Octava reunión del Grupo de Trabajo de la CEA sobre Clasificadores Internacionales (GTCI)
Tegucigalpa, Honduras - 5-8 Agosto 2019.

Actualización del Sistema de Codificación Informatizada -SiCI-

Coordinación de Clasificadores y Nomenclaturas – **CCyN**
Dirección Nacional de Metodología Estadística

Instituto Nacional de Estadística y Censos
República Argentina

Actualización del Sistema de Codificación Informatizada - SiCI

Aspectos de la Actualización

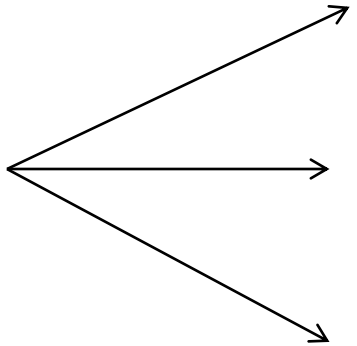
1. Incorporación de los nuevos clasificadores
2. Migración de Clipper a SAS
3. Cambio de clasificador de rama de actividad

Clasificaciones de Rama de Actividad Económica del INDEC

- Clasificación de Actividades Económicas para encuestas Sociodemografica del MERCOSUR (CAES)

- Clasificación Nacional de Actividades Económicas (ClaNAE)

Tres elementos básicos



Procesos Diccionarios

Procesos lingüísticos

Procesos de codificación

Diccionarios

Listados inventariados de palabras o frases que se originan en las respuestas empíricas

Dos tipos

a) Diccionario para el tratamiento de las palabras

b) Diccionario de codificación

Diccionarios

- Diccionario de palabras espurias (E)
- Diccionario de anuladas (A)
- Diccionario de conectores (C)
- Diccionario de excepciones (X)
- Diccionario corrector (R)
- Diccionario de palabras correctas (D)
- Diccionario de lectura (L) E + C + X + D
- Diccionario de codificación D + X

Procesos lingüísticos:

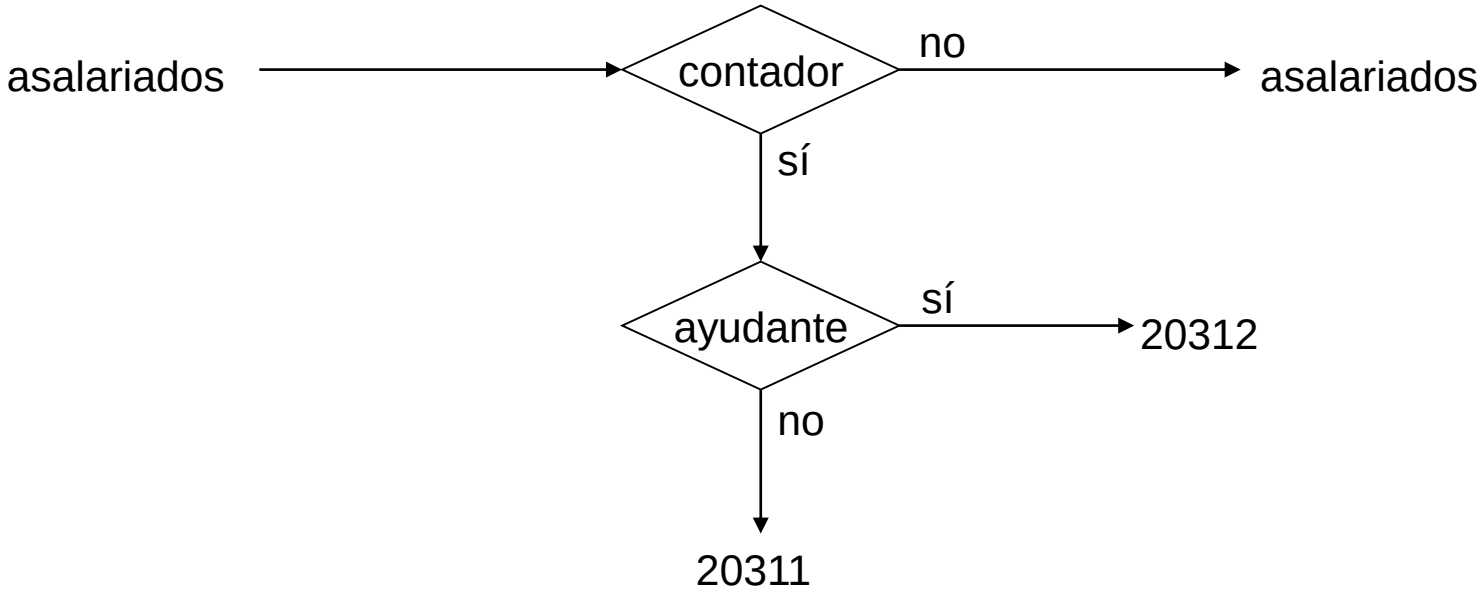
Modifican los literales de las frases a codificar simplificando el vocabulario

- **Normalizado**
- **Campo semántico o familiarizado**
- **Estandarizado**

Procesos de codificación

- **Microprocesos**
- **Frase única**

Ejemplo de microproceso



“Frase única”:

utiliza el diccionario de codificación formado por frases que ofrecen una única alternativa de código y son independientes de otras variables.

Acciones para la actualización del SiCI

- Incorporación ClaNAE 2010
- Recodificación de las 30000 “frases únicas” de CAES
- Revisión y recodificación bases de EPH
- Revisión de la “familiarización” de frases de CAES para ClaNAE

ClaNAE		CAES
47	**COMERCIO AL POR MENOR, EXCEPTO EL COMERCIO DE VEHÍCULOS AUTOMOTORES Y MOTOCICLETAS**	**48**
47111	Venta al por menor en hipermercados	**4808**
47112	Venta al por menor en supermercados	**4808**
47113	Venta al por menor en minimercados	**4808**
47119	Venta al por menor en kioscos, polirrubros y comercios no especializados n.c.p.	**4808**
47190	Venta al por menor en comercios no especializados, sin predominio de productos alimenticios y bebidas	**4809**
47211	Venta al por menor principalmente de fiambres, quesos y productos lácteos	**4803**
47212	Venta al por menor de productos de almacén y dietética	**4803**
47213	Venta al por menor de carnes rojas, menudencias y chacinados frescos	**4803**
47214	Venta al por menor de huevos, carne de aves y productos de granja y de la caza	**4803**
47215	Venta al por menor de pescados y productos de la pesca	**4803**
47216	Venta al por menor de frutas, legumbres y hortalizas frescas	**4803**
47217	Venta al por menor de pan y productos de panadería y confitería	**4803**
47219	Venta al por menor de productos alimenticios n.c.p. en comercios especializados	**4803**

- **Dos nuevos métodos**
  1. Contrastación de palabras y frases
  2. Árboles de codificación

Contrastación de palabras y frases

- Requiere
  - Fuente de datos o **archivo input**
  - Archivos con textos ya codificados

Serán contrastados mediante algoritmos de text mining

Se elaboran diccionarios

- Funcionan como diccionarios:
  - a.- Censo Nacional Económico (CNE) 2004
  - b.- ClaNAE 2010
  - c.- CPC 2.1
  - d.- Bases de la EPH

Archivos input y diccionarios

- Se realiza una partición de la base en función de los verbos que utiliza:
  - p: producción
  - v: venta
  - s: servicio
  - t: transporte

Se contrastan archivos

Score basado en similitudes de palabras y textos y su frecuencia

CÓDIGO CLANAE

Árbol de Clasificación

Dos archivos fuente

- Base ya codificada (EPH) (archivo de entrenamiento)
- Archivo fuente (archivo de testing)

Algoritmo → parte de un nodo

↓

determina el espacio de búsqueda

Búsqueda por secuencia binaria

positiva

negativa

define código binario

- Origen del método
  - Leer registros del archivo de entrenamiento
    - Se depliega el árbol

- **Proceso testing**
  - Se clasifica por cada código
  - por SÍ ó por NO
    - Se confirma con el árbol de entrenamiento

Coordinación de Clasificadores y Nomenclaturas – CCyN

Dirección Nacional de Metodología Estadística

Muchas gracias