can be codified automatically and without errors.[3]

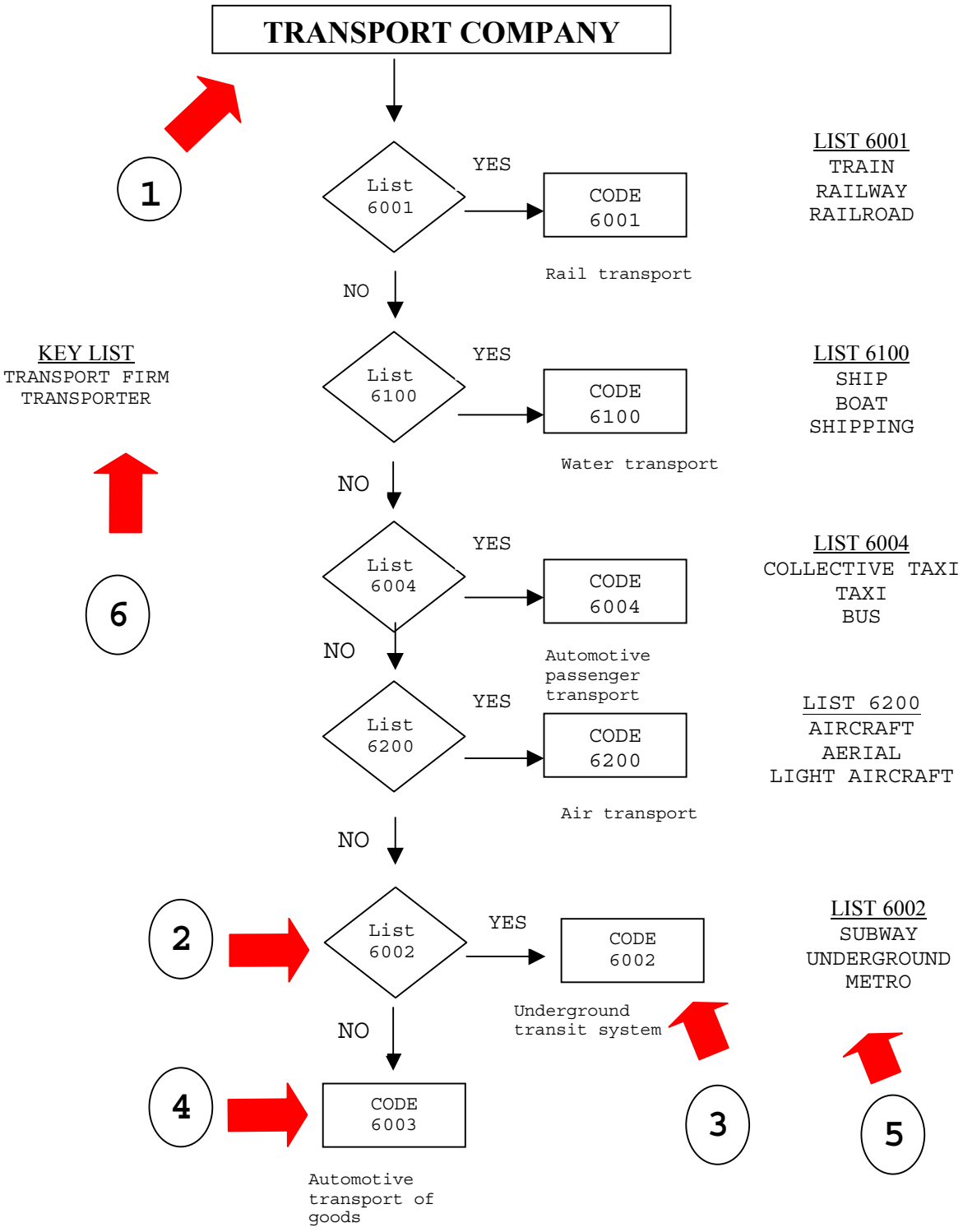
Microprocedures: are a set of rules for decision-making designed by specialists in different sectors and used in the SiCI so that through key words and other variables (for example the number of people employed by the establishment) a code can be assigned to an "activity" or "occupation" phrase which has many coding alternatives. The microprocedures are designed to "make decisions" in an automatic way based on the information contained in other variables which complement the responses of the variable to be codified. For example:

"Activity" phrase----------Transport enterprise
Codes assigned by the codifiers-------60-61-62
(two-digit)

If information provided under "occupation" or "task" shows that the enterprise is concerned with trains, automotive transport or aeroplanes, the case is resolved.

Thus, taking the above example, a microprocedure is designed with the following form:

[3] Unless there is an error in the dictionary, which, as mentioned earlier, should be perfect.


TRANSPORT COMPANY
1
List 6001
YES
CODE 6001
Rail transport
NO
List 6100
YES
CODE 6100
Water transport
NO
List 6004
YES
CODE 6004
Automotive passenger transport
NO
List 6200
YES
CODE 6200
Air transport
NO
2
List 6002
YES
CODE 6002
Underground transit system
NO
4
CODE 6003
Automotive transport of goods
KEY LIST
TRANSPORT FIRM
TRANSPORTER
6
LIST 6001
TRAIN
RAILWAY
RAILROAD
LIST 6100
SHIP
BOAT
SHIPPING
LIST 6004
COLLECTIVE TAXI
TAXI
BUS
LIST 6200
AIRCRAFT
AERIAL
LIGHT AIRCRAFT
LIST 6002
SUBWAY
UNDERGROUND
METRO
3
5

(1) KEY PHRASE: is the phrase with which the record appears for codification by the microprocedures method. Example: when the word “transporter” or “transport” or “transport enterprise” appears in the activity description, these cases will be codified according to the above diagram.

(2) RESTRICTION: determines whether a certain piece of information is available which would help to assign the code.

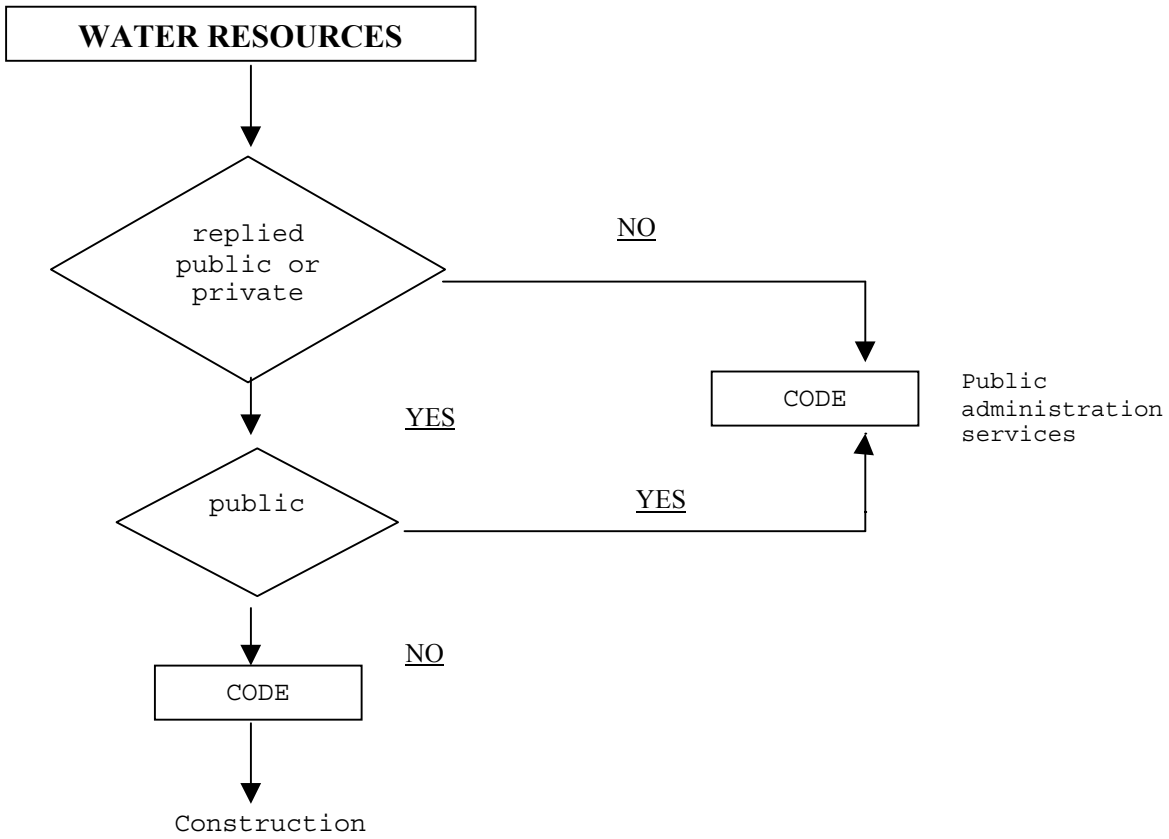
(3) CAES CODE WHICH WOULD BE APPROPRIATE IF THERE ARE RESTRICTIONS.

(4) GENERAL CODE IF THERE ARE NO RESTRICTIONS, which is the code assigned if there was no YES response to any of the restrictions.

(5) WORDS FROM THE LIST OF EXCEPTIONS: set of words which function as key data for the automatic designation of a code.

(6) KEY LIST: is a list of phrases which have the same meaning as the key phrase and which should go through the same microprocedure.

The size of the establishment could also be considered, or whether the company for which the person works is a public or private enterprise. For example:



Scores: is a method which combines the two elements. On the one hand it measures the "specificity" of each word in relation to the different codes. For example the word "milk" is "more specific" than "fabrication" as the first word is associated with a small number of codes while the second word is in more general use in all branches of industry. This is an analytical activity within the codification dictionary. The specificity of each word in the dictionary is measured by the so-called "heuristic weight" which also forms part of the dictionaries together with the texts and the codes. On the other hand the score also analyzes the relationship between the phrases in the dictionary and the phrases to be codified. Given a phrase to codify, the score makes it possible to choose possible phrases within the range offered by the dictionary. These candidates are chosen taking into account the greater number of common words between the phrase to be codified and the phrases of the dictionary. The more coincidence between both types of phrase the higher the "score".

Score 10: occurs when the phrase to be codified corresponds to a phrase in the dictionary that contains the same words, regardless of the order. However the autophrase procedure has already eliminated from the database to be codified those phrases that have the same order.

Score < 10: in this case the phrases to be codified do not correspond exactly to the candidates.

Example of phrase to be codified: **Fabrication of confectionery and biscuits**

(i) Example of a phrase with a score of 10: Fabrication of biscuits and confectionery

(ii) Example of a phrase with a score of <10:

Fabrication of confectionery

Autoword: is a method of automatic or direct codification which makes it possible to assign a single code without the intervention of the codifiers. For this, a codification dictionary is used which consists exclusively of words and associated codes, depending on the phrase that they come from. In practice, this method has been abandoned as the degree of error is about 50 per cent and it was used in only a very small percentage of cases, as a method to be applied only when other methods could not be used.

6. - Pilot test of the SiCI in the Pergamino Experimental Census

At the end of 1999, as part of the scheduled activities for the 2001 census, an experimental census was carried out in the area of Pergamino which produced a database of 35567 records to codify. In this operation a test of the reading dictionaries was carried out and the spelling correction, which adds on a daily basis to the reading dictionaries initially produced, was also tried out. This helped to improve the procedures until they reached their current form. The pre-codification stage, which consists of manipulating the phrases using the correcting, deleting and spurious word dictionaries, together with the procedures of standardization, application of the semantic fields or family grouping and

reduction to unique phrases, achieved the following results: the base to be codified was reduced by 67% from 35 567 to 11 745 records; this is then the starting point for the base to be codified.

Two tests of the codification were carried out, one which was completed in June 2000 and a second which has just been completed. A summary of the results obtained is as follows:

Automatic codification for activities:

Base: Pergamino, October 1999

Records to be codified: 35 567

The codification was carried out on the basis of the Classification of Economic Activities for Sociodemographic Surveys of Mercosur (CAES), to 4 or 2 digits and the tabulation category (letter).

Method	**Autophrase**	**Score 10**	**Micro-procedures**	**Scores between 8.5 and 10***	**Autowords**
Number of phrases codified	5699	799	9908	8971	1026
Percentage	16,02	2,25	27,86	25,22	2,88
Average error	0 %		<6%	30 %	50%

* Method not yet calibrated.

Conclusions: the microprocedures method still has potential for further development that would increase the codification percentage while reducing further the percentage of error. In any case, the maximum error that can be tolerated has not yet been defined. The scores < 10 method, although there is a high degree of error, close to 30 per cent, will certainly be retained to work with critical values of scores by branch of activity and to define a minimum dispersion value for the scores of candidate phrases. Given the errors produced by the method of autowords and the small contribution that they make to codification, that method has been discarded.

Automatic codification for occupations: after the first computerized codification test for the open question on occupation (report of June 2000), two new steps were taken:

(1) corrections to the procedures with key words

(2) standardization of words prior to applying the above procedures

Corrections and re-running of the programme: based on the results obtained in the first test, in relation to the quantity and quality of the computerized codification, the main errors encountered were corrected and the field of codification was expanded by creating new procedures. The correction consisted of both adding to and subtracting from or modifying the lists of words and restrictions associated with the procedures. In order to expand the computerized codification field, new procedures were created which had not been considered during the first test. That was either because at first we had focused on the procedures which we considered most important, or because the analysis of the cases where codification had not been possible had shown how new procedures could be created. Then the programme was run again for the same reference database. The comparative results between the first and the second test are as follows:

	Test 1	Test 2
Cases with assigned code	44,4%	60,8%
Codified cases	35,0%	48,0%
Provisionally codified cases	9,4%	12,8%
Non-codified cases	55,6%	39,2%
Total number of cases	35.567	

It is clear that there is a very significant increase in the number of codes assigned: about 33 per cent more than in the first test. And if the percentage increment is approximately the same in the two areas which make up this item (codified cases and provisionally codified cases), the greater proportion of the first (in a ratio of 4 to 1 in relation to the second) implies that in absolute values the new results are very significant.

Standardization of words prior to application of the procedures: word standardization is a method to extend the range of application of the procedures. It basically consists of reducing the words to their root (the words both from the empirical information and from the key words of the procedures and their lists of associated restrictions). As part of the second test of computerized codification, a codification was carried out using standardization. The results were as follows:

Cases with assigned code	66.8%
Codified cases	54.8%
Provisionally codified cases	12.0%
Non-codified cases	33.2%
Total number of cases	35,567

As we can see, the percentage of cases with assigned code rises from 60.8 per cent (without standardization) to 66.8% (standardized), that is, an increase of 6 percentage points (10 per cent in relative terms) **using the same procedures** for codification with key words. The only difference between the two situations is the standardization of the

words. Also, the increase is entirely associated with the number of cases codified directly, as the number of provisionally codified cases is actually less. This indicates a good method to follow.

Quality control: for Test 1,[4] quality control was carried out on the cases with assigned code (15,788 cases). The results of this control are different in nature according to whether the errors found are among the codified cases or among the provisionally codified cases:

- x the errors among the codified cases are "definitive": a single code is assigned by an automatic procedure and this will be the final code unless other verification and control procedures are applied;

- x the errors in the provisionally codified cases are "non-definitive": the code assigned by the automatic procedure is provisional, and intended as a guide for the process of assisted codification; if the generic code assigned is erroneous ("misleading") this can still be rectified by the codifier who assigns the final code, either by assigning the correct code or directing the case to another codification process.

For the quality control a 10 per cent sample was revised. This sample included at least 1,579 cases of assigned code. The results were as follows:

Cases with incorrectly assigned code	7.5%
"Definitive" errors	5.5%
"Non-definitive" errors	2.0%
Cases with correctly assigned code	92.5%

[4] Quality control is now being conducted for Test 2.