

SÓLO PARA PARTICIPANTES

DOCUMENTO DE REFERENCIA
DDR/27
27 de febrero de 2001

ORIGINAL: ESPAÑOL

CEPAL

Comisión Económica para América Latina y el Caribe

Primera reunión de la Conferencia Estadística de las Américas
de la Comisión Económica para América Latina y el Caribe

Santiago de Chile, 9 al 11 de mayo de 2001

COORDINACIÓN DEL SISTEMA NACIONAL DE NOMENCLATURAS

Sistema de Codificación Informatizada (SICI) para operativos
económicos y sociodemográficos

Este documento fue preparado por Mariano Lanne y Mara Riestra, coordinadores del proyecto SICI, Instituto Nacional de Estadística y Censos (INDEC) de Argentina. Las opiniones expresadas en este documento, que no ha sido sometido a revisión editorial, son de la exclusiva responsabilidad de los autores y pueden no coincidir con las de la Organización.

01-2-160

Introducción

La creación de un sistema de codificación "informatizada" surgió como una inquietud de la Coordinación del Sistema Nacional de Nomenclaturas (SiNN) del INDEC, a fines de 1998. Hasta ese momento, el área estaba preocupada en obtener clasificadores más adecuados a las necesidades nacionales. También se desarrollaron notas explicativas y diccionarios que permitirían agilizar los procesos de clasificación y a la vez documentar las decisiones que se tomaban ante diferentes consultas. Sin embargo, aún persistía la preocupación de que muchos de los procesos de codificación manual resultaban tediosos, no aportaban demasiada experiencia a los codificadores, eran largas jornadas para codificar lo mismo y se producían divergencias entre los criterios aplicados por diferentes codificadores. Además, ello dejaba poco tiempo para la discusión de casos de difícil resolución. En el caso particular de los operativos masivos como los censos, estos inconvenientes se traducían en una elevada demanda de recursos humanos, monetarios y de periodos de codificación extremadamente largos, con lo cual la información tardaba en estar en manos de los usuarios.

Fue justamente la coordinación del Censo 2001 el marco necesario para que la mencionada inquietud encontrara eco. A partir de abril de 1999 se conformó, en el ámbito de la metodología de trabajo propuesta por el SiNN, el Grupo de Aplicación de Nomenclaturas (GAN) 1, formado por integrantes de Actividades y Productos (SiNN-AyP), el Programa de Medición y Análisis de la Estructura Ocupacional (SiNN-ProMAEO) e integrantes de otras áreas como la Dirección de Metodología Estadística, el Departamento de Cartografía, la Dirección de Informática y el Equipo del Censo. Creemos que es este el principal hecho que permitió avanzar y obtener los resultados a los que hoy día hemos podido arribar. Es en esta unión multidisciplinaria que se pudo lograr mentar un sistema que lejos está de ser complejo. La mayor demanda está centrada en obtener una alta calidad en la red de diccionarios de los que se alimenta el sistema

El proyecto consistió básicamente en promover un estudio minucioso de la metodología de codificación aplicada en el procesamiento manual de cada una de las variables a codificar. Para ello se diseñó un sistema de trabajo que provocaran en el "codificador" del SiNN, un trabajo ordenado, pautado y que, a la vez de "explicitar" cada uno de los pasos que lo llevaban a un código, se obtuvieran los instrumentos necesarios para diseñar el SiCI (Sistema de Codificación Informática), es decir los diccionarios.

El documento que se presenta, resume la experiencia adquirida hasta el momento. El trabajo se divide en tres partes. La primera es una introducción al SiCI, luego se mencionan los objetivos y en un tercer apartado se presentan los programas de trabajo sobre los cuales se aplica el mismo. En la segunda parte, se hace referencia al modelo conceptual, en donde se definen los conceptos fundamentales y se explican las etapas que lo conforman. En la tercera y última parte, se exponen los resultados que se obtuvieron en las pruebas del SiCI en el Censo Experimental de Pergamino.

1.- Acerca de la Codificación Informatizada

Como ya se mencionó en la introducción, el proyecto de diseño, desarrollo e implementación del SiCI involucra un gran esfuerzo de inicio ya que el primer paso para todo sistema de este tipo es poder "modelizar" el proceso que se quiere sistematizar. En tal sentido, el SiCI es un sistema que recrea a través de diferentes métodos, todo el conjunto de procesos "intelectuales" que el codificador realiza cuando lee, interpreta, analiza y coloca el código a la frase que tiene delante de él.

Si observamos en detalle la forma en la que nos referimos al proceso de codificación notamos que no es lo mismo hablar de codificación automática que de informatizada, ya que esta última es más amplia. Nos referimos con el término de **automática** a aquella en la que es posible determinar un código sin la intervención de ninguna persona¹; mientras que la codificación **informatizada** incluye a la anterior, pudiendo llegar a poner un código en forma automática, asistida o semi-manual. No siempre los casos que se nos presentan son de resolución masiva, es más, algunos ni siquiera se presentan en forma frecuente ya que solo los encontramos en censos por barrido. Por ello, aquellos casos que no se pueden "modelizar" requieren de la codificación semi-manual. Una vez solucionada la codificación puede resolverse en forma automática para relevamientos futuros dependiendo ello del contexto en que la respuesta esté incluida.

Con relación al proceso de codificación automática podríamos decir que ésta se basa en la aplicación de un conjunto de frases anteriormente codificadas, de tal forma que aquellos casos que se repitan, se resuelvan de la misma manera. Para ello se requiere una herramienta básica del proceso llamado DCCIONARIO, que es un conjunto de casos previamente codificados. Vamos a ver que en realidad, no existe solo un diccionario sino un conjunto de ellos, que interactúan en el proceso de codificación.

2.- Objetivos

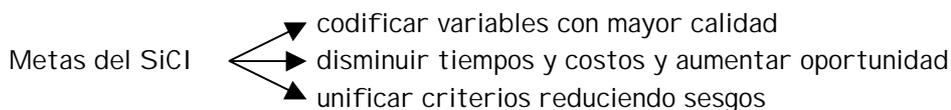
El SiCI tiene como objetivo principal, la codificación de diferentes variables de uso estadístico. Las variables a codificar son las llamadas respuestas "abiertas" es decir, aquellas en las que no existe una precodificación en el formulario y donde el informante responde con sus palabras a la pregunta del cuestionario, lo cual implica que distintas personas que tienen igual ocupación, realizan igual tarea y trabajan en la misma empresa, pueden responder la actividad y su ocupación de diferentes formas. La codificación de variables "cerradas" no requerirán, por lo general, demasiado esfuerzo, puesto que su relevamiento es en sí un tipo de codificación. Sin embargo, en muchos casos, las variables cerradas se utilizarán como complemento para la codificación de las variables abiertas. A título de ejemplo, algunas de las variables a codificar son:

- actividades
- nombre de la ocupación
- descripción de la tarea
- variables geográficas
- carreras universitarias

¹ En realidad la codificación fue realizada por el equipo del SiNN durante la etapa de generación de los diccionarios

Podemos realizar una importante diferencia entre las variables de actividades y ocupaciones por un lado, y las geográficas y carreras universitarias por otro. Esta diferencia se basa en la infinidad de posibles respuestas que se pueden obtener en el primer conjunto; mientras que en el segundo, las posibilidades de que las respuestas se expresen en formas diferentes son mucho más limitadas. Por tal motivo, la codificación informatizada de las variables del primer grupo puede ser más complicadas que las variables del segundo conjunto. En este documento nos referiremos principalmente a la codificación informatizada de actividades y ocupaciones, pero las conclusiones que se obtengan podrán ser utilizadas, en general, para la codificación de las otras variables.

El SiCI no sólo tiene como objetivo la codificación de estas variables, sino también disminuir los tiempos de codificación y unificar criterios de interpretación. La diferencia de criterios adoptados por cada uno de los codificadores es uno de los problemas que lleva a disminuir la calidad de los resultados de una encuesta o un censo. Sin embargo, mediante la aplicación de criterios uniformes adoptados por un sistema informático es posible considerar todos los casos semejantes, bajo una misma óptica; a la vez que permite una rápida recodificación si se requiere cambiar el criterio. Esta misma tarea, si se realizara manualmente, requeriría demasiado esfuerzo y costo. En un sistema informatizado, los casos que permiten diferentes interpretaciones pueden ser, o bien codificado bajo un criterio adoptado, o agrupados automáticamente para su posterior codificación. En síntesis:



3.- Aplicaciones

3.1.- Censo Nacional de Población y Vivienda 2001

Para este censo se calcula que habrá aproximadamente 37 millones de personas a censar y 12 millones de ocupados. Las cifras son grandes de más como para pensar en una codificación manual, pues se debería concentrar gran cantidad de codificadores, o bien, realizar como en el censo de 1991 una muestra, o se debería prescindir de la obtención de resultados en forma oportuna, o peor aún eliminar parte de las preguntas. Ante tales circunstancias, surge como una necesidad la incorporación de la codificación informatizada. Sin embargo, el censo 2001 no es el único fin del Sistema de Codificación Informatizada (SiCI), ya que en el horizonte de planeamiento de utilización del mismo se prevé la incorporación del mismo a distintos programas de trabajo como ser la codificación de la Encuesta Permanente de Hogares, el Censo Nacional Económico y el Directorio Nacional de Unidades Económicas entre los principales destinatarios.

La tarea de coordinar la codificación del Censo no se refiere sólo a producir un software apto para codificarlo, sino también a un conjunto de tareas que se relacionan con la codificación y que influyen en buena medida en la calidad de los datos. Esto implica mantener una interacción permanente con varias áreas de trabajo relacionadas con el censo, a saber:

Con la empresa encargada de la lectura óptica de los formularios. El SiCI provee los diccionarios de palabras, manteniéndose una actualización de los mismos en forma diaria durante el período de lectura.

Con el área de Metodología Estadística. Otra tarea relacionada a la codificación es ayudar en la determinación del método para medir la calidad de los datos, tarea que se realiza en forma conjunta con el área de metodología estadística. Esta es a su vez, es la encargada de desarrollar el método de codificación denominado "scores", cuya explicación se realiza mas adelante.

Con los codificadores. Ella se origina en la necesidad de entrenar al personal, tanto en el uso del software como en el manejo de los instrumentos clasificatorios y en los criterios de codificación propiamente dichos. Es necesario trabajar en forma conjunta, para detectar errores en la interpretación y distribuir la tarea de manera más eficiente.

Con el área de análisis y consistencia. En algunos casos la codificación no es posible sin la consistencia previa de ciertos datos y en otros la misma consistencia implica una codificación previa. Ejemplo de ello es que las variables geográficas requieren ser codificadas previo a la consistencia.

Con Informática. Finalmente, la obtención de un software no sería posible sin la permanente interacción con el área informática, tanto en la etapa de desarrollo del sistema como durante la codificación del censo para poder permitir una actualización continua de ciertos diccionarios, la retroalimentación y calibrado del sistema.

3.2. - Censo Nacional Económico (CNE)

Otro de los grandes operativos que se presenta como desafío para la aplicación del SiCI es la codificación de las variables de actividades y productos. Por un lado porque si bien se reduce el número de respuestas a alrededor de 1.500.000 de casos, el nivel de desagregación con que se requiere la codificación aumenta sustantivamente. Esto implica necesariamente aumentar el nivel de detalle de los diccionarios y conjuntamente la velocidad de procesamiento y codificación ya que los resultados deberían estar disponibles dentro de los cuatro meses siguientes al operativo censal.

Para esta aplicación la fuente primordial para alimentar los diccionarios de codificación, además del Censo Nacional Económico 1994, cobran importancia los literales relevados por el Directorio y las encuestas relevadas por otros sectores del Sistema Estadístico Nacional, tales como la encuesta industrial, el Registro Industrial de la Nación y las encuestas de la Secretaría de Agricultura, Pesca y Alimentación, por citar algunas.

Siguiendo en el ámbito de la clasificación de actividades, en este operativo, el nombre de la empresa es importante dato a la hora de definir los códigos, ya que podría llegar a pensarse en una posible pre-codificación de las empresas previo a la salida a campo, con lo cual el espectro de casos a codificar se reduciría notablemente.

Por último, ante la existencia de preguntas sobre los productos se crea una nueva demanda para el SiCI que es la incorporación de los clasificadores de productos, lo que lleva a desarrollar nuevos diccionarios.

3.3. - Encuesta Permanente a los Hogares (EPH)

La EPH es la principal fuente de descriptores para la casi totalidad de las variables que se proyecta codificar. Sin embargo, hoy es uno de nuestros principales "usuarios". La implementación de un relevamiento continuo, provocará un uso permanente del sistema de codificación. Además de los beneficios que ello reporta en términos de la pronta disponibilidad de los resultados de la EPH, ello permitirá el calibrado del sistema, sentando los antecedentes necesarios para ser luego utilizado durante el operativo censal. Es la continuidad también la que permitirá ir "amortizando" los esfuerzos realizados en este par de años. Actualmente se está realizando una prueba piloto del SiCI sobre la última EPH disponible. Se espera tener resultados para la segunda quincena de febrero.

3.4. - Directorio Nacional de empresas (DiNUE)

El DiNUE es, junto con el Censo Nacional Económico 1994, la principal fuente de literales para procesos de codificación de actividades provenientes de relevamientos de índole económica. El SiCI permitirá al DiNUE relevar información sobre actividades y productos, y en la medida en que esa información esté en los diccionarios, la codificación se realizará en forma automática. Caso contrario, se activará el proceso de codificación asistida y semi-manual, lo cual redundará en una mejora en los diccionarios de codificación del SiCI a la vez que aumenta la calidad de codificación de dichas variables en el DiNUE, reduciendo así las tareas de supervisión.

4. - Definiciones

El SiCI ha sido una creación "original" en el sentido que dada la escasa bibliografía existente, hubo que desarrollar un sistema desde cero incluyendo la terminología utilizada. Es por eso, que pese al alcance del documento, fue preciso incorporar este apartado sobre las definiciones que se encontraron a lo largo del texto.

SiCI: red de diccionarios de diversa índole, interconectados a través de procesos lingüísticos y de codificación. Por medio de este sistema el conjunto de registros que contienen literales originales de las variables a codificar, son transformados en descriptores a los cuales se les aplica diferentes métodos de codificación tendientes a asignar a cada uno el código correspondiente en forma unívoca.

De esta definición surgen los tres elementos básicos del SiCI:

- Diccionarios
- Procesos lingüísticos
- Procesos de codificación

4.1. Diccionarios: son listados inventariados de palabras o frases que conforman los instrumentos fundamentales del SiCI y que se originan en las respuestas empíricas relevadas en cada uno de los operativos que sirvieron de fuente. En el sistema conviven dos tipos de diccionarios: los que sirven para la manipulación de las palabras y los diccionarios de codificación.

Diccionario de palabras espurias (E): conjunto de palabras que si bien poseen significado literal, a efectos de la codificación no son relevantes. Ejemplos de palabras que conforman este diccionario son: números, nombres propios (excepto de empresas que puedan definir un código); nombres de lugares geográficos; adjetivos que no son relevantes a efectos de la codificación como ser colores, tamaños, adjetivos relativos a lugares o formas, etc.; letras sueltas y números romanos; y otras palabras que tengan significado pero son prescindibles para la codificación. Son generalmente de baja frecuencia.

Diccionario de anuladas (A): conjunto de palabras que carecen de significado. Son originadas en errores de tipeo, lectura y/o redacción y no se les puede atribuir ninguna palabra para realizar una corrección. Se forma generalmente por la partición de palabras. Ejemplo: supongamos que la palabra "computadora" aparece cortada al medio: "compu" "tadora". La primera parte puede ser útil para intuir que es algo relativo a la computación por lo que no formará parte del diccionario de anuladas. La segunda parte: "tadora" no puede relacionarse con nada específico o con muchas cosas, por lo tanto forma parte del diccionario de palabras anuladas. Al contrario de las espurias, estas "cuasi-palabras " no tienen significado y por lo tanto no forman parte del diccionario de lectura.

Diccionario de conectores (C): conjunto de artículos, preposiciones, y otras palabras que se utilizan para dar forma a una oración, pero no son relevantes a efectos de la codificación. Ejemplos de conectores son: **y, la, los, con, por**, etc. Por el contrario, son relevantes a efectos de la codificación los conectores **no, para** y **excepto** los cuales no forman parte de este diccionario.

Diccionario de excepciones (X): conjunto de conectores cuya presencia en una frase puede alterar la codificación de la misma y por tanto no forman parte del diccionario de conectores. En la actualidad está formado por tres palabras a saber: **no, para** y **excepto**.

Diccionario corrector (R): conjunto de relaciones entre palabras incorrectas y correctas. Las palabras incorrectas pueden ser generadas por errores de tipeo, ortográficos, abreviaturas u otra clase pero **siempre se puede relacionar con una y solo una palabra correcta**. Ejemplo: la palabra incorrecta "alimenticios" tendrá su par equivalente con la palabra correcta "alimenticios"; la palabra incorrecta "gral" será reemplazada por la palabra correcta "general". Un caso en que no se puede mantener una relación palabra incorrecta-correcta se da con "art" pues, si bien puede venir de un contexto en donde se entiende que es la abreviatura de "artículo", en otras ocasiones se puede tratar de las ART (Aseguradoras de Riesgo de Trabajo).

Diccionario de palabras correctas (D): conjunto de palabras correctamente escritas, que son relevantes para la codificación y por lo tanto no se incluyen en ninguno de los diccionarios anteriores.

Diccionario de lectura (L): está compuesto por la unión de los siguientes diccionarios: espurias, conectores, excepciones, y palabras correctas.

$$L = E+C+X+D$$

Diccionario de codificación: es el listado de frases y palabras asociados a cada código y sobre el cual se calcula el peso heurístico que sirve de base al método de scores, más

adelante expuesto. Los elementos que lo componen son: el diccionario de palabras correctas (D) y el de excepciones (X).

4.2.- Procesos lingüísticos: son aquellos que modifican los literales de las frases a codificar, permitiendo una simplificación del vocabulario y de la cantidad de palabras involucradas. Con **literales o descriptores** nos referimos a la frase que representa la respuesta original brindada por el informante, sea esta en representación de una persona o una unidad económica (empresa y local entre otras). Entre los procesos lingüísticos que operan sobre los literales nos encontramos con:

Proceso de normalizado: consiste en sacar los caracteres no válidos que se encuentran en las frases de la base recibida con las tres variables (actividad, ocupación y tarea) y se convierten a mayúscula.

Campos semánticos o familiarizado: consiste en asignar a una palabra tomada como referencia (denominada padre), una lista de palabras que serán tomadas como sinónimos (denominado hijos).

Proceso de estandarizado: consiste en tratar todas las palabras del diccionario por número, género y truncamiento, según lo que sea mas apropiado, a los efectos de lograr un diccionario de términos únicos (no repetitivos)

Sin estandarizado: no se realiza el proceso anterior.

Es importante destacar que los procesos lingüísticos no son procesos de codificación.

4.3.- Procesos de codificación: estos actúan de diferente forma según el caso a resolver. Son procesos que surgen de la "modelización" de las procesos analíticos que los codificadores realizan en el momento de asignar un código.

Macroproceso: es un conjunto de instrucciones que se modelizan a través de sentencias informáticas y que permiten dividir al universo a codificar en grandes grupos. Esta división permite luego acotar el rango de códigos posible. Son ejemplo los macroprocesos "patrón" en ocupaciones y "ventas" en actividad, como se explicará en su correspondiente apartado.

Microproceso o autoproceso: es un conjunto de instrucciones que se modelizan a través de sentencias informáticas y que permiten arribar a la codificación de un determinado literal sin la intervención de codificadores. A diferencia de los macroprocesos, son métodos de codificación propiamente dicho.

3ra Generación: forma de codificación, elemento que me permite determinar que variables se utilizaron para codificar una variable específica.

AutoFrase: es un método de codificación automático o directo que permite la asignación de un código único sin intervención de los codificadores. Para esto, utiliza un diccionario de codificación formado exclusivamente por frases que ofrecen una única alternativa de código y que son independientes de las restantes variables del cuestionario.

Scores: es un método que combina dos elementos. Por un lado la "especificidad" que cada palabra tiene respecto a los distintos códigos. Por ejemplo la palabra leche es "más específica" que "fabricación" pues le aparecen a la primera una limitada cantidad de códigos asociados mientras que la segunda es de uso más difundido en todas las ramas de la industria. La especificidad de cada palabra del diccionario se mide a través del llamado "**peso heurístico**" que también forman parte de los diccionarios junto con los literales y los códigos. Por el otro lado, el score también analiza la relación entre las frases del diccionario y las frases a codificar. Dada una frase a codificar, el "score" permite elegir "frases candidatas" dentro de la "oferta" que da el diccionario. Esas "candidatas" se eligen teniendo en cuenta el mayor número de palabras comunes entre la frase a codificar y las frases del diccionario. Cuanto mayor coincidencia se de entre ambos tipos de frases mayor será el "score".

Autopalabra: es un método de codificación automático o directo que permite la asignación de un código único sin intervención de los codificadores. Para esto, utiliza un diccionario de codificación formado exclusivamente por palabras y los códigos asociados según la frase de la cual provengan.

Asistido: es un método de codificación indirecto que permite la asignación de un código único con intervención de los codificadores. En este caso, el SiCI da la posibilidad de elegir entre un limitado número de alternativas propuestas automáticamente.

Semimanual: es un método de codificación que permite la asignación de un código único con intervención de los codificadores. En este caso, dada la gran cantidad de alternativas de elección, el SiCI ofrece elementos de ayuda para el codificador sin realizar propuestas automáticas.

5.- ETAPAS DEL SiCI

Para simplificar la explicación de las distintas etapas del SiCI se ha tomado como ejemplo a la codificación del Censo Nacional de Población y Vivienda 2001. Por lo tanto, el esquema general que figura en la próxima página puede modificarse levemente cuando se procesan otros operativos, como la EPH o el Censo Nacional Económico. A continuación se realiza una brevíssima descripción integral del SiCI, para luego pasar al detalle a partir del punto 5.1.

En la implementación del SiCI hay dos grandes períodos de trabajo bien diferenciados:

Primer periodo: desarrollo de los diccionarios. Es el momento previo a la realización del operativo a ser codificado. Una de las características de este sistema es la importancia que se le da a la corrección ortográfica. Para ello se generan los distintos diccionarios con la única finalidad de corregir las descripciones que provienen del operativo de campo a ser codificado. Es natural que se repitan errores de ortografía, abreviaturas, siglas y formas de escritura que hacen que las frases que dicen lo mismo no sean perfectamente iguales y por lo tanto no puedan ser codificadas automáticamente. El sistema incorpora todas aquellas correcciones que se repiten con el ánimo de poder interpretar mejor las descripciones. Esta etapa es una de las más tediosas pues consiste básicamente en codificar, corregir y relacionar la mayor cantidad de palabras y frases provenientes de distintas fuentes, para obtener a final de todo el proceso los DICCIONARIOS.

Segundo periodo: preparación de las bases a codificar y codificación. A partir de la recepción del archivo del operativo a codificar se inicia una serie de etapas basadas en la necesidad de acondicionar las bases para su codificación. Dado que para codificar se utilizan los diccionarios obtenidos en la primera etapa, es preciso llevar a cabo un conjunto de tareas tendientes a otorgarle al archivo a codificar, las mismas características de los diccionarios. Por ello se suceden los siguientes pasos, que se corresponde con el gráfico de la página siguiente:

1 Se identifican las palabras no contenidas en el diccionario de lectura. Esto tiene por finalidad enviar a la empresa que lee las cédulas el **diccionario de lectura** actualizado.

2 Se verifica que la estructura de las bases sea la que requiere el SiCI y se normalizan las frases. Esto consiste en eliminar caracteres extraños o no válidos para la codificación (comillas, dobles espacios, puntos, etc.).

3 Se corrige automáticamente la ortografía mediante la utilización de los **diccionarios de corrección y anulador**.

4 Se corrige manualmente la ortografía que no pudo ser corregida automáticamente, empezando por las de mayor frecuencia.

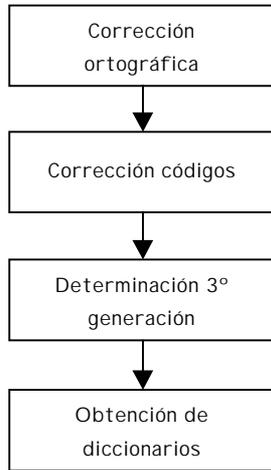
5 Si es necesario, se crean nuevos campos semánticos y estandarizados, sobre las palabras nuevas.

6 Se arma el archivo con las frases corregidas en las etapas descriptas y comienza la etapa de "Acondicionamiento de bases a codificar"

7 Se eliminan las palabras que no son útiles a los efectos de la codificación utilizando el **diccionario de palabras espurias**.

ETAPA I: PERIODO DE ELABORACION DE DICCIONARIOS

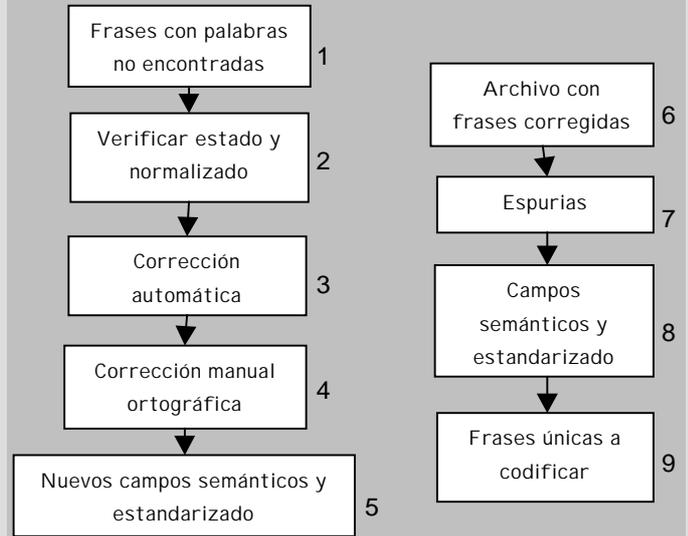
Proceso Diccio



ETAPA II: PREPARACION DE BASES

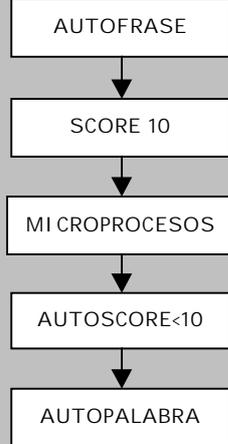
Corrección ortográfica

Acondicionamiento de bases

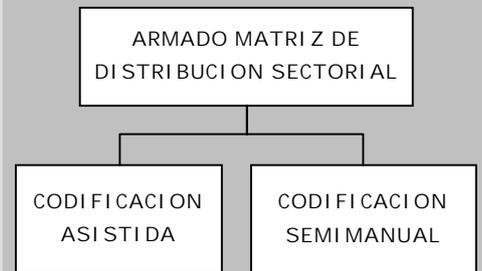


ETAPA III: CODIFICACION

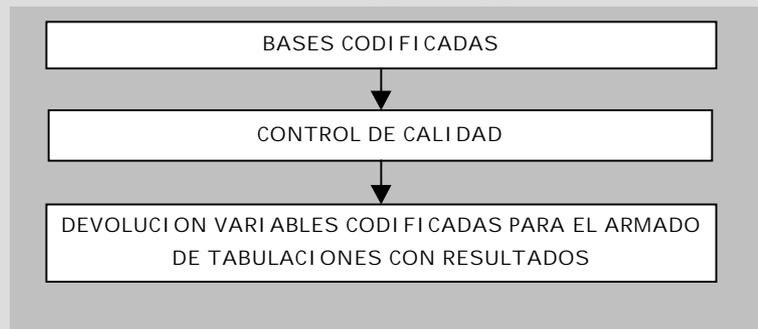
Métodos Automáticos



Métodos No Automáticos



ETAPA IV: RESULTADOS



8 Se modifican las frases a codificar simplificando sus palabras. Se aplican los campos semánticos, es decir se llevan distintas palabras que significan lo mismo a efectos de la codificación (palabras hijos) a una palabra en común (palabra padre). Se aplica también el estandarizado, es decir, se eliminan los géneros y números de las palabras dejando solamente la raíz de las mismas. Por último si dentro de la frase, como consecuencia de lo anterior, quedaron palabras repetidas, se procede a realizar una simplificación (ver apartado I I.2, para mayores detalles).

9 Muchas de las frases a codificar van a ser iguales por causa de los procesos antes mencionados. Se obtiene entonces un archivo con frases únicas (sin repeticiones de frases) a codificar. En la práctica se ha logrado reducir la base de actividades en alrededor de 67%. Es decir que de 35.000 frases recibidas se redujeron a 11.000.

Una vez obtenida la base de frases únicas se inicia la etapa de codificación propiamente dicha. La misma consiste en aplicar secuencialmente los distintos métodos de codificación, cuya explicación se realiza en forma detallada a partir del punto 5.3. El hecho de ser secuencial implica no solo un orden sino que además si un método logró colocar un código único, esa frase ya queda codificada y no se vuelve a codificar. Es decir cada método codifica la base residual que recibe del método anterior.

Por último llegar al autoscore < 10 el sistema puede llegar a codificar la frase con más de un código. Si se coloca tres o menos se pasa al método codificación "asistida"; si es mayor a tres se envía la base a la codificación "semimanual".

Como síntesis de todas las etapas se obtiene la base codificada, la cual se somete al control de calidad y finalmente se transmiten los datos a la oficina de procesamiento del censo.

SOLO SE PRODUCE EL HECHO CODIFICATORIO EN LA ETAPA III
--

5.1.- Etapa I: elaboración de los diccionarios

Los diccionarios son la base del sistema de codificación informatizada ya que los utiliza en todas las etapas de su funcionamiento. Este capítulo se refiere principalmente al modo de crear esos diccionarios en una forma sistematizada.

Un error en un diccionario se reflejará en la codificación de una actividad u ocupación, tantas veces como estas aparezcan para ser codificadas. De ahí proviene la necesidad de poseer un diccionario sin errores. Pero es a través de los diccionarios que se pueden aplicar criterios únicos de codificación evitando distintas interpretaciones. Un error en el diccionario se multiplica en la codificación automática. De todos modos, ya sea por un error en el diccionario o por la decisión de tomar un criterio distinto al que figura en el diccionario, se puede modificar el mismo y correr nuevamente el sistema de codificación para aquellos registros que se pretenden modificar. En un comienzo se mencionó que no existe un solo diccionario, sino un conjunto de ellos; veamos como surgen.

5.1.1 Fuentes para la creación de los diccionarios

Al utilizar un diccionario para la codificación, se busca que la base a codificar y la del diccionario se asemejen lo más posible. Es necesario, entonces, armar los diccionarios con registros provenientes de campo, que se respondan en forma similar a la base que se pretenda codificar. Por ejemplo, si se quiere codificar una encuesta sociodemográfica es conveniente utilizar principalmente registros que provengan de programas del mismo ámbito. Ello no implica que no se puedan utilizar registros que provengan de encuestas del área económica, y mucho menos que éstas no aporten nada al diccionario, sino que es más probable que las respuestas de dos encuestas de la misma área sean más parecidas. Además, se pueden realizar diccionarios paralelos, es decir, crear un diccionario que utiliza registros de una fuente para codificar cierto tipo de encuestas y otro que se compone de registros de otras fuentes para codificar otro tipo de encuestas. Las áreas que sirven de fuentes para la creación de los diccionarios son:

- Encuesta Permanente de Hogares (EPH)
- Tercera y cuarta Prueba Piloto del Censo 2000
- Muestra del Censo de Población y Viviendas de 1991
- Directorio Nacional de Unidades Económicas (DiNUE)

5.1.2 Proceso Diccio

Para proceder a armar los diccionarios sobre las fuentes antes mencionadas, se desarrolló una pantalla que permite sistematizar la codificación, la corrección ortográfica y la aplicación de la tercera generación². Esta pantalla no forma parte de la codificación informática propiamente dicha, sino que es parte de una etapa anterior a la codificación. Esta etapa fue denominada "Proceso Diccio" y es mediante este proceso en donde se invierte tiempo y recursos en la corrección y codificación de registros que luego formarán parte de los diccionarios. Esta etapa es una de las más largas y tediosas, puesto que implica la revisión o codificación de las bases que se elijan para que formen parte de los diccionarios. Las decisiones que aquí se tomen le servirán al SiCI como guía para saber como actuar en determinados casos.

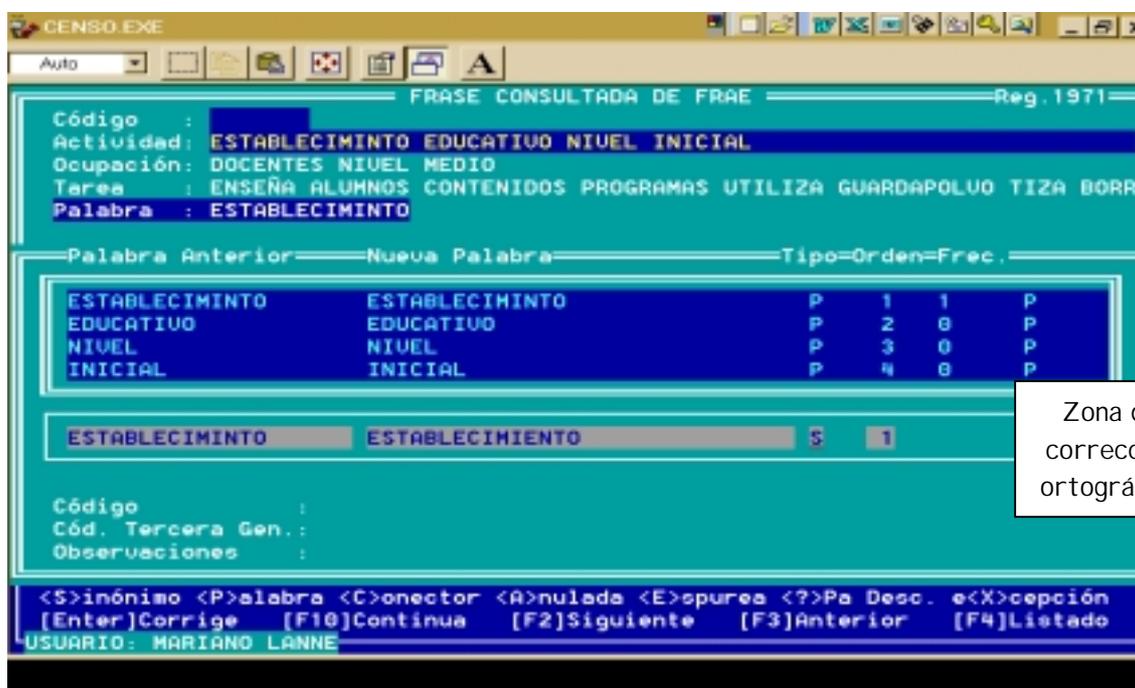
El trabajo realizado a través de esta pantalla (Proceso Diccio) se podría hacer en cualquier tabla o archivo, sin embargo, es conveniente realizarlo a través de una pantalla en donde se sistematizan las tareas efectuadas y automáticamente realiza la gestión de las bases que formarán los diccionarios. A la vez que permite una mayor seguridad en el manejo de las bases. En la próxima página se muestra un modelo de pantalla del módulo de corrección.

5.1.2.1 Corrección ortográfica

La corrección ortográfica busca obtener diccionarios "correctos", sin embargo al corregir los registros que luego formarán parte de los diccionarios, estos se diferenciarán de los registros a codificar que tengan errores ortográficos. Es por eso, que se crea el primer diccionario, al que denominaremos "Diccionario Corrector". Este se compone de un conjunto de pares ordenados de palabras (palabra incorrecta y palabra correcta), que se

² La tercera generación permite indicar para cada caso codificado, que elementos se tuvieron en cuenta para arribar al código como se verá en el punto 5.1.2.3

obtienen de la experiencia de la corrección ortográfica. La corrección ortográfica es importante para reducir el tamaño de los diccionarios de codificación.



Es importante distinguir las correcciones que se pueden hacer porque el contexto de la frase permite definir una palabra correcta asociada a la palabra incorrecta, de aquellas en las que existe una única relación palabra incorrecta - palabra correcta sin considerar el contexto de la frase. Por ejemplo, en el caso anterior en que se relaciona la palabra "establecimiento" con "establecim~~in~~to" no quedan dudas de dicha correspondencia y por lo tanto se puede generalizar y permitir que se corrija automáticamente en todos los casos en los que aparezca. Pero si aparece la palabra "art" en algunos casos se la puede relacionar gracias al contexto de la frase con la palabra "artículos". Sin embargo, en otros casos es una palabra correcta pues se refiere a las "ART, Administradoras de Riesgo de Trabajo". El diccionario corrector solo debe estar compuesto de los casos en que se puede generalizar la corrección de una palabra.

El diccionario anulador es el que se compone de palabra anuladas, que son aquellas que no tienen ningún significado porque efectivamente no existen. No incluyen aquellas palabras que poseen errores ortográficos y que no se pueden incluir en el diccionario corrector por tener mas de una palabra correcta anulada.

5.1.2.2 Codificación

Las bases fuente utilizadas para crear los diccionarios, en muchos casos ya vienen codificadas desde las áreas de trabajo que las proveen. Es necesario entonces revisar la calidad de la codificación recibida para garantizar la veracidad de cada código existente en los diccionarios y solucionar los problemas provenientes de diferencias de criterio. Esta etapa de corrección de la codificación se realiza en la misma pantalla antes mostrada y a continuación de la corrección ortográfica.

Dado que estamos hablando del Censo de Población, los clasificadores utilizados son la ClaNAE-97 (Clasificación Nacional de Actividades Económicas 1997) para actividad y el Clasificador Nacional de Ocupaciones para las ocupaciones. Luego a los efectos de cumplir con los compromisos adoptados en el marco del acuerdo Mercosur, a través de las respectivas tabla de correspondencia se obtuvieron los diccionarios en CAES (Clasificación de Actividades Económicas para Encuestas Sociodemográficas del MERCOSUR) y en las agrupaciones de la Clasificación Internacional Uniforme de Ocupaciones (CIUO).

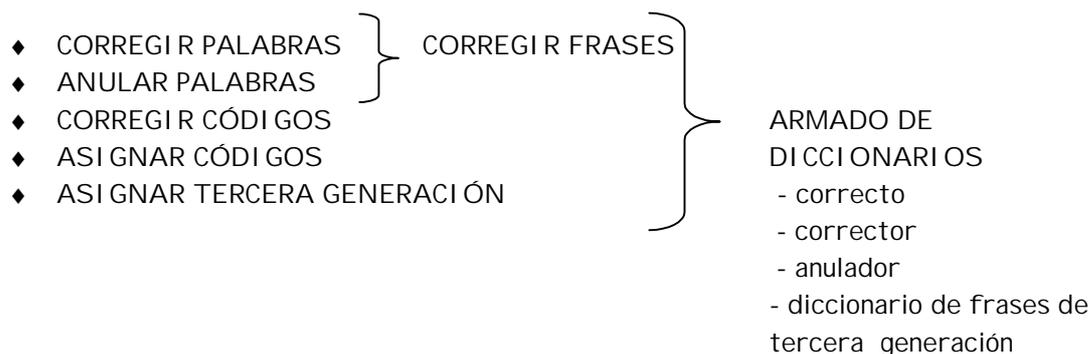
La pauta de trabajo para la codificación de las variables que luego forman los diccionarios de trabajo, es asignar a cada "leyenda" el mayor número de dígitos posible. Una vez colocado el código, el sistema verifica la existencia del mismo.

5.1.2.3. Tercera generación

La tercera generación es un código en si mismo, que indica cómo fue codificada esa variable. Es decir, si fue necesario leer solamente la variable a codificar o existe en alguna otra información complementaria para determinar un código. Para ilustrar lo ante dicho, se presenta a continuación los códigos posibles de tercera generación para actividades:

- A = el código fue puesto con la información de la variable actividad
- O = el código fue puesto con la información de la variable actividad más la ocupación
- T = el código fue puesto con la información de la variable actividad más la descripción de la tarea
- Ch = este es un caso específico para distinguir a las personas que realizan changas (changador), sin importar en donde se leyó la información (en la variable de actividad, ocupación o tarea)
- Am = este es un caso específico para distinguir a las personas que realizan ventas ambulantes, sin importar en donde se leyó la información (en la variable de actividad, ocupación o tarea)
- ? = información insuficiente (no se puede determinar un código)

Los códigos de tercera generación permiten crear otro conjunto diferente de diccionarios. Una frase de actividad a codificar idéntica a una frase del diccionario que posee tercera generación "A" se puede codificar sin problemas en forma totalmente automática; pero una frase de actividad a codificar idéntica a una frase de diccionario que posee tercera generación "O" indica que para poner el código debe observarse la ocupación. Por lo tanto, para una misma frase de actividad se presentan diferentes códigos posibles. En resumen, el módulo de corrección nos permite realizar las siguientes tareas:



5.2.- Etapa II: preparación de las bases

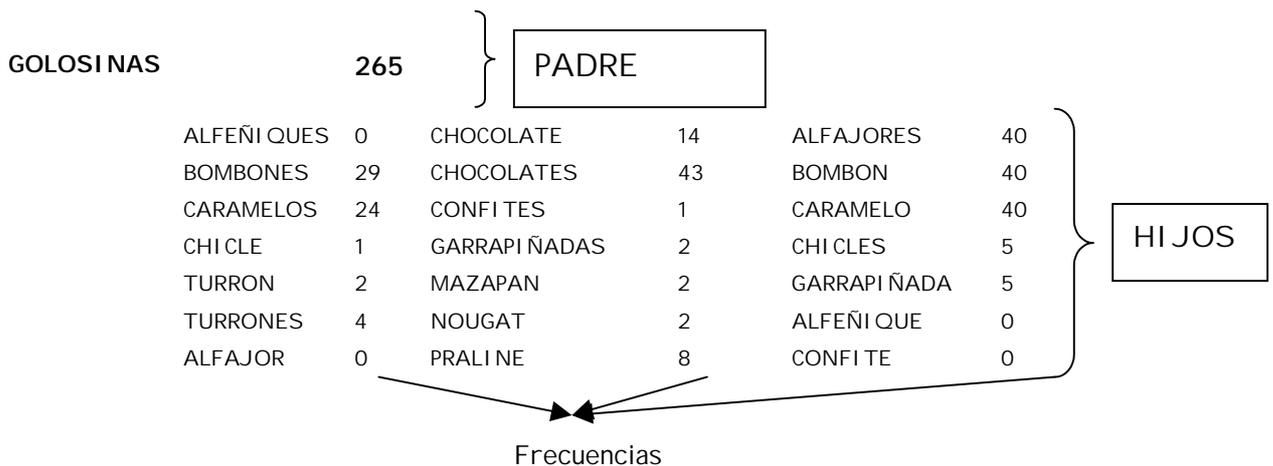
Lectura de los formularios: si bien la lectura de los formularios no forma parte de las tareas a realizar por el SiCI, este interviene de cierta forma a través de la creación del diccionario de lectura. En un principio se había pensado en utilizar un diccionario de uso corriente donde nos aseguraríamos que se encontrarían todas las palabras que se utilizan en el idioma español. Esta idea fue descartada pues un diccionario de tal dimensión demoraría la lectura de los formularios. Con la creación de los diccionarios de codificación, se fueron recolectando las palabras más usuales que se utilizan para responder las variables en consideración que a la fecha no superan los 10.000 vocablos. Sin embargo, dado que durante la lectura aparecen palabras nuevas estas se irán incorporando a medida que se consideren que son correctas.

Normalizado: una vez obtenido el archivo de frases que no han podido ser interpretadas durante la lectura óptica, lo primero que se hace es verificar el estado de las bases. Esto consiste en determinar si la estructura de las bases es compatible con lo establecido por el Sistema. Un segundo paso es el de normalización que consiste en sacar todos los caracteres que no son de utilidad para la codificación. A título de ejemplo

"=" reemplaza por " " (espacio vacío)
 ")" reemplaza por " "
 "1°" reemplaza por ""

Acondicionamiento de bases

El eje principal en esta etapa es el proceso lingüístico que denominamos en forma equivalente como "campo semántico" o "familiarizado". Solo a los efectos de aclarar los puntos se puede resumir que un campo semántico es un conjunto de palabras (denominados hijos) semánticamente diferentes pero que a los efectos codificatorios son reducibles a un solo vocablo (denominado padre). Por ejemplo:



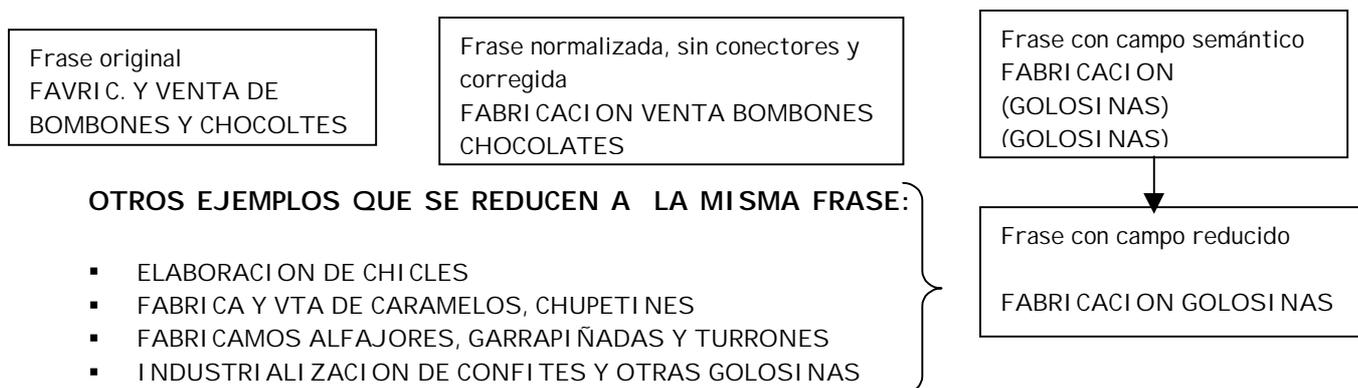
Mediante la aplicación de los campos semánticos se puede obtener una mayor frecuencia de la palabra padre, pues esta aparecerá reemplazando a cualquiera de sus hijos. Del mismo modo, se aumenta la frecuencia de las frases que contengan la palabra padre. Ello es muy importante a la hora del cálculo de los pesos heurísticos y los scores.

La familia involucra vocablos hijos cuya raíz es igual al vocablo padre. Ejemplo:

Padre: FABRICACION Hijos: FABRICA, FABRICAN, FABRICACIONES, FABRICO.

Por lo tanto la familia no es mas que un caso particular de "campo semántico". Aclarado esto, el proceso de campo semántico actúa en la base a codificar de la siguiente forma:

Ejemplo completo de la etapa de preparación de bases



Para tener idea de los efectos que este proceso lingüístico tiene se presenta un cuadro de resultados.

Resultados de la pre-codificación

OPERATIVO	Cantidad de frases originales	Cantidad de frases únicas luego de la Pre-codificación
Censo experimental	35.567 (100%)	11.745 (33%)
EPH (octubre 98)	34.047 (100%)	9.984 Frase a codificar (29%)

5.3.- Etapa III: Codificación

La etapa de codificación es la más importante en términos de que es aquí donde se encuentra la solución a las necesidades de codificar en forma rápida y precisa. Es por ello que se han "ingeniado" diversas estrategias de codificación uniendo en las mismas tres disciplinas:

- Normativa clasificatoria (marco normativo de la clasificación pericia sectorial y práctica codificatoria).
- Informática (lógica y desarrollo de sistema).
- Metodología estadística (scores y control de calidad)

Así el sistema de codificación resultante abarca los siguientes métodos:

Autofrase: de manera muy simple podría decirse que si en los diccionarios de frases codificadas se tiene una frase como por ejemplo "Fabricación de leche", que por no tener necesidad de mirar otras variables tiene 3° generación "A", quiere decir que todo operativo

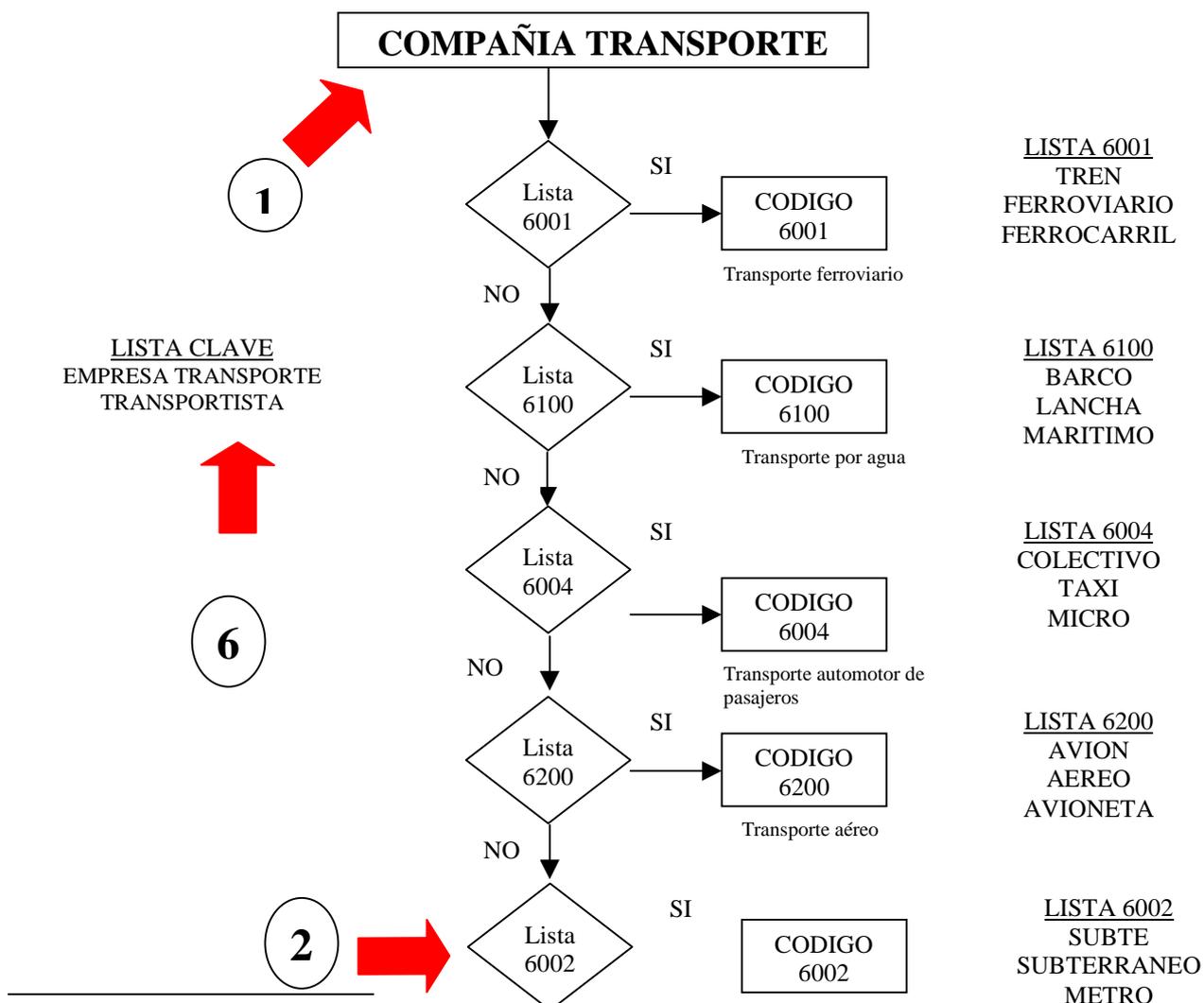
que traiga la descripción "Fabricación de leche"(o su equivalente en términos de campos semánticos) se codificará en forma automática y sin errores posibles³.

Microprocesos: son un conjunto de reglas de decisión diseñadas por los sectorialistas y utilizadas en el SiCI para que mediante palabras claves u otras variables (por ejemplo cantidad de ocupados del establecimiento) se le pueda asignar un código a una frase de actividad u ocupación que presente múltiples alternativas de codificación. Los microprocesos están dirigidos a "tomar decisiones" en forma automática a partir de la información contenida en otras variables que complementan las respuestas de la variable a codificar. Por ejemplo:

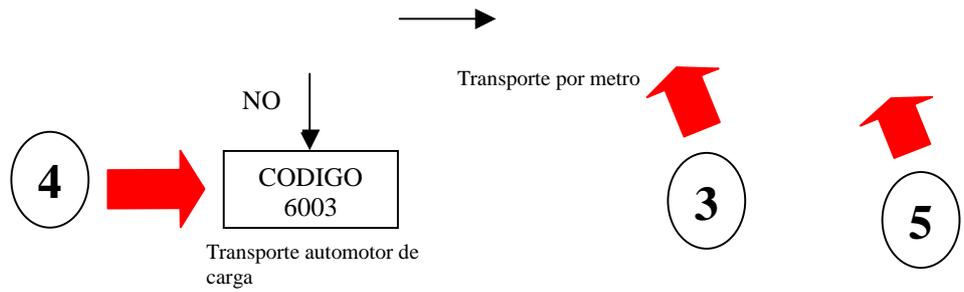
Frase de actividad ----- Empresa de transporte
 Códigos asignados por los codificadores----- 60-61-62
 (a dos dígitos)

Si en ocupación o tarea surge información que lo conecte a trenes, transporte automotor o avión, el caso se resuelve.

Así el diseño de un microproceso tomando el ejemplo anterior que daría de la siguiente forma:



³ Salvo error en los diccionarios que como ya se dijo deben ser perfectos.



1) FRASE CLAVE: es la frase por la cual el registro pasa a la codificación por el método de microprocesos. Ejemplo: cuando en la descripción de actividad aparezca transportista o transporte o empresa transporte, esos casos serán codificados según el esquema presentado.

2) RESTRICCION: determina si existe algún dato para definir un código.

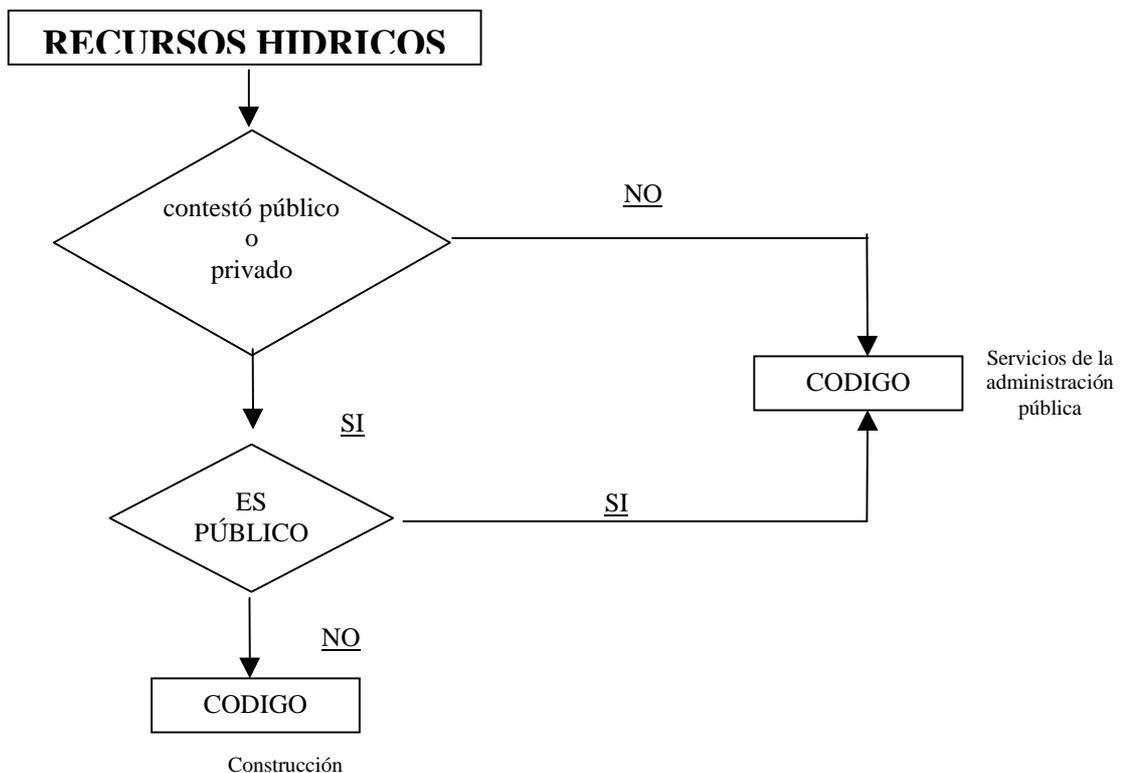
3) CODIGO CAES QUE CORRESPONDERIA SI HAY RESTRICCIONES

4) CODIGO GENERAL SI NO SE ENCUENTRA RESTRICCIONES, es decir que si no entró a ninguna respuesta SI de las restricciones queda como código asignado.

5) PALABRAS DE LAS LISTAS DE EXCEPCIONES: conjunto de palabras que funcionan como dato clave para designar un código automáticamente.

6) LISTA CLAVE: es una lista de frases que tienen igual significación que la frase clave y que deberían tener el mismo diseño de microproceso.

También puede tomarse como tamaño del establecimiento o si la empresa en la que trabaja la persona es una empresa pública o privada. Por ejemplo:



Scores: es un método que combina dos elementos. Por un lado la “**especificidad**” que cada palabra tiene respecto a los distintos códigos. Por ejemplo la palabra leche es “más específica” que “fabricación” pues a la primera se le asocia una limitada cantidad de códigos mientras que la segunda es de uso más difundido en todas las ramas de la industria. Esto es un movimiento analítico dentro del diccionario de codificación. La especificidad de cada palabra del diccionario se mide a través del llamado “**peso heurístico**” que también forman parte de los diccionarios junto con los literales y los códigos. Por el otro lado, el score también analiza la relación entre las frases del diccionario y las frases a codificar. Dada una frase a codificar, el “score” permite elegir “frases candidatas” dentro de la “oferta” que da el diccionario. Esas “candidatas” se eligen teniendo en cuenta el mayor número de palabras comunes entre la frase a codificar y las frases del diccionario. Cuanto mayor coincidencia se da entre ambos tipos de frases mayor será el “score”.

Score 10: se produce cuando la frase a codificar encuentra una “frase con la misma palabra” en el diccionario independientemente de su orden. Sin embargo el autofrase ya eliminó de la base a codificar aquellas que tiene el mismo orden.

Score < 10: en este caso la coincidencia entre frases a codificar y candidatas no es perfecta.

Ejemplo de frase a codificar:

i) Ejemplo frase score 10

ii) Ejemplo de frase score < 10

Fabricación de golosinas y galletitas

Fabricación de galletitas y golosinas

Fabricación de golosinas.

Autopalabra: es un método de codificación automático o directo que permite la asignación de un código único sin intervención de los codificadores. Para esto, utiliza un diccionario de codificación formado exclusivamente por palabras y los códigos asociados según la frase de la cual provengan. En la práctica, este método se ha abandonado pues el grado de error da por encima del 50% y el porcentaje a codificar, dado que es un método que se aplica a los casos que no se pudieron codificar por otros métodos, es extremadamente bajo.

6.- Prueba piloto del SiCI en el Censo Experimental de Pergamino

A fines de 1999, dentro de las actividades previstas en el cronograma del censo 2001, se realizó un censo experimental en la localidad de Pergamino que dio origen a una base a codificar de 35567 registros. En ese operativo se realizó la prueba de los diccionarios de lectura y se ensayó la corrección ortográfica que alimentó en forma diaria los diccionarios de lectura provistos inicialmente. Ello permitió mejorar los procesos previstos hasta adoptar la forma actual. En términos de los resultados de la etapa de pre-codificación, la manipulación de las frases a través de la aplicación de los diccionarios corrector, anulador y espurias más los procesos de estandarizado, aplicación de los campos semánticos o familiarizado y reducción a frase única, la base a codificar se redujo de 35567 a 11745 registros, esto es un 67 %; siendo este el punto de partida de la base a codificar.

Con relación a la codificación se realizaron dos pruebas, una que finalizó en junio de 2000 y una segunda que se acaba e terminar. Haciendo una síntesis de los resultados obtenidos a este momento se tiene:

Codificación Automática para actividades:

Base: Pergamino, octubre de 1999 **Registros a codificar:** 35567

Se codificó sobre la base del Clasificador de actividades económicas para encuestas sociodemográficas para el Mercosur (CAES), a 4 ó 2 dígitos y categoría de tabulación (letra).

Método	Autofrase	Score 10	Micro-procesos	Scores entre 8.5 y 10 *	Autopalabras
Cantidad de codificados	5699	799	9908	8971	1026
Porcentaje	16,02	2,25	27,86	25,22	2,88
Error promedio	0 %		<6%	30 %	50%

*método aún no calibrado

Conclusiones: el método de microprocesos aún tiene un potencial sin explotar que daría un margen para aumentar el porcentaje de codificación tratando a la vez de reducir aún más el porcentaje de error. Queda todavía definir el error máximo a tolerar. El método scores < 10 si bien ha dado errores muy altos, cercanos al 30%, también es cierto que aún queda por trabajar los valores críticos de los scores por rama de actividad y definir un valor de dispersión mínimo entre los scores de las frases candidatas. Dado los errores producidos por el método de autopalabras y el escaso aporte que realiza a la codificación el método queda descartado.

Codificación automática para ocupaciones: luego de la primera prueba de codificación informatizada para la pregunta abierta de ocupación (informe de junio de 2000) se dieron dos nuevos pasos:

- 1) correcciones en los procesos con palabras clave
- 2) aplicación de la estandarización de palabras antes de aplicar los mencionados procesos

Correcciones y nueva corrida del programa: a partir de los resultados obtenidos en la primera prueba, en cuanto a cantidad y calidad de la codificación informatizada, se corrigieron los principales errores encontrados y se amplió el campo de la codificación mediante la creación de nuevos procesos. La corrección consistió tanto en el agregado como en la reducción o modificación de los listados de palabras y restricciones ligados a los procesos. La ampliación del campo de la codificación informatizada supuso la creación de nuevos procesos que no habían sido considerados en ocasión de la primera prueba, ya sea porque en principio nos centramos en los procesos que consideramos más importantes, ya sea porque el análisis de los casos que quedaron sin código asignado revelaron la posibilidad de crear nuevos procesos. Luego se corrió nuevamente el programa sobre la misma base de datos de referencia. Los resultados comparativos entre la primera y la segunda prueba son los siguientes:

	Prueba 1	Prueba 2
<u>Casos con código asignado</u>	<u>44,4%</u>	<u>60,8%</u>
Casos codificados	35,0%	48,0%
Casos embolsados	9,4%	12,8%
<u>Casos no codificados</u>	<u>55,6%</u>	<u>39,2%</u>
Total de casos	35.567	35.567

Como se ve, se produjo un incremento muy sensible en la cantidad de códigos asignados: aproximadamente un 33% más que en la primera prueba. Y si bien el incremento porcentual es aproximadamente el mismo en los dos rubros que integran este ítem (casos codificados y casos embolsados), la mayor proporción del primero (en una relación de 4 a 1 respecto del segundo) implica que en valores absolutos los nuevos resultados sean muy significativos.

Estandarización de palabras y nueva aplicación de los procesos: un método para expandir el rango de aplicación de los procesos es la estandarización de las palabras. Consiste básicamente en reducir las palabras (tanto las de la información empírica como las de las palabras clave de los procesos y sus listas de restricciones asociadas) a su raíz. Como parte de la segunda prueba de codificación informatizada se realizó una codificación utilizando la estandarización. Los resultados fueron los siguientes:

<u>Casos con código asignado</u>	<u>66,8%</u>
Casos codificados	54,8%
Casos embolsados	12,0%
<u>Casos no codificados</u>	<u>33,2%</u>
Total de casos	35.567

Como se ve, los casos con código asignado pasan de 60,8% (sin estandarizar) a 66,8% (estandarizando), es decir, se incrementan 6 puntos (10% relativo) **utilizando los mismos procesos** de codificación con palabras clave. La única diferencia entre ambas situaciones es la estandarización de las palabras. Además, el incremento se produce totalmente en los casos codificados en forma directa, ya que los casos embolsados incluso disminuyen. Esto indica un buen camino a seguir.

Control de calidad: para la Prueba 1⁴ se realizó un control de calidad sobre los casos con código asignado (15.788 casos). Los resultados de este control difieren conceptualmente de acuerdo a si los errores localizados se encuentran entre los casos codificados o entre los casos embolsados:

- los errores entre los casos codificados son "definitivos": un código único es asignado por proceso automático y este será el código final a menos que actúen otras instancias de verificación y control;
- los errores entre los casos embolsados son "no definitivos": el código asignado por proceso automático es provisorio, orientativo para la instancia de codificación asistida; si el código genérico asignado es erróneo ("orienta mal") esto aún puede ser subsanado por el codificador que deba asignar el código final, ya sea asignándole el código correcto o derivando el caso a otra instancia de codificación.

Para el control de calidad se revisó una muestra del 10%. Dicha muestra incluyó por lo tanto 1.579 casos con código asignado. El resultado fue el siguiente:

<u>Casos con código asignado incorrectamente</u>	<u>7,5%</u>
Errores "definitivos"	5,5%
Errores "no definitivos"	2,0%
<u>Casos con código asignado correctamente</u>	<u>92,5%</u>

4 Actualmente se está realizando el control de calidad de la Prueba 2